BUSINESS ANALYTICS

SAMRAT KRISHNA GADDAM
DR.A.SRINAGESH,DR.T.S.RAVI KIRAN

The Author of this book is solely responsible and liable for its content including but not limited to the views, representations, descriptions, statements, information, opinions and references [“Content”]. The Content of this book shall not constitute or be construed or deemed to reflect the opinion or expression of the Publisher or Editor. Neither the Publisher nor Editor endorse or approve the Content of this book or guarantee the reliability, accuracy or completeness of the Content published herein and do not make any representations or warranties of any kind, express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose. The Publisher and Editor shall not be liable whatsoever for any errors, omissions, whether such errors or omissions result from negligence, accident, or any other cause or claims for loss or damages of any kind, including without limitation, indirect or consequential loss or damage arising out of use, inability to use, or about the reliability, accuracy or sufficiency of the information contained in this book.

Made with ❤ on the Notion Press Platform
www.notionpress.com

TO MY PARENTS AND BROTHER

for raising me to believe that anything was possible.

Foreword

The book has been written in an easy-to-understand and student-friendly manner, and includes several illustrative figures and examples, sample codes and project case studies. This book is written while keeping multidisciplinary undergraduates and post-graduates in mind as primary readers. Each chapter begins by defining the learning objectives for each section, recall from previous chapters, introduction and the important terms covered in that chapter.The book is intended as a text for students of Computer Science and Engineering, Information Technology, Master of Computer Applications, M.Sc(Computer Science) and M.Sc(Computational Data Science).

Preface

This book deals with business analytics (BA)—an emerging area in modern business decision making.

BA is a data driven decision making approach that uses statistical and quantitative analysis, information technology, management science (mathematical modeling, simulation), along with data mining and fact-based data to measure past business performance to guide an organization in business planning, predicting the future outcomes, and effective decision making.

BA tools are also used to visualize and explore the patterns and trends in the data to predict future business outcomes with the help of forecasting and predictive modeling.

In this age of technology, companies collect massive amount of data. Successful companies use their data as inputs to take effective decision which helps them to emerge as the best in the market.

Acknowledgements

First and foremost, we are grateful to all the co-authors Dr.A. Sai Nagesh, Professor in CSE Department,RVR & J.C College of Engineering, Chowdavaram, Guntur and Dr.T.S.Ravi Kiran HoD & Associate Professor, P.B.Siddhartha College of Arts & Science,Vijayawada for their time, effort, and understanding during the preparation of the book.It is an extraordinary delight to recognize the help and commitments of numerous people to this exertion. It is an enormous pleasure to acknowledge the support and contributions of many individuals to this effort. First, we would like to thank our editor, Notion Press Publishing for guidance, encouragement, and support. I also thank the management of P.B.Siddhartha College of Arts & Science, Vijayawada, Andhar Pradesh, India for their support.

Samrat Krishna Gaddam,
Assistant Professor,
P.B.Siddhartha College of Arts & Science,
Vijayawada, AP, India.
P.No-9177937461
Email-gsamratkrishna@pbsiddhartha.ac.in

Prologue

Chapter 1 gives an overview of Types of Digital Data: structured, unstructured and semi-structured data. Big data from business, Introduction of big data, Characteristics of big data, Data in the warehouse, Importance of Big data, Big data Use cases: Patterns for Big data deployment –Big data Market Survey.

Chapter 2 explains MongoDB: Why MongoDB-Terms used in RDBMS and MongoDB- Data Types- MongoDB Query Language, Mapper-Reducer-Combiner-Partitioner-Searching-Sorting- Compression

Chapter 3 describes What and Why Business Analytics, Business Analytics Importance, Descriptive Analytics-Data Warehousing, Business Reporting, Visual Analytics, and Business Performance Management, Predictive Analytics-Techniques for Predictive Modeling, Web Analytics, Web Mining, and Social Analytics-Case Study

Chapter 4 ddiscusses Case Study, Model-Based Decision Making: Optimization and Multi-Criteria Systems, Modeling and Analysis: Heuristic Search Methods and Simulation -Case Study.

Chapter 5 defines Opening Vignette, Location-Based Analytics for Organizations, Analytics Applications for Consumers, Web 2.0 and Online Social Networking, Cloud Computing and Bl, Impacts of Analytics in Organizations, Analytics Ecosystem

ONE

Big Data Analytics

1.1 Types of Big Data

Now that we are on track with what is big data, let's have a look at the types of big data:

a) **Structured:** Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

b) **Unstructured:** Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

c) **Semi-structured:** Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tagsthat segregate individual elements within the data. Thus we come to the end of types of data.

1.2 What is Big Data?

According to Gartner, the definition of Big Data – "Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

This definition clearly answers the "What is Big Data?" question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

The History of Big Data Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

1.3 Characteristics of Big Data

Back in 2001, Gartner analyst Doug Laney listed the 3 ‘V’s of Big Data – Variety, Velocity, and Volume. Let’s discuss the characteristics of big data. These characteristics, isolated, are enough to know what big data is. Let’s look at them in depth:

a) **Variety:** Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

b) **Velocity:** Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

c) **Volume:** Volume is one of the characteristics of big data. We already know that Big Data indicates huge ‘volumes’ of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus comes to the end of characteristics of big data.

1.4 Why is Big Data Important?

The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

1. **Cost Savings:** Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are

to be stored and these tools also help in identifying more efficient ways of doing business.

2. **Time Reductions:** The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.

3. **Understand the market conditions:** By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

4. **Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

5. **Using Big Data Analytics to Boost Customer Acquisition and Retention:** The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.

6. **Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights:** Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.

7. **Big Data Analytics As a Driver of Innovations and Product Development:** Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

1.5 Business Intelligence vs Big Data

Although Big Data and Business Intelligence are two technologies used to analyze data to help companies in the decision-making process, there are differences between both of them. They differ in the way they work as much as in the type of data they analyze.

Traditional BI methodology is based on the principle of grouping all business data into a central server. Typically, this data is analyzed in offline mode, after storing the information in an environment called Data Warehouse. The data is structured in a conventional relational database with an additional set of indexes and forms of access to the tables (multidimensional cubes).

A Big Data solution differs in many aspects to BI to use. These are the main differences between Big Data and Business Intelligence:

1. In a Big Data environment, information is stored on a distributed file system, rather than on a central server. It is a much safer and more flexible space.

2. Big Data solutions carry the processing functions to the data, rather than the data to the functions. As the analysis is centered on the information, it´s easier to handle larger amounts of information in a more agile way.

3. Big Data can analyze data in different formats, both structured and unstructured. The volume of unstructured data (those not stored in a traditional database) is growing at levels much higher than the structured data. Nevertheless, its analysis carries different challenges. Big Data solutions solve them by allowing a global analysis of various sources of information.

4. Data processed by Big Data solutions can be historical or come from real-time sources. Thus, companies can make decisionsthat affect their business in an agile and efficient way.

5. Big Data technology uses parallel mass processing (MPP) concepts, which improves the speed of analysis. With MPP many instructions are executed simultaneously, and since the various jobs are divided into several parallel execution parts, at the end the overall results are reunited and presented. This allows you to analyze large volumes of information quickly.

1.6 Big Data vs Data Warehouse

Big Data has become the reality of doing business for organizations today. There is a boom in the amount of structured as well as raw data that floods every organization daily. If this data is managed well, it can lead to powerful insights and quality decision making.

Big data analytics is the process of examining large data sets containing a variety of data types to discover some knowledge in databases, to identify interesting patterns and establish relationships to solve problems, market trends, customer preferences, and other useful information. Companies and businesses that implement Big Data Analytics often reap several business benefits. Companies implement Big Data Analytics because they want to make more informed business decisions.

A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the Data Warehouse through the processes of extraction, transformation and loading (ETL tools). Data analysis tools, such as business intelligence software, access the data within the warehouse.

1.7 Hadoop Environment Big Data Analytics

Hadoop is changing the perception of handling Big Data especially the unstructured data. Let’s know how Apache Hadoop software library, which is a framework, plays a vital role in handling Big Data. Apache Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers, as both versions might be vulnerable to failures.

1.7.1 Hadoop Community Package Consists of

- File system and OS level abstractions
- A MapReduce engine (either MapReduce or YARN)
- The Hadoop Distributed File System (HDFS)

- Java ARchive (JAR) files
- Scripts needed to start Hadoop
- Source code, documentation and a contribution section

1.7.2 Activities performed on Big Data

- Store – Big data need to be collected in a seamless repository, and it is not necessary to store in a single physical database.
- Process – The process becomes more tedious than traditional one in terms of cleansing,enriching, calculating, transforming, and running algorithms.
- Access – There is no business sense of it at all when the data cannot be searched, retrieved easily, and can be virtually showcased along the business lines.

1.8 Classification of analytics

1.8.1 Descriptive analytics

Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.

Data aggregation and **data mining** are two techniques used in descriptive analytics to discover historical data. Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts.

Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning. Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.

Advantages:

- Quickly and easily report on the Return on Investment (ROI) by showing how performance achieved business or target goals.
- Identify gaps and performance issues early **-** before they become problems.
- Identify specific learners who require additional support, regardless of how many students or employees there are.

- Identify successful learners in order to offer positive feedback or additional resources.
- Analyze the value and impact of course design and learning resources.

1.8.2 Predictive analytics

Predictive Analytics is a statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors.

The software for predictive analytics has moved beyond the realm of statisticians and is becoming more affordable and accessible for different markets and industries, including the field of learning & development.

For online learning specifically, predictive analytics is often found incorporated in the Learning Management System (LMS), but can also be purchased separately as specialized software.

For the learner, predictive forecasting could be as simple as a dashboard located on the main screen after logging in to access a course. Analyzing data from past and current progress, visual indicators in the dashboard could be provided to signal whether the employee was on track with training requirements.

Advantages:

- Personalize the training needs of employees by identifying their gaps, strengths, and weaknesses; specific learning resources and training can be offered to support individual needs.
- Retain Talent by tracking and understanding employee career progression and forecasting what skills and learning resources would best benefit their career paths. Knowing what skills employees need also benefits the design of future training.
- Support employees who may be falling behind or not reaching their potential by offering intervention support before their performance puts them at risk.
- Simplified reporting and visuals that keep everyone updated when predictive forecasting is required.

1.8.3 Prescriptive analytics

Prescriptive analytics is a statistical method used to generate recommendations and make decisions based on the computational findings of algorithmic models.

Generating automated decisions or recommendations requires specific and unique algorithmic models and clear direction from those utilizing the analytical technique. A recommendation cannot be generated without knowing what to look for or what problem is desired to be solved. In this way, prescriptive analytics begins with a problem.

Example: A Training Manager uses predictive analysis to discover that most learners without a particular skill will not complete the newly launched course. What could be done? Now prescriptive analytics can be of assistance on the matter and help determine options for action. Perhaps an algorithm can detect the learners who require that new course, but lack that particular skill, and send an automated recommendation that they take an additional training resource to acquire the missing skill.

The accuracy of a generated decision or recommendation, however, is only as good as the quality of data and the algorithmic models developed. What may work for one company's training needs may not make sense when put into practice in another company's training department. Models are generally recommended to be tailored for each unique situation and need.

Descriptive vs Predictive vs Prescriptive Analytics

Descriptive Analytics is focused solely on historical data.

You can think of Predictive Analytics asthen using this historical data to develop statistical models that will then forecast about future possibilities.

Prescriptive Analytics takes Predictive Analytics a step further and takes the possible forecasted outcomes and predicts consequences for these outcomes

1.9 What are Big Data Analytics Challenges

1. Need For Synchronization Across Disparate Data Sources As data sets are becoming bigger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is overlooked, it will create gaps and lead to wrong messages and insights.

2. Acute Shortage Of Professionals Who Understand Big Data Analysis The analysis of data is important to make this voluminous amount of data being

produced in every minute, useful. With the exponential rise of data, a huge demand for big data scientists and Big Data analysts has been created in the market. It is important for business organizations to hire a data scientist having skills that are varied as the job of a data scientist is multidisciplinary. Another major challenge faced by businesses is the shortage of professionals who understand Big Data analysis. There is a sharp shortage of data scientists in comparison to the massive amount of data being produced.

3. Getting Meaningful Insights Through The Use Of Big Data Analytics It is imperative for business organizations to gain important insights from Big Data analytics, and also it is important that only the relevant department has accessto this information. A big challenge faced by the companies in the Big Data analytics is mending this wide gap in an effective manner.

4. Getting Voluminous Data Into The Big Data Platform It is hardly surprising that data is growing with every passing day. This simply indicates that business organizations need to handle a large amount of data on daily basis. The amount and variety of data available these days can overwhelm any data engineer and that is why it is considered vital to make data accessibility easy and convenient for brand owners and managers.

5. Uncertainty Of Data Management Landscape With the rise of Big Data, new technologies and companies are being developed every day. However, a big challenge faced by the companies in the Big Data analytics is to find out which technology will be best suited to them without the introduction of new problems and potential risks.

6. Data Storage And Quality Business organizations are growing at a rapid pace. With the tremendous growth of the companies and large business organizations, increases the amount of data produced. The storage of this massive amount of data is becoming a real challenge for everyone. Popular data storage options like data lakes/ warehouses are commonly used to gather and store large quantities of unstructured and structured data in its native format. The real problem arises when a data lakes/ warehouse try to combine unstructured and inconsistent data from diverse sources, it encounters errors. Missing data, inconsistent data, logic conflicts, and duplicates data all result in data quality challenges.

7. Security And Privacy Of Data Once business enterprises discover how to use Big Data, it brings them a wide range of possibilities and opportunities. However, it also involves the potential risks associated with big data when it comes to the privacy and the security of the data. The Big Data tools used for analysis and storage utilizes the data disparate sources. This eventually leads to a high risk of exposure of the data, making it vulnerable. Thus, the rise of voluminous amount of data increases privacy and security concerns.

1.10 Terminologies Used In Big Data Environments

- **As-a-service infrastructure:** Data-as-a-service, software-as-a-service, platform-as-a-service – all refer to the idea that rather than selling data, licences to use data, or platforms for running Big Data technology, it can be provided "as a service", rather than as a product. This reduces the upfront capital investment necessary for customers to begin putting their data, or platforms, to work for them, as the provider bears all of the costs of setting up and hosting the infrastructure. As a customer, as-a-service infrastructure can greatly reduce the initial cost and setup time of getting Big Data initiatives up and running.
- **Data science:** Data science is the professional field that deals with turning data into value such as new insights or predictive models. It brings together expertise from fields including statistics, mathematics, computer science, communication as well as domain expertise such as business knowledge. Data scientist has recently been voted the No 1 job in the U.S., based on current demand and salary and career opportunities.
- **Data mining:** Data mining is the process of discovering insights from data. In terms of Big Data, because it is so large, this is generally done by computational methods in an automated way using methods such as decision trees, clustering analysis and, most recently, machine learning. This can be thought of as using the brute mathematical power of computers to spot patterns in data which would not be visible to the human eye due to the complexity of the dataset.

Hadoop Hadoop is a framework for Big Data computing which has been released into the public domain as open source software, and so can freely be used by anyone. It consists of a number of modules all tailored for a different vitalstep ofthe Big Data process – from file storage (Hadoop File System – HDFS) to database (HBase) to carrying out data operations (Hadoop MapReduce – see below). It has become so popular due to its power and flexibility that it has developed its own industry of retailers (selling tailored versions), support service providers and consultants.

- **Predictive modelling:** At its simplest, this is predicting what will happen next based on data about what has happened previously. In the Big Data age, because there is more data around than ever before, predictions are becoming more and more accurate. Predictive modelling is a core component of most Big Data initiatives, which are formulated to help us choose the course of action which will lead to the most desirable outcome. The speed of modern computers and the volume of data available means that predictions can be made based on a huge number of variables, allowing an ever-increasing number of variables to be assessed for the probability that it will lead to success.
- **MapReduce:** MapReduce is a computing procedure for working with large datasets, which was devised due to difficulty of reading and analysing really Big Data using conventional computing methodologies. As its name suggest, it consists of two procedures – mapping (sorting information into the format needed for analysis – i.e. sorting a list of people according to their age) and reducing (performing an operation, such checking the age of everyone in the dataset to see who is over 21).
- **NoSQL:** NoSQL refers to a database format designed to hold more than data which is simply arranged into tables, rows, and columns, as isthe case in a conventional relational database. This database format has proven very popular in Big Data applications because Big Data is often messy, unstructured and does not easily fit into traditional database frameworks.
- **Python:** Python is a programming language which has become very popular in the Big Data space due to its ability to work very well with large,

unstructured datasets (see Part II for the difference between structured and unstructured data). It is considered to be easier to learn for a data science beginner than other languages such as R (see also Part II) and more flexible.

- **R Programming:** R is another programming language commonly used in Big Data, and can be thought of as more specialised than Python, being geared towards statistics. Its strength lies in its powerful handling of structured data. Like Python, it has an active community of users who are constantly expanding and adding to its capabilities by creating new libraries and extensions.
- **Recommendation engine:** A recommendation engine is basically an algorithm, or collection of algorithms, designed to match an entity (for example, a customer) with something they are looking for. Recommendation engines used by the likes of Netflix or Amazon heavily rely on Big Data technology to gain an overview of their customers and, using predictive modelling, match them with products to buy or content to consume. The economic incentives offered by recommendation engines has been a driving force behind a lot of commercial Big Data initiatives and developments over the last decade.
- **Real-time:** Real-time means "as it happens" and in Big Data refers to a system or process which is able to give data-driven insights based on what is happening at the present moment. Recent years have seen a large push for the development of systems capable of processing and offering insights in real-time (or near-real-time), and advances in computing power as well as development of techniques such as machine learning have made it a reality in many applications today.
- **Reporting:** The crucial "last step" of many Big Data initiative involves getting the right information to the people who need it to make decisions, at the right time. When this step is automated, analytics is applied to the insights themselves to ensure that they are communicated in a way that they will be understood and easy to act on. This will usually involve creating multiple reports based on the same data or insights but each intended for a different audience (for example, in-depth technical analysis for engineers, and an overview of the impact on the bottom line for c-level executives).

- **Spark:** Spark is another open source framework like Hadoop but more recently developed and more suited to handling cutting-edge Big Data tasks involving real time analytics and machine learning. Unlike Hadoop it does not include its own filesystem, though it is designed to work with Hadoop's HDFS or a number of other options. However, for certain data related processes it is able to calculate at over 100 times the speed of Hadoop, thanks to its in-memory processing capability. This means it is becoming an increasingly popular choice for projects involving deep learning, neural networks and other compute-intensive tasks.
- **Structured Data:** Structured data is simply data that can be arranged neatly into charts and tables consisting of rows, columns or multi-dimensioned matrixes. This is traditionally the way that computers have stored data, and information in this format can easily and simply be processed and mined for insights. Data gathered from machines is often a good example ofstructured data, where various data points – speed, temperature, rate of failure, RPM etc. – can be neatly recorded and tabulated for analysis.
- **Unstructured Data:** Unstructured data is any data which cannot easily be put into conventional charts and tables. This can include video data, pictures, recorded sounds, text written in human languages and a great deal more. This data has traditionally been far harder to draw insight from using computers which were generally designed to read and analyze structured information. However, since it has become apparent that a huge amount of value can be locked away in this unstructured data, great efforts have been made to create applications which are capable of understanding unstructured data – for example visual recognition and natural language processing.
- **Visualization:** Humans find it very hard to understand and draw insights from large amounts of text or numerical data – we can do it, but it takes time, and our concentration and attention is limited. For this reason effort has been made to develop computer applications capable of rendering information in a visual form – charts and graphics which highlight the most important insights which have resulted from our Big Data projects. A subfield of reporting (see above), visualizing is now often an automated process, with

visualizations customized by algorithm to be understandable to the people who need to act or take decisions based on them.

