Data Analytics (IT-3006)

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

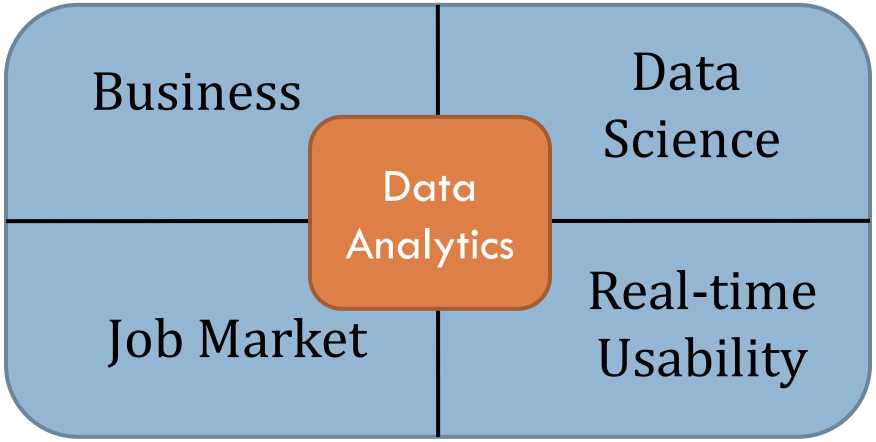
School of Computer Engineering

3 Credit

Lecture Note

Importance of the Course

- The data analytics is indeed a revolution in the field of computer.
- The use of data analytics by the companies is enhancing every year and the primary focus of the companies is on customers.
- Many organizations are actively looking out for the right talent to analyze vast amounts of data.
- Following four perspectives leads to importance of data analytics.



Why Learn Data Analytics?

- A priority for top organizations.
- Gain problem solving skills.
- High demand
  - Increasing job opportunities.
  - Increasing pay.
  - Various job titles from which to choose (Metrics and Analytics Specialist, Data Analyst, Big Data Engineer, Data Analytics Consultant)
- Analytics is everywhere.
- It's only becoming more important.
- It represents perfect freelancing opportunities.
- Develop new revenue streams

Course Contents

Sr #	Major and Detailed Coverage Area	Hrs
1	**Introduction to Big Data**	9
	Introduction to Data, Big Data Characteristics, Types of Big Data, Challenges of Traditional, Systems, Web Data, Evolution of Analytic Scalability, OLTP, MPP, Grid Computing, Cloud Computing, Fault Tolerance, Analytic Processes and Tools, Analysis Versus Reporting, Statistical Concepts, Types of Analytics.	
2	**Data Analysis**	12
	Introduction to Data Analysis, Importance of Data Analysis, Data Analytics Applications, Regression Modelling Techniques: Linear Regression, Multiple Linear Regression, Non Linear Regression, Logistic Regression, Bayesian Modelling, Basian Networks, Support Vector Machines, Time Series Analysis, Rule Induction, Sequential Cover Algorithm.	
3	**Mining Data Streams**	10
	Introduction to Mining Data Streams, Data Stream Management Systems, Data Stream Mining, Examples of Data Stream Applications, Stream Queries, Issues in Data Stream Query, Processing, Sampling in Data Streams, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Querying on Windows – Counting Ones in a Window, Decaying Windows, Real-Time Analytics Platform (RTAP).	

Course Contents continue…

Sr #	Major and Detailed Coverage Area	Hrs
4	**Frequent Itemsets and Clustering**	10
	Introduction to Frequent Itemsets, Market-Basket Model, Algorithm for Finding Frequent, Itemsets, Association Rule Mining, Apriori Algorithm, Introduction to Clustering, Overview of Clustering Techniques, Hierarchical Clustering, Partitioning Methods, K- Means Algorithm, Clustering High-Dimensional Data.	
5	**Frameworks and Visualization**	8
	Introduction to framework and Visualization, Introduction to Hadoop, Core Components of Hadoop, Hadoop Ecosystem, Physical Architecture, Hadoop Limitations, Hive, MapReduce and The New Software Stack, MapReduce, Algorithms Using MapReduce, NOSQL, NoSQL Business Drivers, NoSQL Case Studies, NoSQL Data Architectural Patterns, Variations of NoSQL, Architectural Patterns, Using NoSQL to Manage Big Data, Visualizations	

Course Outcome

CO #	CO	Unit
CO1	Understand and classify the characteristics, concepts and principles of big data.	Introduction to Big Data
CO2	Apply the data analytics techniques and models.	Data Analysis
CO3	Implement and analyze the data analysis techniques for mining data streams.	Mining Data Streams
CO4	Examine the techniques of clustering and frequent item sets.	Frequent Itemsets and Clustering
CO5	Analyze and evaluate the framework and visualization for big data analytics.	Frameworks and Visualization
CO6	Formulate the concepts, principles and techniques focusing on the applications to industry and real world experience.	Applications of all units

Prerequisites

- NIL

Books

Textbook

- Data Analytics, Radha Shankarmani,M. Vijayalaxmi, Wiley India Private Limited, ISBN: 9788126560639.

Reference Books

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (Editor), Wiley, 2014
- Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with advanced analystics, John Wiley & sons, 2012.
- Glenn J. Myatt, Making Sense of Data, John Wiley & Sons, 2007 Pete Warden, Big
- Data Glossary,O'Reilly, 2011.
- Jiawei Han, MichelineKamber "Data Mining Concepts and Techniques", Second Edition, Elsevier, Reprinted 2008.
- Stephan Kudyba, Thomas H. Davenport, Big Data, Mining, and Analytics, Components of Strategic Decision Making, CRC Press, Taylor & Francis Group. 2014
- Big Data, Black Book, DT Editorial Services, Dreamtech Press, 2015

Evaluation

Grading:

- Internal assessment – 30 marks
  - 2 quizzes = 2.5 X 2 = 5 marks
  - 5 group assignments = 2 X 5 = 10 marks
  - Class participation = 5 marks
  - Mini Project = 10 marks
- Mid-Term exam - 20 marks
- End-Term exam - 50 marks

