Big Data (CS-3032)

Kalinga Institute of Industrial Technology
Deemed to be University
Bhubaneswar-751024

School of Computer Engineering

3 Credit

Lecture Note

Motivating Quotes

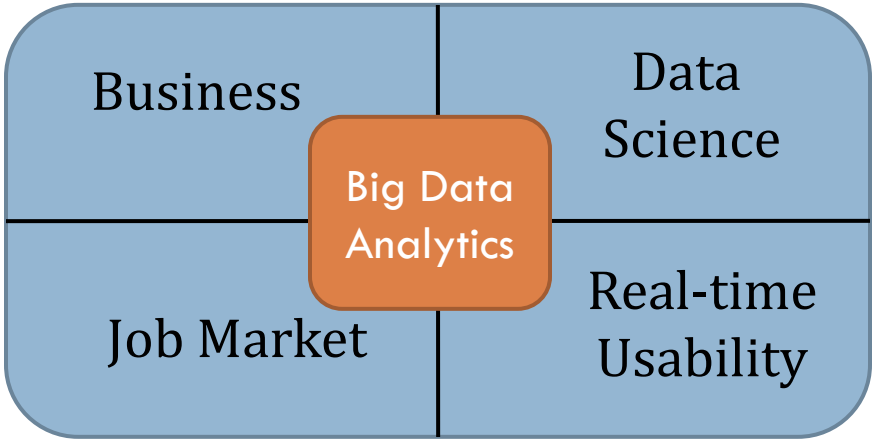
- "The world is one big data platform." - Andrew McAfee, co-director of the MIT Initiative on the Digital Economy, and the associate director of the Center for Digital Business at the MIT Sloan School of Management.
- "Errors using inadequate data are much less than those using no data at all." - Charles Babbage, inventor and mathematician.
- "The most valuable commodity I know of is information." - Gordon Gekko, fictional character in the 1987 film Wall Street and its 2010 sequel Wall Street: Money Never Sleeps, played by Michael Douglas.
- "Big data will replace the need for 80% of all doctors" - Vinod Khosla, Indian-born American engineer and businessman.
- "Thanks to big data, machines can now be programmed to the next thing right. But only humans can do the next right thing." - Dov Seidman, American author, attorney, columnist and businessman

Motivating Quotes cont'd

- "With data collection, 'the sooner the better' is always the best answer." - Marissa Mayer, former president and CEO of Yahoo!
- "Data is a precious thing and will last longer than the systems themselves." - Tim Berners-Lee, inventor of the World Wide Web.
- "Numbers have an important story to tell. They rely on you to give them a voice." - Stephen Few, Information Technology innovator, teacher, and consultant.
- "When we have all data online it will be great for humanity. It is a prerequisite to solving many problems that humankind faces" - Vinod Khosla, Indian-born American engineer and businessman.
- "Thanks to big data, machines can now be programmed to the next thing right. But only humans can do the next right thing." - Robert Cailliau, Belgian informatics engineer and computer scientist who, together with Tim Berners-Lee, developed the World Wide Web.

Importance of the Course

- The Big Data is indeed a revolution in the field of Information Technology.
- The use of big data by the companies is enhancing every year and the primary focus of the companies is on customers. the field is flourishing specifically in Business to Consumer (B2C) applications.
- Many organizations are actively looking out for the right talent to analyze vast amounts of data.
- Following four perspectives leads to importance of big data analytics.



Further study: https://www.whizlabs.com/blog/big-data-analytics-importance/

Why Learn Big Data?

To get an answer to why you should learn Big Data? Let's start with what industry leaders say about Big Data:

- Gartner – Big Data is the new Oil.
- IDC – Its market will be growing 7 times faster than the overall IT market.
- IBM – It is not just a technology – it's a business strategy for capitalizing on information resources.
- IBM – Big Data is the biggest buzz word because technology makes it possible to analyze all the available data.
- McKinsey – There will be a shortage of 1500000 Big Data professionals by the mid of 2020.

Industries today are searching new and better ways to maintain their position and be prepared for the future. According to experts, Big Data analytics provides leaders a path to capture insights and ideas to stay ahead in the tough competition.

Course Objective

- To understand the concepts and principles of big data.
- To explore the big data stacks and the technologies associated with it.
- To evaluate the different NOSQL databases and frameworks required to handle the big data.
- To apply the techniques for analysis of big data using R tool.
- To formulate the concepts, principles and techniques focusing on the applications to industry and real world experience.

How?

Blend of Theory and Practical

Prerequisites

- Database Management System

Course Contents

Sr #	Major and Detailed Coverage Area	Hrs
1	**Introduction to Big Data**	6
	Importance of Data, Characteristics of Data Analysis of Unstructured Data, Combining Structured and Unstructured Sources. Introduction to Big Data Platform – Challenges of conventional systems – Web data – Evolution of Analytic scalability, analytic processes and tools, Analysis vs reporting – Modern data analytic tools, Types of Data, Elements of Big Data, Big Data Analytics, Data Analytics Lifecycle.	
2	**Big Data Technology Foundations**	8
	Exploring the Big Data Stack, Data Sources Layer, Ingestion Layer, Storage Layer, Physical Infrastructure Layer, Platform Management Layer, Security Layer, Monitoring Layer, Analytics Engine, Visualization Layer, Big Data Applications, Virtualization. Introduction to Streams Concepts – Stream data model and architecture – Stream Computing, Sampling data in a stream – Filtering streams, Counting distinct elements in a stream.	

Course Contents continue…

Sr #	Major and Detailed Coverage Area	Hrs
3	**Big Data Tools**	
	NOSQL, MapReduce – Hadoop, HDFS, Hive, MapR – Hadoop -YARN - Pig and PigLatin, Jaql - Zookeeper - HBase, Cassandra- Oozie, Lucene- Avro, Mahout. Hadoop Distributed file systems.	8
4	**Data Analysis through R**	8
	Exploring R: Exploring Basic Features of R, Programming Features, Packages, Exploring RStudio, Handling Basic Expressions in R, Basic Arithmetic in R, Mathematical Operators, Calling Functions in R, Working with Vectors, Creating and Using Objects, Handling Data in R Workspace, Creating Plots, Using Built-in Datasets in R, Reading Datasets and Exporting Data from R, Manipulating and Processing Data in R	
5	**Frameworks and Visualization**	6
	Distributed and Parallel Computing for Big Data, Visualizations – Visual data analysis techniques, interaction techniques; Systems and applications. Exploring the Use of Big Data in Business Context, Use of Big Data in Social Networking, Business Intelligence, Product Design and Development	

Books

Textbook

- Big Data, Black Book, DT Editorial Services, Dreamtech Press, 2016

Reference Books

- Big Data and Analytics, Seema Acharya, Subhashini Chellappan, Infosys Limited, Publication: Wiley India Private Limited,1st Edition 2015.
- Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (Editor), Wiley, 2014
- Stephan Kudyba, Thomas H. Davenport, Big Data, Mining, and Analytics, Components of Strategic Decision Making, CRC Press, Taylor & Francis Group. 2014
- Norman Matloff , THE ART OF R PROGRAMMING, No Starch Press, Inc.2011
- Big Data For Dummies, Judith Hurwitz et al. Wiley 2013
- Glenn J. Myatt, Making Sense of Data, John Wiley & Sons, 2007 Pete Warden, Big Data Glossary, O'Reilly, 2011.

Course Outcomes

By the end of this course, students will be able to

- Identity the basic characteristics of big data and deploy a structured life cycle approach.
- Classify and examine the data under big data stack and associated technologies.
- Evaluate big data technologies to analyze big data and create models.
- Compose efficient data analysis techniques using R tools.
- Contextually integrate and correlate large amounts of information to gain faster insights for real time scenarios.

Data

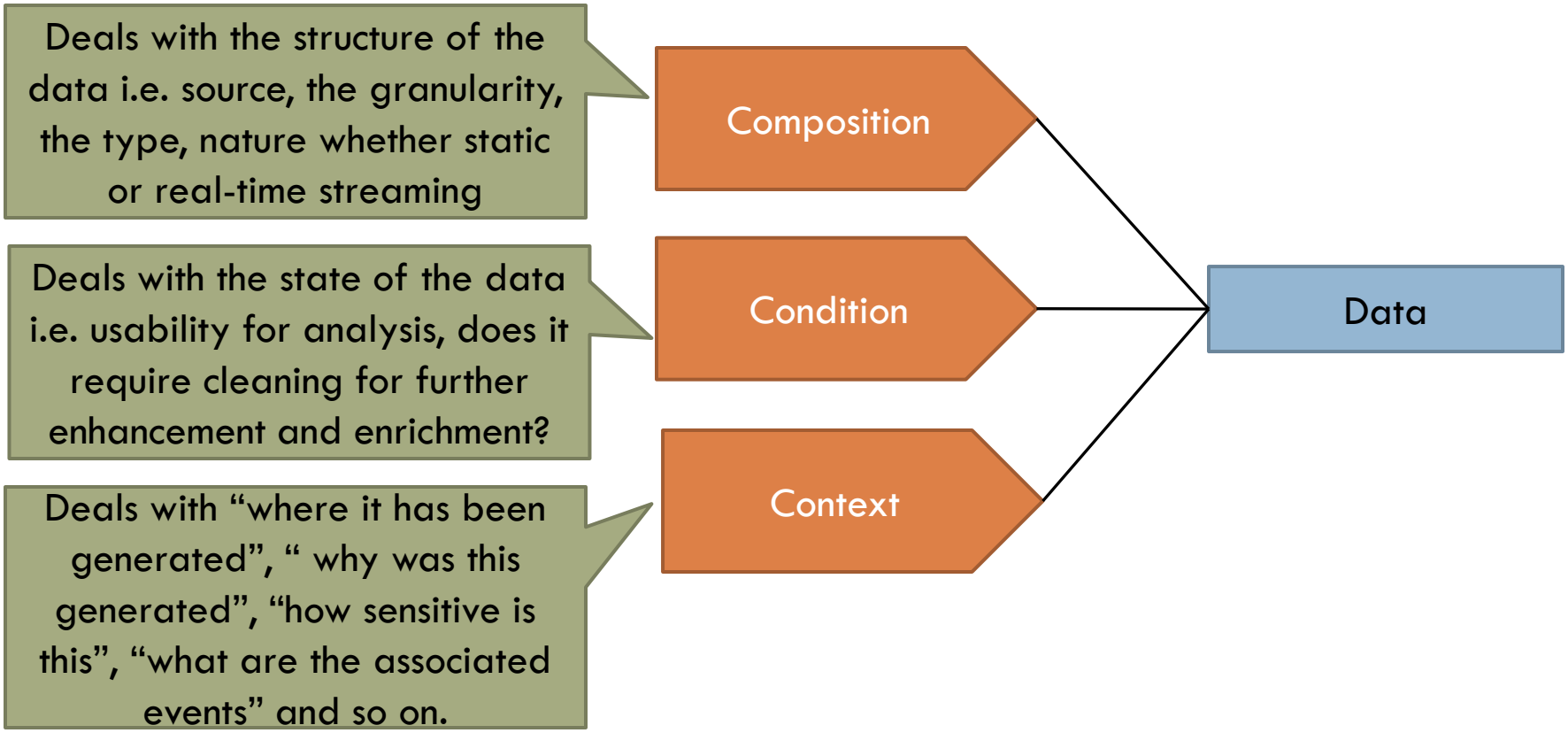
- A representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalized manner.
- Data is either intended to be processed, is being processed, or has been processed.
- It can be in any form stored internally in a computer system or computer network or in a person's mind.
- Since the mid-1900s, people have used the word **data** to mean computer information that is transmitted or stored.
- Data is the plural of datum (a Latin word meaning something given), a single piece of information. In practice, however, people use data as both the singular and plural form of the word.
- It must be interpreted, by a human or machine to derive meaning.
- It is presents in homogeneous sources as well as heterogeneous sources.
- The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights.

Data → Information → Knowledge → Actionable Insights

Importance of Data

- The ability to analyze and act on data is increasingly important to businesses. It might be part of a study helping to cure a disease, boost a company's revenue, understand and interpret market trends, study customer behavior and take financial decisions
- The pace of change requires companies to be able to react quickly to changing demands from customers and environmental conditions. Although prompt action may be required, decisions are increasingly complex as companies compete in a global marketplace.
- Managers may need to understand high volumes of data before they can make the necessary decisions
- Relevant data creates strong strategies - Opinions can turn into great hypotheses, and those hypotheses are just the first step in creating a strong strategy. It can look something like this: "Based on X, I believe Y, which will result in Z"
- Relevant data strengthens internal teams
- Relevant data quantifies the purpose of the work

Characteristics of Data



Human vs. Machine Readable data

- Human-readable refers to information that only humans can interpret and study, such as an image or the meaning of a block of text. If it requires a person to interpret it, that information is human-readable.
- Machine-readable refers to information that computer programs can process. A program is a set of instructions for manipulating data. Such data can be automatically read and processed by a computer, such as CSV, JSON, XML, etc.

Non-digital material (for example printed or hand-written documents) is by its non-digital nature not machine-readable. But even digital material need not be machine-readable. For example, a PDF document containing tables of data. These are definitely digital but are not machine-readable because a computer would struggle to access the tabular information - even though they are very human readable. The equivalent tables in a format such as a spreadsheet would be machine readable.

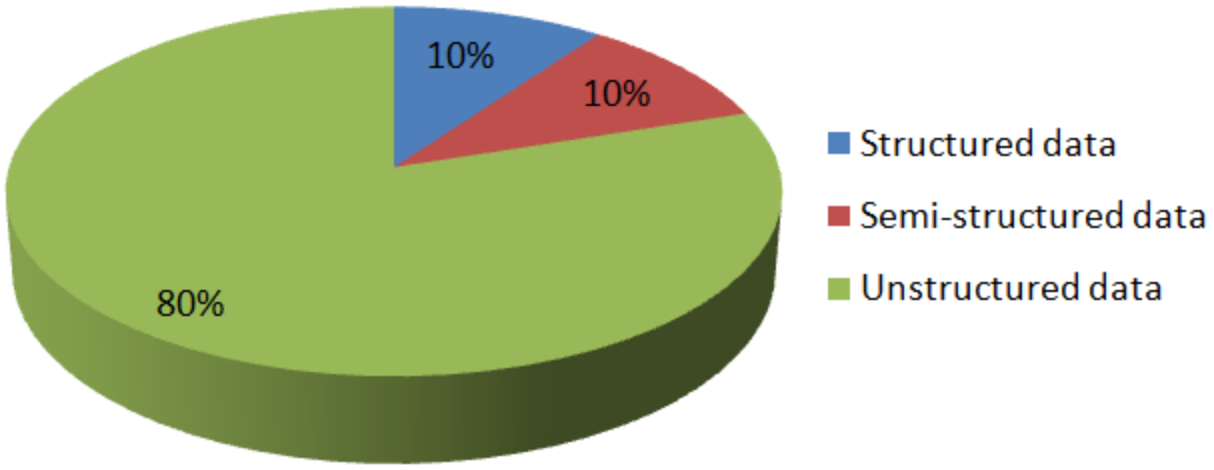
Another example scans (photographs) of text are not machine-readable (but are human readable!) but the equivalent text in a format such as a simple ASCII text file can machine readable and processable.