Basic availability, Soft state and Eventual consistency

Basic availability implies continuous system availability despite network failures and tolerance to temporary inconsistency.

Soft state refers to state change without input which is required for eventual consistency.

Eventual consistency means that if no further updates are made to a given updated database itemfor long enough period of time , all users will see the same value for the updated item.

1.11 Top Analytics Tools

*** R** is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility.
- It provides a suite of operators for calculations on arrays, in particular, matrices.
- It provides coherent, integrated collection of big data tools for data analysis.
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy.

*** Apache Spark** is a powerful open source big data analytics tool. It offers over 80 high-level operators that make it easy to build parallel apps. It is used at a wide range of organizations to process large datasets.

Features:

- It helps to run an application in Hadoop cluster, up to 100 times faster in memory, and ten times faster on disk.
- It offers lighting Fast Processing.
- Support for Sophisticated Analytics.
- Ability to Integrate with Hadoop and Existing Hadoop Data.

* **Plotly** is an analytics tool that lets users create charts and dashboards to share online.

Features:

- Easily turn any data into eye-catching and informative graphics.
- It provides audited industries with fine-grained information on data provenance.
- Plotly offers unlimited public file hosting through its free community plan.

* **Lumify** is a big data fusion, analysis, and visualization platform. It helps users to discover connections and explore relationships in their data via a suite of analytic options.

Features:

- It provides both 2D and 3D graph visualizations with a variety of automatic layouts.
- It provides a variety of options for analyzing the links between entities on the graph.
- It comes with specific ingest processing and interface elements for textual content, images, and videos.
- It spaces feature allows you to organize work into a set of projects, or workspaces.
- It is built on proven, scalable big data technologies.

* **IBM SPSS** Modeler is a predictive big data analytics platform. It offers predictive models and delivers to individuals, groups, systems and the enterprise. It has a range of advanced algorithms and analysis techniques.

Features:

- Discover insights and solve problems faster by analyzing structured and unstructured data
- Use an intuitive interface for everyone to learn
- You can select from on-premises, cloud and hybrid deployment options

- Quickly choose the best performing algorithm based on model performance

*** MongoDB** is a NoSQL, document-oriented database written in C, C++, and JavaScript. It is free to use and is an open source tool that supports multiple operating systems including Windows Vista (and later versions), OS X (10.7 and later versions), Linux, Solaris, and FreeBSD. Its main features include Aggregation, Adhoc-queries, Uses BSON format, Sharding, Indexing, Replication, Server-side execution of javascript, Schemaless, Capped collection, MongoDB management service (MMS), load balancing and file storage.

Features:

- Easy to learn.
- Provides support for multiple technologies and platforms.
- No hiccups in installation and maintenance.
- Reliable and low cost.

TWO

Introduction to MongoDB and MapReduce Programming

2.1 INTRODUCTION TO MONGODB AND MAPREDUCE PROGRAMMING

MongoDB is a cross-platform, document-oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

Database

Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases.

Collection

Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.

Document

A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection’s documents may hold different types of data.

Sample Document

Following example shows the document structure of a blog site, which is simply a comma separated key value pair.

```
{
_id: ObjectId(7df78ad8902c) title: ‘MongoDB Overview’,
description: ‘MongoDB is no sql database’, by: ’tutorials point‘,
url: ’http://www.tutorialspoint.com‘, tags: [’mongodb‘, ’database’, ‘NoSQL’], likes: 100,
comments: [
{
user:‘user1’,
message: ‘My first comment’,
dateCreated: new Date(2011,1,20,2,15), like: 0
},
{
user:‘user2’,
message: ‘My second comments’, dateCreated: new Date(2011,1,25,7,45), like: 5
}
]
}
```

_id is a 12 bytes hexadecimal number which assures the uniqueness of every document. You can provide _id while inserting the document. If you don’t provide then MongoDB provides a unique id for every document. These 12 bytes first 4 bytes for the current timestamp, next 3 bytes for machine id, next 2 bytes for process id of MongoDB server and remaining 3 bytes are simple incremental VALUE.

Any relational database has a typical schema design that shows number of tables and the relationship between these tables. While in MongoDB, there is no concept of relationship.

2.2 Advantages of MongoDB over RDBMS

Schema less − MongoDB is a document database in which one collection holds different documents. Number of fields, content and size of the document can differ from one document to another.

Structure of a single object is clear.

No complex joins.

Deep query-ability. MongoDB supports dynamic queries on documents using a document- based query language that's nearly as powerful as SQL.

Tuning.

Ease of scale-out − MongoDB is easy to scale.

Conversion/mapping of application objects to database objects not needed.

Uses internal memory for storing the (windowed) working set, enabling faster access of data.

2.3 Why Use MongoDB?

- Document Oriented Storage − Data is stored in the form of JSON style documents.
- Index on any attribute
- Replication and high availability
- Auto-Sharding
- Rich queries
- Fast in-place updates

Professional support by MongoDB Where to Use MongoDB?

- Big Data
- Content Management and Delivery
- Mobile and Social Infrastructure
- User Data Management
- Data Hub

MongoDB supports many datatypes. Some of them are −

- String − This is the most commonly used datatype to store the data. String in MongoDB must be UTF-8 valid.
- Integer − This type is used to store a numerical value. Integer can be 32 bit or 64 bit depending upon your server.
- Boolean − This type is used to store a boolean (true/ false) value.
- Double − This type is used to store floating point values.
- Min/ Max keys − This type is used to compare a value against the lowest and highest BSON elements.
- Arrays − This type is used to store arrays or list or multiple values into one key.
- Timestamp − ctimestamp. This can be handy for recording when a document has been modified or added.
- Object − This datatype is used for embedded documents.
- Null − This type is used to store a Null value.
- Symbol − This datatype is used identically to a string; however, it's generally reserved for languages that use a specific symbol type.
- Date − This datatype is used to store the current date or time in UNIX time format. You can specify your own date time by creating object of Date and passing day, month, year into it.
- Object ID − This datatype is used to store the document's ID.
- inary data − This datatype is used to store binary data.
- Code − This datatype is used to store JavaScript code into the document.
- Regular expression − This datatype is used to store regular expression

The find() Method

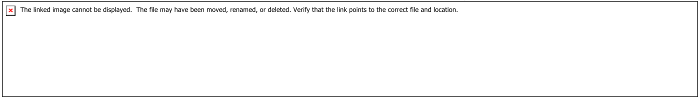
To query data from MongoDB collection, you need to use MongoDB's find() method. Syntax

The basic syntax of find() method is as follows −

```
>db.COLLECTION_NAME.find()
```

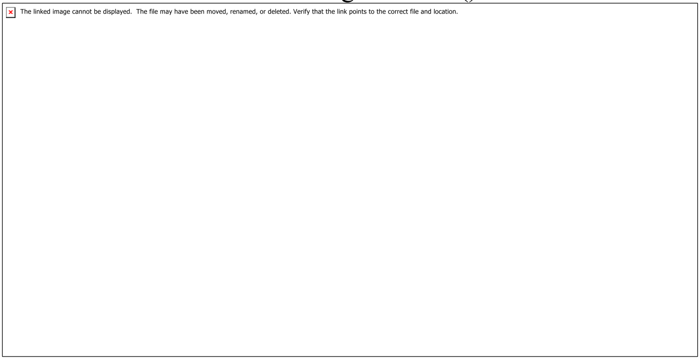
find() method will display all the documents in a non-structured way. Example

Assume we have created a collection named mycol as −



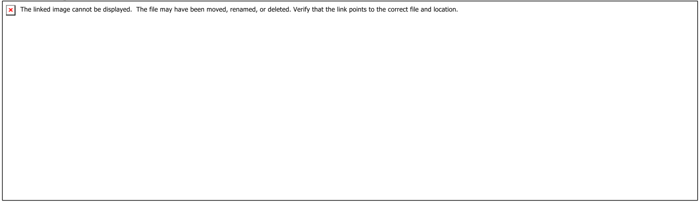
The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

And inserted 3 documents in it using the insert() method as shown below −



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Following method retrieves all the documents in the collection −



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

The pretty() Method

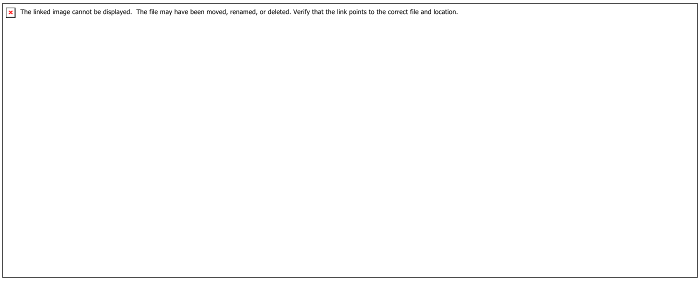
To display the results in a formatted way, you can use pretty() method.

Syntax

```
>db.COLLECTION_NAME.find().pretty()
```

Example

Following example retrieves all the documents from the collection named mycol and arranges them in an easy-to-read format.



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

The findOne() method

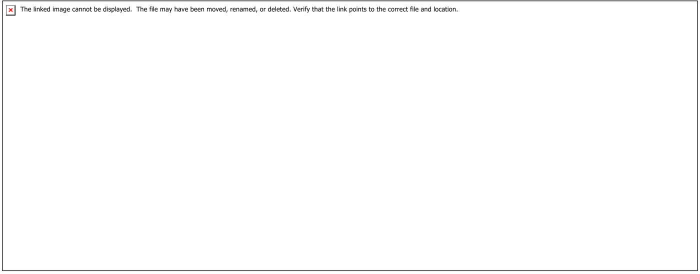
Apart from the find() method, there is findOne() method, that returns only one document.

Syntax

>db.COLLECTIONNAME.findOne()

Example

Following example retrieves the document with title MongoDB Overview.



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

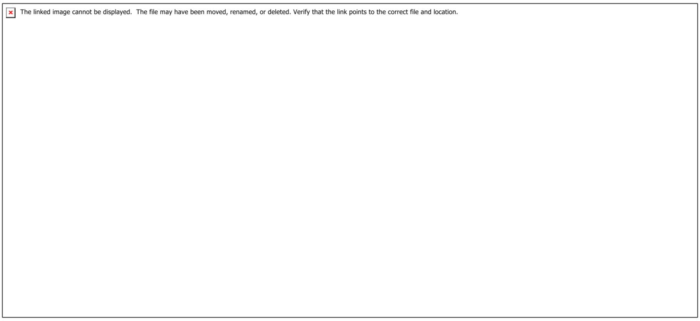
AND in MongoDB

Syntax

To query documents based on the AND condition, you need to use $and keyword. Following is the basic syntax of AND −

>db.mycol.find({ $and: [{<key1>:<value1>}, { <key2>:<value2>}] }) Example

Following example will show all the tutorials written by ’tutorials point‘ and whose title is ’MongoDB Overview‘.


The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

For the above given example, equivalent where clause will be ’ where by = ’tutorials point‘ AND title = ’MongoDB Overview‘ ’. You can pass any number of key, value pairs in find clause.

OR in MongoDB

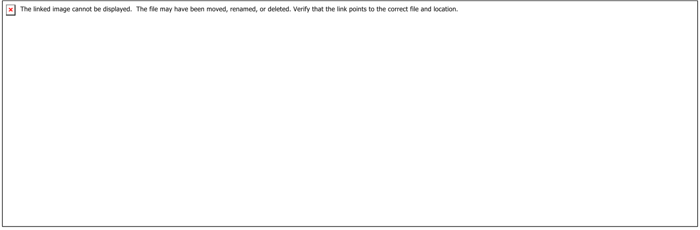
Syntax

To query documents based on the OR condition, you need to use $or keyword. Following is the basic syntax of OR −

```
>db.mycol.find(
{
$or: [
{key1: value1}, {key2:value2}
]
}
).pretty()
```

Example

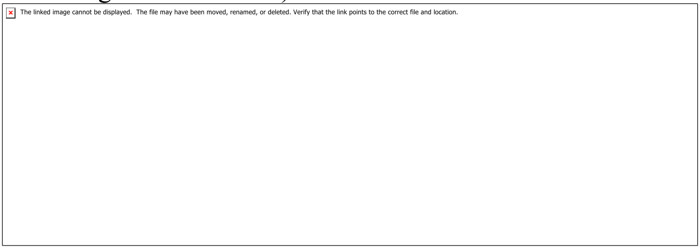
Following example will show all the tutorials written by ’tutorials point‘ or whose title is ’MongoDB Overview‘.


The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Using AND and OR Together

Example

The following example will show the documents that have likes greater than 10 and whose title is either ’MongoDB Overview‘ or by is ’tutorials point’. Equivalent SQL where clause is ‘where likes>10 AND (by = ’tutorials point’ OR title = ‘MongoDB Overview’)'


The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

NOR in MongoDB Syntax

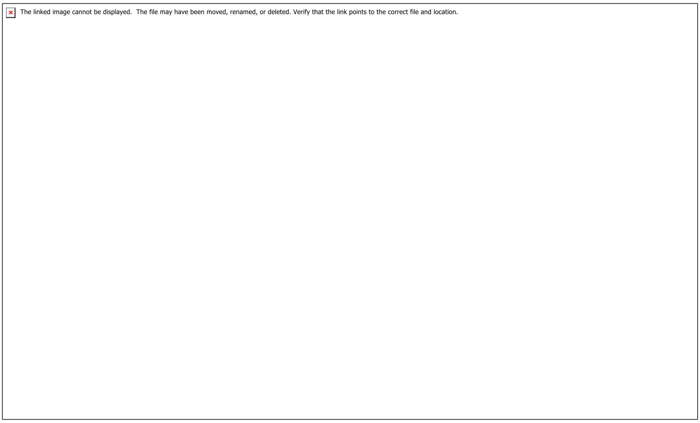
To query documents based on the NOT condition, you need to use $not keyword. Following is the basic syntax of NOT −

```
>db.COLLECTION_NAME.find(
{
}
)
```

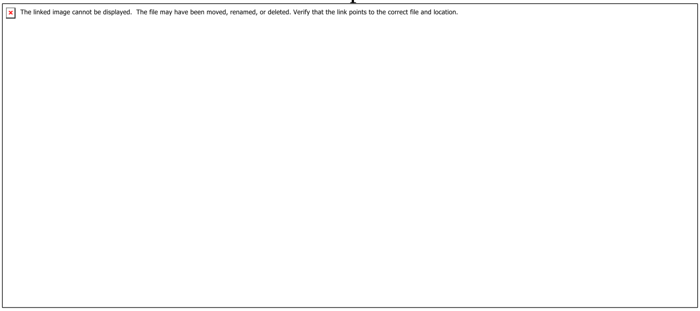
Example

```
$not: [
]
{key1: value1}, {key2:value2}
```

Assume we have inserted 3 documents in the collection empDetails as shown below −


The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Following example will retrieve the document(s) whose first name is not "Radhika" and last name is not "Christopher"


The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

NOT in MongoDB Syntax

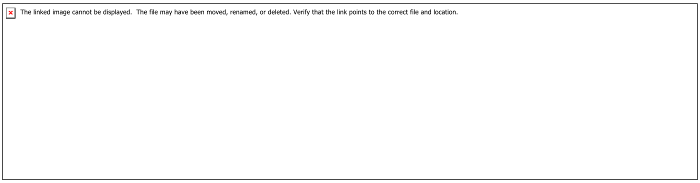
To query documents based on the NOT condition, you need to use $not keyword following is the basic syntax of NOT −

```
>db.COLLECTION_NAME.find(
{
}
).pretty()
```

Example

```
$NOT: [
]
{key1: value1}, {key2:value2}
```

Following example will retrieve the document(s) whose age is not greater than 25



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

2.4 MapReduce:

MapReduce addresses the challenges of distributed programming by providing an abstraction that isolates the developer from system-level details (e.g., locking of data structures, data starvation issues in the processing pipeline, etc.). The programming model specifies simple and well-defined interfaces between a small number of components, and therefore is easy for the programmer to reason about. MapReduce maintains a separation of what computations are to be performed and how those computations are actually carried out on a cluster of machines. The first is under the control of the programmer, while the second is exclusively the responsibility of the execution framework or "runtime". The advantage is that the execution framework only needs to be designed once and verified for correctness—thereafter, as long as the developer expresses computations in the programming model, code is guaranteed to behave as expected. The upshot is that the developer is freed from having to worry about system-level details (e.g., no more debugging race conditions and addressing lock contention) and can instead focus on algorithm or application design.

Which often has multiple cores). Why is MapReduce important? In practical terms, it provides a very effective tool for tackling large-data problems. But beyond that, MapReduce is important in how it has changed the way we organize computations at a massive scale. MapReduce represents the first widely-adopted step away from the von Neumann model that has served as the foundation of computer science over the last half plus century. Valiant called this a bridging model [148], a conceptual bridge between the physical implementation of a machine and the software that is to be executed on that machine. Until recently, the von Neumann model has served us well: Hardware designers focused on efficient implementations of the von Neumann model and didn't have to think much about the actual software that would run on the machines. Similarly, the software industry developed software targeted at the model without worrying about the hardware details. The result was extraordinary growth: chip designers churned out successive generations of increasingly powerful processors, and software engineers were able to develop applications in high-level languages that exploited those processors.

MapReduce can be viewed as the first breakthrough in the quest for new abstractions that allow us to organize computations, not over individual machines, but over entire clusters. As Barroso puts it, the datacenter is the computer. MapReduce is certainly not the first model of parallel computation that has been proposed. The most prevalent model in theoretical computer science, which dates back several decades, is the PRAM. MAPPERS AND REDUCERS Key-value pairs form the basic data structure in MapReduce. Keys and values may be primitives such as integers, floating point values, strings, and raw bytes, or they may be arbitrarily complex structures (lists, tuples, associative arrays, etc.). Programmers typically need to define their own custom data types, although a number of libraries such as Protocol Buffers,5 Thrift,6 and Avro7 simplify the task. Part of the design of MapReduce algorithms involves imposing the key-value structure on arbitrary datasets. For a collection of web pages, keys may be URLs and values may be the actual HTML content. For a graph, keys may represent node ids and values may contain the adjacency lists of those nodes. In some algorithms, input keys are not particularly meaningful and are simply ignored during processing, while in other cases input keys are used to uniquely identify a datum (such as a record id). In Chapter 3, we discuss the role of complex keys and values in the design of various algorithms. In MapReduce, the programmer defines a mapper and a reducer with the following signatures: map: (k1, v1) → [(k2, v2)] reduce: (k2, [v2]) → [(k3, v3)] The convention [. . .] is used throughout this book to denote a list. The input to a MapReduce job starts as data stored on the underlying distributed file system (see Section 2.5). The mapper is applied to every input key-value pair (split across an arbitrary number of files) to generate an arbitrary number of intermediate key-value pairs. The reducer is applied to all values associated with the same intermediate key to generate output key-value pairs.8 Implicit between the map and reduce phases is a distributed “group by” operation on intermediate keys.