Data

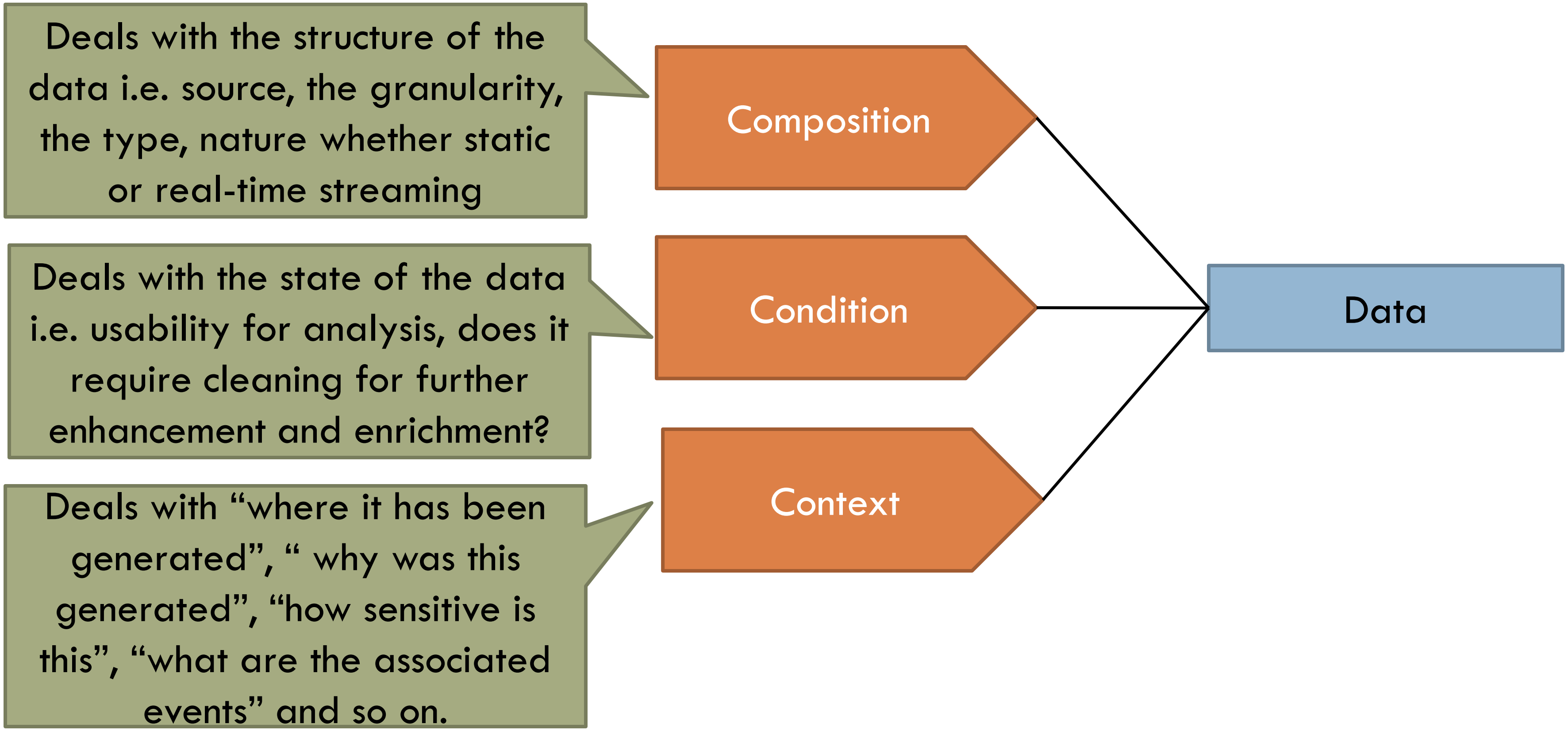
- A representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalized manner.
- Data is either intended to be processed, is being processed, or has been processed.
- It can be in any form stored internally in a computer system or 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



Composition

Deals with the structure of the data i.e. source, the granularity, the type, nature whether static or real-time streaming

Condition

Deals with the state of the data i.e. usability for analysis, does it require cleaning for further enhancement and enrichment?

Context

Deals with "where it has been generated", " why was this generated", "how sensitive is this", "what are the associated events" and so on.

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.

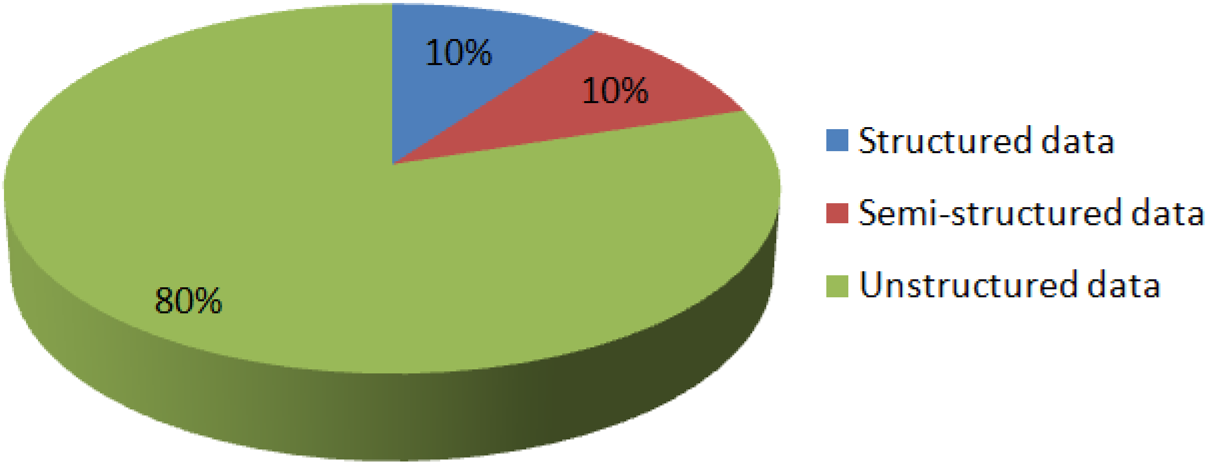
Another example scans (photographs) of text are not machine-readable (but are human readable!) but the equivalent text in a format such as a simple ASCII text file can machine readable and processable.