Classification of Digital Data

Digital data is classified into the following categories:

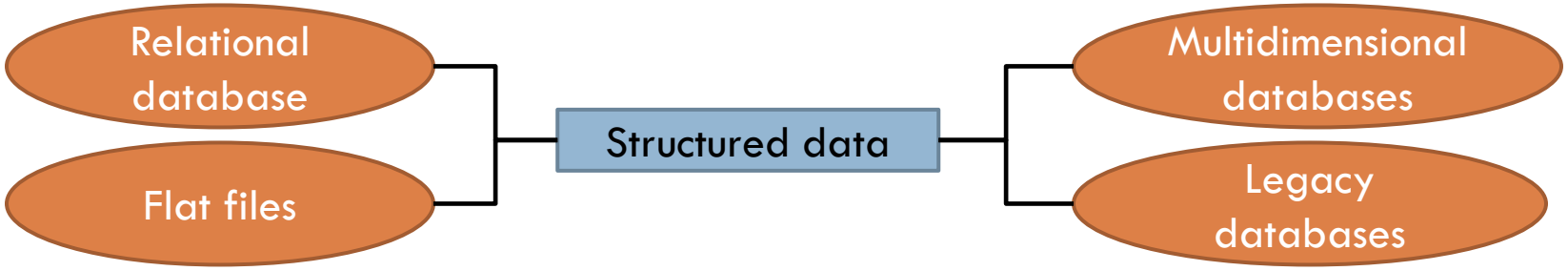
- Structured data
- Semi-structured data
- Unstructured data

Approximate percentage distribution of digital data

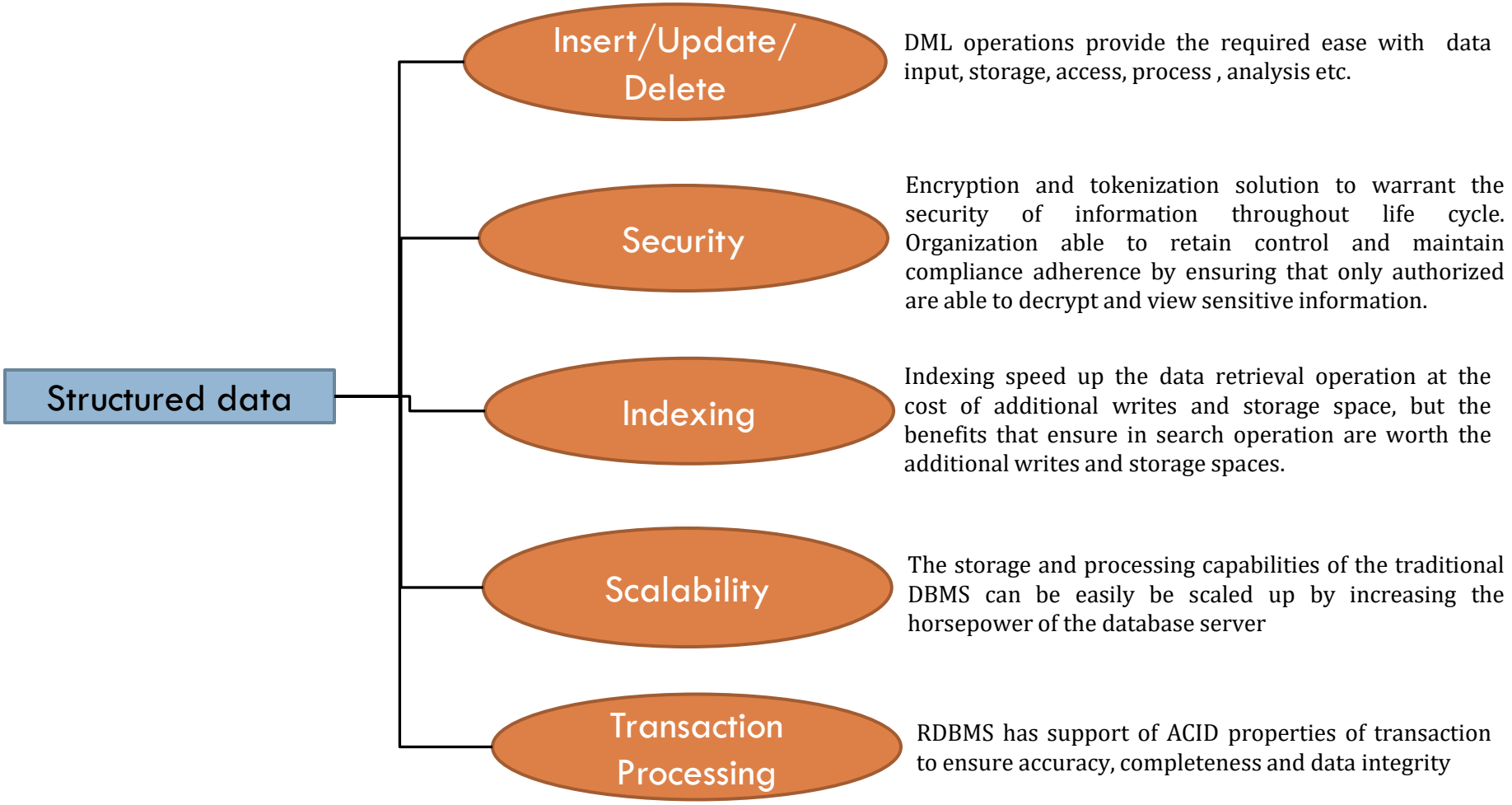


Structured Data

- It is defined as the data that has a defined repeating pattern and this pattern makes it easier for any program to sort, read, and process the data.
- This is data is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- Relationships exist between entities of data.
- Structured data:
  - Organize data in a pre-defined format
  - Is stored in a tabular form
  - Is the data that resides in a fixed fields within a record of file
  - Is formatted data that has entities and their attributes mapped
  - Is used to query and report against predetermined data types
- Sources:

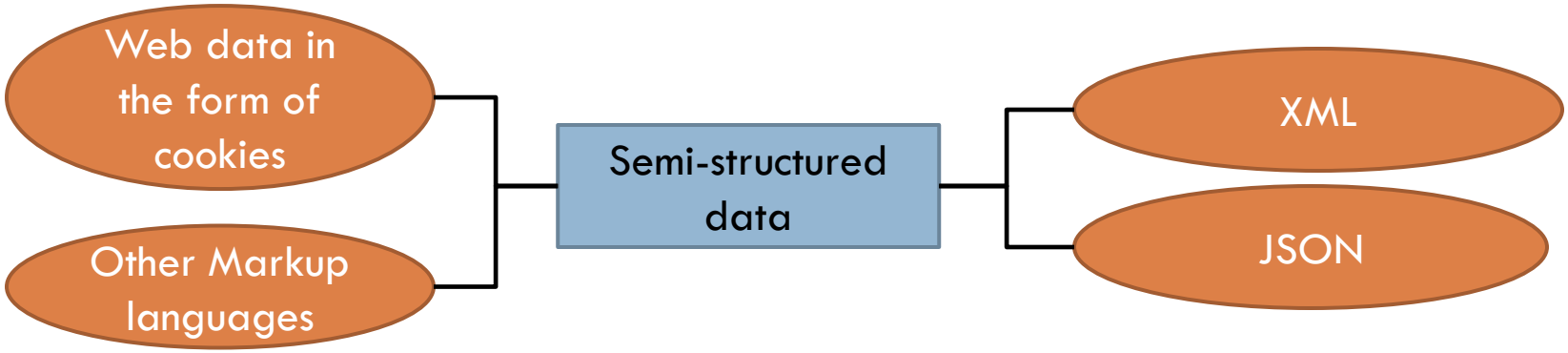


Ease with Structured Data



Semi-structured Data

- Semi-structured data, also known as having a schema-less or self-describing structure, refers to a form which does not conform to a data model as in relational database but has some structure.
- In other words, data is stored inconsistently in rows and columns of a database.
- However, it is not in a form which can be used easily by a computer program.
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
- Sources:



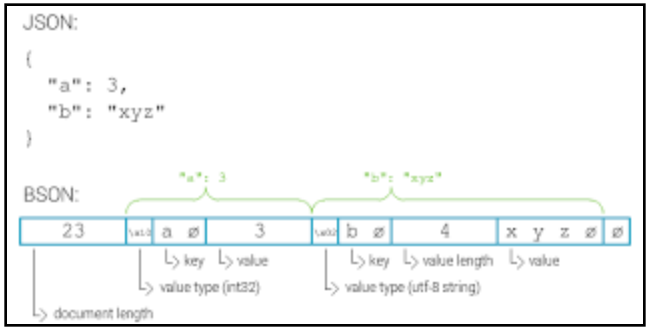
XML, JSON, BSON format

XML

```
<Node>
  <id>10002</id>
  <Name>john</Name>
</Node>
<Node>
  <id>10003</id>
  <Name>Scott</Name>
</Node>
<Node>
  <id>10004</id>
  <Name>Mohan</Name>
</Node>
<Node>
  <id>10001</id>
  <Name>Deepak </Name>
</Node>
```

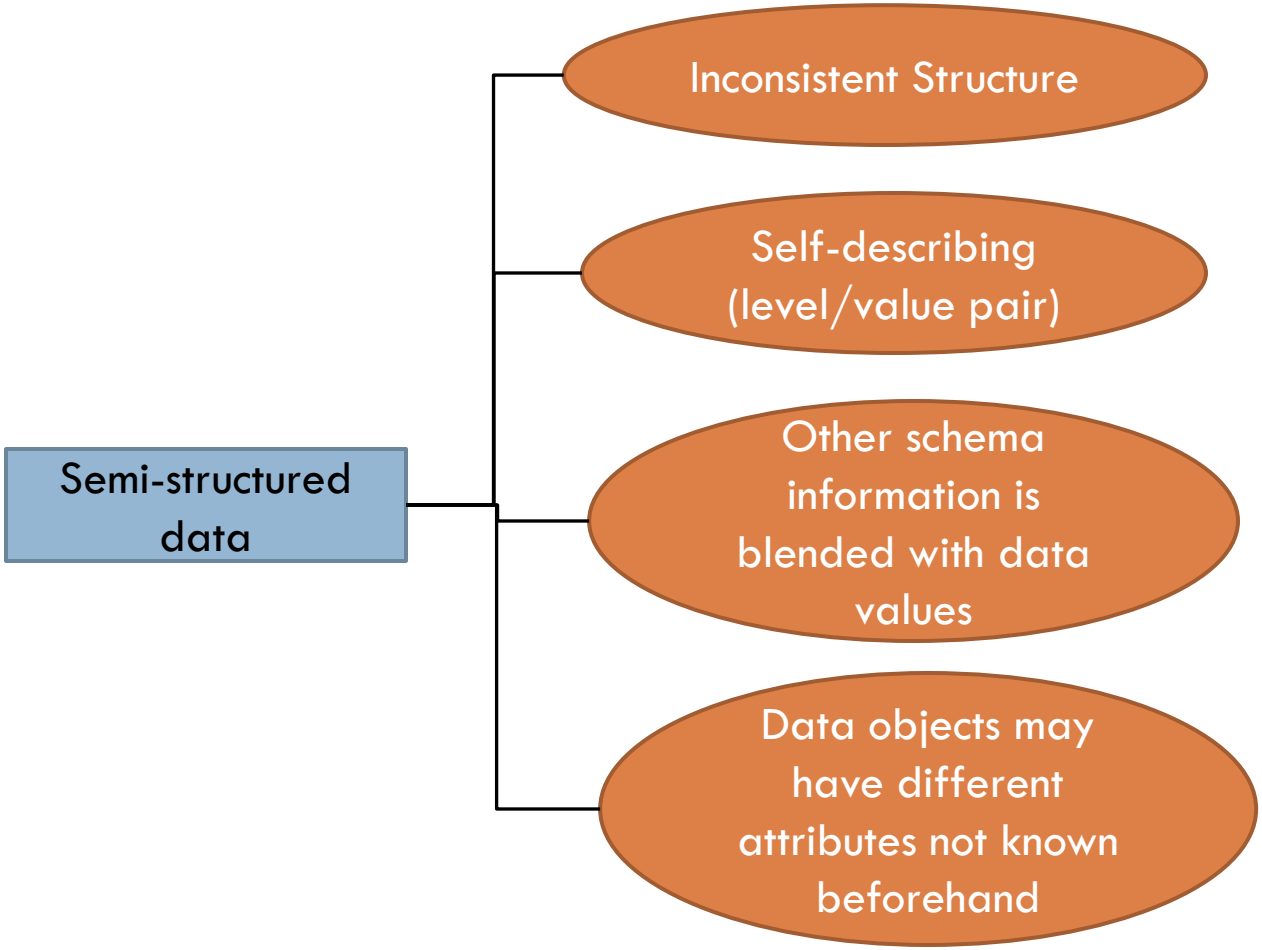
JSON

```
[
    {
        "id":10002,
        "name":"john"
    },
    {
        "id":10003,
        "name":"Scott"
    },
    {
        "id":10004,
        "name":"Mohan"
    },
    {
        "id":10001,
        "name":"Deepak"
    }
]
```



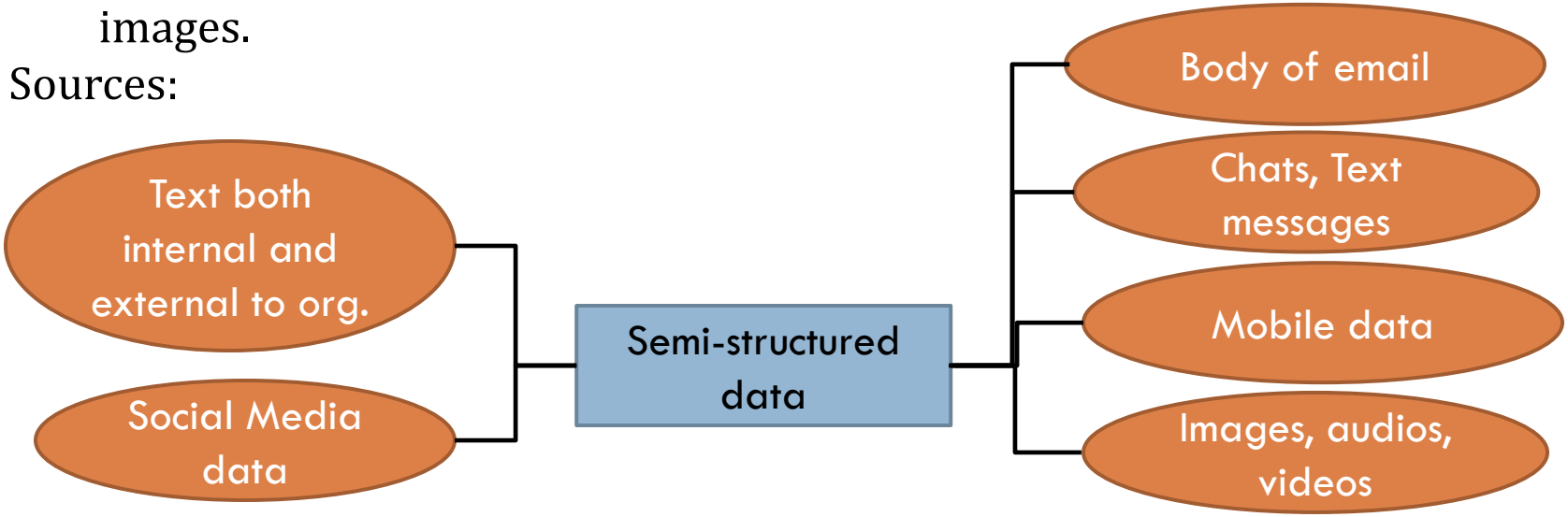
Source (XML & JSON): http://sqllearnergroups.blogspot.com/2014/03/how-to-get-json-format-through-sql.html
Source (JSON & BSON): http://www.expert-php.fr/mongodb-bson/

Characteristics of Semi-structured Data



Unstructured Data

- Unstructured data is a set of data that might or might not have any logical or repeating patterns and is not recognized in a pre-defined manner.
- About 80 percent of enterprise data consists of unstructured content.
- Unstructured data:
  - Typically consists of metadata i.e. additional information related to data.
  - Comprises of inconsistent data such as data obtained from files, social media websites, satellites etc
  - Consists of data in different formats such as e-mails, text, audio, video, or images.
- Sources:



Challenges associated with Unstructured data

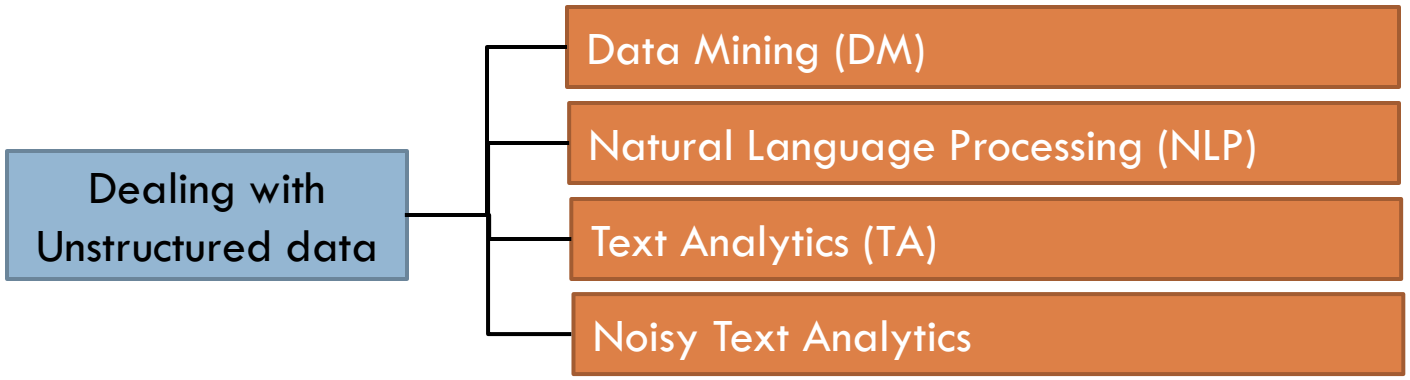
Working with unstructured data poses certain challenges, which are as follows:

- Identifying the unstructured data that can be processed
- Sorting, organizing, and arranging unstructured data indifferent sets and formats
- Combining and linking unstructured data in a more structured format to derive any logical conclusions out of the available information
- Costing in terms of storage space and human resources need to deal with the exponential growth of unstructured data

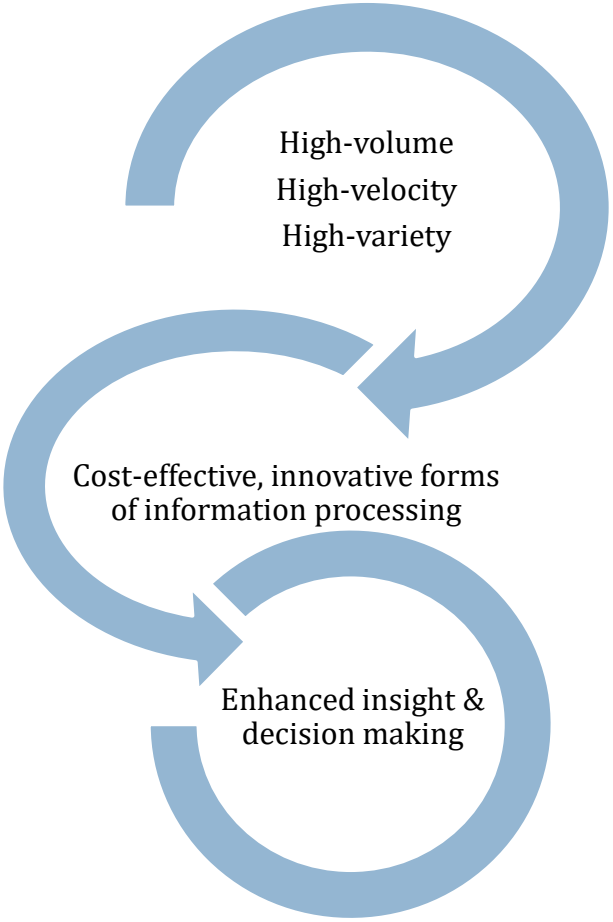
Data Analysis of Unstructured Data

The complexity of unstructured data lies within the language that created it. Human language is quite different from the language used by machines, which prefer structured information. Unstructured data analysis is referred to the process of analyzing data objects that doesn't follow a predefine data model and/or is unorganized. It is the analysis of any data that is stored over time within an organizational data repository without any intent for its orchestration, pattern or categorization.

Dealing with Unstructured data



Definition of Big Data



Big Data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

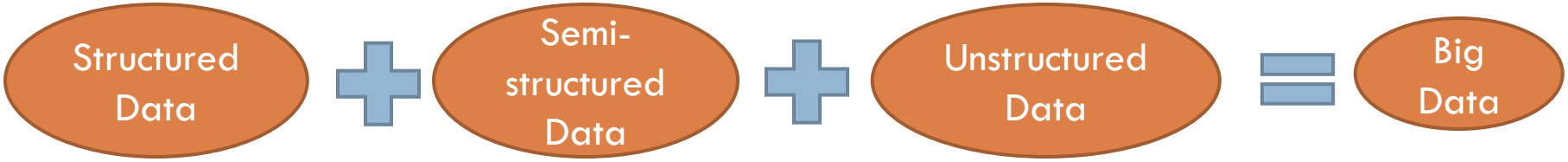
Source: Gartner IT Glossary

What is Big Data?

Think of following:

- Every second, there are around 822 tweets on Twitter
- Every minutes, nearly 510 comments are posted, 293 K statuses are updated, and 136K photos are uploaded in Facebook
- Every hour, Walmart, a global discount departmental store chain, handles more than 1 million customer transactions.
- Everyday, consumers make around 11.5 million payments by using PayPal.

In the digital world, data is increasing rapidly because of the ever increasing use of the internet, sensors, and heavy machines at a very high rate. The sheer volume, variety, velocity, and veracity of such data is signified the term '**Big Data**'.



Challenges of Conventional Systems

The main challenge in the traditional approach for computing systems to manage 'Big Data' because of immense speed and volume at which it is generated. Some of the challenges are:

- Traditional approach cannot work on unstructured data efficiently
- Traditional approach is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. This approach will not adequate for big data
- Traditional approach is batch oriented and need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained
- Traditional data management, warehousing, and analysis systems fizzle to analyze this type of data. Due to it's complexity, big data is processed with parallelism. Parallelism in a traditional system is achieved through costly hardware like MPP (Massively Parallel Processing) systems
- Inadequate support of aggregated summaries of data

Challenges of Conventional Systems cont'd

Other challenges can be categorized as:

- Data Challenges:
  - Volume, velocity, veracity, variety
  - Data discovery and comprehensiveness
  - Scalability
- Process challenges
  - Capturing Data
  - Aligning data from different sources
  - Transforming data into suitable form for data analysis
  - Modeling data(Mathematically, simulation)
- Management Challenges:
  - Security
  - Privacy
  - Governance
  - Ethical issues

Elements of Big Data

In most big data circles, these are called the four V's: **v**olume, **v**ariety, **v**elocity, and **v**eracity. (One might consider a fifth V, **v**alue.)

Volume - refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, M2M sensors, photographs, video, etc. The vast amounts of data have become so large in fact it can no longer store and perform data analysis using traditional database technology. So using distributed systems, where parts of the data is stored in different locations and brought together by software.

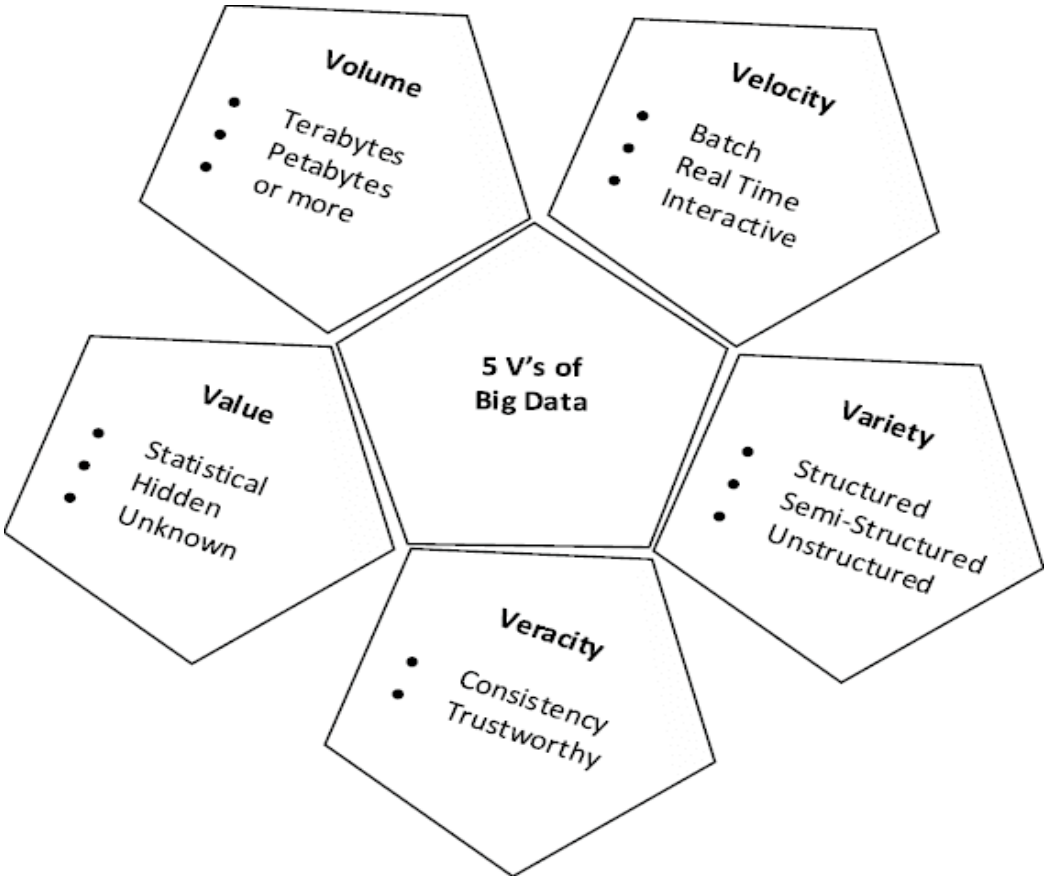
Variety - defined as the different types of data the digital system now use. Data today looks very different than data from the past. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.

Velocity - refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every second of every day data is increasing. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access. Big data technology allows to analyze the data while it is being generated, without ever putting it into databases.

Veracity - is the quality or trustworthiness of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content.

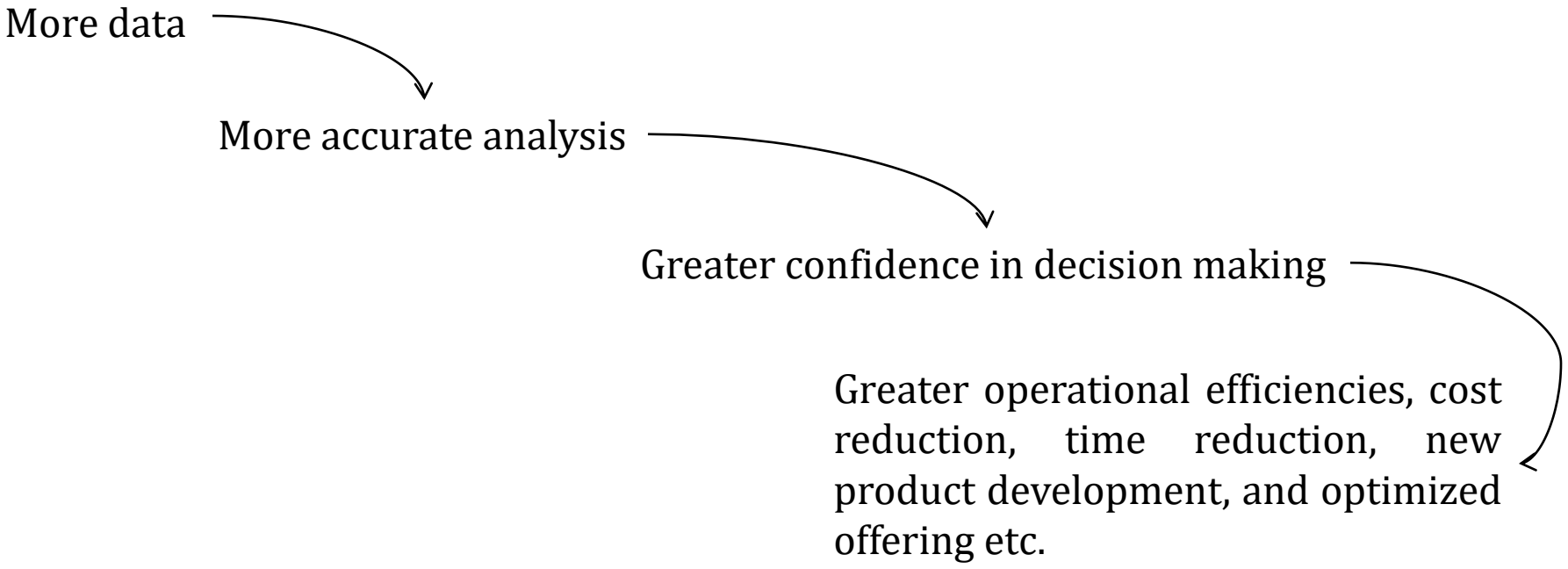
Elements of Big Data cont'd

Value - refers to the ability to transform a tsunami of data into business. Having endless amounts of data is one thing, but unless it can be turned into value it is useless.



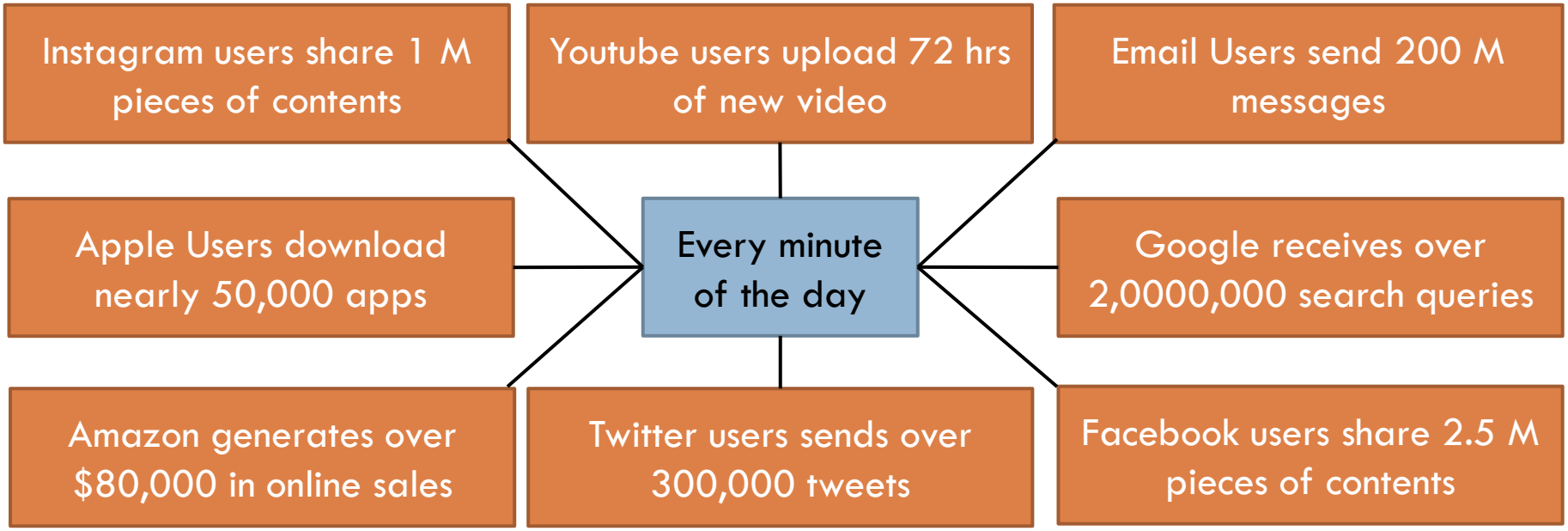
Why Big Data?

More data for analysis will result into **greater analytical accuracy** and greater **confidence in the decisions** based on the analytical findings. This would entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services and optimizing existing services.