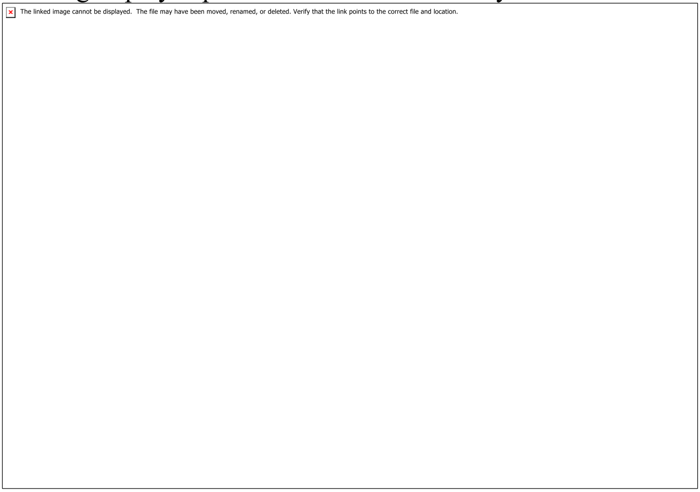

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 2.2 Simplified View of Map Reduce

2.4 MAPREDUCE BASICS

Distribution over words in a collection). Input key-values pairs take the form of (docid, doc) pairs stored on the distributed file system, where the former is a unique identifier for the document, and the latter is the text of the document itself. The mapper takes an input key-value pair, tokenizes the document, and emits an intermediate key-value pair for every word: the word itself serves as the key, and the integer one serves as the value (denoting that we’ve seen the word once). The MapReduce execution framework guarantees that all values associated with the same key are brought together in the reducer. Therefore, in our word count algorithm, we simply need to sum up all counts (ones) associated with each word. The reducer does exactly this, and emits final keyvalue pairs with the word as the key, and the count as the value. Final output is written to the distributed file system, one file per reducer. Words within each file will be sorted by alphabetical order, and each file will contain roughly the same number of words. The partitioner, which we discuss later in Section 2.4, controls the assignment of words to reducers. The output can be examined by the programmer or used as input to another MapReduce program.

There are some differences between the Hadoop implementation of MapReduce and Google's implementation.9 In Hadoop, the reducer is presented with a key and an iterator over all values associated with the particular key. The values are arbitrarily ordered. Google's implementation allows the programmer to specify a secondary sort key for ordering the values (if desired)—in which case values associated with each key would be presented to the developer's reduce code in sorted order. Later in Section 3.4 we discuss how to overcome this limitation in Hadoop to perform secondary sorting. Another difference: in Google's implementation the programmer is not allowed to change the key in the reducer. That is, the reducer output key must be exactly the same as the reducer input key. In Hadoop, there is no such restriction, and the reducer can emit an arbitrary number of output key-value pairs (with different keys).

To provide a bit more implementation detail: pseudo-code provided in this book roughly mirrors how MapReduce programs are written in Hadoop. Mappers and reducers are objects that implement the Map and Reduce methods, respectively. In Hadoop, a mapper object is initialized for each map task (associated with a particular sequence of key-value pairs called an input split) and the Map method is called on each key-value pair by the execution framework. In configuring a MapReduce job, the programmer provides a hint on the number of map tasks to run, but the execution framework . The situation is similar for the reduce phase: a reducer object is initialized for each reduce task, and the Reduce method is called once per intermediate key. In contrast with the number of map tasks, the programmer can precisely specify the number of reduce tasks. We will return to discuss the details of Hadoop job execution in Section 2.6, which is dependent on an understanding of the distributed file system . To reiterate: although the presentation of algorithms in this book closely mirrors the way they would be implemented in Hadoop, our focus is on algorithm design and conceptual

understanding—not actual Hadoop programming. For that, we would recommend Tom White's book [154]. What are the restrictions on mappers and reducers? Mappers and reducers can express arbitrary computations over their inputs. However, one must generally be careful about use of external resources since multiple mappers or reducers may be contending for those resources. For example, it may be unwise for a mapper to query an external SQL database, since that would introduce a scalability bottleneck on the number of map tasks that could be run in parallel (since they might all be simultaneously querying the database).10 In general, mappers can emit an arbitrary number of intermediate key-value pairs, and they need not be of the same type as the input key-value pairs. Similarly, reducers can emit an arbitrary number of final key-value pairs, and they can differ in type from the intermediate key-value pairs. Although not permitted in functional programming, mappers and reducers can have side effects. This is a powerful and useful feature: for example, preserving state across multiple inputs is central to the design of many MapReduce algorithms . Such algorithms can be understood as having side effects that only change state that is internal to the mapper or reducer. While the correctness of such algorithms may be more difficult to guarantee (since the function's behavior depends not only on the current input but on previous inputs), most potential synchronization problems are avoided since internal state is private only to individual mappers and reducers. In other cases , it may be useful for mappers or reducers to have external side effects, such as writing files to the distributed file system. Since many mappers and reducers are run in parallel, and the distributed file system is a shared global resource, special care must be taken to ensure that such operations avoid synchronization conflicts. One strategy is to write a temporary file that is renamed upon successful completion of the mapper or reducer .

In addition to the "canonical" MapReduce processing flow, other variations are also possible. MapReduce programs can contain no reducers, in which case mapper output is directly written to disk (one file per mapper). For embarrassingly parallel problems, e.g., parse a large text collection or independently analyze a large number of images, this would be a common pattern. The converse—a MapReduce program with no mappers—is not possible, although in some cases it is useful for the mapper to implement the identity function and simply pass input key-value pairs to the reducers. This has the effect of sorting and regrouping the input for reduce-side processing. Similarly, in some cases it is useful for the reducer to implement the identity function, in which case the program simply sorts and groups mapper output. Finally, running identity mappers and reducers has the effect of regrouping and resorting the input data (which is sometimes useful).

Although in the most common case, input to a MapReduce job comes from data stored on the distributed file system and output is written back to the distributed file system, any other system that satisfies the proper abstractions can serve as a data source or sink. With Google's MapReduce implementation, BigTable [, a sparse, distributed, persistent multidimensional sorted map, is frequently used as a source of input and as a store of MapReduce output. HBase is an open-source BigTable clone and has similar capabilities. Also, Hadoop has been integrated with existing MPP (massively parallel processing) relational databases, which allows a programmer to write MapReduce jobs over database rows and dump output into a new database table. Finally, in some cases MapReduce jobs may not consume any input at all (e.g., computing π) or may only consume a small amount of data (e.g., input parameters to many instances of processorintensive simulations running in parallel).

2.5 PARTITIONERS AND COMBINERS

We have thus far presented a simplified view of MapReduce. There are two additional elements that complete the programming model: partitioners and combiners. Partitioners are responsible for dividing up the intermediate key space and assigning intermediate key-value pairs to reducers. In other words, the partitioner specifies the task to which an intermediate key-value pair must be copied. Within each reducer, keys are processed in sorted order (which is how the "group by" is implemented). The simplest partitioner involves computing the hash value of the key and then taking the mod of that value with the number of reducers. This assigns approximately the same number of keys to each reducer (dependent on the quality of the hash function). Note, however, that the partitioner only considers the key and ignores the value—therefore, a roughly-even partitioning of the key space may nevertheless yield large differences in the number of key-values pairs sent to each reducer (since different keys may have different numbers of associated values). This imbalance in the amount of data associated with each key is relatively common in many text processing applications due to the Zipfian distribution of word occurrences.

Combiners are an optimization in MapReduce that allow for local aggregation before the shuffle and sort phase. We can motivate the need for combiners by considering the word count algorithm in Figure 2.3, which emits a key-value pair for each word in the collection. Furthermore, all these key-value pairs need to be copied across the network, and so the amount of intermediate data will be larger than the input collection itself. This is clearly inefficient. One solution is to perform local aggregation on the output of each mapper, i.e., to compute a local count for a word over all the documents processed by the mapper. With this modification (assuming the maximum amount of local aggregation possible), the number of intermediate key-value pairs will be at most the number of unique words in the collection times the number of mappers (and typically far smaller because each mapper may not encounter every word).

smaller because each mapper may not encounter every word). The combiner in MapReduce supports such an optimization. One can think of combiners as "mini-reducers" that take place on the output of the mappers, prior to the shuffle and sort phase. Each combiner operates in isolation and therefore does not have access to intermediate output from other mappers. The combiner is provided keys and values associated with each key (the same types as the mapper output keys and values). Critically, one cannot assume that a combiner will have the opportunity to process all values associated with the same key. The combiner can emit any number of key-value pairs, but the keys and values must be of the same type as the mapper output (same as the reducer input).12 In cases where an operation is both associative and commutative (e.g., addition or multiplication), reducers can directly serve as combiners. In general, however, reducers and combiners are not interchangeable.

In many cases, proper use of combiners can spell the difference between an impractical algorithm and an efficient algorithm which focuses on various techniques for local aggregation. It suffices to say for now that a combiner can significantly reduce the amount of data that needs to be copied over the network, resulting in much faster algorithms. The complete MapReduce model is shown in Figure 2.3. Output of the mappers are processed by the combiners, which perform local aggregation to cut down on the number of intermediate key- value pairs. The partitioner determines which reducer will be responsible for processing a particular key, and the execution framework uses this information to copy the data to the right location during the shuffle and sort phase.13 Therefore, a complete MapReduce job consists of code for the mapper, reducer, combiner, and partitioner, along with job configuration parameters. The execution framework handles everything else.

☒ The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 2.3 Complete View of Map reduce

2.6 SECONDARY SORTING

MapReduce sorts intermediate key-value pairs by the keys during the shuffle and sort phase, which is very convenient if computations inside the reducer rely on sort order (e.g., the order inversion design pattern described in the previous section). However, what if in addition to sorting by key, we also need to sort by value? Google's MapReduce implementation provides built-in functionality for (optional) secondary sorting, which guarantees that values arrive in sorted order. Hadoop, unfortunately, does not have this capability built in.

Consider the example of sensor data from a scientific experiment: there are m sensors each taking readings on continuous basis, where m is potentially a large number. A dump of the sensor data might look something like the following, where rx after each timestamp represents the actual sensor readings (unimportant for this discussion, but may be a series of values, one or more complex records, or even raw bytes of images).

(t1, m1, r80521) (t1, m2, r14209) (t1, m3, r76042) ... (t2, m1, r21823) (t2, m2, r66508)

(t2, m3, r98347)

Suppose we wish to reconstruct the activity at each individual sensor over time. A MapReduce program to accomplish this might map over the raw data and emit the sensor id as the intermediate key, with the rest of each record as the value:

m1 → (t1, r80521)

This would bring all readings from the same sensor together in the reducer. However, since MapReduce makes no guarantees about the ordering of values associated with the same key, the sensor readings will not likely be in temporal order. The most obvious solution is to buffer all the readings in memory and then sort by timestamp before additional processing. However, it should be apparent by now that any in-memory buffering of data introduces a potential scalability bottleneck. What if we are working with a high frequency sensor or sensor readings over a long period of time? What if the sensor readings themselves are large complex objects? This approach may not scale in these cases—the reducer would run out of memory trying to buffer all values associated with the same key.

This is a common problem, since in many applications we wish to first group together data one way (e.g., by sensor id), and then sort within the groupings another way (e.g., by time). Fortunately, there is a general purpose solution, which we call the "value-to-key conversion" design pattern. The basic idea is to move part of the value into the intermediate key to form a composite key, and let the MapReduce execution framework handle the sorting. In the above example, instead of emitting the sensor id as the key, we would emit the sensor id and the timestamp as a composite key: (m1, t1) → (r80521).

The sensor reading itself now occupies the value. We must define the intermediate key sort order to first sort by the sensor id (the left element in the pair) and then by the timestamp (the right element in the pair). We must also implement a custom partitioner so that all pairs associated with the same sensor are shuffled to the same reducer. Properly orchestrated, the key-value pairs will be presented to the reducer in the correct sorted order: (m1, t1) → [(r80521)] (m1, t2) → [(r21823)].

(m1, t3) → [(r146925)] . . .

However, note that sensor readings are now split across multiple keys. The reducer will need to preserve state and keep track of when readings associated with the current sensor end and the next sensor begin.9 The basic tradeoff between the two approaches discussed above (buffer and inmemory sort vs. value-to-key conversion) is where sorting is performed. One can explicitly implement secondary sorting in the reducer, which is likely to be faster but suffers from a scalability bottleneck.10 With value-to-key conversion, sorting is offloaded to the MapReduce execution framework. Note that this approach can be arbitrarily extended to tertiary, quaternary, etc. sorting. This pattern results in many more keys for the framework to sort, but distributed sorting is a task that the MapReduce runtime excels at since it lies at the heart of the programming model.

2.7 INDEX COMPRESSION

We return to the question of how postings are actually compressed and stored on disk. This chapter devotes a substantial amount of space to this topic because index compression is one of the main differences between a “toy” indexer and one that works on real-world collections. Otherwise, MapReduce inverted indexing algorithms are pretty straightforward.

Let us consider the canonical case where each posting consists of a document id and the term frequency. A na¨ıve implementation might represent the first as a 32-bit integer9 and the second as a 16-bit integer. Thus, a postings list might be encoded as follows: [(5, 2),(7, 3),(12, 1),(49, 1),(51, 2), . . .]

where each posting is represented by a pair in parentheses. Note that all brackets, parentheses, and commas are only included to enhance readability; in reality the postings would be represented as a long stream of integers. This na¨ıve implementation would require six bytes per posting. Using this scheme, the entire inverted index would be about as large as the collection itself. Fortunately, we can do significantly better. The first trick is to encode differences between document ids as opposed to the document ids themselves. Since the postings are sorted by document ids, the differences (called d-gaps) must be positive integers greater than zero. The above postings list, represented with d-gaps, would be: [(5, 2),(2, 3),(5, 1),(37, 1),(2, 2)

Of course, we must actually encode the first document id. We haven't lost any information, since the original document ids can be easily reconstructed from the d-gaps. However, it's not obvious that we've reduced the space requirements either, since the largest possible d-gap is one less than the number of documents in the collection. This is where the second trick comes in, which is to represent the d-gaps in a way such that it takes less space for smaller numbers. Similarly, we want to apply the same techniques to compress the term frequencies, since for the most part they are also small values. But to understand how this is done, we need to take a slight detour into compression techniques, particularly for coding integers.

Compression, in general, can be characterized as either lossless or lossy: it's fairly obvious that loseless compression is required in this context. To start, it is important to understand that all compression techniques represent a time–space tradeoff. That is, we reduce the amount of space on disk necessary to store data, but at the cost of extra processor cycles that must be spent coding and decoding data. Therefore, it is possible that compression reduces size but also slows processing. However, if the two factors are properly balanced (i.e., decoding speed can keep up with disk bandwidth), we can achieve the best of both worlds: smaller and faster.

2.8 POSTINGS COMPRESSION

Having completed our slight detour into integer compression techniques, we can now return to the scalable inverted indexing algorithm shown in Figure 4.4 and discuss how postings lists can be properly compressed. As we can see from the previous section, there is a wide range of choices that represent different tradeoffs between compression ratio and decoding speed. Actual performance also depends on characteristics of the collection, which, among other factors, determine the distribution of d-gaps. B¨uttcher et al. [30] recently compared the performance of various compression techniques on coding document ids. In terms of the amount of compression that can be obtained (measured in bits per docid), Golomb and Rice codes performed the best, followed by γ codes, Simple-

9, varInt, and group varInt (the least space efficient). In terms of raw decoding speed, the order was almost the reverse: group varInt was the fastest, followed by varInt.14 Simple-9 was substantially slower, and the bit-aligned codes were even slower than that. Within the bit-aligned codes, Rice codes were the fastest, followed by γ, with Golomb codes being the slowest (about ten times slower than group varInt).

Let us discuss what modifications are necessary to our inverted indexing algorithm if we were to adopt Golomb compression for d-gaps and represent term frequencies with γ codes. Note that this represents a space-efficient encoding, at the cost of slower decoding compared to alternatives. Whether or not this is actually a worthwhile tradeoff in practice is not important here: use of Golomb codes serves a pedagogical purpose, to illustrate how one might set compression parameters.

Coding term frequencies with γ codes is easy since they are parameterless. Compressing d-gaps with Golomb codes, however, is a bit tricky, since two parameters are required: the size of the document collection and the number of postings for a particular postings list (i.e., the document frequency, or df). The first is easy to obtain and can be passed into the reducer as a constant. The df of a term, however, is not known until all the postings have been processed—and unfortunately, the parameter must be known before any posting is coded. At first glance, this seems like a chicken-and-egg problem. A two-pass solution that involves first buffering the postings (in memory) would suffer from the memory bottleneck we've been trying to avoid in the first place.

To get around this problem, we need to somehow inform the reducer of a term's df before any of its postings arrive. The solution is to have the mapper emit special keys of the form ht, ∗i to communicate partial document frequencies. That is, inside the mapper, in addition to emitting intermediate key-value pairs of the following form:

(tuple ht, docidi,tf f)

we also emit special intermediate key-value pairs like this:

(tuple ht, ∗i, df e)

to keep track of document frequencies associated with each term. In practice, we can accomplish this by applying the in-mapper combining design pattern (see

Section 3.1). The mapper holds an in-memory associative array that keeps track of how many documents a term has been observed in (i.e., the local document frequency of the term for the subset of documents processed by the mapper). Once the mapper has processed all input records, special keys of the form ht, ∗i are emitted with the partial df as the value.

To ensure that these special keys arrive first, we define the sort order of the tuple so that the special symbol ∗ precedes all documents (part of the order inversion design pattern). Thus, for each term, the reducer will first encounter the ht, ∗i key, associated with a list of values representing partial df values originating from each mapper. Summing all these partial contributions will yield the term's df, which can then be used to set the Golomb compression parameter b. This allows the postings to be incrementally compressed as they are encountered in the reducer—memory bottlenecks are eliminated since we do not need to buffer postings in memory.

Once again, the order inversion design pattern comes to the rescue. Recall that the pattern is useful when a reducer needs to access the result of a computation (e.g., an aggregate statistic) before it encounters the data necessary to produce that computation. For computing relative frequencies, that bit of information was the marginal. In this case, it's the document frequency.

2.9 PARALLEL BREADTH-FIRST SEARCH

One of the most common and well-studied problems in graph theory is the single-source shortest path problem, where the task is to find shortest paths from a source node to all other nodes in the graph (or alternatively, edges can be associated with costs or weights, in which case the task is to compute lowest-cost or lowest-weight paths). Such problems are a staple in undergraduate

algorithm courses, where students are taught the solution using Dijkstra's algorithm. However, this famous algorithm assumes sequential processing—how would we solve this problem in parallel, and more specifically, with MapReduce?