Classification of Digital Data

Digital data is classified into the following categories:

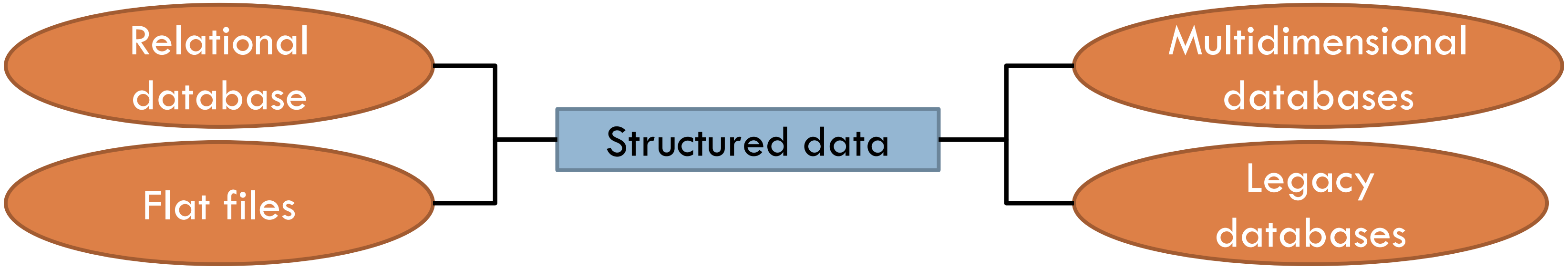
- ❑ Structured data
- ❑ Semi-structured data
- ❑ Unstructured data

Approximate percentage distribution of digital data

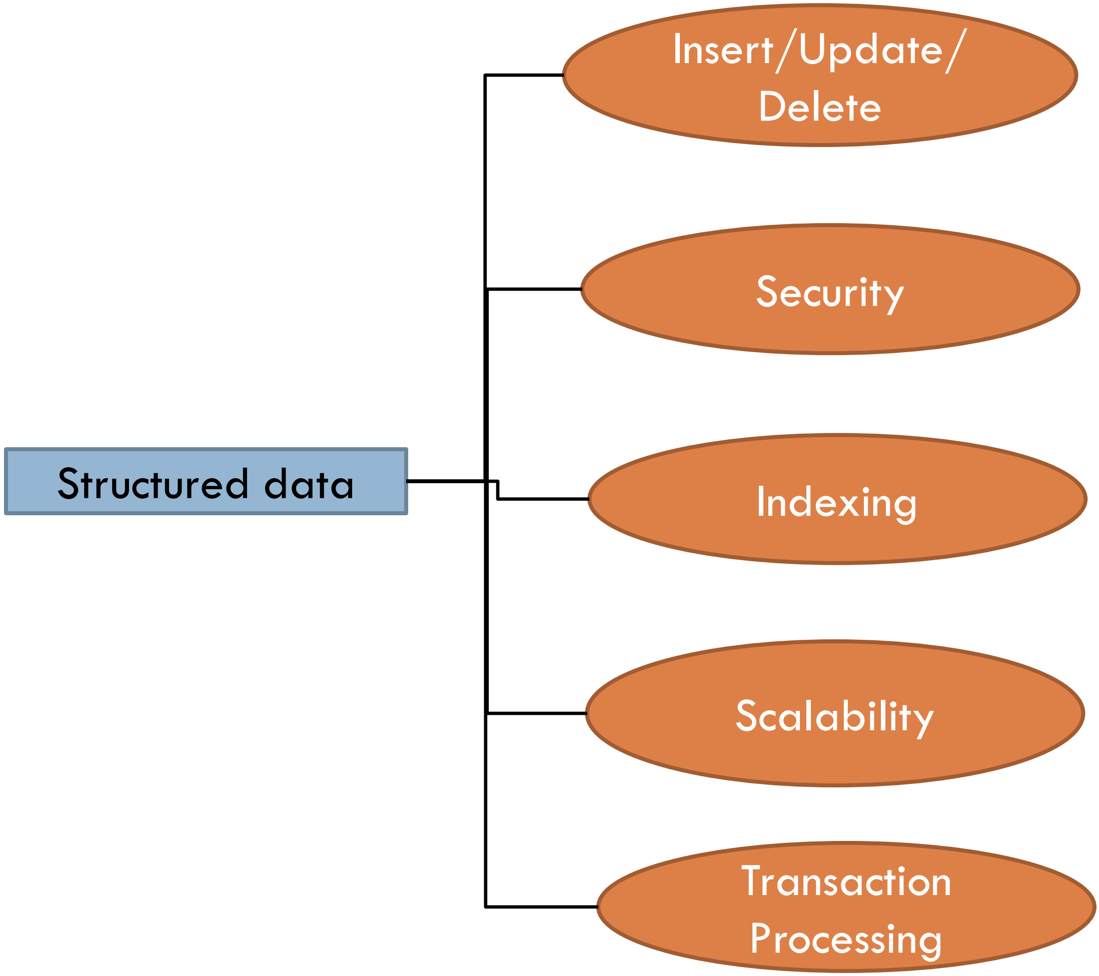


Structured Data

- 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



DML operations provide the required ease with data input, storage, access, process , analysis etc.

Encryption and tokenization solution to warrant the security of information throughout life cycle. Organization able to retain control and maintain compliance adherence by ensuring that only authorized are able to decrypt and view sensitive information.

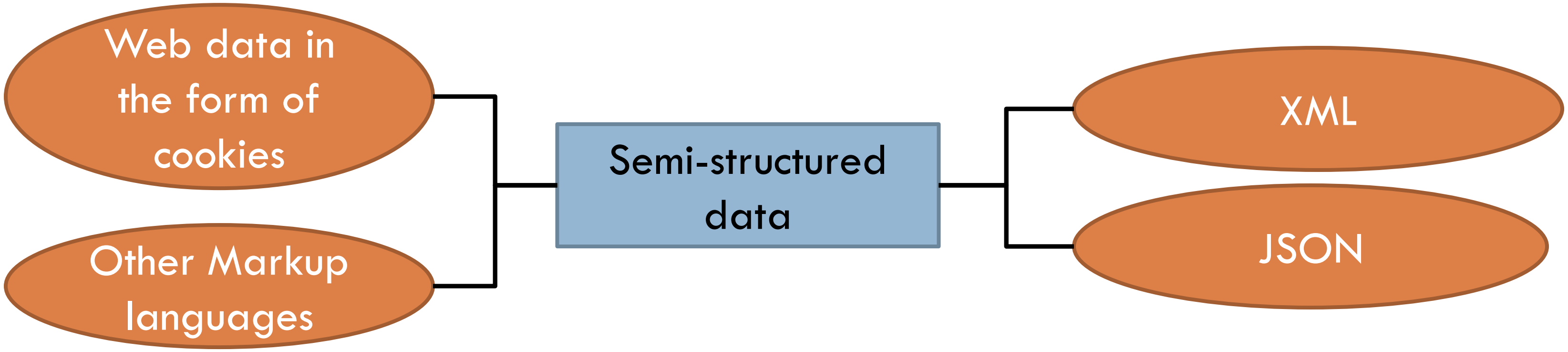
Indexing speed up the data retrieval operation at the cost of additional writes and storage space, but the benefits that ensure in search operation are worth the additional writes and storage spaces.