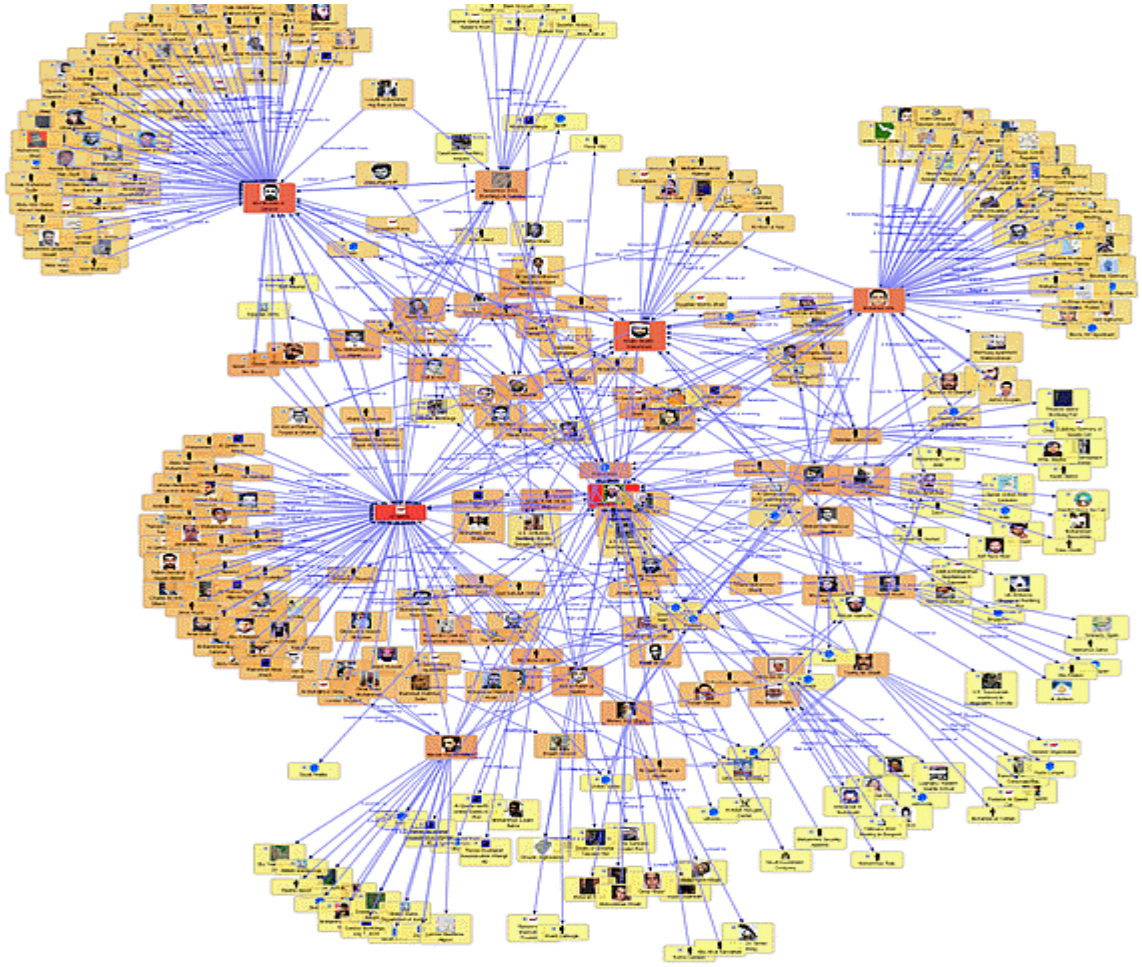


Use of Big Data in Social Networking

A human being lives in a social environment and gains knowledge and experience through communication. Today, communication is not restricted to meeting in person. Internet and mobile have made communication and sharing of data possible across the globe. Some social networking sites such as Twitter, Facebook, and LinkedIn produces data from people. **Social network Analysis (SNA)** is the analysis performed on the data obtained from social media. As such data is generated in huge volume, it results in the formation of a Big Data pool.



Social Network – MNO perspective



Source: http://www.fmsasg.com/socialnetworkanalysis/

The data captured by mobile network operator (MNO) in a day such as the cell phone calls, test messages, and other related details of all its customers is very huge in volume. Such type of data is used daily for different purposes. The company must study the data of the people whom the customer called and also for the people in the customer's network who called back the customer. Such network is called a **social network** and is depicted. The analysis can go deeper and deeper within the network to get a complete picture of the social network. As the analysis goes deeper, the volume of data to be analyzed also become massive.

Decision Making Process Influencers by Social Media data

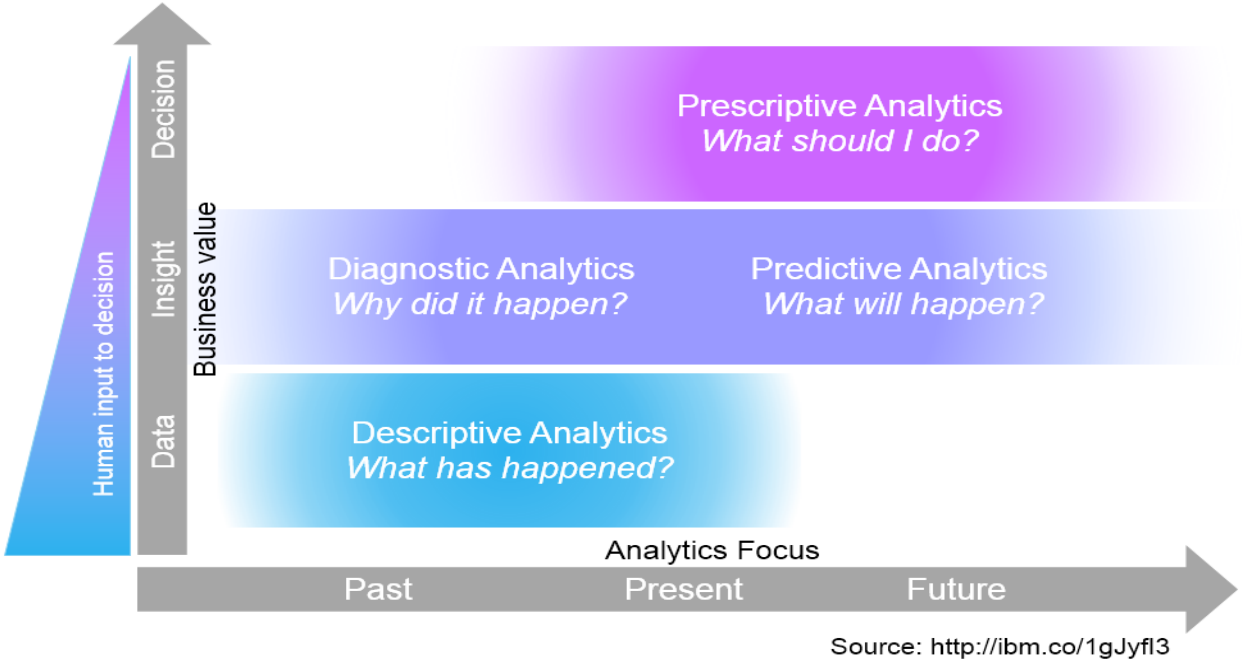
The following are the areas in which decision-making process are influenced by social network data:

- **Business Intelligence:** It is a data analysis process to convert a raw dataset to meaningful information by using different techniques and tools for boosting business performances. This system allows a company to collect, store, access, and analyze data for adding value to the decision making.
- **Product design and development:** With the increasing popularity of all social media and growing volume of data every second, organizations competing to make a big in the market must not only identify and extracts the information relevant to their company, products, and services but also comprehend and respond to the information on a continuous basis. By listening to what customers want, by understanding where the gap in the offering is, and so on, organizations can make the right directions in the direction of their product development and offerings. In this way, social network data can help organizations to improve product development and services, making sure that the customers ultimately get the products and services they want.

Data Analytics

Data analytics is the process of extracting useful information by analysing different types of data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful information for the benefit of faster decision making.

There are 4 types of analytics:

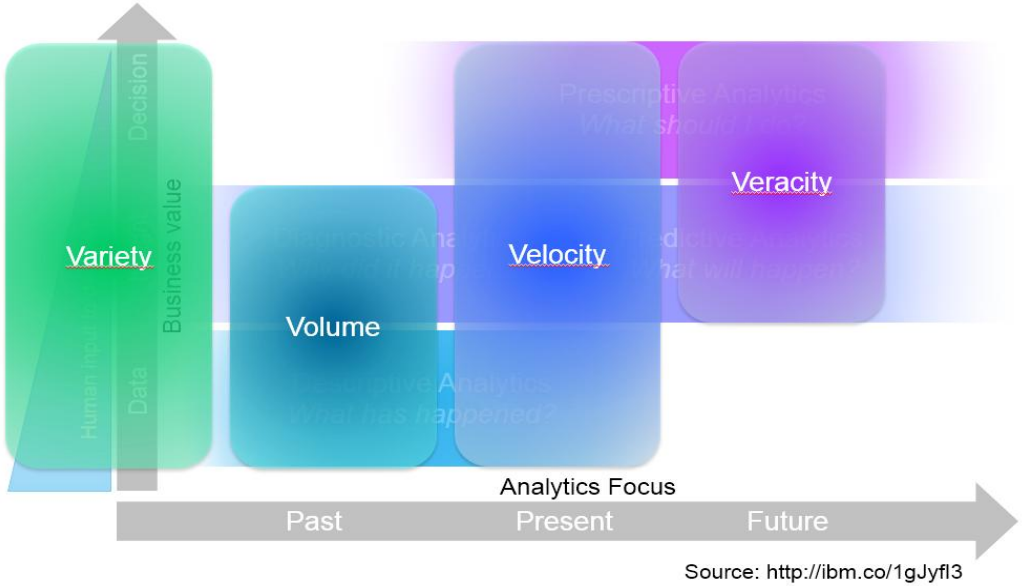


Source: http://ibm.co/1gJyfl3

Analytics Approach – What is the data telling?

Approach	Explanation
Descriptive	What's happening in my business? • Comprehensive, accurate and historical data • Effective Visualisation
Diagnostic	Why is it happening? • Ability to drill-down to the root-cause • Ability to isolate all confounding information
Predictive	What's likely to happen? • Decisions are automated using algorithms and technology • Historical patterns are being used to predict specific outcomes using algorithms
Prescriptive	What do I need to do? • Recommended actions and strategies based on champion/challenger strategy outcomes • Applying advanced analytical algorithm to make specific recommendations

Mapping of Big Data's Vs to Analytics Focus



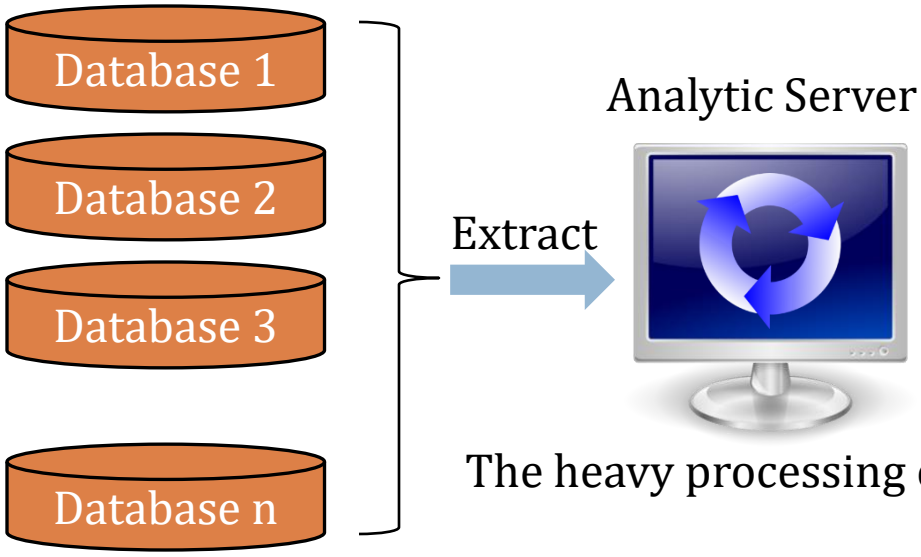
Source: http://ibm.co/1gJyfl3

History data can be quite large. There might be a need to process huge amount of data many times a day as it gets updated continuously. Therefore volume is mapped to history. Variety is pervasive. Input data, insights, and decisions can span a variety of forms, hence it is mapped to all three. High velocity data might have to be processed to help real time decision making and plays across descriptive, predictive, and prescriptive analytics when they deal with present data. Predictive and prescriptive analytics create data about the future. That data is uncertain, by nature and its veracity is in doubt. Therefore veracity is mapped to prescriptive and predictive analytics when it deal with future.

Evolution of Analytics Scalability

It goes without saying that the world of big data requires new levels of scalability. As the amount of data organizations process continues to increase, the same old methods for handling data just won't work anymore. Organizations that don't update their technologies to provide a higher level of scalability will quite simply choke on big data. Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.

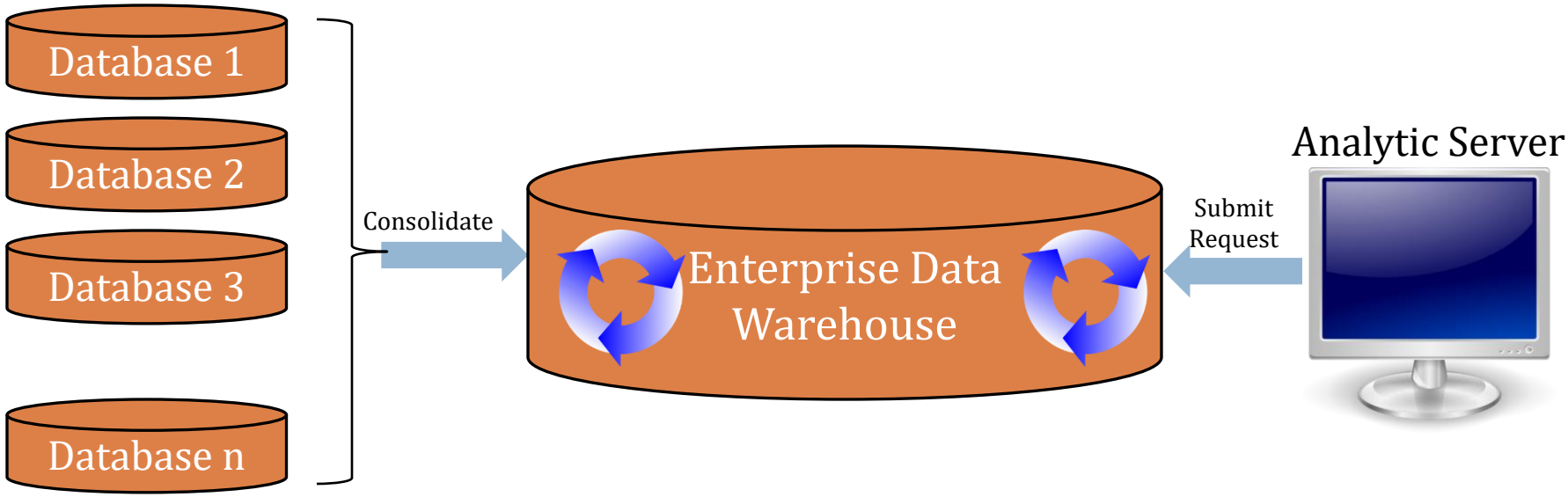
Traditional Analytics Architecture



The heavy processing occurs in the analytic environment. This may even a PC

Evolution of Analytics Scalability cont'd

Modern In-Database Analytics Architecture



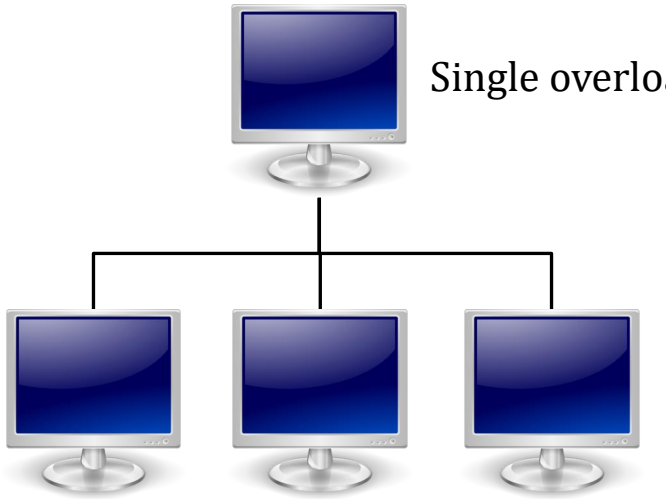
In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

Evolution of Analytics Scalability cont'd

MPP Database Analytics Architecture

Massively parallel processing (MPP) database systems is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data. An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources. Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house. The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.

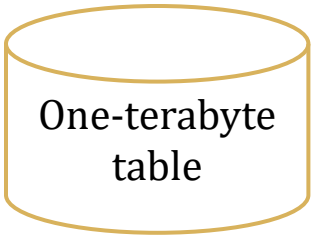
In stead of single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU.