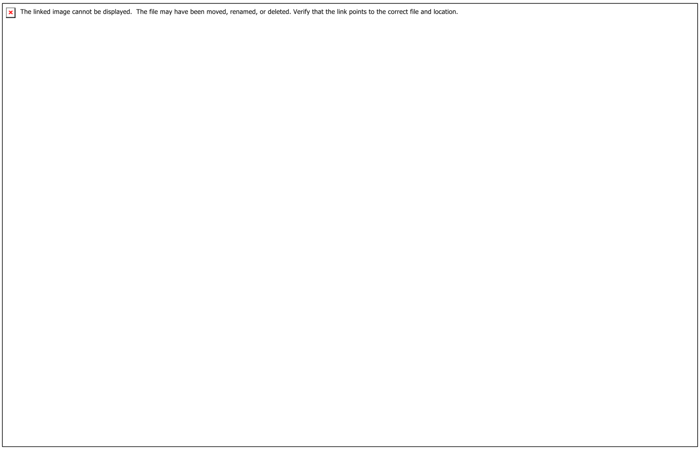
```
Dijkstra(G, w, s)
$d[s] \leftarrow 0$
for all vertex $v \in V$ do $d[v] \leftarrow \infty$
$Q \leftarrow \{V\}$
```

```
while $Q$ 6= $\emptyset$ do
    $u \leftarrow$ ExtractMin($Q$)
    for all vertex $v \in u$.Adjacency List do
        if $d[v] > d[u] + w(u, v)$ then
            $d[v] \leftarrow d[u] + w(u, v)$
```

Pseudo-code for Dijkstra’s algorithm, which is based on maintaining a global priority queue of nodes with priorities equal to their distances from the source node. At each iteration, the algorithm expands the node with the shortest distance and updates distances to all reachable nodes. As a refresher and also to serve as a point of comparison, Dijkstra’s algorithm is shown in Figure 5.2, adapted from Cormen, Leiserson, and Rivest’s classic algorithms textbook [41] (often simply known as CLR). The input to the algorithm is a directed, connected graph $G = (V, E)$ represented with adjacency lists, w containing edge distances such that $w(u, v) \geq 0$, and the source node s. The algorithm begins by first setting distances to all vertices $d[v]$, $v \in V$ to ∞, except for the source node, whose distance to itself is zero. The algorithm maintains Q, a global priority queue of vertices with priorities equal to their distance values d.

Dijkstra’s algorithm operates by iteratively selecting the node with the lowest current distance from the priority queue (initially, this is the source node). At each iteration, the algorithm “expands” that node by traversing the adjacency list of the selected node to see if any of those nodes can be reached with a path of a shorter distance. The algorithm terminates when the priority queue Q is empty, or equivalently, when all nodes have been considered. Note that the algorithm as presented in Figure 5.2 only computes the shortest distances. The actual paths can be recovered by storing “backpointers” for every node indicating a fragment of the shortest path.

A sample trace of the algorithm running on a simple graph is shown in Figure 5.3 (example also adapted from CLR). We start out in (a) with n1 having a distance of zero (since it’s the source) and all other nodes having a distance of ∞. In the first iteration (a), n1 is selected as the node to expand (indicated by the thicker border). After the expansion, we see in (b) that n2 and n3 can be reached at a distance of 10 and 5, respectively. Also, we see in (b) that n3 is the next node selected for expansion. Nodes we have already considered for expansion are shown in black. Expanding n3, we see in (c) that the distance to n2 has decreased because we've found a shorter path. The nodes that will be expanded next, in order, are n5, n2, and n4. The algorithm terminates with the end state shown in (f), where we've discovered the shortest distance to all nodes.



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 2.4 Dijkstra's algorithm Applied for Simple graph

The key to Dijkstra's algorithm is the priority queue that maintains a globallysorted list of nodes by current distance. This is not possible in MapReduce, as the programming model does not provide a mechanism for exchanging global data. Instead, we adopt a brute force approach known as parallel breadth-first search. First, as a simplification let us assume that all edges have unit distance (modeling, for example, hyperlinks on the web). This makes the algorithm easier to understand, but we'll relax this restriction later.

The intuition behind the algorithm is this: the distance of all nodes connected directly to the source node is one; the distance of all nodes directly connected to those is two; and so on. Imagine water rippling away from a rock dropped into a pond— that's a good image of how parallel breadth-first search works. However, what if there are multiple paths to the same node? Suppose we wish to compute the shortest distance to node n. The shortest path must go through one of the nodes in M that contains an outgoing edge to n: we need to examine all m ∈ M to find ms, the node with the shortest distance. The shortest distance to n is the distance to ms plus one.

Pseudo-code for the implementation of the parallel breadth-first search algorithm is provided in Figure 5.4. As with Dijkstra's algorithm, we assume a connected, directed graph represented as adjacency lists. Distance to each node is directly stored alongside the adjacency list of that node, and initialized to ∞ for all nodes except for the source node. In the pseudo-code, we use n to denote the node id (an integer) and N to denote the node's corresponding data structure (adjacency list and current distance). The algorithm works by mapping over all nodes and emitting a key-value pair for each neighbor on the node's adjacency list. The key contains the node id of the neighbor, and the value is the current distance to the node plus one. This says: if we can reach node n with a distance d, then we must be able to reach all the nodes that are connected to n with distance d + 1.

After shuffle and sort, reducers will receive keys corresponding to the destination node ids and distances corresponding to all paths leading to that node. The reducer will select the shortest of these distances and then update the distance in the node data structure.

h iteration corresponds to a MapReduce job. The first time we run the algorithm, we "discover" all nodes that are connected to the source. The second iteration, we discover all nodes connected to those, and so on. Each iteration of the algorithm expands the "search frontier" by one hop, and, eventually, all nodes will be discovered with their shortest distances (assuming a fully-connected graph). Before we discuss termination of the algorithm, there is one more detail required to make the parallel breadth-first search algorithm work. We need to "pass along" the graph structure from one iteration to the next. This is accomplished by emitting the node data structure itself, with the node id as a key (Figure 5.4, line 4 in the mapper). In the reducer, we must distinguish the node data structure from distance values (Figure 5.4, lines 5–6 in the reducer), and update the minimum distance in the node data structure before emitting it as the final value. The final output is now ready to serve as input to the next iteration.

So how many iterations are necessary to compute the shortest distance to all nodes? The answer is the diameter of the graph, or the greatest distance between any pair of nodes. This number is surprisingly small for many real-world problems: the saying “six degrees of separation” suggests that everyone on the planet is connected to everyone else by at most six steps (the people a person knows are one step away, people that they know are two steps away, etc.). If this is indeed true, then parallel breadthfirst search on the global social network would take at most six MapReduce iterations.

```
class Mapper
    method Map(nid n, node N)
        d ← N.Distance
        Emit(nid n, N) . Pass along graph structure
        for all nodeid m ∈ N.AdjacencyList do
            Emit(nid m, d + 1) .
```

Emit distances to reachable nodes

```
class Reducer
    method Reduce(nid m, [d1, d2, . . .])
        dmin ← ∞
        M ← ∅
        for all d ∈ counts [d1, d2, . . .] do
            if IsNode(d) then
                M ← d . Recover graph structure
            else if d < dmin then . Look for shorter distance
                dmin ← d
        M.Distance ← dmin . Update shortest distance
        Emit(nid m, node M)
```

Pseudo-code for parallel breath-first search in MapReduce: the mappers emit distances to reachable nodes, while the reducers select the minimum of those distances for each destination node. Each iteration (one MapReduce job) of the algorithm expands the “search frontier” by one hop.

For more serious academic studies of “small world” phenomena in networks. In practical terms, we iterate the algorithm until there are no more node distances that are ∞. Since the graph is connected, all nodes are reachable, and since all edge distances are one, all discovered nodes are guaranteed to have the shortest distances (i.e., there is not a shorter path that goes through a node that hasn’t been discovered).

The actual checking of the termination condition must occur outside of MapReduce. Typically, execution of an iterative MapReduce algorithm requires a nonMapReduce “driver” program, which submits a MapReduce job to iterate the algorithm, checks to see if a termination condition has been met, and if not, repeats. Hadoop provides a lightweight API for constructs called “counters”, which, as the name suggests, can be used for counting events that occur during execution, e.g., number of corrupt records, number of times a certain condition is met, or anything that the programmer desires. Counters can be defined to count the number of nodes that have distances of ∞: at the end of the job, the driver program can access the final counter value and check to see if another iteration is necessary.

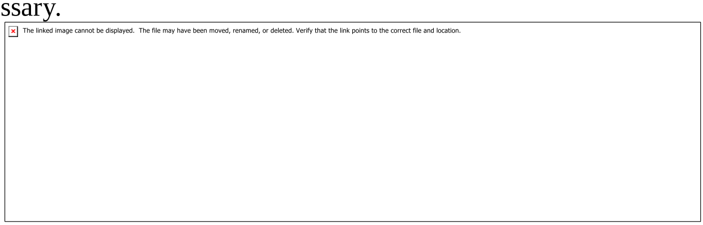

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 2.5 Parallel BFS

Finally, as with Dijkstra’s algorithm in the form presented earlier, the parallel breadth-first search algorithm only finds the shortest distances, not the actual shortest paths. However, the path can be straightforwardly recovered. Storing “backpointers” at each node, as with Dijkstra’s algorithm, will work, but may not be efficient since the graph needs to be traversed again to reconstruct the path segments. A simpler approach is to emit paths along with distances in the mapper, so that each node will have its shortest path easily accessible at all times. The additional space requirements for shuffling these data from mappers to reducers are relatively modest, since for the most part paths (i.e., sequence of node ids) are relatively short.

Up until now, we have been assuming that all edges are unit distance. Let us relax that restriction and see what changes are required in the parallel breadth-first search algorithm. The adjacency lists, which were previously lists of node ids, must now encode the edge distances as well. In line

6 of the mapper code in Figure 5.4, instead of emitting d + 1 as the value, we must now emit d + w where w is the edge distance. No other changes to the algorithm are required, but the termination behavior is very different. To illustrate, consider the graph fragment in Figure 5.5, where s is the source node,

and in this iteration, we just “discovered” node r for the very first time. Assume for the sake of argument that we’ve already discovered the shortest distance to node p, and that the shortest distance to r so far goes through p. This, however, does not guarantee that we’ve

discovered the shortest distance to r, since there may exist a path going through q that we haven’t encountered yet (because it lies outside the search frontier).6 However, as the search frontier expands, we’ll eventually cover q and all other nodes along the path from p to q to r—which means that with sufficient iterations, we will discover the shortest distance to r. But how do we know that we’ve found the shortest distance to p? Well, if the shortest path to p lies within the search frontier, we would have already discovered it. And if it doesn’t, the above argument applies. Similarly, we can repeat the same argument for all nodes on the path from s to p. The conclusion is that, with sufficient iterations, we’ll eventually discover all the shortest distances.

So exactly how many iterations does “eventually” mean? In the worst case, we might need as many iterations as there are nodes in the graph minus one. In fact, it is not difficult to construct graphs that will elicit this worse-case behavior: Figure 5.6 provides an example, with n1 as the source. The parallel breadth-first search algorithm would not discover that the shortest path from n1 to n6 goes through n3, n4, and n5 until the fifth iteration. Three more iterations are necessary to cover the rest of the graph. Fortunately, for most real-world graphs, such extreme cases are rare, and the number of iterations necessary to discover all shortest distances is quite close to the diameter of the graph, as in the unit edge distance case.

In practical terms, how do we know when to stop iterating in the case of arbitrary edge distances? The algorithm can terminate when shortest distances at every node no longer change. Once again, we can use counters to keep track of such events. Every time we encounter a shorter distance in the reducer, we increment a counter. At the end of each MapReduce iteration, the driver program reads the counter value and determines if another iteration is necessary.

Compared to Dijkstra’s algorithm on a single processor, parallel breadth-first search in MapReduce can be characterized as a brute force approach that “wastes” a lot of time performing computations whose results are discarded. At each iteration, the algorithm attempts to recompute distances to all nodes, but in reality only useful work is done along the search frontier: inside the search frontier, the algorithm is simply repeating previous computations.[7] Outside the search frontier, the algorithm hasn’t discovered any paths to nodes there yet, so no meaningful work is done. Dijkstra’s algorithm, on the other hand, is far more efficient. Every time a node is explored, we’re guaranteed to have already found the shortest path to it. However, this is made possible by maintaining a global data structure (a priority queue) that holds nodes sorted by distance—this is not possible in MapReduce because the programming model does not provide support for global data that is mutable and accessible by the mappers and reducers. These inefficiencies represent the cost of parallelization.

The parallel breadth-first search algorithm is instructive in that it represents the prototypical structure of a large class of graph algorithms in MapReduce. They share in the following characteristics:

The graph structure is represented with adjacency lists, which is part of some larger node data structure that may contain additional information (variables to store intermediate output, features of the nodes). In many cases, features are attached to edges as well (e.g., edge weights).

The graph structure is represented with adjacency lists, which is part of some larger node data structure that may contain additional information (variables to store intermediate output, features of the nodes). In many cases, features are attached to edges as well (e.g., edge weights).

In addition to computations, the graph itself is also passed from the mapper to the reducer. In the reducer, the data structure corresponding to each node is updated and written back to disk.

THREE

Descriptive and Predictive Analytics

3.1 INTRODUCTION BUSINESS ANALYTICS

WHAT IS BUSINESS ANALYTICS?

Business analytics is the process of using quantitative methods to derive meaning from data in order to make informed business decisions.

There are three primary methods of business analysis:

Descriptive: The interpretation of historical data to identify trends and patterns

Predictive: The use of statistics to forecast future outcomes

Prescriptive: The application of testing and other techniques to determine which outcome will yield the best result in a given scenario

Deciding which method to employ is dependent on the business situation at hand.

To better understand how data insights can drive organizational performance, here's a look at some of the ways firms have benefitted from using business analytics.

3.2 BENEFITS OF BUSINESS ANALYTICS

- Enable data-driven decision making that has the potential to increase profits and improve efficiency
- With predictive analytics, allow businesses to plan for the future in ways that were previously impossible
- Helps a company make informed business decisions
- By modeling the outcomes and understanding the past, guesswork is minimized
- Present meaningful, clear data to support decision making and convince stakeholder

3.3 BUSINESS ANALYTICS EXAMPLES AND TOOLS

There are a host of business analytics tools that can perform these advanced data analytics functions automatically, requiring few of the special analytical skills or deep knowledge of programming languages necessary in data science.

These tools help businesses organize and make use of the massive amount of data that modern enterprise cloud applications produce. These applications may include supply chain management (SCM), enterprise resource planning (ERP) and customer relationship management (CRM) tools.

Below are some popular business analytics tools:

- Qlik, which has data visualization and automated data association features.

- Splunk, which is especially popular for small and medium-sized businesses because of its intuitive user interface and data visualization features.
- Sisense, which is known for its dynamic text analysis features and data warehousing
- KNIME, which is known for its high-performance data pipelining and machine learning
- Dundas BI, which is popular because of its automated trend forecasting and its user-friendly, drag-and-drop interface features.
- TIBCO Spotfire, which is considered one of the more advanced BA tools and offers powerful automated statistical and unstructured text analysis.
- Tableau Big Data Analytics, which is also highly regarded for its advanced unstructured text analysis and natural language processing (NLP) capabilities.

3.4 DESCRIPTIVE ANALYTICS

Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.

For learning analytics, this is a reflective analysis of learner data and is meant to provide insight into historical patterns of behaviors and performance in online learning environments.

For example, in an online learning course with a discussion board, descriptive analytics could determine how many students participated in the discussion, or how many times a particular student posted in the discussion foru How does descriptive analytics work?

Data aggregation and data mining are two techniques used in descriptive analytics to discover historical data. Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts.

Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning. Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.

What can descriptive analytics tell us?

The kind of information that descriptive analytics can provide depends on the learning analytic capability of the learning management system (LMS) being used and what the system is reporting on specifically.

Some common indicators that can be identified include learner engagement and learner performance. With learner engagement, analysts can detect the participation level of learners in the course and how and when course resources were accessed.

Performance data provides analysts with insight into how well learners succeeded on the course; this information could come from data taken from assessments or assignments. It’s important to note that insights learned from descriptive analysis are not used for making inferences or predictions about a learner’s future performance.

The analytical method is meant to provide strategic insight into where learners, or a specific learner, may have needed more support. It can also help course designers improve the design of learning by providing insight into what went well and what did not go well on the course.

Examples of descriptive analytics

Many LMS platforms and learning systems offer descriptive analytical reporting with the aim of help businesses and institutions measure learner performance to ensure that training goals and targets are met.

The findings from descriptive analytics can quickly identify areas that require improvement - whether that be improving learner engagement or the effectiveness of course delivery.

Here are some examples of how descriptive analytics is being used in the field of learning analytics:

- Tracking course enrollments, course compliance rates,
- Recording which learning resources are accessed and how often
- Summarizing the number of times a learner posts in a discussion board
- Tracking assignment and assessment grades
- Comparing pre-test and post-test assessments
- Analyzing course completion rates by learner or by course
- Collating course survey results

- Identifying length of time that learners took to complete a course

Advantages of descriptive analytics

When learners engage in online learning, they leave a digital trace behind with every interaction they have in the learning environment.

This means that descriptive analytics in online learning can gain insight into behaviors and performance indicators that would otherwise not be known.

Here are some advantages to utilizing this information:

- Quickly and easily report on the Return on Investment (ROI) by showing how performance achieved business or target goals.
- Identify gaps and performance issues early - before they become problems.
- Identify specific learners who require additional support, regardless of how many students or employees there are.
- Identify successful learners in order to offer positive feedback or additional resources.
- Analyze the value and impact of course design and learning resources.

3.5 DATA WAREHOUSE

How to Use Data Warehouses in Business Intelligence

Business intelligence, as we know it today, would not be possible without the data warehouse.

At its core, business intelligence is the ability to answer complex questions about your data and use those answers to make informed business decisions. In order to do this well, you need a data warehouse, which not only provides a safe way to centralize and store all your data but also a method to quickly find the answers you need, when you need them.

And that’s a pretty important role. By 2025, it’s estimated humanity will have produced a total of 175 zetta bytes of data. For context, that’s 175,000,000,000 terabytes.

Where does all of this information go? Well, most of it goes in the data warehouses.

Companies use data warehouses to manage transactions, understand their data, and keep it all organized. In short, data warehouses make large amounts of information more usable for organizations of all sizes and types.

This has made them a linchpin of data pipelines and business intelligence systems the world over. And understanding how data warehouses work can help you fulfill the full potential of business intelligence (it's not as complex as it may seem).

What Is a Data Warehouse?

A data warehouse is a data management system that stores large amounts of data for later use in processing and analysis. You can think of it as a large warehouse where trucks (i.e., source data) unload their data. That data is then sorted into rows and rows of well-organized shelves that make it easy to find exactly what you're looking for later.

The biggest innovation data warehouses introduced at their inception, according to DW 2.0: The Architecture for the Next Generation of Data Warehousing, was the ability to store "integrated granular historical data."

Breaking that down into human terms, this means data warehouses excel at storing data that's:

Integrated: They combine data from many databases and data sources.

Granular: The data they house is highly detailed and can be used in many different ways.

Historical: They can host a continuous record of data over years and years.

You can store this data in three different ways: on-premise data warehouses, cloud data warehouses, and hybrid data warehouses.

On-premise data warehouses run on physical servers that your company owns and manages. Cloud data warehouses are fully online, and you pay for space on servers that another company manages, like Amazon Redshift. Hybrid data warehouses are a mix of both on-premise and cloud, and companies making the transition to the cloud over a period of time use this option.

With all the data stored in one place, data warehouses use a specific approach to process data called online analytical processing (OLAP), which is specifically designed for complex queries.

One way to think about it is that when you go to your data warehouse to ask a question about the relationship between one set of data and another, OLAP is a way of organizing and moving among the rows and rows of shelves to quickly find that information.

This is great for business intelligence because the questions you ask about your data in order to make decisions are rarely simple. Because data warehouses use OLAP, they make finding answers to these complex questions very efficient. As a result, they've become a foundation for many successful business intelligence systems.

What Is the Role of Data Warehousing in Business Intelligence?

In business intelligence, data warehouses serve as the backbone of data storage. Business intelligence relies on complex queries and comparing multiple sets of data to inform everything from everyday decisions to organization-wide shifts in focus.

To facilitate this, business intelligence is comprised of three overarching activities: data wrangling, data storage, and data analysis. Data wrangling is usually facilitated by extract, transform, load (ETL) technologies, which we'll explain in detail below, and data analysis is done using business intelligence tools, like Chartio.