The storage and processing capabilities of the traditional DBMS can be easily be scaled up by increasing the horsepower of the database server

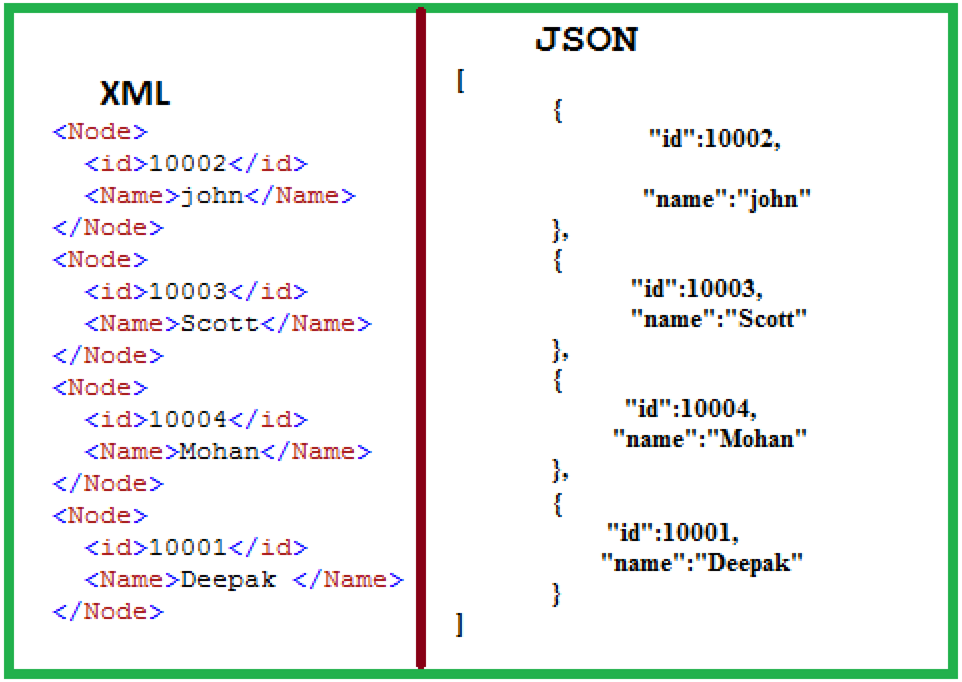
RDBMS has support of ACID properties of transaction to ensure accuracy, completeness and data integrity

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



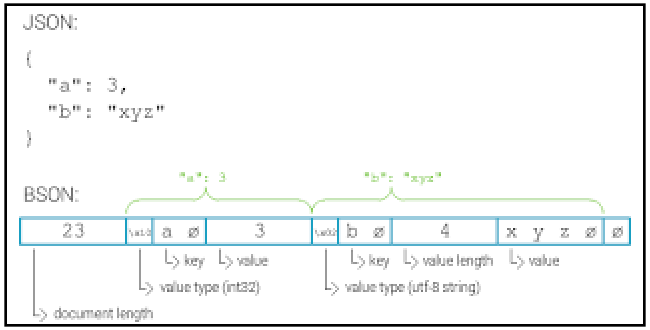
XML

```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

```json
[
      {
            "id":10002,

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



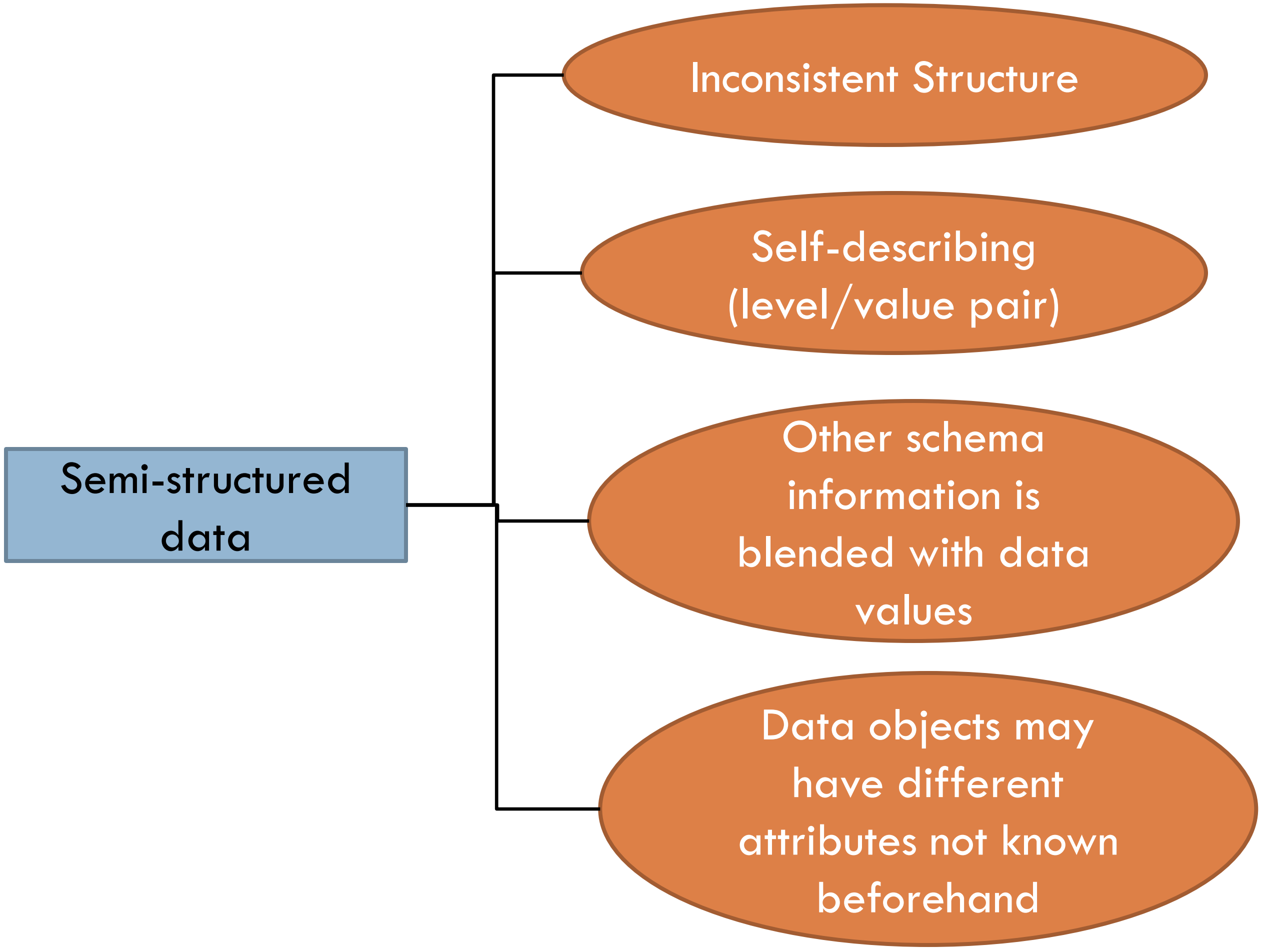
JSON:

```json
{
  "a": 3,
  "b": "xyz"
}
```