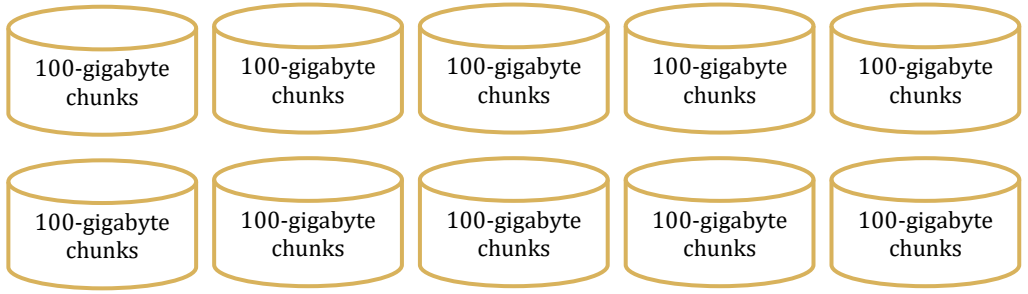
Single overloaded server

Multiple lightly loaded server

MPP Database Example



A Traditional database will query a one-terabyte table one row at time



10 simultaneous 100-gigabyte queries

MPP database is based on the principle of **SHARE THE WORK!**

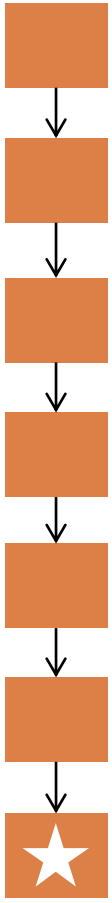
A MPP database spreads data out across multiple sets of CPU and disk space. Think logically about dozens or hundreds of personal computers each holding a small piece of a large set of data. This allows much faster query execution, since many independent smaller queries are running simultaneously instead of just one big query

If more processing power and more speed are required, just bolt on additional capacity in the form of additional processing units

MPP systems build in redundancy to make recovery easy and have resource management tools to manage the CPU and disk space

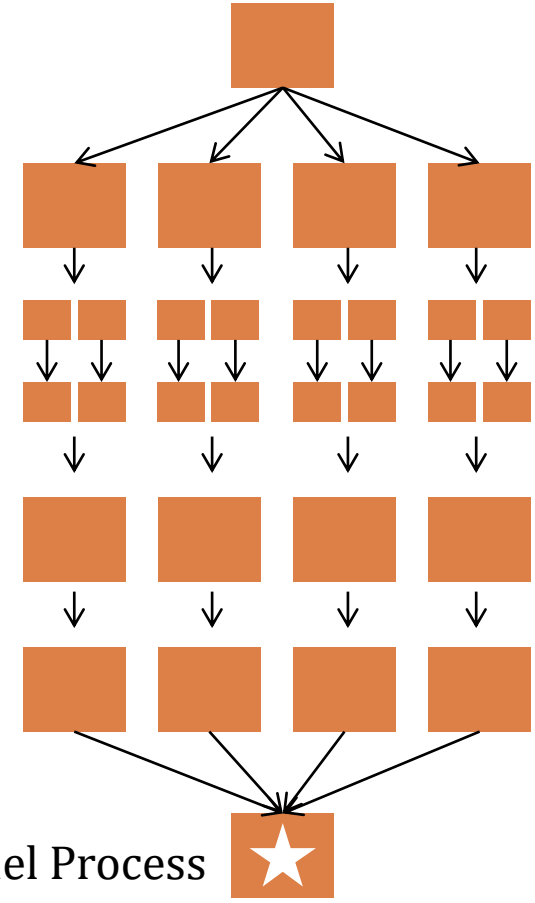
MPP Database Example cont'd

An MPP system allows the different sets of CPU and disk to run the process concurrently



Single Threaded Process

An MPP system breaks the job into pieces



Parallel Process

Analysis vs. Reporting

Reporting - The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

Analysis: The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

Difference b/w Reporting and Analysis:

- Reporting translates raw data into information. Analysis transforms data and information into insights.
- Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges. Good reporting should raise questions about the business from its end users. The goal of analysis is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.
- In summary, **reporting shows you what is happening** while **analysis focuses on explaining why it is happening and what you can do about it**.

Big Data Analytics

Big data analytics is the process of extracting useful information by analysing different types of big data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful info for the benefit of faster decision making.

Big Data Application in different Industries

Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply–chain management and analytics
- Event– and behavior–based targeting
- Market and consumer segmentations

Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media

- Large–scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click–fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences

- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply–chain management
- Drug discovery and development analysis

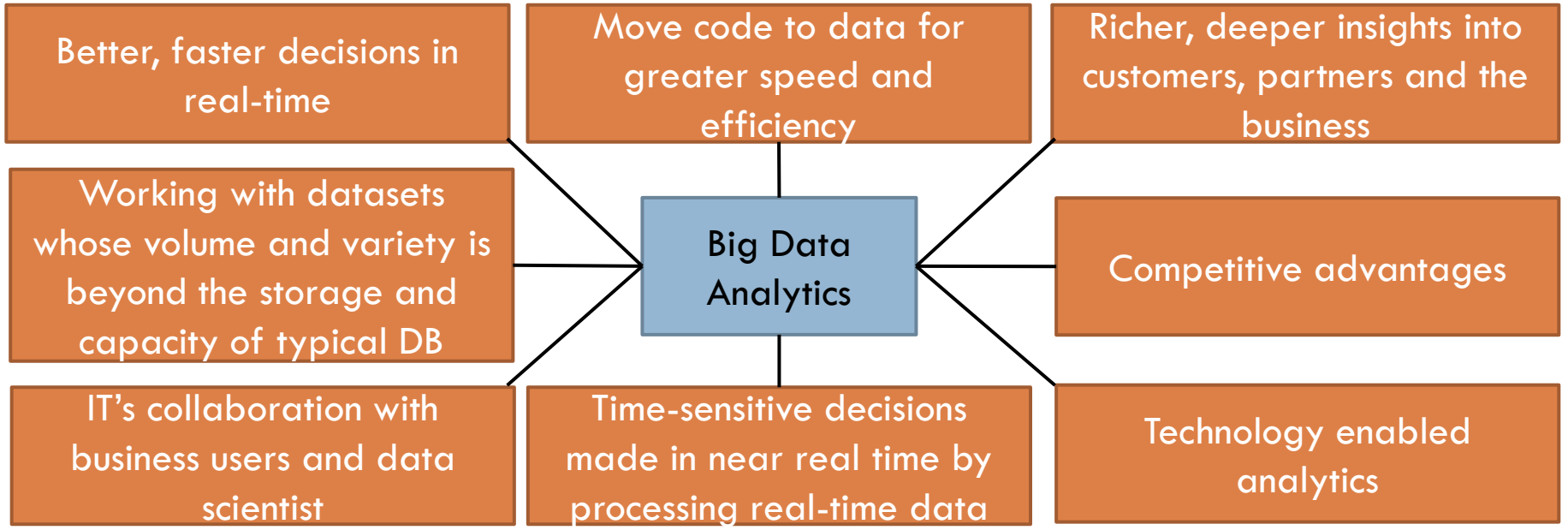
Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

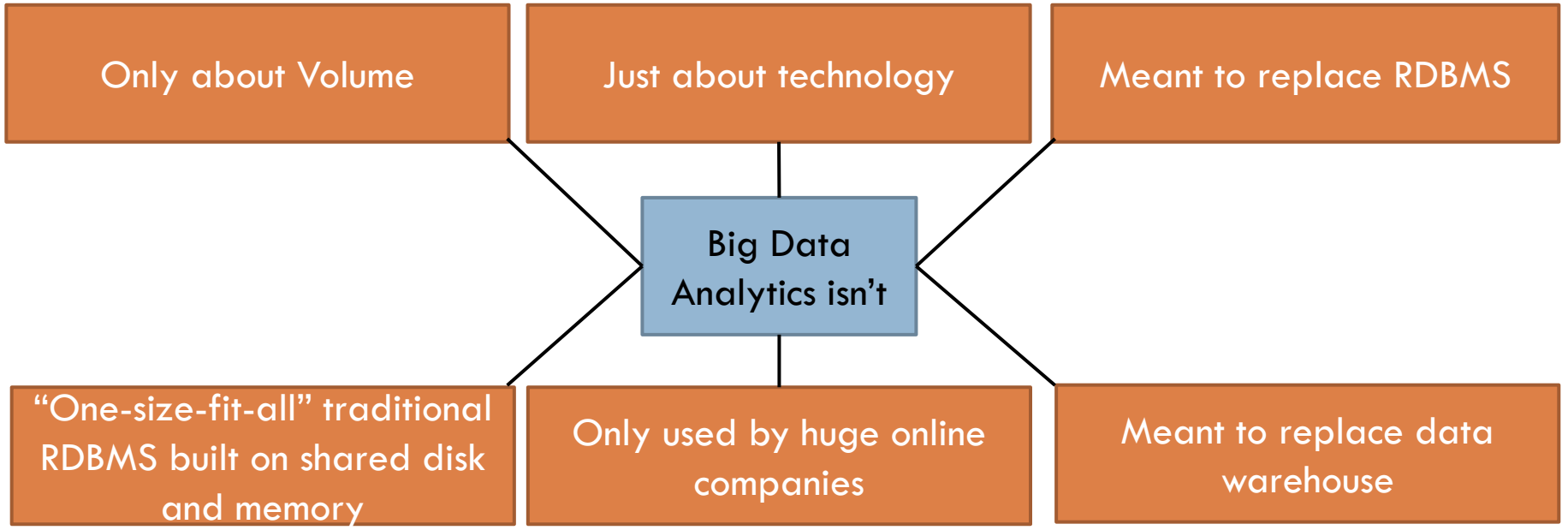
Ecommerce & customer service

- Cross–channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action

What is Big Data Analytics ?



What is Big Data Analytics isn't?



Challenges that prevent business from capitalizing on Big Data

1. Obtaining executive sponsorships for investments in big data and its related activities such as training etc.
2. Getting the business units to share information across organizational silos.
3. Fining the right skills that can manage large amounts of structured, semi-structured, and unstructured data and create insights from it.
4. Determining the approach to scale rapidly and elastically. In other words, the need to address the storage and processing of large volume, velocity and variety of big data.
5. Deciding whether to use structured or unstructured, internal or external data to make business decisions.
6. Determining what to do with the insights created from big data.
7. Choosing the optimal way to report findings and analysis of big data for the presentations to make the most sense.

Top challenges facing Big Data

1. **Scale**: Storage is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should scale vertically or horizontally?
2. **Security**: Most of the NoSQL (Not only SQL) big data platforms have poor security mechanism (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data.
3. **Schema**: Rigid schema have no place. The need of the hour is dynamic schema and static (pre-defined) schemas are passed.
4. **Data Quality**: How to maintain data quality – data accuracy, completeness, timeliness etc. Is the appropriate metadata in place?
5. **Partition Tolerant**: How to build partition tolerant systems that can take care of both hardware and software failures?
6. **Continuous availability:** The question is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.

Kind of Technologies to help meet the challenges posed by Big Data

1. Cheap and abundant storage
2. Faster processors to help with quicker processing of big data
3. Affordable open-source, distributed big data platforms
4. Parallel processing, clustering, visualisation, large grid environments, high connectivity, and high throughputs rather than low latency
5. Cloud computing and other flexible resource allocation agreements

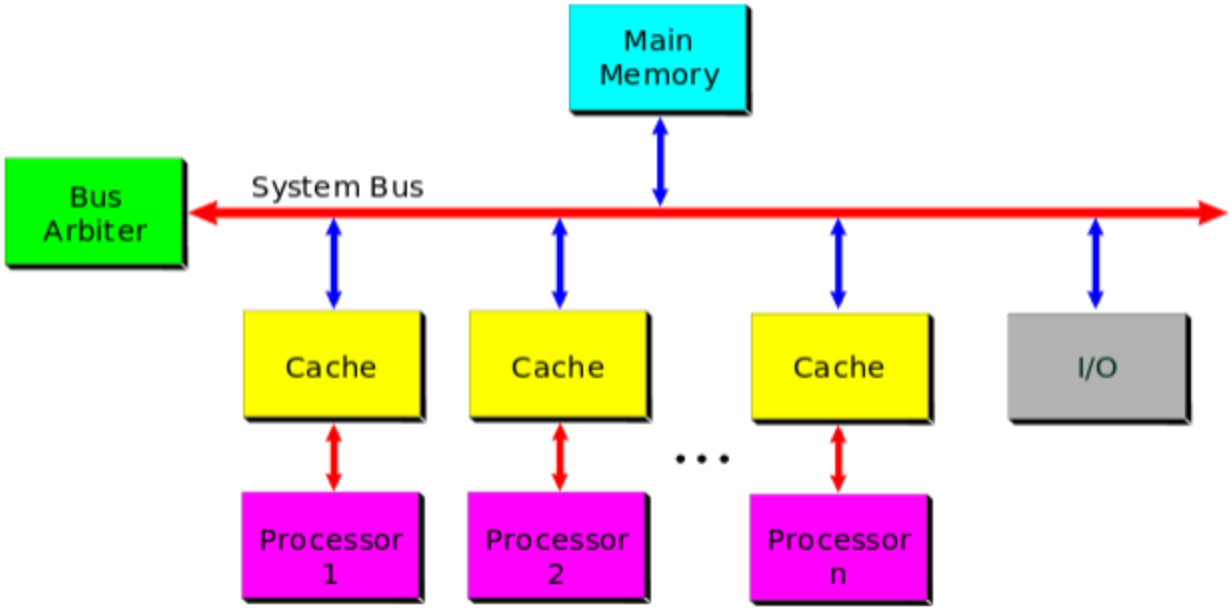
Terminologies used in Big Data

In-Memory Analytics: Data access from non-volatile storage such as hard disk is a slow process. The more the data is required to be fetched from hard disk or secondary storage, the slower the process gets. The problem can be addressed using in-memory analytics. All the relevant data is stored in RAM or primary storage thus eliminating the need to access the data from hard disk. The advantage is faster access, rapid deployment, better insights and minimal IT involvement. In-memory Analytics makes everything **Instantly Available** due to lower cost of RAM or Flash Memory, and data can be stored and processed at lightening speed.

In-Database Processing: Also called as In-Database analytics. It works by fusing data warehouses with analytical systems. Typically the data from various enterprise Online Transaction Processing (OLTP) systems after cleaning up (de-duplication, scrubbing etc.) through the process of ETL is stored in the Enterprise Data Warehouse or data marts. The huge datasets are then exported to analytical programs for complex and extensive computations.

Terminologies used in Big Data cont'd

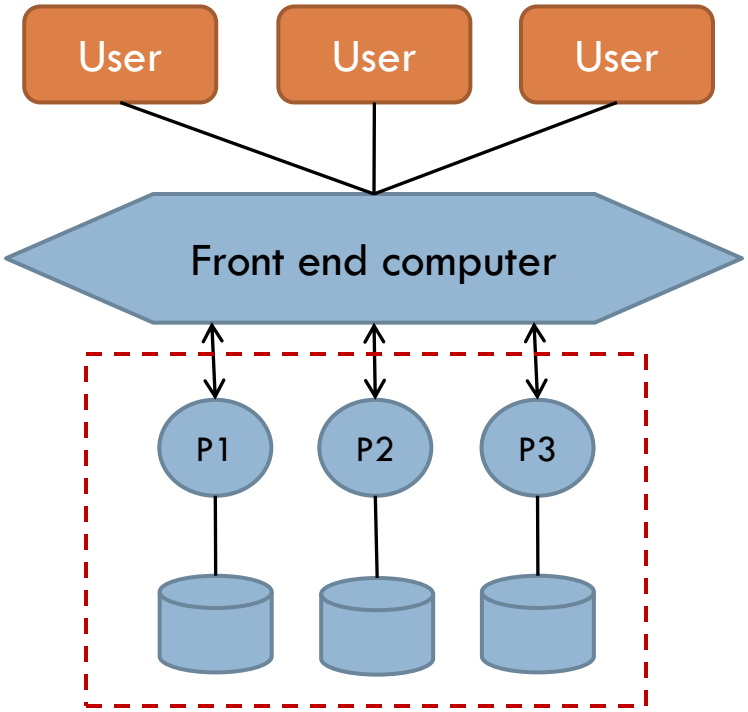
Symmetric Multiprocessor System (SMP): In SMP, there is a single common main memory that is shared by two or more identical processors. The processors have full access to all I/O devices and are controlled by a single operating system instance. Each processor has its own high-speed memory, called cache memory and are connected using a system bus.