The glue holding this process together is data warehouses, which serve as the facilitator of data storage using OLAP. They integrate, summarize, and transform data, making it easier to analyze.

Even though data warehouses serve as the backbone of data storage, they're not the only technology involved in data storage. Many companies go through a data storage hierarchy before reaching the point where they absolutely need a data warehouse.

3.6 BUSINESS REPORTING

Business reports are valuable and essential tools for any enterprise regardless of size or industry. They provide a means to track and analyze the performance and overall health of the business while identifying areas for improvement and opportunities for growth.

Some business reporting is necessary as part of a regulatory requirement. For example, financial reports are a legal necessity for all businesses as determined by the Government of the country in which the business is based. Regular business reporting and monitoring are also necessary for many organizations to keep senior management, board members and other stakeholders advised on what is happening within the organization.

3.6.1 The purpose of business reporting

The aim of a business report is to provide critical analysis of how the business is tracking in all areas of the organization. Business reports are important tools to guide decision-making and to allow business owners and senior managers the opportunity to investigate and solve any identified issues.

Reporting is done through the process of compiling and reviewing the information within a specific functional area such as finance, sales, operations, inventory control or any area of the business where performance is monitored and measured.

Once information is gathered and reviewed, conclusions can then be drawn and recommendations made. The outcome of the report may explain why an issue has occurred or may identify performance problems and generally will recommend a course of action.

3.6.2 The importance of business reporting

Business reports provide useful insights for management such as information on spending, profits and growth. Reports will provide important detail that can be used to help develop future forecasts, marketing plans, guide budget planning and improve decision-making.

Managers also use business reports to track progress and growth, identify trends or any irregularities that may need further investigation. In addition to helping guide important decisions, business reports help to build an audit trail of business activities including reports that document annual budgets, sales, and meetings and planning initiatives.

Business reporting promotes transparency and for many public companies, an annual report is a legal requirement to provide shareholders, the government and others with financial data and ownership information about the business. Additionally, regular reporting throughout the business year enables businesses within the same sector to measure and compare their performances against others.

3.6.3 Business reports in action

Different reports will provide distinct value for all functional areas of an organization. Examples of some common reports include market analysis, trend analysis and financial analysis as well as operational and performance reports.

3.6.4 Inventory stock reports

Inventory stock records report on the movement of inventory into and out of the warehouse. They help a business identify any problems affecting performance such as product loss, obsolescence or dead stock.

3.6.5 Market analysis reports

They help business owners decide how to allocate their resources. For example, when an analysis of the market concludes that the ensuing business year will see accelerated growth, companies can increase their marketing budget to take advantage of this.

3.6.6 Trend analysis reports

These reports support long-term business development by examining statistical trends such as consumer preferences and the demographic groups that are experiencing the quickest growth rate. The objective of a trend analysis report is to identify growth opportunities to enable businesses to build market share ahead of competitors.

3.6.7 Financial reports

Financial reports are generally prepared on a regular basis by most companies and help to keep them on track toward achieving revenue and profit objectives. These reports highlight any variances in the financial results compared to forecasts in the annual business plan and will explain the reason for any significant negative variance.

3.6.8 Operational analysis reports

These reports show how efficiently a company is operating and will recommend ways to further improve productivity. An analysis of inventory control might indicate that the company experiences periodic shortages of key raw materials that prevent timely order fulfillment. The report may recommend that the company look for back-up suppliers of essential items to ensure availability when needed.

3.6.9 Performance reports

Monitoring performance trends help the company to set KPIs, benchmarks and business goals based on the most important aspects of the business. Performance reporting allows the business to compare performance over different timeframes and report objectives should always align with KPIs to demonstrate if these have been met or even exceeded.

3.6.10 Business reporting for business success

Business reports document the progress of your businesses and the data collected serves several important purposes. It guides strategic decision making, helping business leaders to formulate budget and planning activities for the ensuing year using the report data to back choices and provide justification for each decision.

Monitoring and reporting over time will not only highlight problems but can also identify opportunities for growth or expansion? Reports also work as a means of recording previous activities and help to define future growth opportunities by identifying already proven successes or what else could be done moving forward.

3.7 VISUAL ANALYTICS

Visual analytics is the use of sophisticated tools and processes to analyze datasets using visual representations of the data. Visualizing the data in graphs, charts, and maps helps users identify patterns and thereby develop actionable insights. These insights help organizations make better, data-driven decisions.

Sometimes confused with data visualization, visual analytics isn’t simply a matter of representing data graphically. Rather, modern, interactive visual analytics makes it easy to combine data from multiple sources and deeply analyze the data directly within the visualization itself. Further, AI and machine learning algorithms can offer recommendations to help guide the user’s exploration.

Visual Analytics Benefits

- The overall benefit of visual analytics is that it helps turn massive data sets into business insights which can have a major positive impact on your organization.

- Share findings and track progress: Interactive reports and dashboards help users track, organize and share key performance indicators across an organization.
- Make faster decisions: Users can understand data insights much more quickly by seeing and working with data sets when they are in a visual format.
- Explore data more easily: Self-service analytics tools which allow users to interact with data in a visual context allow them to discover hidden relationships and patterns in the data without relying on help from IT.
- Promote data literacy: Making data easier to work with and understand democratizes data analytics, getting more people across and organization involved.

3.8 BUSINESS PERFORMANCE MANAGEMENT

Business Performance Management (BPM) or Corporate Performance Management (CPM) is a term that is used to describe the methodologies, processes or workflows, metrics and systems that an organization uses to manage and optimize business performance. Business Performance Management also refers to robust software applications that enable the data collection, reporting and execution of the components of business performance management.

Business Performance Management is focused on business processes such as planning and forecasting. It helps organizations discover efficient use of their business units, financial, human and material resources. It involves consolidation of data from various sources, querying, and analysis of the data, and then putting the results into practice. Continuous and real-time reviews help to identify and eliminate problems before they grow. BPM’s forecasting abilities help the company take corrective action in time to meet earnings projections. Forecasting is characterized by a high degree of predictability, which is put into good use to answer what-if scenarios. BPM is useful in risk analysis and predicting outcomes of merger and acquisition scenarios, and coming up with a plan to overcome potential problems. Business Performance Management provides key performance indicators (KPI) that help companies monitor efficiency of projects and employees against operational targets.

Organizations have begun to make data more readily available and as such tools facilitate the work. Common tool categories of a business performance management system include:

- Online Analytical Processing (OLAP)
- Data warehouses
- Document warehouses
- Text mining
- Data mining
- BPM
- Scorecarding
- Dashboarding and data visualization
- Executive information systems
- Decision support systems
- Management information systems

These tools enable planning, budgeting, forecasting, performance analysis, analytics, dashboards, balanced scorecard analyses, and process management.

Business Performance Management software provides a range of functionality around business intelligence that benefits organizations. These systems promote greater visibility and effectively align employee goals and corporate strategies. BPM integrates the company’s processes with customer relationship management or Enterprise Resource Planning. Companies become able to gauge customer satisfaction, control customer trends and influence shareholder value.

3.9 PREDICTIVE ANALYTICS

Predictive Analytics is a statistical method that utilizes algorithms and machine learning to identify trends in data and predict future behaviors.

With increasing pressure to show a return on investment (ROI) for implementing learning analytics, it is no longer enough for a business to simply show how learners performed or how they interacted with learning content. It is now desirable to go beyond descriptive analytics and gain insight into whether training initiatives are working and how they can be improved.

Predictive Analytics can take both past and current data and offer predictions of what could happen in the future. This identification of possible risks or opportunities enables businesses to take actionable intervention in order to improve future learning initiatives.

How does Predictive Analytics work?

The software for predictive analytics has moved beyond the realm of statisticians and is becoming more affordable and accessible for different markets and industries, including the field of learning & development.

For online learning specifically, predictive analytics is often found incorporated in the Learning Management System (LMS), but can also be purchased separately as specialized software.

For the learner, predictive forecasting could be as simple as a dashboard located on the main screen after logging in to access a course. Analyzing data from past and current progress, visual indicators in the dashboard could be provided to signal whether the employee was on track with training requirements.

At the business level, an LMS system with predictive analytic capability can help improve decision-making by offering in-depth insight to strategic questions and concerns. This could range from anything to course enrolment, to course completion rates, to employee performance.

Predictive analytic models

Because predictive analytics goes beyond sorting and describing data, it relies heavily on complex models designed to make inferences about the data it encounters. These models utilize algorithms and machine learning to analyze past and present data in order to provide future trends.

Each model differs depending on the specific needs of those employing predictive analytics. Some common basic models that are utilized at a broad level include:

- **Decision trees** use branching to show possibilities stemming from each outcome or choice.
- **Regression techniques** assist with understanding relationships between variables.

- **Neural networks** utilize algorithms to figure out possible relationships within data sets.

What are the benefits of using predictive analytics?

Here are a few key benefits that businesses can expect to find when incorporating predictive analytics into their overall learning analytics strategy:

- Personalize the training needs of employees by identifying their gaps, strengths, and weaknesses; specific learning resources and training can be offered to support individual needs.
- Retain Talent by tracking and understanding employee career progression and forecasting what skills and learning resources would best benefit their career paths. Knowing what skills employees need also benefits the design of future training.
- Support employees who may be falling behind or not reaching their potential by offering intervention support before their performance puts them at risk.
- Simplified reporting and visuals that keep everyone updated when predictive forecasting is required.

3.10 WEB ANALYTICS

What Is Web Analytics

Web analytics is the measurement and analysis of data to inform an understanding of user behavior across web pages.

Analytics platforms measure activity and behavior on a website, for example: how many users visit, how long they stay, how many pages they visit, which pages they visit, and whether they arrive by following a link or not.

Businesses use web analytics platforms to measure and benchmark site performance and to look at key performance indicators that drive their business, such as purchase conversion rate.

Why Web Analytics Are Important

There’s an old business adage that whatever is worth doing is worth measuring.

Website analytics provide insights and data that can be used to create a better user experience for website visitors. Understanding customer behavior is also key to optimizing a website for key conversion metrics.

For example, web analytics will show you the most popular pages on your website, and the most popular paths to purchase.

With website analytics, you can also accurately track the effectiveness of your online marketing campaigns to help inform future efforts.

Web Analytics Examples

The most popular web analytics tool is Google Analytics, although there are many others on the market offering specialized information such as real-time activity or heat mapping.

The following are some of the most commonly used tools:

- Google Analytics - the ‘standard’ website analytics tool, free and widely used
- Piwik - an open-source solution similar in functionality to Google and a popular alternative, allowing companies full ownership and control of their data
- Adobe Analytics - highly customizable analytics platform (Adobe bought analytics leader Omniture in 2009)
- Kissmetrics - can zero in on individual behavior, i.e. cohort analysis, conversion and retention at the segment or individual level
- Mixpanel - advanced mobile and web analytics that measure actions rather than pageviews
- Parse.ly - offers detailed real-time analytics, specifically for publishers
- CrazyEgg - measures which parts of the page are getting the most attention using ‘heat mapping’

With a wide variety of analytics tools on the market, the right vendors for your company’s needs will depend on your specific requirements. Luckily, Optimizely integrates with most of the leading platforms to simplify your data analysis.

3.11 SOCIAL ANALYTICS

In simple words, the term "social analytics" refers to the act of gathering and then directly interpreting social media related data so that you can:

1. Gain a better view of the current market conditions and do proper market research
2. Understand where the target consumer stands and gain their unique insights
3. Learn more about the preferences of your ideal audience
4. Collect various types of feedback that can be used to arrive at better, more refined business decisions

According to Gohfar F. Khan, author of "Seven Layers of Social Media" book, social analytics is defined as…

"The art and science of extracting valuable hidden insights from vast amounts of semistructured and unstructured social media data to……enable informed and insightful decision making." Importance of Social Analytics

The use of social analytics is critical in today's times for brands that want to make a mark.

Not listening to social media conversations can be costly because it can lead to expensive mistakes

. But due to the overwhelming nature of a large amount of available data, that's exactly what's happening with many businesses leveraging social media.

By using the right tools to not only capture, but also interpret conversations in the social media arena, brands can make real sense of their data.

They can actually extract valuable insights by focusing on the signals, and avoiding the noise surrounding the social media world.

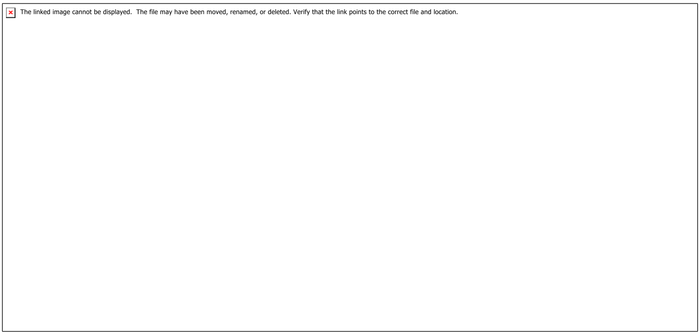

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 3.1 Social Analytics

To put it broadly, social analytics not only helps businesses (of any size or type) gather the right kind of data.

But they also make it easy for businesses to measure what level of impact their marketing efforts are having on their followers/fans.

Which is necessary at all times.

Social Analytics: 6 Key Metrics Businesses Can Keep Track of

Businesses can use social analytics in a number of ways (some of which we’ll be discussing in this post).

But how they are used differs from business to business. Because what your brand wants to track and measure may not be the same as mine.

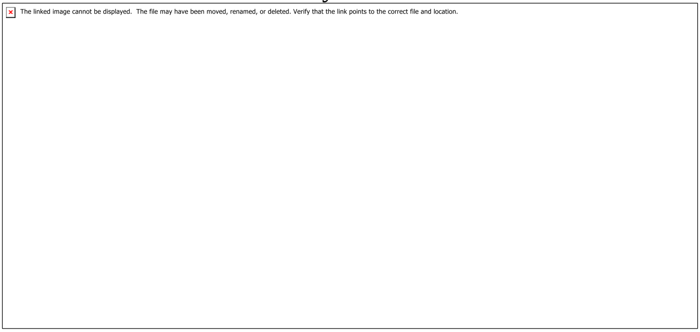

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 3.2 Social Media Metrics

Let’s quickly go through the key metrics that you get access to when you use social analytics:

1. Engagement: It tells you about the level of engagement your social media post receives in terms of likes, shares, etc.
2. Mentions: It gives you information about how often your brand name or……a particular word/phrase has been mentioned on various social media platforms.
3. Visual Mentions: It lets you know the amount of times your brand logo appears in visuals posted on a particular social network.
4. Sentiment: It sheds light on the kind of sentiment social media users have towards your industry, brand and even your competitors.
5. Virality: It basically shows how "viral" your social media post is or how quickly it is being shared across a network.
6. Share of Voice: It separates and tells you the percentage of mentions related to your bran(within your industry) compared to your competitor's percentage of mentions.

FOUR

Prescriptive Analytics

4.1 Prescriptive Analytics

Prescriptive analytics is a process that analyzes data and provides instant recommendations on how to optimize business practices to suit multiple predicted outcomes. In essence, prescriptive analytics takes the "what we know" (data), comprehensively understands that data to predict what could happen, and suggests the best steps forward based on informed simulations.

Benefits of prescriptive analytics:

Prescriptive analytics affords organizations the ability to:

- Effortlessly map the path to success. Prescriptive analytic models are designed to pull together data and operations to produce the roadmap that tells you what to do and how to do it right the first time. Artificial intelligence takes the reins of business intelligence to apply simulated actions to a scenario to produce the steps necessary to avoid failure or achieve success.
- Inform real-time and long-term business operations. Decision makers can view both real-time and forecasted data simultaneously to make decisions

that support sustained growth and success. This streamlines decision making by offering specific recommendations.

- Spend less time thinking and more time doing. The instant turnaround of data analysis and outcome prediction lets your team spend less time finding problems and more time designing the perfect solutions. Artificial intelligence can curate and process data better than your team of data engineers and in a fraction of the time.
- Reduce human error or bias. Through more advanced algorithms and machine learning processes, predictive analytics provides an even more comprehensive and accurate form of data aggregation and analysis than descriptive analytics, predictive analytics, or even individuals.the objective of this part is to simply illustrate what is possible and how it has been implemented in some real settings.

4.2 Model-Based Decision Making: Optimization and Multi-Criteria Systems, Modeling and Analysis:

OPENING VIGNETTE: Midwest ISO Saves Billions by Better Planning of Power Plant Operations and Capacity Planning:

INTRODUCTION

Midwest ISO (MISO) operates in 13 U.S. states as well as the province of Manitoba in Canada. It manages 35 transmission owners and 100 non-transmission owners, ensuring that all members of the organization have equal access to high-voltage power lines.

Together, the United States and the province of Manitoba constitute one of the largest energy markets in the world, with yearly energy transactions amounting to about

$23 billion. Before Midwest ISO existed, each transmission company operated independently. Now, after a company joins MISO, it still maintains control of its power plants and transmission lines, and shares in the responsibility of supplying and buying energy in a wholesale electricity market to meet demand. MISO, however, has the responsibility of deciding when and how much energy to produce and administer to the market in such a way as to increase benefit to society.

PRESENTATION OF PROBLEM

Individually, the companies had to make extra investments to manage risk. Their mode of operation resulted in inefficient use of transmission lines. Deregulation policies were introduced by Congress and were implemented by the Federal Energy Regulatory Commission (FERC) for the wholesale electricity industry. When MISO was formed, it first started an energy-only market in 2005 that ensured unbiased access to transmission lines. In 2009, it added ancillary services (regulation and contingency reserves) to its operations. Regulation was supposed to ensure that the frequency did not deviate from 60 hertz. Contingency reserves were supposed to help ensure that in the event of unexpected power loss, demand was met within 10 minutes of the power loss. Operations research methods were considered as means to provide the level of performance demanded by the ancillary services.

METHODOLOGY/SOLUTION

Sequentially, two optimization algorithms were used. These were the commitment algorithm and the dispatch algorithm. The commitment algorithm committed power plants to be either on or off. The dispatch algorithm determined the level of a power plant’s output and price. With these two algorithms, facilities were given constraints on how much electricity to cany within their physical limits in order to avoid overload and damage to expensive equipment. The commitment problem for the energy-only market made use of the Lagrangian relaxation method. As mentioned earlier, it determined when each plant should turn on or off. The dispatch problem was solved with a linear programming model. It helped decide how much output should be produced by each power plant. It also helped determine the price of energy based on the location of the power plant. Even though these methods were just fine , they were not appropriate for the ancillary service market commitment problem. Rather, a mixed integer programming model was used as a result of its superior modeling capacity.

RESULTS/BENEFITS

Based on the improvements made , reliability of the transmission grid improved. Also, a dynamic transparent pricing structure was created . Value proposition n studies show that Midwest ISO achieved about $2.1 billion and $3 billion dollars in net cumulative savings between 2007 and 2010. Future savings are expected to accrue to about $6.1 billion.

QUESTIONS FOR THE OPENING VIGNETTE

In what ways were the individual companies in Midwest ISO better off being part of MISO as opposed to operating independently?

The dispatch problem was solved with a linear programming method. Explain the need of such method in light of the problem discussed in the case.

What were the two main optimization algorithms used? Briefly explain the use of each algorithm.

LESSONS WE CAN LEARN FROM THIS VIGNETTE

Operations research (OR) methods were used by Midwest ISO to provide efficient and cheaper sources of energy for states in the midwestern region of the United States. A combination of linear programming and the Lagrangian relaxation methods was used to determine an optimized approach to generate and supply power.