BSON:

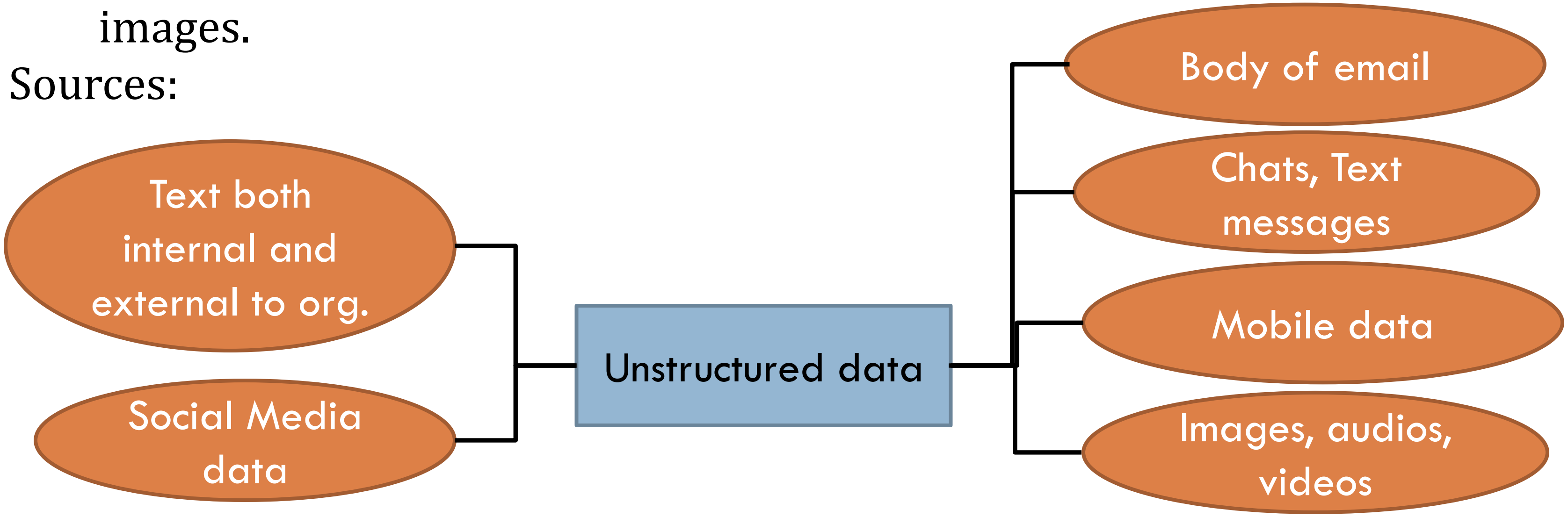
Source (XML & JSON): http://sqllearnergroups.blogspot.com/2014/03/how-to-get-json-format-through-sql.html
Source (JSON & BSON): http://www.expert-php.fr/mongodb-bson/

Characteristics of Semi-structured Data


Semi-structured data
Inconsistent Structure
Self-describing (level/value pair)
Other schema information is blended with data values
Data objects may have different attributes not known beforehand

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

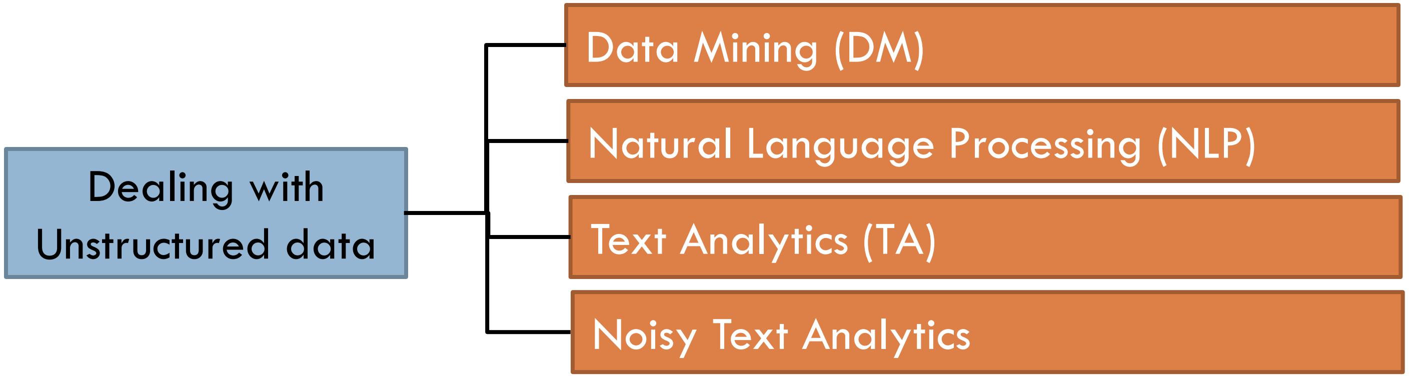
Working with unstructured data poses certain challenges, which are as follows:

- 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



Dealing with Unstructured data

Data Mining (DM)