Source: https://en.wikipedia.org/wiki/Symmetric_multiprocessing

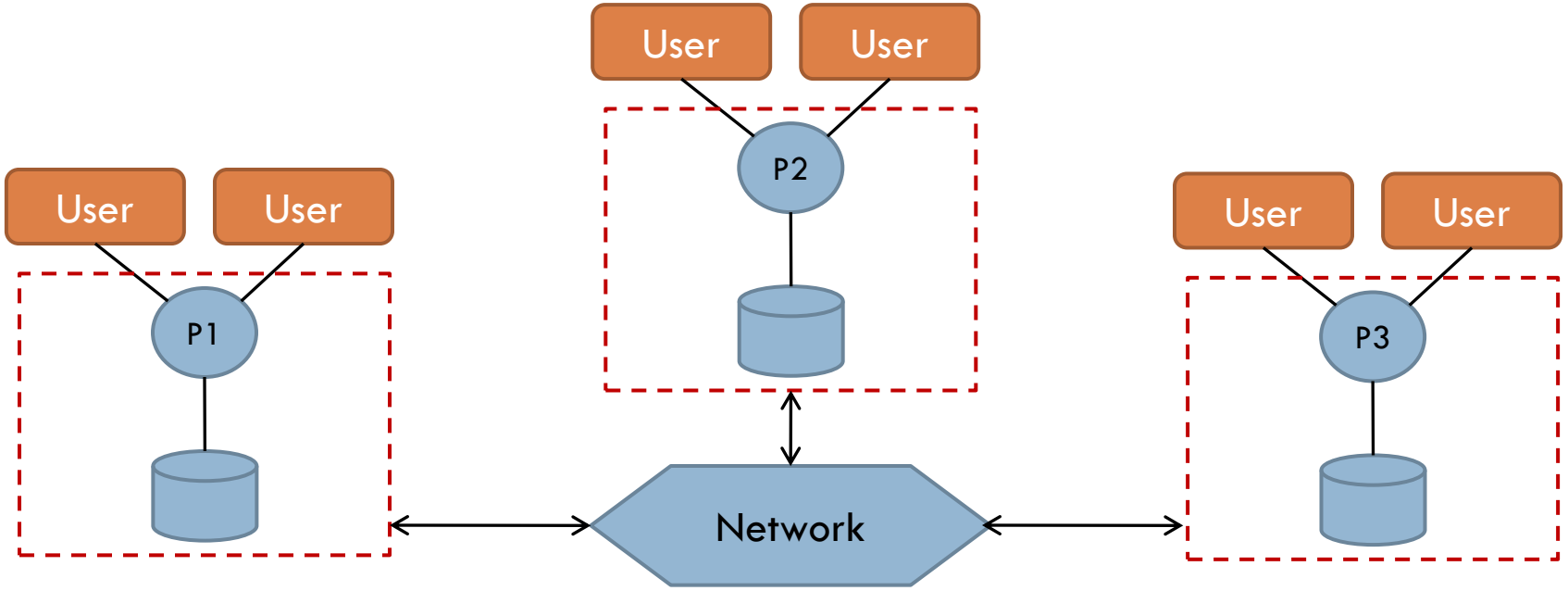
Terminologies used in Big Data cont'd

Parallel Systems: A parallel database system is a tightly coupled system. The processors co-operate for query processing. The user is unaware of the parallelism since he/she has no access to a specific processor of the system.



Terminologies used in Big Data cont'd

Distributed Systems: Known to be loosely coupled and are composed of individual machines. Each of the machine can run their individual application and serve their own respective users. The data is usually distributed across several machines, thereby necessitating quite a number of machines to be accessed to answer a user query.

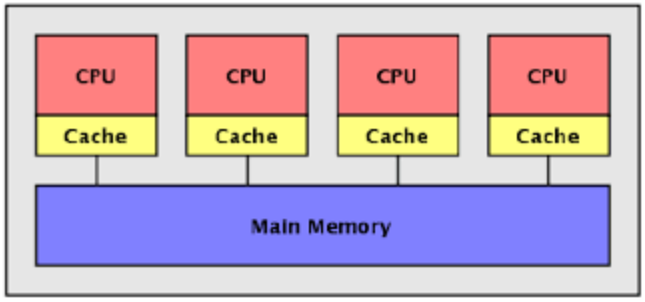


Distributed vs. Parallel Computing

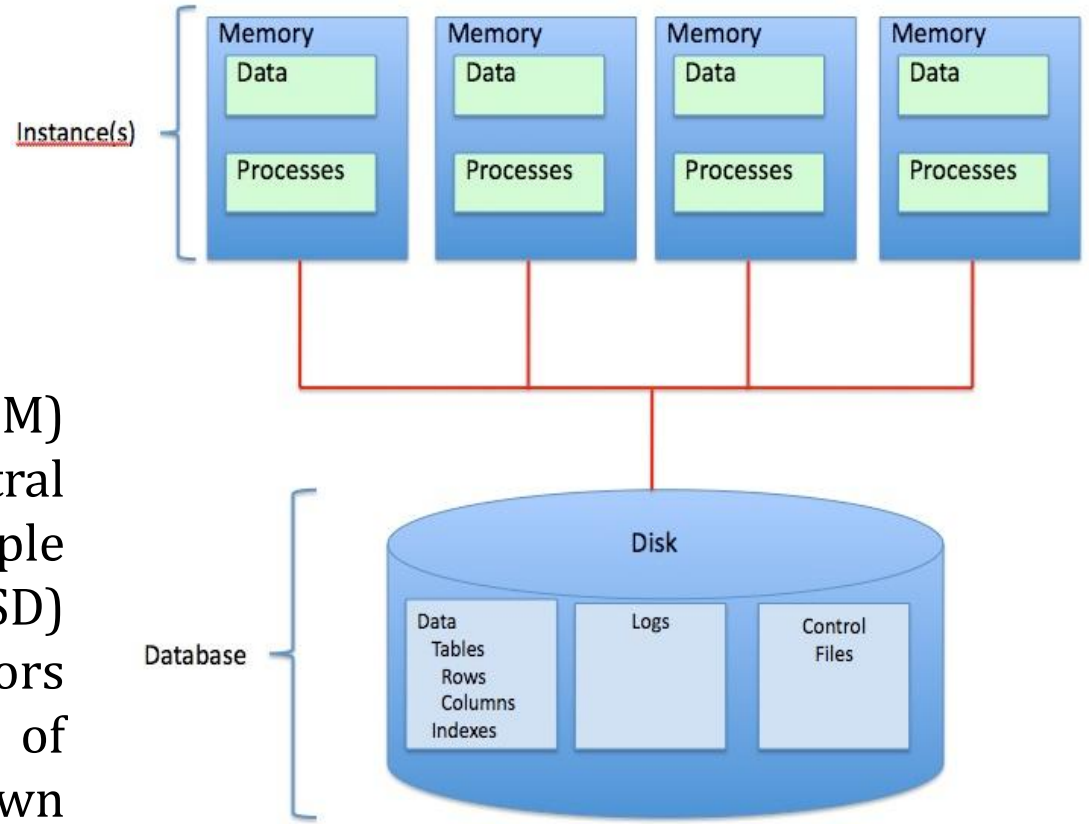
Parallel Computing	Distributed Computing
Shared memory system	Distributed memory system
Multiple processors share a single bus and memory unit	Autonomous computer nodes connected via network
Processor is order of Tbps	Processor is order of Gbps
Limited Scalability	Better scalability and cheaper
	Distributed computing in local network (called **cluster computing**). Distributed computing in wide-area network (**grid computing**)

Terminologies used in Big Data cont'd

SM



SD



In a shared memory (SM) architecture, a common central memory is shared by multiple processors. In a shared disk (SD) architecture, multiple processors share a common collection of disks while having their own private memory.

Terminologies used in Big Data cont'd

In a shared nothing (SN) architecture, neither memory nor disk is shared among multiple processors.

Advantages:

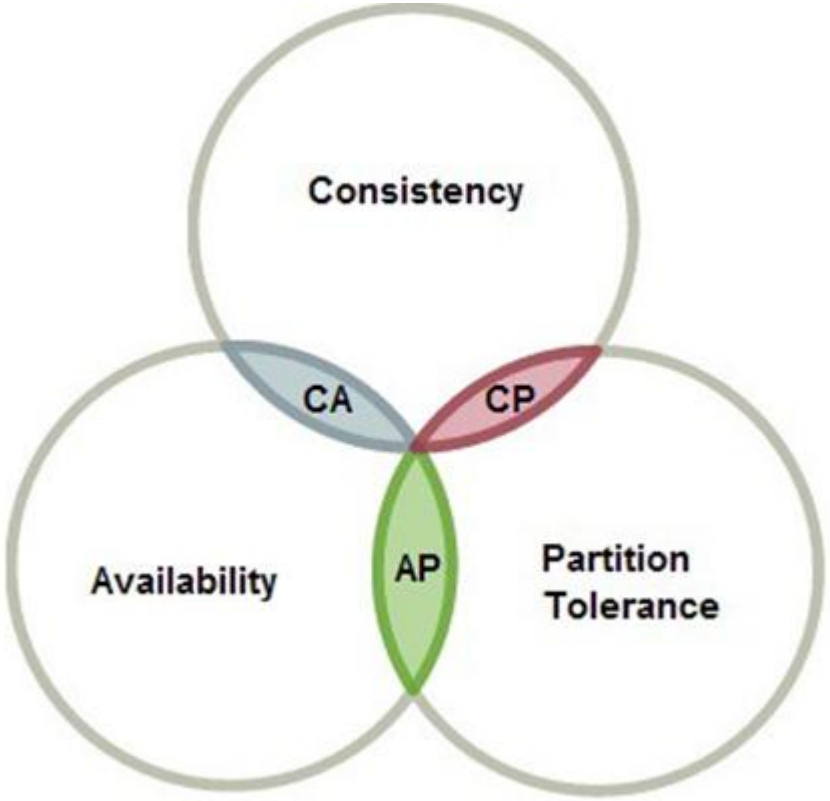
- Fault Isolation: provides the benefit of isolating fault. A fault in a single machine or node is contained and confined to that node exclusively and exposed only through messages.
- Scalability: If the disk is a shared resource, synchronization will have to maintain a consistent shared state and it means that different nodes will have to take turns to access the critical data. This imposes a limit on how many nodes can be added to the distributed shared disk system, this compromising on scalability.

Terminologies used in Big Data cont'd

CAP Theorem: In the past, when we wanted to store more data or increase our processing power, the common option was to scale vertically (get more powerful machines) or further optimize the existing code base. However, with the advances in parallel processing and distributed systems, it is more common to expand horizontally, or have more machines to do the same task in parallel. However, in order to effectively pick the tool of choice like Spark, Hadoop, Kafka, Zookeeper and Storm in Apache project, a basic idea of CAP Theorem is necessary. The CAP theorem is called the **Brewer's Theorem**. It states that a distributed computing environment can only have 2 of the 3: **C**onsistency, **A**vailability and **P**artition Tolerance – one must be sacrificed.

- **Consistency** implies that every read fetches the last write
- **Availability** implies that reads and write always succeed. In other words, each non-failing node will return a response in a reasonable amount of time
- **Partition Tolerance** implies that the system will continue to function when network partition occurs

CAP Theorem cont'd



Source: Towards Data Science

The CAP theorem categorizes systems into three categories:

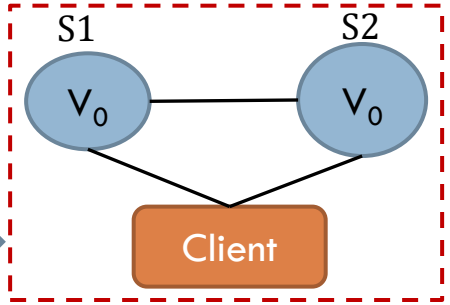
CP (Consistent and Partition Tolerant) - a system that is consistent and partition tolerant but never available. CP is referring to a category of systems where availability is sacrificed only in the case of a network partition.

CA (Consistent and Available) - CA systems are consistent and available systems in the absence of any network partition. Often a single node's DB servers are categorized as CA systems. Single node DB servers do not need to deal with partition tolerance and are thus considered CA systems.

AP (Available and Partition Tolerant) - These are systems that are available and partition tolerant but cannot guarantee consistency.

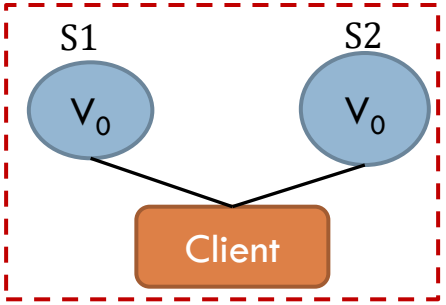
CAP Theorem Proof

Let's consider a very simple distributed system. Our system is composed of two servers, S1 and S2. Both of these servers are keeping track of the same variable, v, whose value is initially v_0. S1 and S2 can communicate with each other and can also communicate with external client. Here's what the system looks like.

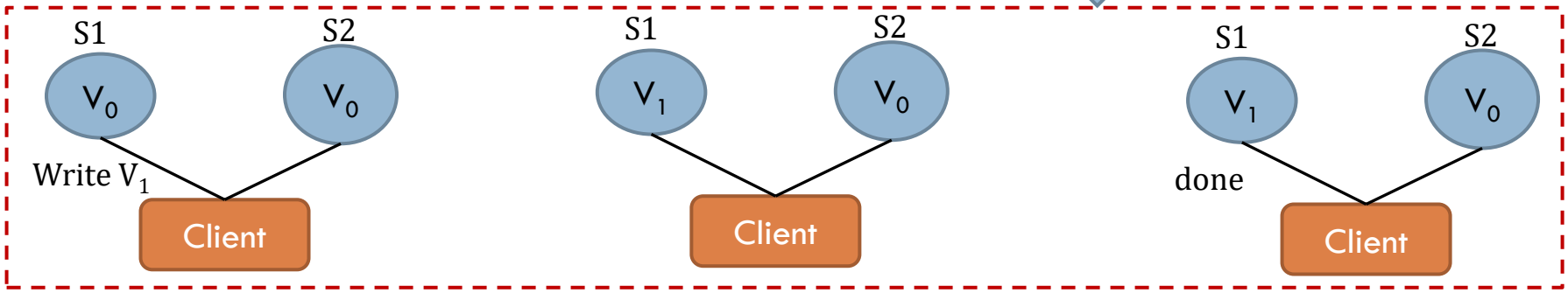


Assume for contradiction that the system is consistent, available, and partition tolerant.

The first thing we do is partition our system. It looks like this.

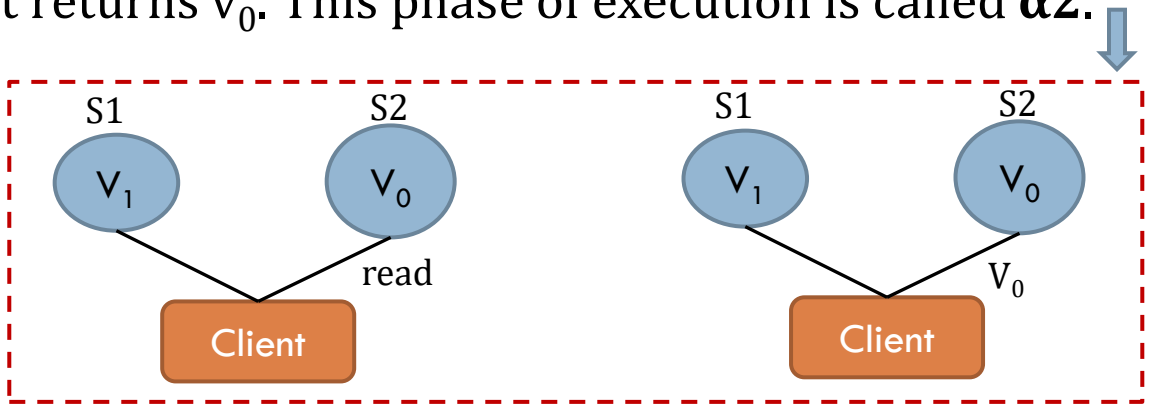


Next, the client request that v_1 be written to S1. Since the system is available, S1 must respond. Since the network is partitioned, however, S1 cannot replicate its data to S2. This phase of execution is called **α1**.