By extension, this methodology could be used by both government agencies and the private sector to optimize the cost and provision of services such as healthcare and education .

4.3 DECISION SUPPORT SYSTEMS MODELING:

Many readily accessible applications describe how the models incorporated in DSS contribute to organizational success. These include Pillowtex, Fiat, Procter and Gamble and other forms.

Simulation models can enhance an organization’s decision-making process and enable it to see the impact of its future choices. Fiat (Pro Model, 2006) saves $1 million annually in manufacturing costs through simulation. IBM has predicted the behavior of the 230-mile-long Guadalupe River and its many tributaries. The prediction can be made several days before the imminent flood of the river. This is important as it would allow for enough time for disaster management and preparation. IBM used a combination of weather and sensor data to build a river system simulation application that could simulate thousands of river branches at a time. Besides flood prediction, the application could also be used for irrigation planning in such a way as to avoid the impact of droughts and surplus water. Even companies under financial stress need to invest in such solutions to squeeze more efficiency out of their limited resources-maybe even more so. Pillowtex, a $2 billion company that manufactures pillows, mattress pads, and comforters, had filed for bankruptcy and needed to reorganize its plants to maximize net profits from the company's operations. It employed a simulation model to develop a new lean manufacturing environment that would reduce the costs and increase throughput.

The company estimated that the use of this model resulted in over $12 million savings immediately.

Modeling is a key element in most DSS and a necessity in a model-based DSS. There are many classes of models, and there are often many specialized techniques for solving each one. Simulation is a common modeling approach, but there are several others. Applying models to real-world situations can save millions of dollars or generate millions of dollars in revenue. Christiansen Tal. describe the applications of such models in shipping company operations. They describe applications of Turbo Router, a DSS for ship routing and scheduling. They claim that over the course of just a 3-week period, a company used this model to better utilize its fleet, generating additional profit of $1-2 million in just a short time.

CASE:1 Optimal Transport for ExxonMobil Downstream Through a DSS:

ExxonMobil, a petroleum and natural gas company, operates in several countries worldwide. It provides several ranges of petroleum products including clean fuels, lubricants, and high-value products and feedstock to several customers. This is completed through a complex supply chain between its refineries and customers. One of the main products ExxonMobil transports is vacuum gas oil (VGO). ExxonMobil transports several shiploads of vacuum gas oil from Europe to the United States. In a year, it is estimated that ExxonMobil transports about 60- 70 ships of VGO across the Atlantic Ocean. Hitherto, both ExxonMobil-managed vessels and third-party vessels were scheduled to transport VGO across the Atlantic through a cumbersome manual process. The whole process required the collaboration of several individuals across the supply chain organization. Several customized spreadsheets with special constraints, requirements, and economic trade-offs were used to determirle the transportation schedule of the vessels. Some of the constraints included:

- Constantly varying production and demand projections
- Maximum and minimum inventory constraints
- A pool of heterogeneous vessels (e.g., ships with varying speed, cargo size)
- Vessels that load and discharge at multiple ports.
- Both ExxonMobil-managed and third-party supplies and ports.
- Complex transportation cost that includes variable overage and demurrage costs.
- Vessel size and draft limits for different ports.

The manual process could not determine the actual routes of vessels, the timing of each vessel, and the quantity of VGO loaded and discharged. Additionally, consideration of the production and consumption data at several locations rendered the manual process burdensome and inefficient. Methodology/Solution A decision support tool that suppo1ted schedulers in planning an optimal schedule for ships to load, transport, and discharge VGO to and from multiple locations was developed. The problem was formulated as a mixed-integer linear programming problem. The solution had to satisfy requirements for routing, transportation, scheduling, and inventory management vis-a-vis varying production and demand profiles. A mathematical programming language, GAMS, was used for the problem formulation and Microsoft Excel was used as the user interface . When the solver (ILOG CPLEX) is run, an optimal solution is reached at a point when the objective value of the incumbent solution stops improving. This stopping criterion is determined by the user during each program run.

Results/Benefits:

It was expected that using the optimization model will lead to reduced shipping cost and less demurrage expenses. These would be achieved because the tool would be able to support higher utilization of ships and help make ship selection and design more optimal routing schedules. The researchers expected to extend the research by exploring other alternate mathematical methods to solve the scheduling problem. They also intended to give the DSS tool the capability to consider multiple products for a pool of vessels.

Current Modeling Issues: major modeling issues, such as problem identification and environmental analysis, variable identification, forecasting, the use of multiple models, model categories (or appropriate selection), model management, and knowledge-based modeling.

4.3.1 IDENTIFICATION OF THE PROBLEM AND ENVIRONMENTAL ANALYSIS:

One very important aspect of it is environmental scanning and analysis, which is the monitoring, scanning, and interpretation of collected information . No decision is made in a vacuum. It is important to analyze the scope of the domain and the forces and dynamics of the environment. A decision maker needs to identify the organizational culture and the corporate decision-making processes (e.g., who makes decisions, degree of centralization). It is entirely possible that environmental factors have created the current problem. Bl/business analytics (BA) tools can help identify problems by scanning for them. The problem must be understood and everyone involved should share the same frame of understanding, because the problem will ultimately be represented by the model in one form or another. Otherwise, the model will not help the decision maker.

VARIABLE IDENTIFICATION: Identification of a model's variables (e.g. decision, result, uncontrollable) is critical, as are the relationships among the variables. Influence diagrams, which are graphical models of mathematical models, can facilitate the identification process. A more general form of an influence diagram, a cognitive map , can help a decision maker develop a better understanding of a problem, especially of variables and their interactions.

FORECASTING (PREDICTIVE ANALYTICS): Forecasting is predicting the future. This form of predictive analytics is essential for construction and manipulating models, because when a decision is implemented the results usually occur in the future . Whereas DSS are typically designed to determine what will be, traditional MIS report what is or what was. There is no point in running a what-if (sensitivity) analysis on the past, because decisions made then have no impact on the future . Forecasting is getting easier as software vendors automate many of the complications of developing such models.

E-commerce has created an immense need for forecasting and an abundance of available information for performing it. E-commerce activities occur quickly, yet information about purchases is gathered and should be analyzed to produce forecasts. Part of the analysis involves simply predicting demand; however, forecasting models can use product life-cycle needs and information about the marketplace and consumers to analyze the entire situation, ideally leading to additional sales of products and services. Many organizations have accurately predicted demand for products and services, using a variety of qualitative and quantitative methods. But until recently, most companies viewed their customers and potential customers by categorizing them into only a few, time-tested groupings. Today, it is critical not only to consider customer characteristics, but also to consider how to get the right product(s) to the right customers at the right price at the right time in the right format/ packaging. The more accurately a firm does this, the more profitable the firm is. In addition, a firm needs to recognize when not to sell a pa1ticular product or bundle of products to a particular set of customers. Part of this effort involves identifying lifelong customer profitability. These customer relationship management (CRM) system and revenue management system (RMS) approaches rely heavily on forecasting techniques, which are typically described as predictive analytics. These systems attempt to predict who their best (i.e., most profitable) customers (and worst ones as well) are and focus on identifying products and services at appropriate prices to appeal to them.

4.4 STRUCTURE OF MATHEMATICAL MODELS FOR DECISION SUPPORT:

These include the components and the structure of models.

The Components of Decision Support Mathematical Models All quantitative models are typically made up of four basic components 4 result (or outcome) variables, decision variables, uncontrollable variables (and/ or parameters), and intermediate result variables. Mathematical relationships link these components together. In non-quantitative models, the relationships are symbolic or qualitative. The results of decisions are determined based on the decision made (i.e. , the values of the decision variables), the factors that cannot be controlled by the decision maker and the relationships among the variables. The modeling process involves identifying the variables and relationships among them. Solving a model determines the values of these and the result variable(s).

4.4.1 RESULT (OUTCOME) VARIABLES Result (outcome) variables reflect the level of effectiveness of a system; that is, they indicate how well the system performs or attains its goal(s). These variables are outputs. Examples of Result variables are considered dependent variables. Intermediate result variables are sometimes used in modeling to identify intermediate outcomes. In the case of a dependent variable, another event must occur first before the event described by the variable can occur. Result variables depend on the occurrence of the decision variables and the uncontrollable variables.

4.4.2 DECISION VARIABLES Decision variables describe alternative courses of action. The decision maker controls the decision variables. For example, for an investment problem, the amount to invest in bonds is a decision variable. In a scheduling problem, the decision variables are people, times, and schedules.

4.4.3 UNCONTROLLABLE VARIABLES, OR PARAMETERS In any decision-making situation, there are factors that affect the result variables but are not under the control of the decision maker. Either these factors can be fixed, in which case they are called uncontrollable variables, or parameters, or they can vary, in which case they are called variables. Examples of factors are the prime interest rate, a city’s building code, tax regulations, and utilities costs. Most of these factors are uncontrollable because they are in and determined by elements of the system environment in which the decision maker works. Some of these variables limit the decision maker and therefore form what are called the constraints of the problem.

4.4.4 INTERMEDIATE RESULT VARIABLES Intermediate result variables reflect intermediate outcomes in mathematical models. For example, in determining machine scheduling, spoilage is an intermediate result variable, and total profit is the result variable (i.e., spoilage is one determinant of total profit). Another example is employee salaries. This constitutes a decision variable for management: It determines employee satisfaction (i.e. , intermediate outcome), which, in turn, determines the productivity level (i.e., final result).

4.4.5 **The Structure of Mathematical Models**

The components of a quantitative model are linked together by mathematical (algebraic) expressions- equations or inequalities. A very simple financial model is P=R-C where P = profit, R = revenue, and C = cost. This equation describes the relationship among the variables. Another well-known financial model is the simple present-value cash flow model, where P = present value, F = a future single payment in dollars, i = interest rate (percentage), and n = number of years. With this model, it is possible to determine the present value of a payment of $100,000 to be made 5 years from today, at a 10 percent (0.1) interest rate, as follows: P = 100,000/(l + 0.1)5 = 62,092

4.4.6 CERTAINTY, UNCERTAINTY, AND RISK1:

Decision situations are often classified on the basis of what the decision maker knows (or believes) about the forecasted results. We customarily classify this knowledge into three categories ranging from complete knowledge to complete ignorance:

- Certainty
- Risk
- Uncertainty.

Decision Making Under Certainty In decision making under certainty, it is assumed that complete knowledge is available so that the decision maker knows exactly what the outcome of each course of action will be (as in a deterministic environment). It may not be true that the outcomes are 100 percent known, nor is it necessary to really evaluate all the outcomes, but often this assumption simplifies the model and makes it tractable. The decision maker is viewed as a perfect predictor of the future because it is assumed that there is only one outcome for each alternative. For example, the alternative of investing in U.S. Treasury bills is one for which there is complete availability of info1mation about the future return on the investment if it is held to maturity. A situation involving decision making under certainty occurs most often with structured problems with short time horizons (up to 1 year). Certainty models are relatively easy to develop and solve, and they can yield optimal solutions. Many financial models are constructed under assumed certainty, even though the market is anything but 100 percent certain.

Decision Making Under Uncertainty In decision making under uncertainty, the decision maker considers situations in which several outcomes are possible for each course of action. In contrast to the risk situation, in this case, the decision maker does not know, or cannot estimate, the probability of occurrence of the possible outcomes. Decision making under uncertainty is more difficult than decision making under certainty because there is insufficient information. Modeling of such situations involves assessment of the decision maker's (or the organization's) attitude toward risk. Managers attempt to avoid uncertainty as much as possible , even to the point of assuming it away. Instead of dealing with uncertainty, they attempt to obtain more information so that the problem can be treated under certainty (because it can be "almost" certain) or under calculated (i.e ., assumed) risk. If more information is not available, the problem must be treated under a condition of uncertainty, which is less definitive than the other categories.

Decision Making Under Risk (Risk Analysis) A decision made under risk2 (also known as a probabilistic or stochastic decision making situation) is one in which the decision maker must consider several possible outcomes for each alternative, each with a given probability of occurrence.

4.5 Modeling and Analysis: Heuristic Search Methods and Simulation:

The concepts and motivating applications of these advanced techniques are described in which is organized into the following sections:

Opening Vignette: System Dynamics Allows Fluor Corporation to Better Plan for Project and Change Management

1. 1.Problem-Solving Search Methods
2. Genetic Algorithms and Developing GA Applications
3. Simulation Pan .
4. Visu al Interactive Simulation
5. System Dynamics Modeling
6. Agents-Based Mode ling .

OPENING VIGNETTE: System Dynamics Allows Fluor Corporation to Better Plan for Project and Change Management

INTRODUCTION

Fluor is an engineering and construction company with over 36,000 employers spread over several countries worldwide . The company’s net income in 2009 amounted to about $680 million based on total revenue of $22 billion. As part of its operations, Fluor manages varying sizes of projects that are subject to scope changes, design changes, and schedule changes.

PRESENTATION OF PROBLEM

Fluor estimated that changes accounted for about 20 to 30 percent of revenue . Most changes were due to secondary impacts like ripple effects, disruptions, and productivity loss. Previously, the changes were collated and reported at a later period and the burden of cost allocated to the stakeholder responsible. In certain instances when late surprises about cost and project schedule are attributed to clients, it causes friction between clients and Fluor, which eventually affect future business dealings. Sometimes, cost impacts occur in such a time and fashion when it is difficult to take preventive measures. The company determined that to improve on its efficiency, reduce legal ramification with clients, and keep them happy it had to review its method of handling changes to projects. One challenge the company faced was the fact that changes stayed extremely remote from the situation , which warranted the change. In such a case, it is difficult to determine the cause of a change , and it affects subsequent measures to handle related change issues.

METHODOLOGY/SOLUTION

For sure, Fluor knew that one way of combating the issue was to foresee and avoid the events that might lead to changes. However, that alone would not be enough to solve the problem. The company needed to understand the dynamics of the different situations that could warrant changes to project plans. Systems dynamics was used as a base method in a three-part analytical solution for understanding the dynamics between different factors that could cause changes to be made. System dynamics is a methodology and simulation-modeling technique for analyzing complex systems using principles of cause and effect, feedback loops, and time-delayed and nonlinear effects. Building tools for rapidly tailoring a solution to different situations form the next part of the three-part analytical solution. In this part, industry standards and company references are embedded. The project plan is also embedded as an input. The model is then converged to simulate the correct amounts and timing of other factors like staffing, project progress, productivity, and effects on productivity. The last p art of the analytical solution was to deploy the project models to non modelers. Basically, the system takes inputs that are specific to a particular project being worked and its environment, such as the labor market. Some other input parameters, transformed into numerical data, are related to progress curves, expe nses, and labor laws and constraints.

RESULTS/BENEFITS

With this system, customers are able to perform "what-if‛ analysis even before a project is started so the project performance can be gauged. Through diagnostics, the system also helps explain why certain effects are realized based on impact to the project plan. Since its development, Fluor has recorded over 100 extensive uses of their system dynamics model and project simulation system. As an example, the model was used to analyze and save

$10 million in the future impact of changes to a mining project. Also, based on the what-if capability of Fluor’s model, a company saved $10 million when the project team used the model to redesign the process of reviewing changes so that the speed of the company’s definition and approval procedures was increase

QUESTIONS FOR THE OPENING VIGNETTE

- Explain the use of system dynamics as a simulation tool for solving complex problems.
- In what ways was it applied in Fluor Corporation to solve complex problems?
- How does a what-if analysis help a decision maker to save on cost?
- In your own words, explain the factors that might have triggered the use of system dynamics to solve change management problems in Fluor Corporation.
- Pick a geographic region and business domain and list some corresponding relevant factors that would be used as inputs in building such a system

WHAT WE CAN LEARN FROM THIS VIGNETTE

Changes to project plans and timelines are a major contributing factor to upward increase in cost from initial amount budgeted for projects. In this case, Fluor relied on system dynamics to understand what, why, when, and how changes occurred to project plans. The models that the system dynamics model produced helped them correctly quantify the cost of projects even before they started. The vignette demonstrates that system dynamics is still a credible and robust methodology in understanding business processes and creating "what-if" analyses of the impact of both expected and unexpected changes in project plans.

4.6 PROBLEM-SOLVING SEARCH METHODS

We next turn to several well-known search methods used in the choice phase of problem solving. These include analytical techniques, algorithms, blind searching, and heuristic searching. The choice phase of problem solving involves a search for an appropriate course of action (among those identified during the design phase) that can solve the problem. Several major search approaches are possible , depending on the criteria (or criterion) of choice and the type of modeling approach used. These search approaches are shown in Figure 10.1. For normative models, such as mathematical programming-based ones, either an analytical approach is used or a complete, exhaustive enumeration (comparing the outcomes of all the alternatives) is applied. For descriptive models, a comparison of a limited number of alternatives is used, either blindly or by employing heuristics. Usually the results guide the decision maker's search

4.6.1 Analytical Techniques:

Analytical techniques use mathematical formulas to derive an optimal solution directly or to predict a certain result. Analytical techniques are used mainly for solving structured problems, usually of a tactical or operational nature, in areas such as resource allocation or inventory management. Blind or heuristic search approaches generally are employed to solve more complex problems.

4.6.2 Algorithms

Analytical techniques may use algorithms to increase the efficiency of the search. An algorithm is a step-by-step search process for obtaining an optimal solution. Solutions are generated and tested for possible improvements. An improvement is made whenever possible, and the new solution is subjected to an improvement test, based on the principle of choice The process continues until no further improvement is possible. Most mathematical programming problems are solved by using efficient algorithms. Web search engines use various algorithms to speed up searches and produce accurate results.

4.6.3 Blind Searching

Blind search techniques are arbitrary search approaches that are not guided. There are two types of blind searches: a complete enumeration, for which all the alternatives are considered and therefore an optimal solution is discovered; and an incomplete, or partial, search, which continues until a good-enough solution is found. The latter is a form of suboptimization.

4.6.4 Heuristic Searching

Heuristics are the informal, judgmental knowledge of an application area that constitute the rules of good judgment in the field. Through domain knowledge, they guide the problem-solving process. Heuristic programming is the process of using heuristics in problem solving. This is done via heuristic search methods, which often operate as algorithms but limit the solutions examined either by limiting the search space or stopping the method early. Usually, rules that have either demonstrated their success in practice or are theoretically solid are applied in heuristic searching.

Application Case .1

Chilean Government Uses Heuristics to Make Decisions on School Lunch Providers The Junta Nacional de Auxilio Escolar y Because (JUNAEB), an agency of the Chilean government, promotes integration and retention of socially vulnerable children in the country’s school system. JUNAEB’s school meal program provides meals for approximately 10,000 schools. Decisions on meal providers are made through an annual tender using a combinatorial auction, where food industry firms bid on supply contracts, based on a series of disjoint, compact geographical areas called territorial units (TUs). These territorial units consist of districts spanning the country. When the Chilean economy suffered a downturn, many competing meal service providers ceased their operations. Thus, the number of suppliers participating in the combinatorial auction was reduced.

4.7 GENETIC ALGORITHMS AND DEVELOPING GA APPLICATIONS

Genetic algorithms (also known as evolutionary algorithms) demonstrate self-organization and adaptation in much the same way that biological organisms do by following the chief rule of evolution, survival of the fittest. The method improves the solutions by producing offspring (i.e., a new collection of feasible solutions) using the best solutions of the current generation eas "parents." The generation of offspring is achieved by a process modeled after biological reproduction whereby mutation and crossover operators are used to manipulate genes in constructing newer and "better" chromosomes. Notice that a simple analogy between genes and decision variables and between chromosomes and potential solutions underlies the genetic algorithm terminology

Example: The Vector Game To illustrate how genetic algorithms work, we describe the classical Vector game This game is similar to Master Mind. As your opponent gives you clues about how good your guess is (i.e., the outcome of the fitness function), you create a new solution, using the knowledge gained from the recently proposed solutions and their quality.