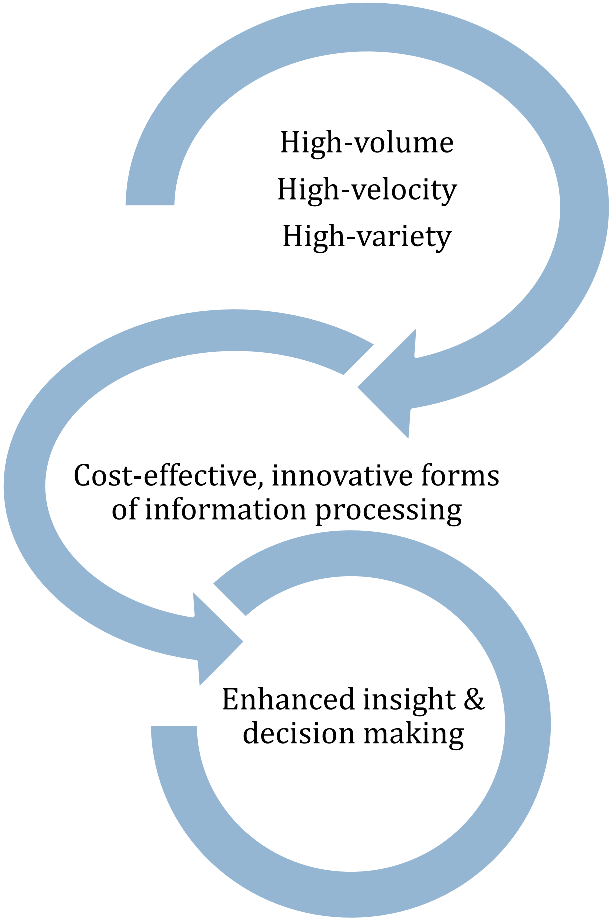
Natural Language Processing (NLP)

Text Analytics (TA)

Noisy Text Analytics

Note: Refer to Appendix for further details.

Definition of Big Data



High-volume
High-velocity
High-variety

Cost-effective, innovative forms of information processing

Enhanced insight & decision making

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.

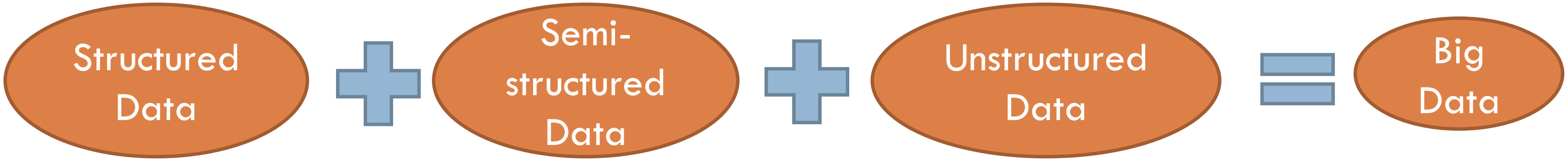
Source: Gartner IT Glossary

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**’.



Structured Data + Semi-structured Data + Unstructured Data = Big Data

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.

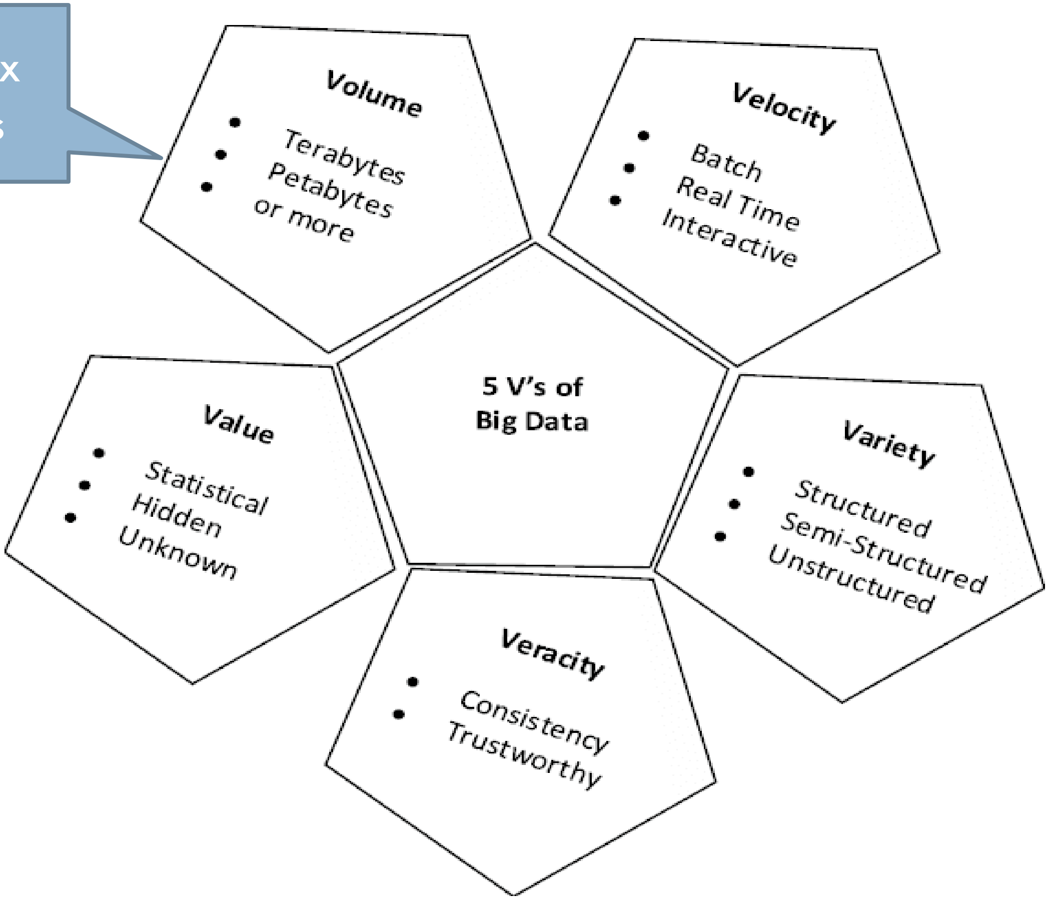
Velocity - refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every second of every day data is increasing. Not only must it be analyzed, but the speed of transmission, and access to the data must also remain instantaneous to allow for real-time access. Big 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