CAP Theorem Proof cont'd

Next, the client issue a read request to S2. Again, since the system is available, S2 must respond and since the network is partitioned, S2 cannot update its value from G1. It returns v_0. This phase of execution is called **α2.**



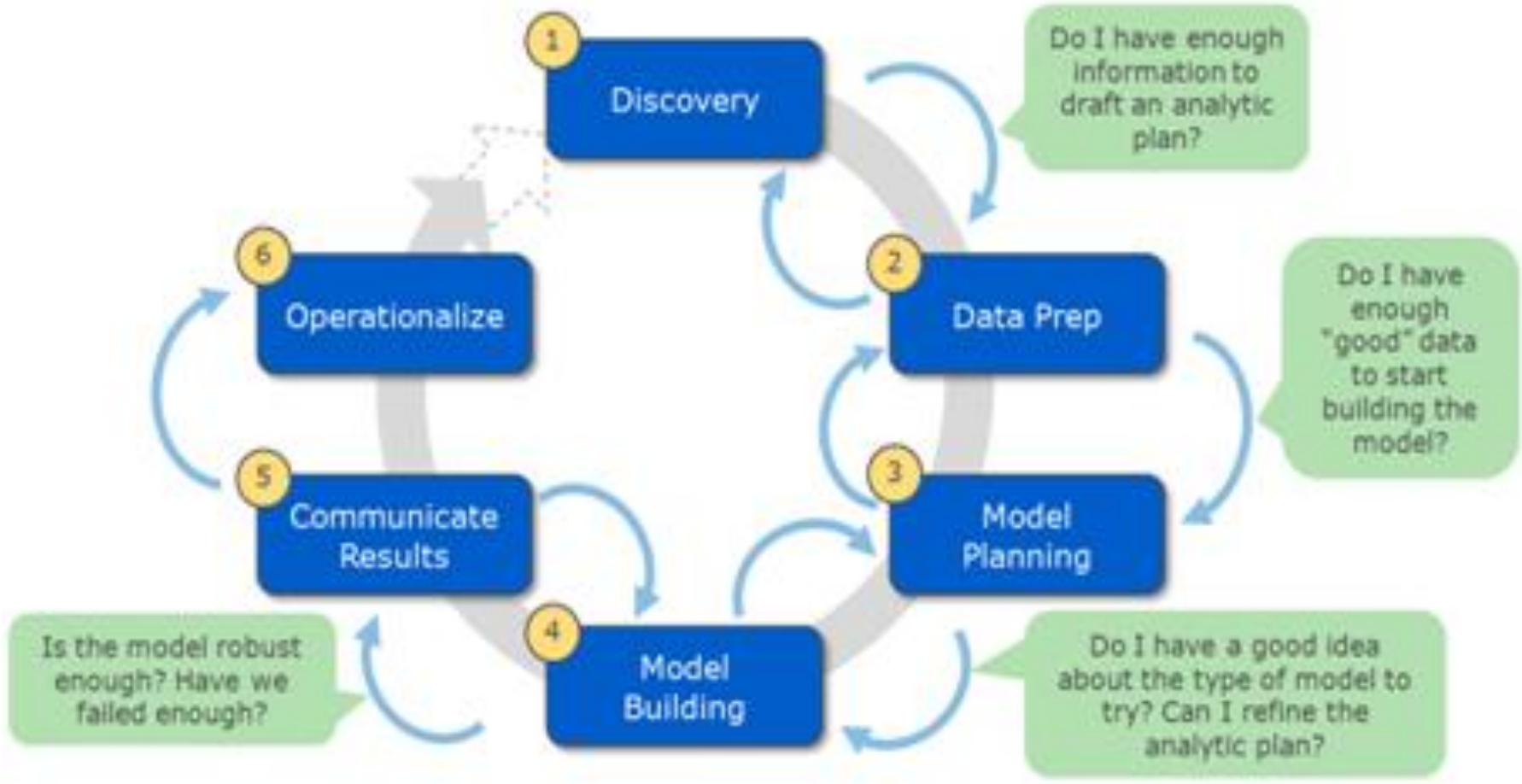
S2 returns v_0 to the client after the client had already written v_1 to G1. This is inconsistent.

We assumed a consistent, available, partition tolerant system existed, but we just showed that there exists an execution for any such system in which the system acts inconsistently. Thus, no such system exists.

Data Analytics Lifecycle

- It is a process to understand the data and apply analytics techniques to get insights for a business objective
- It's primarily defines the analytics process, and the best practices from project discovery to completion
- The data analytic lifecycle is designed for traditional data problems and data science projects
- The cycle is iterative to represent a real project
- Work can return to earlier phases as new information is uncovered
- It is a cyclical life cycle that has iterative parts in each of its six steps:

Data Analytics Lifecycle cont'd



Source: mkhernandez, data-analytics-lifecycle

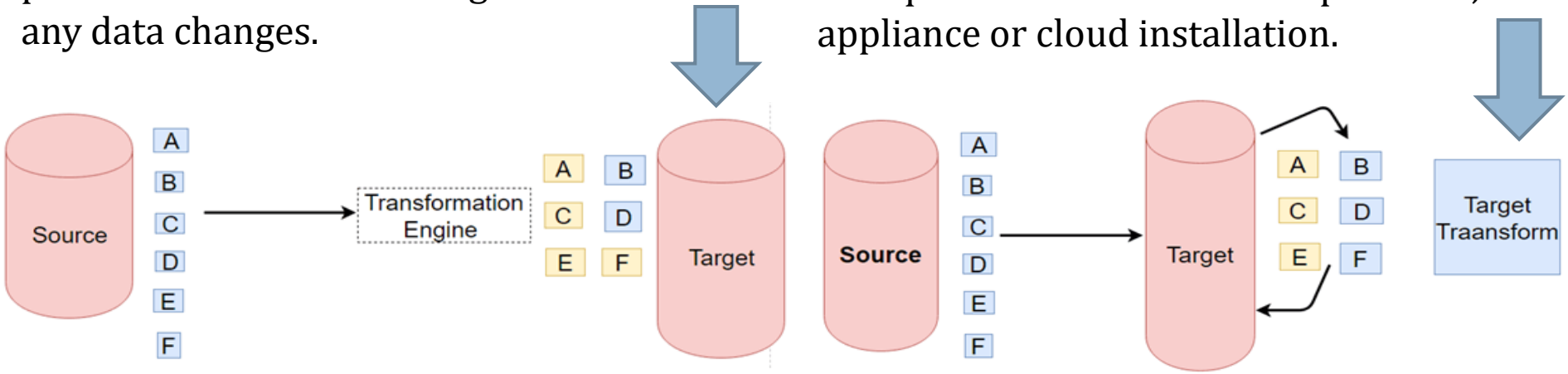
Data Analytics Lifecycle cont'd

1. Step 1 — Discovery: In this step, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this step include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.
2. Step 2— Data preparation: It requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it. In this step, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

ETL vs. ELT

ETL is an abbreviation of Extract, Transform and Load. In this process, an ETL tool extracts the data from different source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into the target system. ETL is used in RDBMS database like Oracle, Microsoft SQL Server etc. In ETL process transformation engine takes care of any data changes.

ELT is an abbreviation of Extract, Load, and Transform. ELT is a different method of looking at the tool approach to data movement. Instead of transforming the data before it's written, ELT lets the target system to do the transformation. The data first copied to the target and then transformed in place. ELT usually used with no-Sql databases like Hadoop cluster, data appliance or cloud installation.



Source: guru99, ETL vs. ELT: Must Know Differences

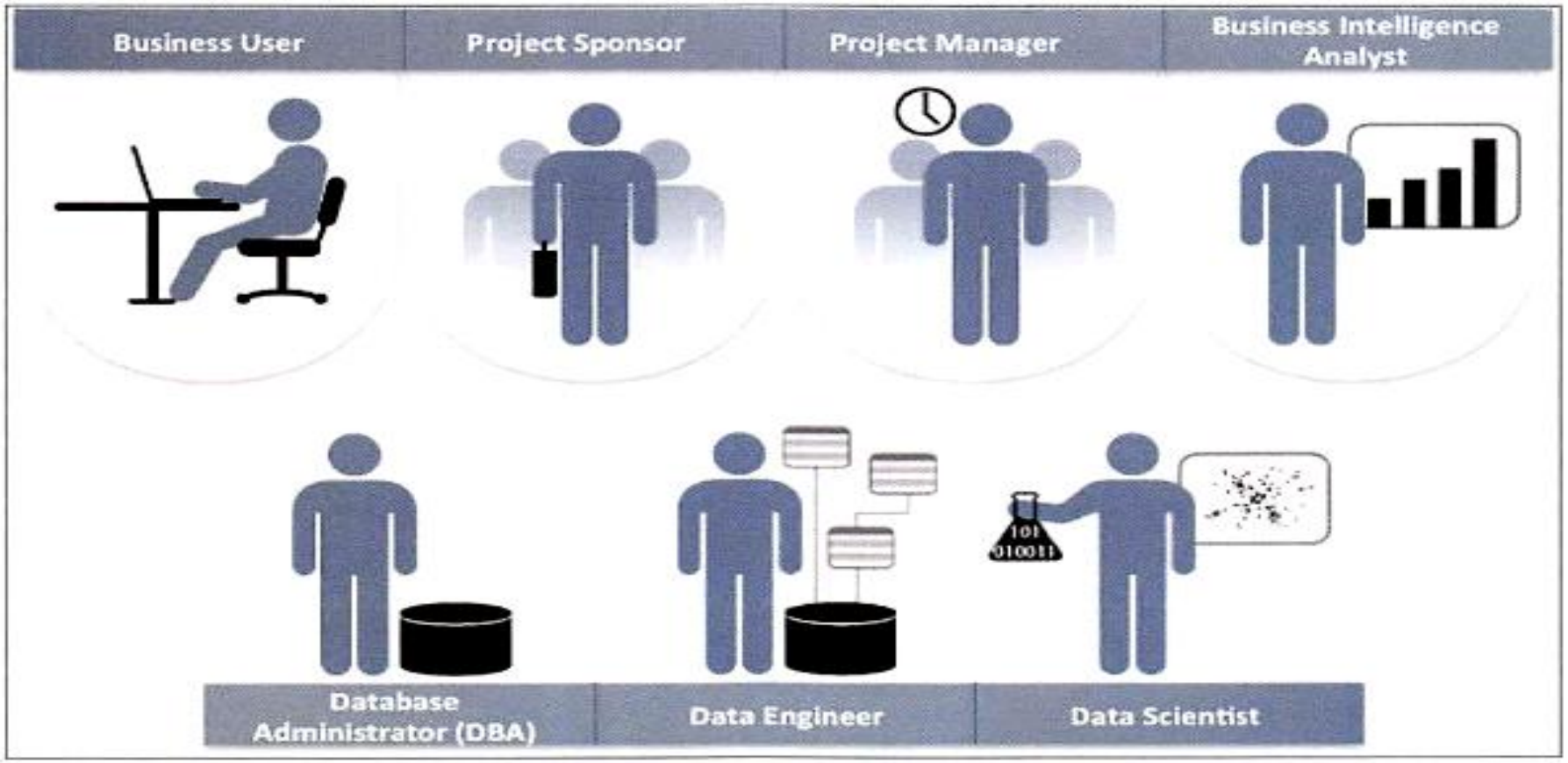
Data Analytics Lifecycle cont'd

3. Step 3 — Model planning: Step 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
4. Step 4 — Model building: In step 4, the team develops datasets for testing, training, and production purposes. In addition, in this step the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing, if applicable).

Data Analytics Lifecycle cont'd

5. Step 5 — Communicate results: In step 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in step 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.
6. Step 6 — Operationalize: In step 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.

Stakeholders in Data Analytics Project



Stakeholders in Data Analytics Project cont'd

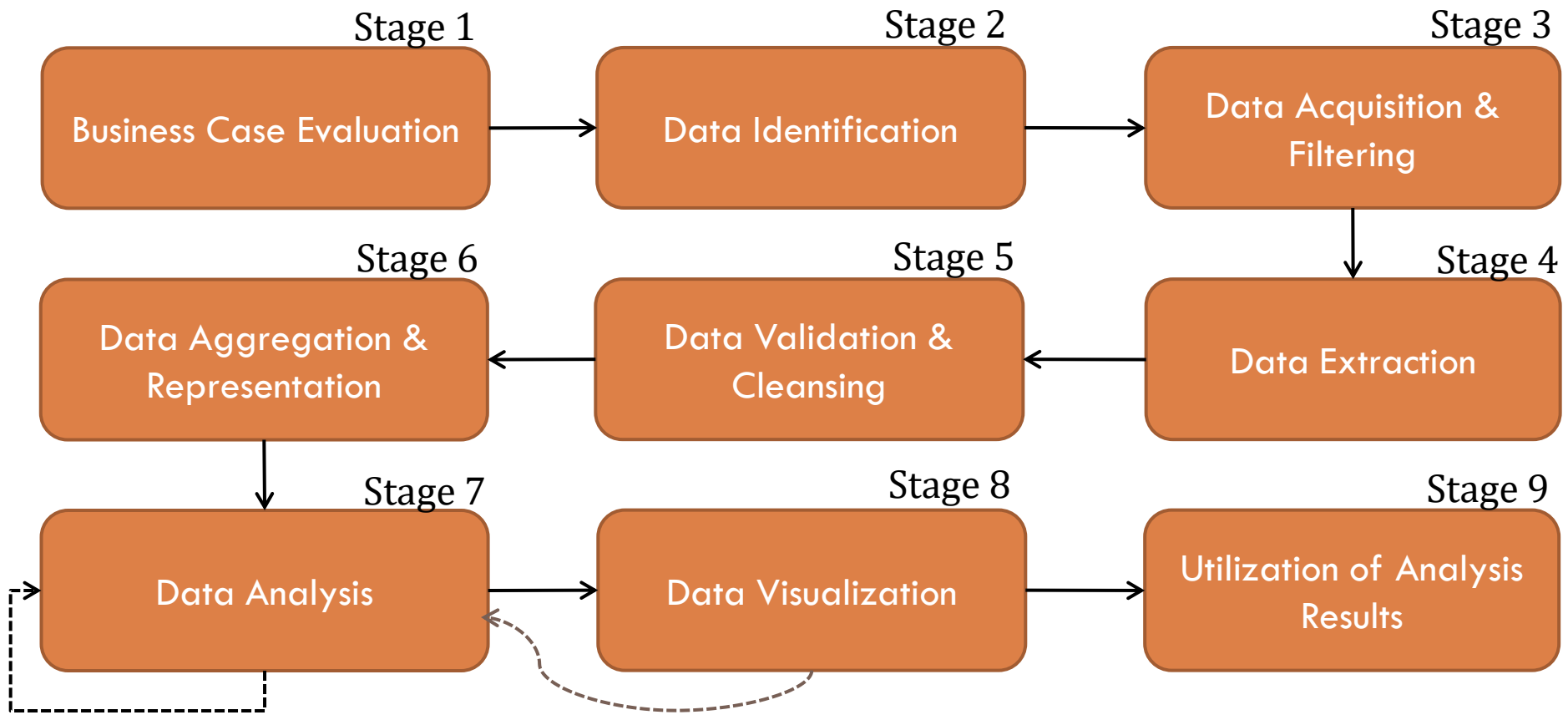
- Business User – understands the domain area
- Project Sponsor – provides requirements
- Project Manager – ensures meeting objectives
- Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data
- Database Administrator (DBA) – creates DB environment
- Data Engineer – provides technical skills, assists data management and extraction, supports analytic sandbox
- Data Scientist – provides analytic techniques and modeling

Big Data Analytics Lifecycle

- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.
- To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.
- From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.
- The Big Data analytics lifecycle can be divided into the following nine stages namely –

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Big Data Analytics Lifecycle cont'd



1. Business Case Evaluation

- Before any Big Data project can be started, it needs to be clear what the business objectives and results of the data analysis should be.
- This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition.
- A preliminary plan is designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.
- Once an overall business problem is defined, the problem is converted into an analytical problem.