Description of The Vector Game Vector is played against an opponent who secretly writes down a string of six digits (in a genetic algorithm, this string consists of a chromosome). Each digit is a decision variable that can take the value of either O or 1. For example, say that the secret number that you are to figure out is 001010. You must try to guess this number as quickly as possible (with the least number of trials). You present a sequence of digits (a guess) to your opponent, and he or she tells you how many of the digits (but not which ones) you guessed are correct (i.e., the fitness function or quality of your guess). For example, the guess 110101 has no correct digits (i.e

., the score = 0). The guess 111101 has only one correct digit (the third one, and hence the score = 1).

Default Strategy: Random Trial and Error There are 64 possible six-digit strings of bina1y numbers. If you pick numbers at random, you will need, on average, 32 guesses to obtain the right answer. Can you do it faster? Yes, if you can interpret the feedback provided to you by your opponent (a measure of the goodness or fitness of your guess). This is how a genetic algorithm works

Improved Strategy: Use of Genetic Algorithms The following are the steps in solving the Vector game with genetic algorithms:

1. Present to your opponent four strings, selected at random. (Select four arbitrarily. Through experimentation, you may find that five or six would be better.) Assume that you have selected these four110100; score = 1 (i.e., one digit guessed correctly) (B) 111101; score= 1(C) 011011; score = 4(D) 101100; score = 3
2. Because none of the strings is entirely correct, continue.
3. Delete (A) and (B) because of their low scores. Call (C) and (D) parents.
4. "Mate" the parents by splitting each number as shown here between the second and third digits (the position of the split is randomly selected): (C) 01:1011 (D) 10:1100 Now combine the first two digits of with the last four of (this is called crossover). The result is , the first offspring: (E) 011100; score = 3 Similarly, combine the first two digits of (D) with the last four of (C).The result is (F), the second offspring: (F) 101011; score= 4 It looks as though the offspring are not doing much better than the parents.
5. Now copy the original (C) and (D).
6. Mate and crossover the new parents, but use a different split. Now you have two new offspring, (G) and (H): (C) 0110:11

(D) 1011:00

(G) 0110:00; score = 4

(H) 1011:11; score = 3 Next, repeat step 2: Select the best "couple" from all the previous solutions to reproduce. You have several options, such as (G) and (C). Select (G) and (F). Now duplicate and crossover. Here are the results:

(F) 1:01011

0:11000 (I) 111000; score = 3 (J) 001011; score= 5 You can also generate more offspring: (F) 101:011 (G) 011:000 (K) 101000; score = 4 (L) 011011; score = 4 Now repeat the processes with

and (K) as parents, and duplicate the crossover: (J) 00101:l

10100:0 (M) 001010; score = 6 That's it! You have reached the solution after 13 guesses. Not bad compared to the expected average of 32 for a random-guess strategy.

4.7.1 Terminology of Genetic Algorithms

A genetic algorithm is an iterative procedure that represents its candidate solutions as strings of genes called chromosomes and measures their viability with a fitness function. The fitness function is a measure of the objective to be obtained (i.e., maximum or minimum). As in biological systems, candidate solutions combine to produce offspring in each algorithmic iteration, called a generation.

Reproduction. Through reproduction, genetic algorithms produce new generations of potentially improved solutions by selecting parents with higher fitness ratings or by giving such parents a greater probability of being selected to contribute to the reproduction process.

Crossover. Many genetic algorithms use a string of binary symbols (each corresponding to a decision variable) to represent chromosomes (potential solutions), as was the case in the Vector game described earlier. Crossover means choosing a random position in the string (e.g. , after the first two digits) and exchanging the segments either to the right or the left of that point with those of another string's segments (generated using the same splitting schema) to produce two new offspring.

Mutation. This genetic operator was not shown in the Vector game example. Mutation is an arbitrary (and minimal) change in the representation of a chromosome. It is often used to prevent the algorithm from getting stuck in a local optimum. The procedure randomly selects a chromosome (giving more probability to the ones with better fitness value) and randomly identifies a gene in the chromosome and inverses its value (from O to 1 or from 1 to 0), thus generating one new chromosome for the next generation. The occurrence of mutation is usually set to a very low probability (0.1 percent).

Elitism. An important aspect in genetic algorithms is to preserve a few of the best solutions to evolve through the generations. That way, you are guaranteed to end up with the best possible solution for the current application of the algorithm. In practice, a few of the best solutions are migrated to the next generation.

4.7.2 **Limitations of Genetic Algorithms** According to Gruppe and Jooste (2004), the following are among the most important limitations of genetic algorithms:

- Not all problems can be framed in the mathematical manner that genetic algorithms demand.
- Development of a genetic algorithm and interpretation of the results require an expert who has both the programming and statistical/mathematical skills demanded by the genetic algorithm technology in use.
- It is known that in a few situations the "genes" from a few comparatively highly fit (but not optimal) individuals may come to dominate the population, causing it to converge on a local maximum. When the population has converged, the ability of the genetic algorithm to continue to search for better solutions is effectively eliminated.
- Most genetic algorithms rely on random-number generators that produce different results each time the model runs. Although there is likely to be a high degree of consistency among the runs, they may vary
- Locating good variables that work for a pa1ticular problem is difficult. Obtaining the data to populate the variables is equally demanding.
- Selecting methods by which to evolve the system requires thought and evaluation. If the range of possible solutions is small, a genetic algorithm will converge too quickly on a solution. When evolution proceeds too quickly, thereby altering good solutions too quickly, the results may miss the optimum solution.

4.7.3 Genetic Algorithm Applications

Genetic algorithms are a type of machine learning for representing and solving complex problems. They provide a set of efficient, domain-independent search heuristics for a broad spectrum of applications, including the following:

- Dynamic process control
- Induction of optimization of rules
- Discove1y of new connectivity topologies

- Simulation of biological models of behavior and evolution
- Complex design of engineering structures
- Pattern recognition
- Scheduling
- Transportation and routing

SIMULATION

Simulation is the appearance of reality. In MSS, simulation is a technique for conducting experiments (e.g., what-if analyses) with a computer on a model of a management system. Typically, real decision-making situations involve some randomness. Because DSS deals with semi structured or unstructured situations, reality is complex, which may not be easily represented by optimization or other models but can often be handled by simulation. Simulation is one of the most commonly used DSS methods.

Application Case 2:

American Airlines Uses Should-Cost Modeling to Assess the Uncertainty of Bids for Shipment Routes Introduction American Airlines, Inc. (AA) is one of the world's largest airlines. Its core business is passenger transportation but it has other vital ancillary functions that include full-truckload (FTL) freight shipment of maintenance equipment and in-flight shipment of passenger service items that could add up to over $1 billion in inventory at any given time. AA receives numerous bids from suppliers in response to request for quotes (RFQs) for inventories. AA's RFQs could total over 500 in any given year. Bid quotes va1y significantly as a result of the large number of bids and resultant complex bidding process. Sometimes, a single contract bid could deviate by about 200 percent. As a result of the complex process, it is common to either overpay or underpay suppliers for their services. To this end, AA wanted a should-cost model that would streamline and assess bid quotes from suppliers in order to choose bid quotes that were fair to both them and their suppliers.

Methodology/Solution :In order to determine fair cost for supplier products and services, three steps were taken:

Primary (e.g., interviews) and secondary (e.g., Internet) sources were scouted for base-case and range data that would inform cost variables that affect an FTL bid.

Cost variables were chosen so that they were mutually exclusive and collectively exhaustive.

The DPL decision analysis software was used to model the uncertainty

Furthermore, Extended Swanson-Megill (ESM) approximation was used to model the probability distribution of the most sensitive cost variables used. This was done in order to account for the high variability in the bids in the initial model.

Results/Benefits :

A pilot test was done on an RFQ that attracted bids from six FTL carriers. Out of the six bids presented, five were within three standard deviations from the mean while one was considered an outlier. Subsequently,

A used the should-cost FTL model on more than 20 RFQs to determine what a fair and accurate cost of goods and services should be.

It is expected that this model will help in reducing the risk of either overpaying or underpaying its suppliers.

FIVE

Business Analytics: Emerging Trends and Future Impacts

5.1 OPENING VIGNETTE:

Oklahoma Gas and Electric Employs Analytics to Promote Smart Energy Use

Oklahoma Gas and Electric (OG&E) serves over 789,000 customers in Oklahoma and Arkansas. OG&E has a strategic goal to delay building new fossil fuel generation plants until the year 2020. OG&E forecasts a daily system demand of 5,864 megawatts in 2020, a reduction of about 500 megawatts.

One of the ways to optimize this demand is to engage the consumers in managing their energy usage. OG&E has completed installation of smart meters and other devices on the electronic grid at the consumer end that enable it to capture large amounts of data. For example, currently it receives about 52 million meter reads per day. Apart from this, OG&E expects to receive close to 2 million event messages per day from its advanced metering infrastructure, data networks, meter alarms, and outage management systems. OG&E employs a three- layer information architecture involving data warehouse, improved and expanded integration and data management, and new analytics and presentation capabilities to support the Big Data flow.

With this data, OG&E has started working on consumer-oriented efficiency programs to shift the customer’s usage out of peak demand cycles. OG&E is targeting customers with its smart hours plan. This plan encourages customers to choose a variety of rate options sent via phone, text, or e-mail. These rate options offer attractive summer rates for all other hours apart from the peak hours of 2 P.M. to 7 P.M. OG&E is making an investment in customers by supplying a communicating thermostat that will respond to the price signals sent by OG&E and help customers in managing their utility consumption. OG&E also educates its customers on their usage habits by providing 5-minute interval data every 15 minutes to the demand-responsive customers.

QUESTIONS FOR THE OPENING VIGNETTE

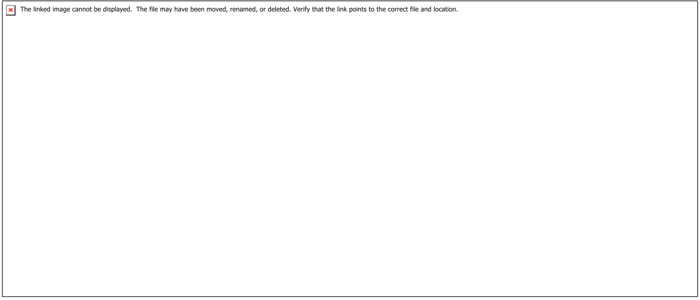
- Why perform consumer analytics?
- What is meant by dynamic segmentation?
- How does geospatial mapping help OG&E?
- What types of incentives might the consumers respond to in changing their energy use?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Many organizations are now integrating the data from the different internal units and turning toward analytics to convert the integrated data into value. The ability to view the operations/customer-specific data using in-database geospatial analytics gives organizations a broader perspective and aids in decision making.

5.2 LOCATION-BASED ANALYTICS FOR ORGANIZATIONS

The potential of new technologies when innovative uses are developed by creative minds. Most of the technologies described in this chapter are nascent and have yet to see widespread adoption. There in lies the opportunity to create the next "killer" application. For example, use of RFID and sensors is growing, with each company exploring its use in supply chains, retail stores, manufacturing, or service operations. The chapter argues that with the right combination of ideas, networking, and applications, it is possible to develop creative technologies that have the potential to impact a company’s operations in multiple ways, or to create entirely new markets and make a major difference to the world. We also study the analytics ecosystem to better understand which companies are the players in this industry.



The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 5.1 Classification of Location-Based Analytics Applications

The traditional location-based analytic techniques using geocoding of organizational locations and consumers hampers the organizations in understanding "true location-based" impacts. Locations based on postal codes offer an aggregate view of a large geographic area. This poor granularity may not be able to pinpoint the growth opportunities within a region. The location of the target customers can change rapidly. An organization’s promotional campaigns might not target the right customers. To address these concerns, organizations are embracing location and spatial extensions to analytics (Gnau, 2010). Addition of location components based on latitudinal and longitudinal attributes to the traditional analytical techniques enables organizations to add a new dimension of "where" to their traditional business analyses, which currently answer questions of "who," "what," "when," and "how much."

Location-based data are now readily available from geographic information systems (GIS). These are used to capture, store, analyze, and manage the data linked to a location using integrated sensor technologies, global positioning systems installed in smartphones, or through radio-frequency identification deployments in retail and healthcare industries.

By integrating information about the location with other critical business data, organizations are now creating location intelligence (LI) (Krivda, 2010). LI is enabling organizations to gain critical insights and make better decisions by optimizing important processes and applications. Organizations now create interactive maps that further drill down to details about any location, offering analysts the ability to investigate new trends and correlate location-specific factors across multiple KPis. Analysts in the organizations can now pinpoint trends and patterns in revenues, sales, and profitability across geographical areas.

5.3 ANALYTICS APPLICATIONS FOR CONSUMERS

The explosive growth of the apps industry for smartphone platforms (iOS, Android, Windows, Blackberry, Amazon, and so forth) and the use of analytics are also creating tremendous opportunities for developing apps that the consumers can use directly. These apps differ from the previous category in that these are meant for direct use by a consumer rather than an organization that is trying to mine a consumer's usage/purchase data to create a profile for marketing specific products or services to them. Predictably, these apps are meant for enabling consumers to do their job better.

Application case: Sense Networks has built a mobile application called CabSense that analyzes large amounts of data from the New York City Taxi and Limousine Commission and helps New Yorkers and visitors in finding the best corners for hailing a taxi based on the person's location, day of the week, and time. CabSense rates the street corners on a 5-point scale by making use of machine-learning algorithms applied to the vast amounts of historical location points obtained from the pickups and drop-offs of all New York City cabs.

Although the app does not give the exact location of cabs in real time, its data-crunching predictions enable people to get to a street corner that has the highest probability of finding a cab. CabSense provides an interactive map based on current user location obtained from the mobile phone's GPS locator to find the best street corners for finding an open cab. It also provides a radar view that automatically points the right direction toward the best street corner. The application also allows users to plan in advance, set up date and time of travel, and view the best corners for finding a taxi. Furthermore, CabSense distinguishes

New York's Yellow Cab services from the for-hire vehicles and readily prompts the users with relevant details of private service providers that can be used in case no Yellow Cabs are available.

5.4 RECOMMENDATION ENGINES

In most decision situations, people rely on recommendations gathered either directly from other people or indirectly through the aggregated recommendations made by others in the form of reviews and ratings posted either in newspapers, product guides, or online. Such information sharing is considered one of the major reasons for the success of online retailers such as Amazon.com. In this section we briefly review the common terms and technologies of such systems as these are becoming key components of any analytic application.

The term recommender systems refers to a Web-based information filtering system that takes the inputs from users and then aggregates the inputs to provide recommendations for other users in their product or service selection choices. Some recommender systems now even try to predict the rating or preference that a user would give for a particular product or service.

The data necessary to build a recommendation system are collected by Web-based systems where each user is specifically asked to rate an item on a rating scale, rank the items from most favorite to least favorite, and/or ask the user to list the attributes of the items that the user likes. Other information such as the user's textual comments, feedback reviews, amount of time that the user spends on viewing an item, and tracking the details of the user's social networking activity provides behavioral information about the product choices made by the user.

Two basic approaches that are employed in the development of recommendation systems are:

1. **Collaborative filtering**
2. **Content filtering**

In collaborative filtering, the recommendation system is built based on the individual user's past behavior by keeping track of the previous history of all purchased items. This includes products, items that are viewed most often, and ratings that are given by the users to the items they purchased. These individual profile histories with item preferences are grouped with other similar user-item profile histories to build a comprehensive set of relations between users and items, which are then used to predict what the user will like and recommend items accordingly.

5.4.1 Collaborative filtering involves aggregating the user-item profiles. It is usually done by building a user-item ratings matrix where each row represents a unique user and each column gives the individual item rating made by the user. The resultant matrix is a dynamic, sparse matrix with a huge dimensionality; it gets updated every time the existing user purchases a new item or a new user makes item purchases. Then the recommendation task is to predict what rating a user would give to a previously unranked item. The predictions that result in higher item rankings are then presented as recommendations to the users. The user-item based approach employs techniques like matrix factorization and low-rank matrix approximation to reduce the dimensionality of the sparse matrix in generating the recommendations.

Collaborative filtering can also take a user-based approach in which the users take the main role. Similar users sharing the same preferences are combined into a group, and recommendations of items to a particular user are based on the evaluation of items by other users in the same group. If a particular item is ranked high by the entire community, then it is recommended to the user. Another collaborative filtering approach is based on the item-set similarity, which groups items based on the user ratings provided by various users. Both of these collaborative filtering approaches employ many algorithms, such as KNN CK-Nearest Neighborhood) and

Pearson Correlation, in measuring user and behavior similarity of ratings among the items.

5.4.2 Content-based recommender systems overcome one of the disadvantages of collaborative filtering recommender systems, which completely rely on the user ratings matrix, by considering specifications and characteristics of items. In the content-based filtering approach, the characteristics of an item are profiled first and then content- based individual user profiles are built to store the information about the characteristics of specific items that the user has rated in the past. In the recommendation process, a comparison is made by filtering the item information from the user profile for which the user has rated positively and compares these characteristics with any new products that the user has not rated yet. Recommendations are made if there are similarities found in the item characteristics.

Content-based filtering involves using information tags or keywords in fetching detailed information about item characteristics and restricts this process to a single user, unlike collaborative filtering, which looks for similarities between various user profiles. This approach makes use of machine-learning and classification techniques like Bayesian classifiers, cluster analysis, decision trees, and artificial neural networks in order to estimate the probability of recommending similar items to the users that match the user's existing ratings for an item.

5.5 WEB 2.0 AND ONLINE SOCIAL NETWORKING

Web 2.0 is the popular term for describing advanced Web technologies and applications, including blogs, wikis, RSS, mashups, user-generated content, and social networks. A major objective of Web 2.0 is to enhance creativity, information sharing, and collaboration.

Representative Characteristics of Web 2.0

- Web 2.0 has the ability to tap into the collective intelligence of users. The more users contribute, the more popular and valuable a Web 2.0 site becomes.
- Data is made available in new or never-intended ways. Web 2.0 data can be remixed or "mashed up," often through Web service interfaces, much the way a dance-club DJ mixes music.
- Web 2.0 relies on user-generated and user-controlled content and data.
- Lightweight programming techniques and tools let nearly anyone act as a Web site developer.
- The virtual elimination of software-upgrade cycles makes everything a perpetual beta or work-in-progress and allows rapid prototyping, using the Web as an application development platform.

- Users can access applications entirely through a browser.
- An architecture of participation and digital democracy encourages users to add value to the application as they use it.
- A major emphasis is on social networks and computing.
- There is strong support for information sharing and collaboration.
- Web 2.0 fosters rapid and continuous creation of new business models.

Other important features of Web 2.0 are its dynamic content, rich user experience, metadata, scalability, open source basis, and freedom (net neutrality).

Most Web 2.0 applications have a rich, interactive, user-friendly interface based on Ajax or a similar framework. Ajax (Asynchronous JavaScript and XML) is an effective and efficient Web development technique for creating interactive Web applications. The intent is to make Web pages feel more responsive by exchanging small amounts of data with the server behind the scenes so that the entire Web page does not have to be reloaded each time the user makes a change. This is meant to increase the Web page's interactivity, loading speed, and usability.