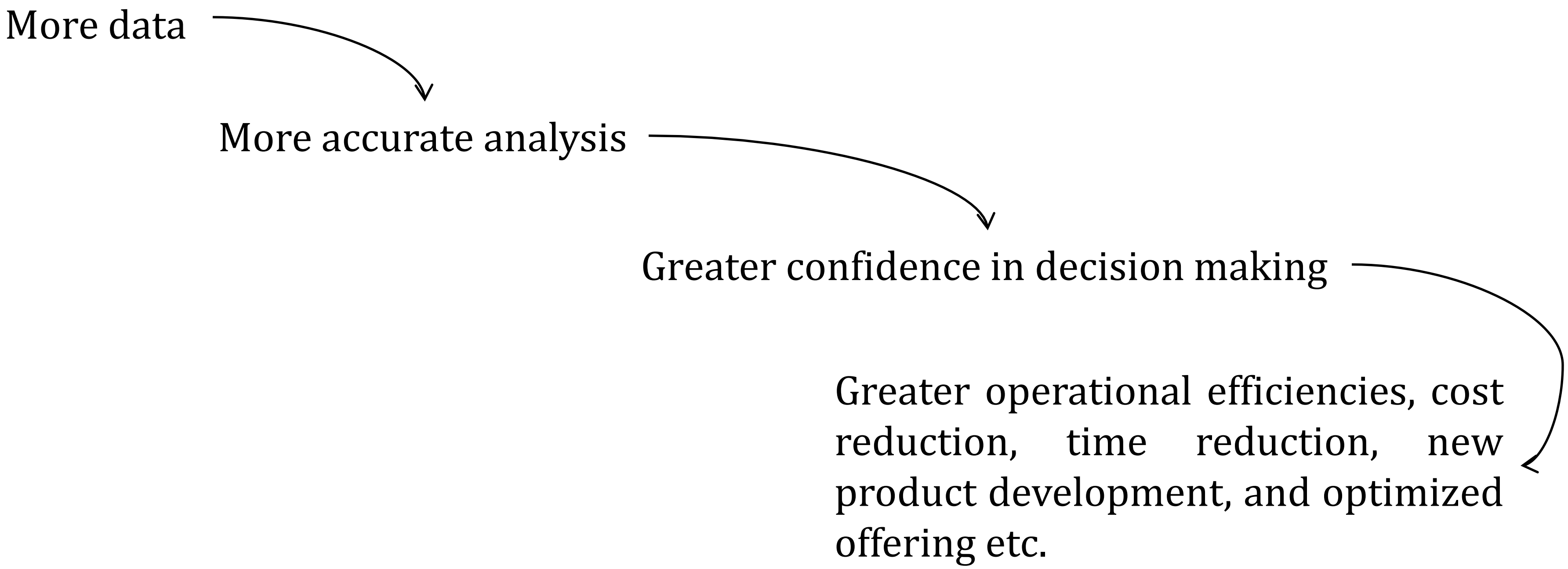
Value - refers to the ability to transform a tsunami of data into business. Having endless amounts of data is one thing, but unless it can be turned into value it is useless.

Refer to Appendix for data volumes



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.



More data

More accurate analysis

Greater confidence in decision making

Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offering etc.

Challenges of Traditional 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 Traditional 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

Web Data

- It refers to the data that is publicly available on the web sites.
- The web data has documents in pdf, doc, docx, plain text as well as images, music, and videos.
- The most widely used and best-known source of big data today is the detailed data collected from web sites.
- The data is unstructured and inappropriate for access by software application, and hence is converted to either semi-structured or structured format that is well suited for both humans and machines.

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**)

EDW, OLTP, MPP

- **Enterprise Data Warehouse:** An enterprise data warehouse (EDW) is a database, or collection of databases, that centralizes a business's information from multiple sources and applications, and makes it available for analytics and use across the organization. EDWs can be housed in an on-premise server or in the cloud. The data stored in this type of digital warehouse can be one of a business's most valuable assets, as it represents much of what is known about the business, its employees, its customers, and more.
- **Online Transactional Processing (OLTP):** It is a category of data processing that is focused on transaction-oriented tasks. OLTP typically involves inserting, updating, and/or deleting small amounts of data in a database. OLTP mainly deals with large numbers of transactions by a large number of users.
- **Massively Parallel Processing (MPP):** It is a storage structure designed to handle the coordinated processing of program operations by multiple processors. This coordinated processing can work on different parts of a program, with each processor using its own operating system and memory. This allows MPP databases to handle massive amounts of data and provide much faster analytics based on large datasets.

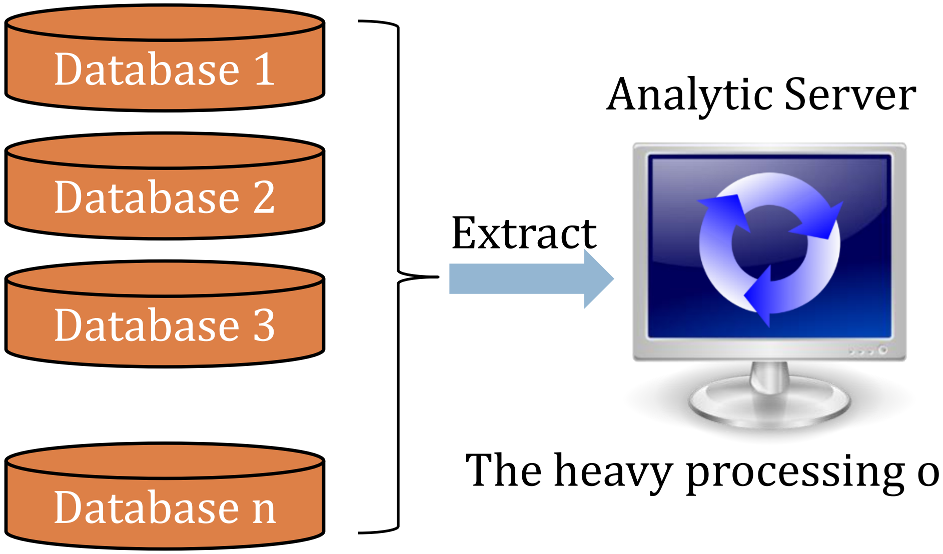
Hadoop

- Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
- It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.
- Importance:
  - Ability to store and process huge amounts of any kind of data, quickly.
  - **Computing power:** It's distributed computing model processes big data fast.
  - **Fault tolerance:** Data and application processing are protected against hardware failure.
  - **Flexibility:** Unlike traditional relational databases, preprocess of data does not require before storing it.
  - **Low cost:** The open-source framework is free and uses commodity hardware to store large quantities of data.
  - Scalability: System can easily grow to handle more data simply by adding nodes. Little administration is required.