2. Data Identification

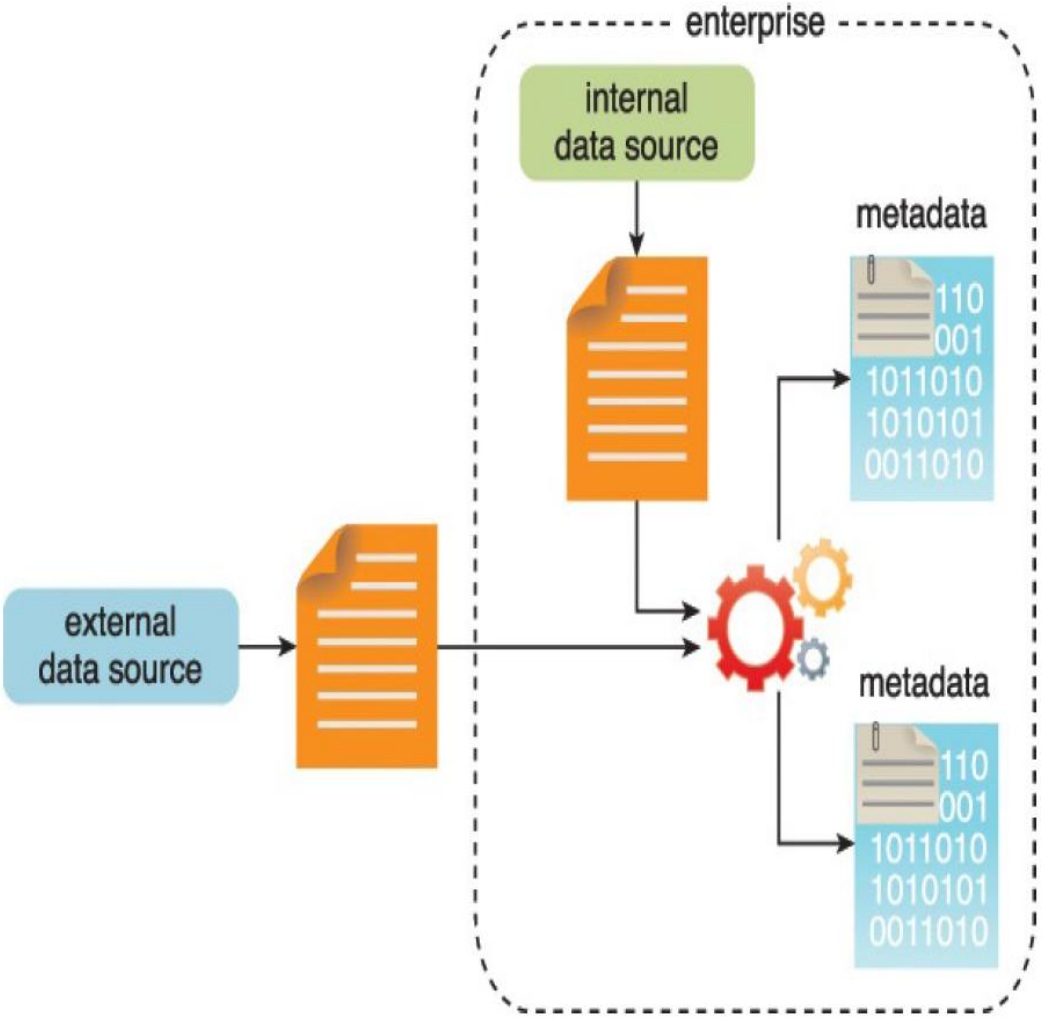
- The Data Identification stage determines the origin of data. Before data can be analysed, it is important to know what the sources of the data will be.
- Especially if data is procured from external suppliers, it is necessary to clearly identify what the original source of the data is and how reliable (frequently referred to as the veracity of the data) the dataset is.
- The second stage of the Big Data Lifecycle is very important, because if the input data is unreliable, the output data will also definitely be unreliable.
- Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations.

3. Data Acquisition and Filtering

- The Data Acquisition and Filtering Phase builds upon the previous stage of the Big Data Lifecycle.
- In this stage, the data is gathered from different sources, both from within the company and outside of the company.
- After the acquisition, a first step of filtering is conducted to filter out corrupt data.
- Additionally, data that is not necessary for the analysis will be filtered out as well.
- The filtering step will be applied on each data source individually, so before the data is aggregated into the data warehouse.
- In many cases, especially where external, unstructured data is concerned, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

3. Data Acquisition and Filtering cont'd

- Data classified as "corrupt" can include records with missing or nonsensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis.
- Metadata can be added via automation to data from both internal and external data sources to improve the classification and querying.
- Examples of appended metadata include dataset size and structure, source information, date and time of creation or collection and language-specific information.

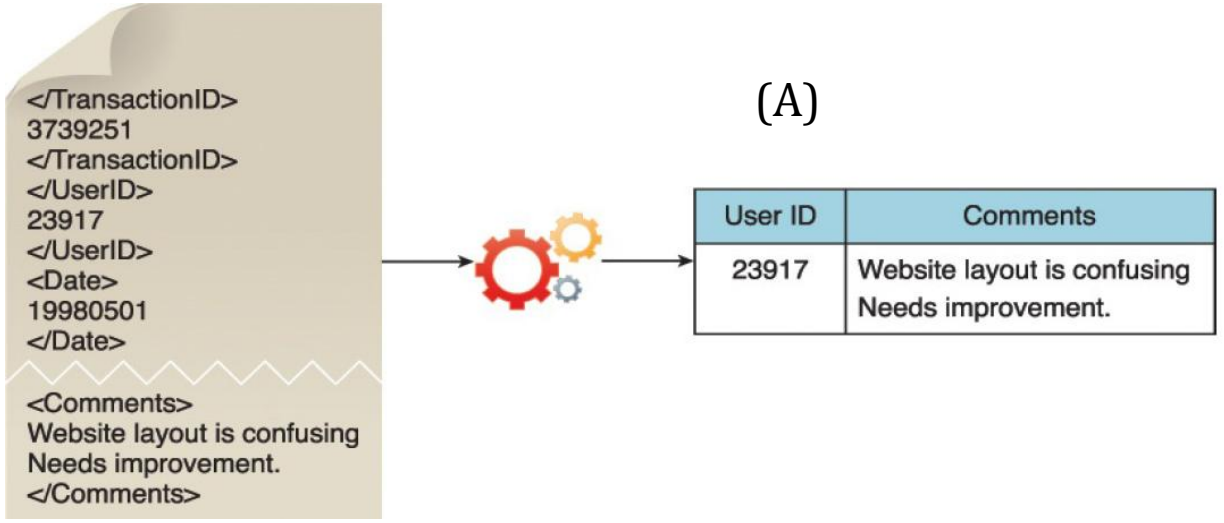


4. Data Extraction

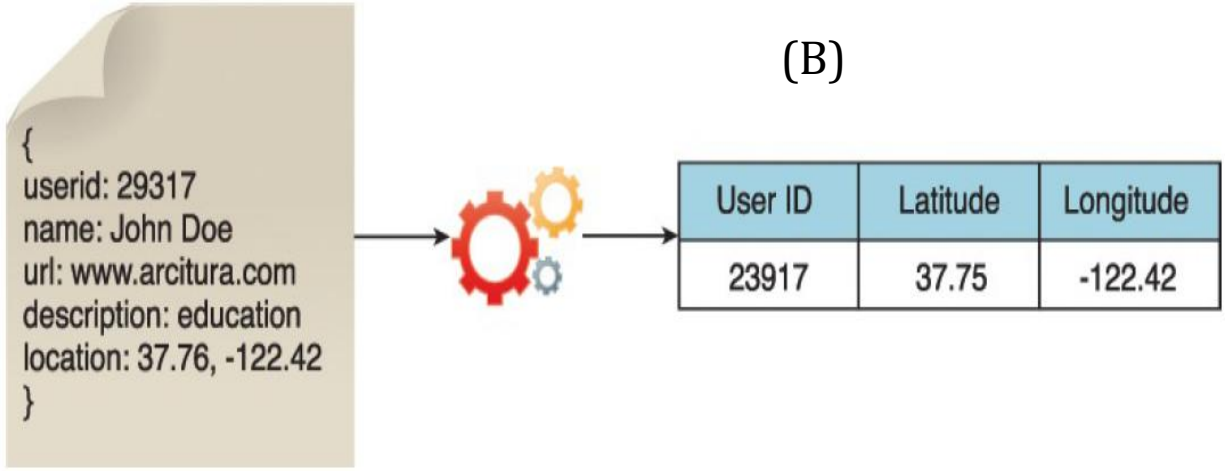
- Some of the data identified in the two previous stages may be incompatible with the Big Data tool that will perform the actual analysis.
- In order to deal with this problem, the Data Extraction stage is dedicated to extracting different data formats from data sets (e.g. the data source) and transforming these into a format the Big Data tool is able to process and analyse.
- The complexity of the transformation and the extent in which is necessary to transform data is greatly dependent on the Big Data tool that has been selected.
- The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

4. Data Extraction cont'd

- (A). Illustrates the extraction of comments and a user ID embedded within an XML document without the need for further transformation.



- (B). Demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

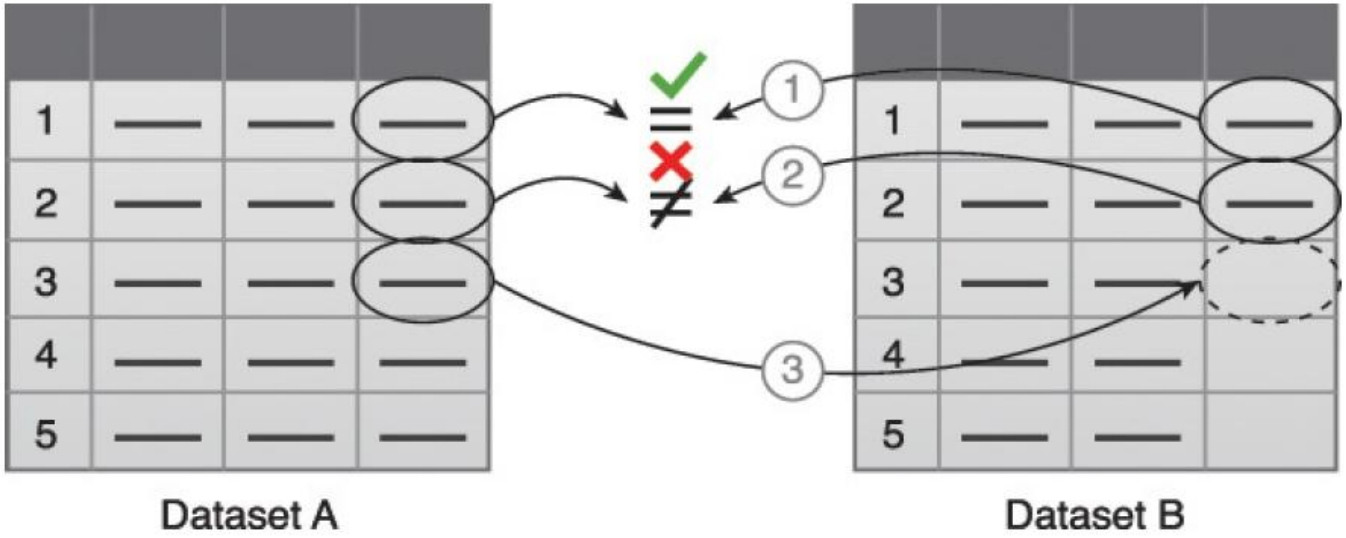


5. Data Validation and Cleansing

- Data that is invalid leads to invalid results. In order to ensure only the appropriate data is analysed, the Data Validation and Cleansing stage of the Big Data Lifecycle is required.
- During this stage, data is validated against a set of predetermined conditions and rules in order to ensure the data is not corrupt.
- An example of a validation rule would be to exclude all persons that are older than 100 years old, since it is very unlikely that data about these persons would be correct due to physical constraints.
- The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.

5. Data Validation and Cleansing cont'd

- For example, as illustrated in below figure, the first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A. If a value is missing, it is inserted from Dataset A.



- Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

6. Data Aggregation and Representation

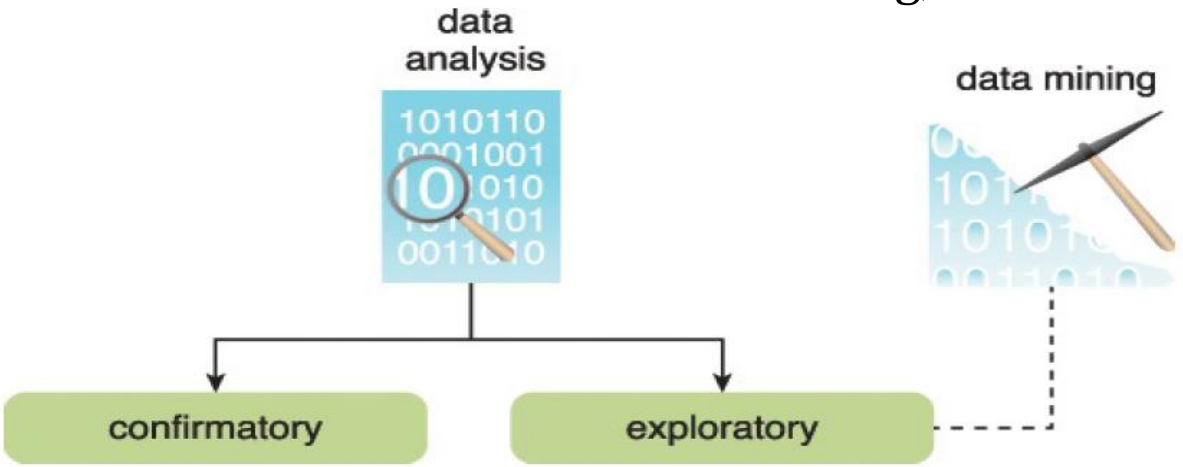
- Data may be spread across multiple datasets, requiring that dataset be joined together to conduct the actual analysis.
- In order to ensure only the correct data will be analysed in the next stage, it might be necessary to integrate multiple datasets.
- The Data Aggregation and Representation stage is dedicated to integrate multiple datasets to arrive at a unified view.
- Additionally, data aggregation will greatly speed up the analysis process of the Big Data tool, because the tool will not be required to join different tables from different datasets, greatly speeding up the process.

7. Data Analysis

- The Data Analysis stage of the Big Data Lifecycle stage is dedicated to carrying out the actual analysis task.
- It runs the code or algorithm that makes the calculations that will lead to the actual result.
- Data Analysis can be simple or really complex, depending on the required analysis type.
- In this stage the 'actual value' of the Big Data project will be generated. If all previous stages have been executed carefully, the results will be factual and correct.
- Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison.
- On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

7. Data Analysis cont'd

- Data analysis can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining, as shown below



- Confirmatory data analysis is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis.
- Exploratory data analysis is an inductive approach that is closely associated with data mining. No hypothesis or predetermined assumptions are generated. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon.

8. Data Visualization

- After the data analysis has been performed an the result have been presented, the final step of the Big Data Lifecycle is to use the results in practice.
- The utilization of Analysis results is dedicated to determining how and where the processed data can be further utilized to leverage the result of the Big Data Project.
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce “models” that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.
- A model may look like a mathematical equation or a set of rules. Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program.

8. Data Visualization cont'd

9. Utilization of Analysis Results

- After the data analysis has been performed an the result have been presented, the final step of the Big Data Lifecycle is to use the results in practice.
- The utilization of Analysis results is dedicated to determining how and where the processed data can be further utilized to leverage the result of the Big Data Project.
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce “models” that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.
- A model may look like a mathematical equation or a set of rules. Models can be used to improve business process logic and application system logic, and they can form the basis of a new system or software program.

Summary

Detailed Lessons

Importance of Data, Characteristics of Data Analysis of Unstructured Data, Combining Structured and Unstructured Sources. Introduction to Big Data Platform – Challenges of conventional systems – Web data – Evolution of Analytic scalability, analytic processes and tools, Analysis vs reporting – Modern data analytic tools, Types of Data, Elements of Big Data, Big Data Analytics, Data Analytics Lifecycle.

How was the journey?

THANK YOU!