5.6 SOCIAL NETWORKING

A Definition and Basic Information

A social network is a place where people create their own space, or homepage, on which they write biogs (Web logs); post pictures, videos, or music; share ideas; and link to other Web locations they find interesting. In addition, members of social networks can tag the content they create and post it with keywords they choose themselves, which makes the content searchable. The mass adoption of social networking Web sites points to an evolution in human social interaction.

Mobile social networing refers to social networking where members converse and connect with one another using cell phones or other mobile devices. Virtually all major social networking sites offer mobile services or apps on smartphones to access their services. The explosion of mobile Web 2.0 services and companies means that many social networks can be based from cell phones and other portable devices, extending the reach of such networks to the millions of people who lack regular or easy access to computers.

Facebook (facebook.com), which was launched in 2004 by former Harvard student Mark Zuckerberg, is the largest social network service in the world, with almost 1 billion users worldwide as of February 2013. A primary reason why Facebook has expanded so rapidly is the network effect**-** more users means more value. As more users become involved in the social space, more people are available to connect with. Initially, Facebook was an online social space for college and high school students that automatically connected students to other students at the same school. Expanding to a global audience has enabled Facebook to become the dominant social network.

Today, Facebook has a number of applications that support photos, groups, events, marketplaces, posted items, games, and notes. A special feature on Facebook is the News Feed, which enables users to track the activities of friends in their social circles. For example, when a user changes his or her profile, the updates are broadcast to others who subscribe to the feed. Users can also develop their own applications or use any of the millions of Facebook applications that have been developed by other users.

Implications of Business and Enterprise Social Networks

Although advertising and sales are the major EC activities in public social networks, there are emerging possibilities for commercial activities in business-oriented networks such as Linkedln and in enterprise social networks.

USING TWITTER TO GET A PULSE OF THE MARKET

Twitter is a popular social networking site that enables friends to keep in touch and follow what others are saying. An analysis of "tweets" can be used to determine how well a product/service is doing in the market. Previous chapters on Web analytics included a significant coverage of social media analytics. This continues to grow in popularity and business use. Analysis of posts on social media sites such as Facebook and Twitter has become a major business. Many companies provide services to monitor and manage such posts on behalf of companies and individuals. One good example is reputation.com.

5.7 CLOUD COMPUTING AND Bl

Another emerging technology trend that business intelligence users should be aware of is cloud computing. Users need not have knowledge of, experience in, or control over the technology infrastructures in the cloud that supports them." This definition is broad and comprehensive. In some ways, cloud computing is a new name for many previous, related trends: utility computing, application service provider, grid computing, on-demand computing, software as a service (SaaS), and even older, centralized computing with dumb terminals. But the term cloud computing originates from a reference to the Internet as a "cloud" and represents an evolution of all of the previously shared/centralized computing trends. The Wikipedia entry also recognizes that cloud computing is a combination of several information technology components as services.

For example, infrastructure as a service (IaaS) refers to providing computing platforms as a service (PaaS), as well as all of the basic platform provisioning, such as management administration, security, and so on. It also includes Saas, which includes applications to be delivered through a Web browser while the data and the application programs are on some other server.

5.7.1 CLOUD COMPUTING SERVICE CATEGORIES

The cloud services are typically based on the end-user or business requirements. Following are the primary cloud services-

5.7.1 Software as a Service (SaaS)

SaaS is often referred to as a software delivery method for providing access to software and its associated functions remotely as a web-based service. Rather than paying for an upfront fee for purchasing the licensed software, SaaS customers pay a recurring fee amount for subscribing to their services. Additionally, they can access the SaaS from any Internet-connected device at any point in time.

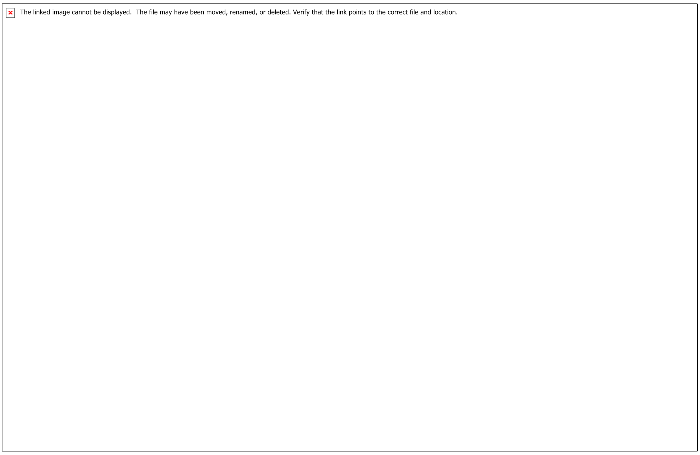

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 5.2 Conceptual Architecture of Service-Oriented DSS

5.7.2 Platform as a Service (PaaS)

In the case of PaaS, the platform is outsourced instead of a company or data centre purchasing and managing its own hardware and software layers. Most of the PaaSes have been designed for developers, and they aim to simplify the process of creating and deploying the software. A web-developer could use PaaS for operating system software, Web server software, etc.

5.7.3 Infrastructure as a Service (IaaS)

The infrastructure-related to computers could include servers, storage, and networking that is delivered as service. IaaS has been popular with enterprises that appreciate the convenience of having a cloud vendor for managing their IT infrastructure. They can even see cost-savings as a result of paying only for the computing resources that are used.

Security in Cloud Computing

Security has remained a primary concern for businesses that contemplate cloud adoption, especially the adoption of the public cloud. The public cloud providers share the underlying hardware infrastructure between a large number of customers, as the public cloud is a multi-tenant environment. This environment asks for copious isolation between the logically-computed resources. Also, access to public cloud storage and compute resources are authenticated only by the account login credentials. Most organizations are bound by the complex obligations, and the governance standards are hesitant in placing data or workloads in the public cloud due to the fear of outrages, loss or theft. With time, this fear has receded as logical isolation has proven to be reliable, and the addition of data encryption along with several identity and access management tools has helped in improving the security of data present within the public cloud.

Use Case of Hybrid Cloud and Business Intelligence Case Discussed: Reducing Data in the Cloud

Today, many companies are working with big data; there are disparate types of data that might be high in volume and velocity. In the public cloud, there can be some examples, including the geospatial data or clickstream data, or even social media data. Companies that deal with big data have decided to processes their external sources in the public cloud and then bring the reduced or analyzed data set-on premises to make it a part of a larger analysis. This proves to be a good use of the hybrid cloud model and the companies carry out the computing work in the public cloud when it doesn't need the source files for anything else. Cloud enables scalability and flexibility of operating on these data sets.

5.7.4 What is Cloud Business Intelligence?

The Cloud Business Intelligence Applications are hosted over a virtual network such as the internet. These are used for providing organizations access to BI-related data such as dashboards, KPIs and other business analytics. The enterprises these days have increasingly turned to cloud-based tools such as Customer Relationship Management (CRM) Software, online file collaboration & storage, as well as the help-desk software. These trends are inclusive of business intelligence tools that embrace the agility and accessibility of the cloud.

Cloud computing and business intelligence give a perfect match that industry verticals need. Business intelligence involves the delivery of the right information to the right set of people at the right time, and cloud computing serves as a lightweight catalyst for accessing the Business Intelligence Applications. The Cloud Business Intelligence Applications can be accessed on multiple devices as well as web browsers.

Benefits of Cloud Business Intelligence

Cloud Business Intelligence solutions are gradually gaining popularity amongst the businesses, as most of the businesses have realized the benefits of data analytics. Businesses are in need of quality insights that are driven by accurate data. The SaaS providers serve the primary interfacing to the business user community, and the concept of Cloud Business Intelligence refers to the delivery of BI capabilities as a service.

Following are key benefits that Cloud Computing for Business Intelligence has-

- **Cost Efficiency** In the case of cloud, the companies don't need a budget for large and an up-front purchase regarding the software packages. The companies treat BI infrastructure as a service and pay only for the computing resources that are needed, and costly asset acquisition is avoided, thus, maintaining a reduced entry threshold barrier.
- **Flexibility and Scalability** The Cloud Business Intelligence solutions have been allowing greater flexibility that can be altered quickly, giving the technical users to access newer data sources by carrying out experimentation with the analytics model. Also, the Cloud resources can automatically and rapidly carry out the scaling in and out by supporting a large number of simultaneous users. Thus, customers can easily increase software usage without any delay in deploying the installation of extra hardware and software.
- **Reliability** Reliability can be improved through the usage of multiple redundant sites, providing reliability and secure the data storage locations. These resources can be spread across several users, making the Cloud Computing an apt option for disaster recovery and business continuity.
- **Enhanced Data Sharing** Cloud computing applications tend to allow data access to be shared remotely and enable an easy cross-location sharing of data capabilities. These are usually deployed by the internet, and it remains outside the company's firewall.
- **No Expenditure of Capita** A low TCO or total cost of ownership is one of the key benefits of the Cloud model. With the help of the cloud, companies only pay for a service that is actually used. Under this policy, Cloud

Computing allows the companies to have better control over the CAPEX and the OPEX that is associated with the non-core activities.

5.8 IMPACTS OF ANALYTICS IN ORGANIZATIONS: AN OVERVIEW

Analytic systems are important factors in the information, Web, and knowledge revolution. This is a cultural transformation with which most people are only now coming to terms. Unlike the slower revolutions of the past, such as the Industrial Revolution, this revolution is taking place very quickly and affecting every facet of our lives. Inherent in this rapid transformation are a host of managerial, economic, and social issues.

Separating the impact of analytics from that of other computerized systems is a difficult task, especially because of the trend toward integrating, or even embedding, analytics with other computer-based information systems. Analytics can have both micro and macro implications. Such systems can affect particular individuals and jobs, and they can also affect the work structures of departments and units within an organization. They can also have significant long-term effects on total organizational structures, entire industries, communities, and society as a whole (i.e., a macro impact). The impact of computers and analytics can be divided into three general categories: organizational, individual, and societal. In each of these, computers have had many impacts.

5.8.1 New Organizational Units

One change in organizational structure is the possibility of creating an analytics department, a BI department, or a knowledge management department in which analytics play a major role. This special unit can be combined with or replace a quantitative analysis unit, or it can be a completely new entity. Some large corporations have separate decision support units or departments. For example, many major banks have such departments in their financial services divisions. Many companies have small decision support or BI/data warehouse units. These types of departments are usually involved in training in addition to consulting and application development activities. Others have empowered a chief technology officer over BI, intelligent systems, and e-commerce applications. Companies such as Target and Walmart have major investments in such units, which are constantly analyzing their data to determine the efficiency of marketing and supply chain management by understanding their customer and supplier interactions.

5.8.2 Restructuring Business Processes and Virtual Teams

In many cases, it is necessary to restructure business processes before introducing new information technologies. For example, before IBM introduced e-procurement, it restructured all related business processes, including decision making, searching inventories, reordering, and shipping. When a company introduces a data warehouse and BI, the information flows and related business processes (e.g., order fulfillment) are likely to change. Such changes are often necessary for profitability, or even survival. Restructuring is especially necessary when major IT projects such as ERP or BI are undertaken. Sometimes an organization-wide, major restructuring is needed; then it is referred to as reengineering. Reengineering involves changes in structure, organizational culture, and processes. In a case in which an entire (or most of an) organization is involved, the process is referred to as business process reengineering (BPR).

5.8.3 The Impacts of ADS Systems

ADS systems, such as those for pricing, scheduling, and inventory management, are spreading rapidly, especially in industries such as airlines, retailing, transportation, and banking. These systems will probably have the following impacts:

Reduction of middle management • Empowerment of customers and business partners • Improved customer service (e.g., faster reply to requests) • Increased productivity of help desks and call centers

The impact goes beyond one company or one supply chain, however. Entire industries are affected. The use of profitability models and optimization are reshaping retailing, real estate, banking, transportation, airlines, and car rental agencies, among other industries.

5.8.4 Job Satisfaction

Although many jobs may be substantially enriched by analytics, other jobs may become more routine and less satisfying. For example, more than 40 years ago, Argyris (1971) predicted that computer-based information systems would reduce managerial discretion in decision making and lead to managers being dissatisfied. In their study about ADS, Davenport and Harris (2005) found that employees using ADS systems, especially those who are empowered by the systems, were more satisfied with their jobs. If the routine and mundane work can be done using an analytic system, then it should free up the managers and knowledge workers to do more challenging tasks.

5.8.5 Job Stress and Anxiety

An increase in workload and/or responsibilities can trigger job stress. Although computerization has benefited organizations by increasing productivity, it has also created an ever-increasing and changing workload on some employees- many times brought on by downsizing and redistributing entire workloads of one employee to another. Some workers feel overwhelmed and begin to feel anxious about their jobs and their performance. These feelings of anxiety can adversely affect their productivity. Management must alleviate these feelings by redistributing the workload among workers or conducting appropriate training.

One of the negative impacts of the information age is information anxiety. This disquiet can take several forms, such as frustration with the inability to keep up with the amount of data present in our lives. Constant connectivity afforded through mobile devices, e-mail, and instant messaging creates its own challenges and stress. Research on e-mail response strategies (iris.okstate.edu/REMS) includes many examples of studies conducted to recognize such stress. Constant alerts about incoming e-mails lead to interruptions, which eventually result in loss of productivity (and then an increase in stress). Systems have been developed to provide decision support to determine how often a person should check his or her e-mail (see Gupta and Sharda, 2009).

5.8.6 Analytics‘ Impact on Managers’ Activities and Their Performance

The most important task of managers is making decisions. Analytics can change the manner in which many decisions are made and can consequently change managers‘ jobs.

The following are some potential impacts of analytics on managers’ jobs:

- Less expertise (experience) is required for making many decisions.

- Faster decision making is possible because of the availability of information and the automation of some phases in the decision-making process.
- Less reliance on experts and analysts is required to provide support to top executives; managers can do it by themselves with the help of intelligent systems. 616 Pan V
- Big Data and Future Directions for Business Analytics
- Power is being redistributed among managers. (The more information and analysis capability they possess, the more power they have.)
- Support for complex decisions makes them faster to make and be of better quality.
- Information needed for high-level decision making is expedited or even self-generated.
- Automation of routine decisions or phases in the decision-making process (e.g., for frontline decision making and using ADS) may eliminate some managers.

5.9 THE ANALYTICS ECOSYSTEM

So, you are excited about the potential of analytics, and want to JOtn this growing industry. Who are the current players, and what do they do? Where might you fit in? The objective of this section is to identify various sectors of the analytics industry, provide a classification of different types of industry participants, and illustrate the types of opportunities that exist for analytics professionals. The section (indeed the book) concludes with some observations about the opportunities for professionals to move across these clusters.

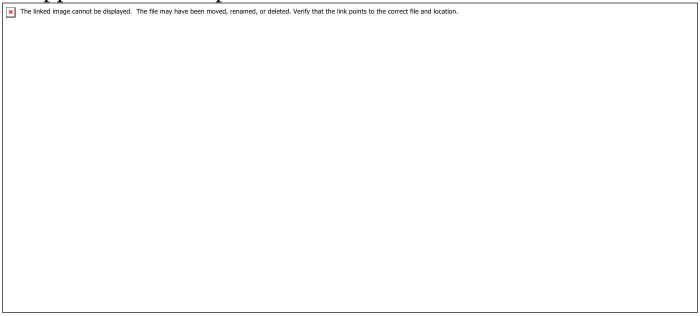

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

Fig 5.3 Analytics Ecosystem

Analytics Industry Clusters

This section is aimed at identifying various analytics industry players by grouping them into sectors. We note that the list of company names included is not exhaustive. These merely reflect our own awareness and mapping of companies' offerings in this space. Additionally, the mention of a company's name or its capability in one specific group does not mean that is the only activity/ offering of that organization. We use these names simply to illustrate our descriptions of sectors. Many other organizations exist in this industry. Our goal is not to create a directory of players or their capabilities in each space, but to illustrate to the students that many different options exist for playing in the analytics industry.

Data Infrastructure Providers

This group includes all of the major players in the data hardware and software industry. These organizations provide hardware and software targeted at providing the basic foundation for all data management solutions. Obvious examples of these would include all major hardware players that provide the infrastructure for database computing-IBM, Dell, HP, Oracle, and so forth. We would also include storage solution providers such as EMC and NetApp in this sector. Many companies provide both hardware and software platforms of their own (e.g., IBM, Oracle, and Teradata).

Data Warehouse Industry

We distinguish between this group and the preceding group mainly due to differences in their focus. Companies with data warehousing capabilities focus on providing integrated data from multiple sources so an organization can derive and deliver value from its data assets. Many companies in this space include their own hardware to provide efficient data storage, retrieval, and processing. Recent developments in this space include performing analytics on the data directly in memory. Companies such as IBM, Oracle, and Teradata are major players in this arena.

Middleware Industry

Data warehousing began with the focus on bringing all the data stores into an enterprisewide platform. By making sense of this data, it becomes an industry in itself. The general goal of this industry is to provide easy-to-use tools for reporting and analytics. Examples of companies in this space include MicroStrategy, Plum, and many others.

Data Aggregators/Distributors

Several companies realized the opportunity to develop specialized data collection, aggregation, and distribution mechanisms. These companies typically focus on a specific industry sector and build upon their existing relationships. For example, Nielsen provides data sources to their clients on retail purchase behavior. Another example is Experian, which includes data on each household in the United States.

Analytics-Focused Software Developers

Companies in this category have developed analytics software for general use with data that has been collected in a data warehouse or is available through one of the platforms identified earlier (including Big Data). It can also include inventors and researchers in universities and other organizations.

Application Developers or System Integrators: Industry Specific or General

The organizations in this group focus on using solutions available from the data infrastructure, data warehouse, middleware, data aggregators, and analytics software providers to develop custom solutions for a specific industry. They also use their analytics expertise to develop specific applications for a user. Thus, this industry group makes it possible for the analytics technology to be truly useful. Of course, such groups may also exist in specific user organizations.

Companies that have traditionally provided application/data solutions to specific sectors have recognized the potential for the use of analytics and are developing industry-specific analytics offerings. For example, Cerner provides electronic medical records (EMR) solutions to medical providers. Their offerings now include many analytics reports and visualizations.

Analytics User Organizations

Clearly, this is the economic engine of the whole analytics industry. If there were no users, there would be no analytics industry. Organizations in every other industry, size, shape, and location are using analytics or exploring use of analytics in their operations. These include the private sector, government, education, and the military. It includes organizations around the world. Examples of uses of analytics in

different industries abound. Others are exploring similar opportunities to try and gain/retain a competitive advantage. We will not identify specific companies in this section. Rather, the goal here is to see what types of roles analytics professionals can play within a user organization.

Analytics Industry Analysts and Influencers

The next cluster includes three types of organizations or professionals. The first group is the set of professional organizations that provides advice to analytics industry providers and users. Their services include marketing analyses, coverage of new developments, evaluation of specific technologies, and development of training/white papers, and so forth. Examples of such players include organizations such as the Gartner Group, The Data Warehousing Institute, and many of the general and technical publications and Web sites that cover the analytics industry. The second group includes professional societies or organizations that also provide some of the same services but are membership based and organized.

Academic Providers and Certification Agencies

In any knowledge-intensive industry such as analytics, the fundamental strength comes from having students who are interested in the technology and choose that industry as their profession. Universities play a key role in making this possible. This cluster, then, represents the academic programs that prepare professionals for the industry. It includes various components of business schools such as information systems, marketing, and management sciences. It also extends far beyond business schools to include computer science, statistics, mathematics, and industrial engineering departments across the world.