Evolution of Analytics Scalability

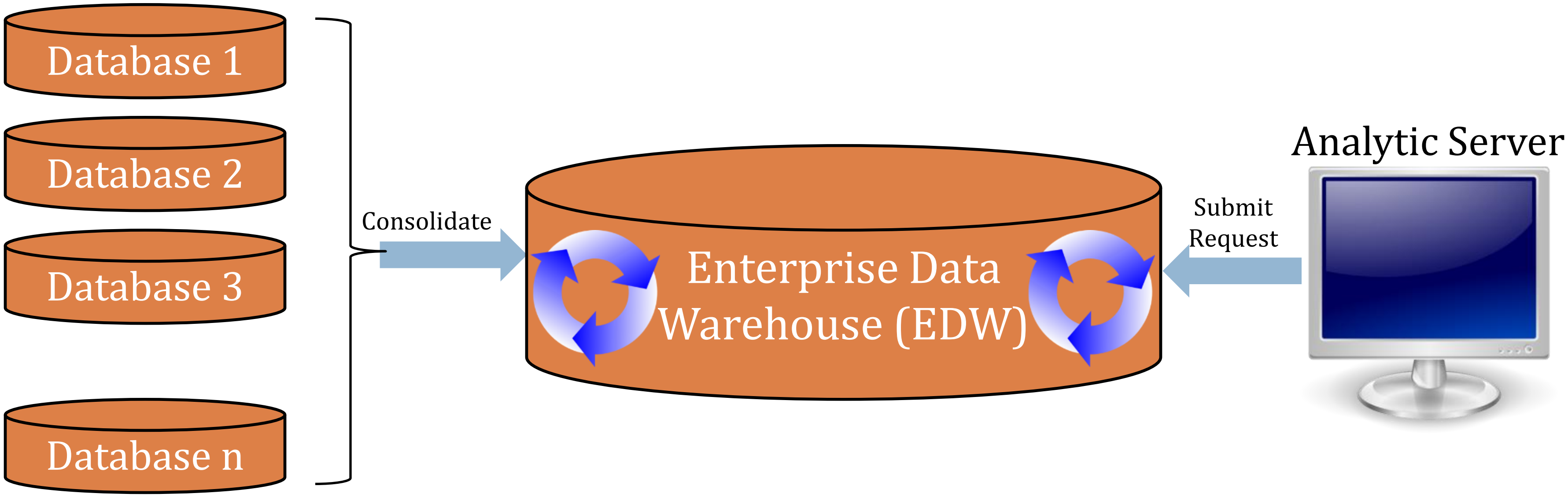
- As the amount of data organizations process continue to increase, the world of big data requires new levels of scalability. Organizations need to update the technology to provide a higher level of scalability.
- Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.
- The technologies are:
  - MPP (massively parallel processing)
  - Cloud computing
  - Grid computing
  - MapReduce

Traditional Analytics Architecture



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

Modern In-Database Analytics Architecture

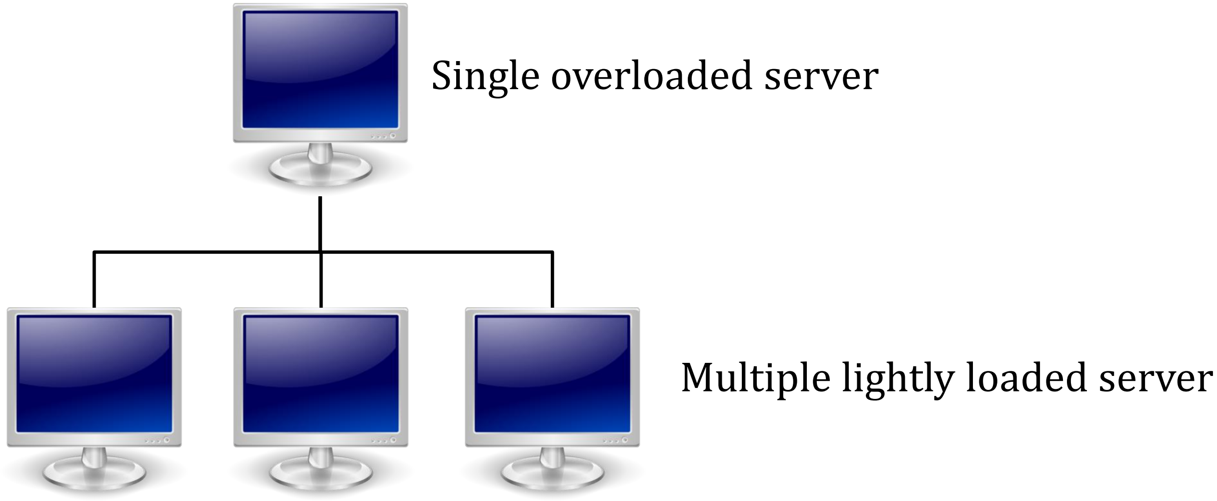


In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

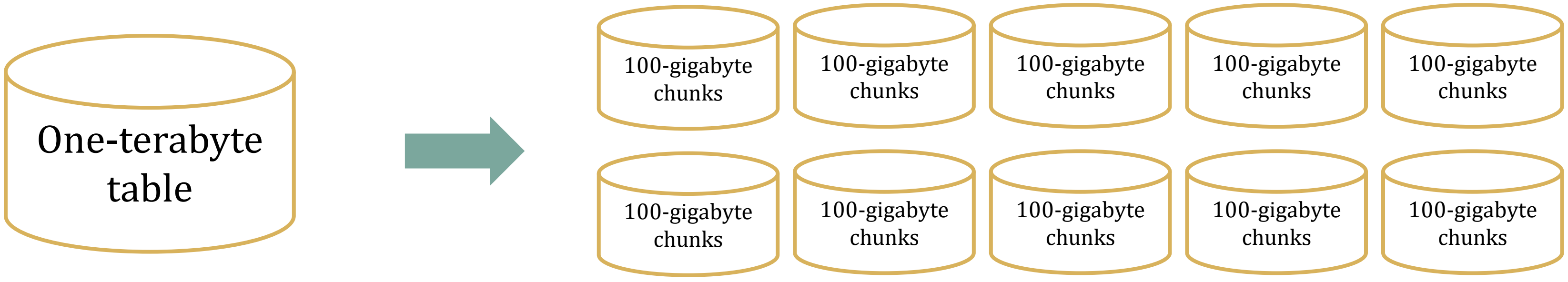
MPP Analytics Architecture

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.



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!**

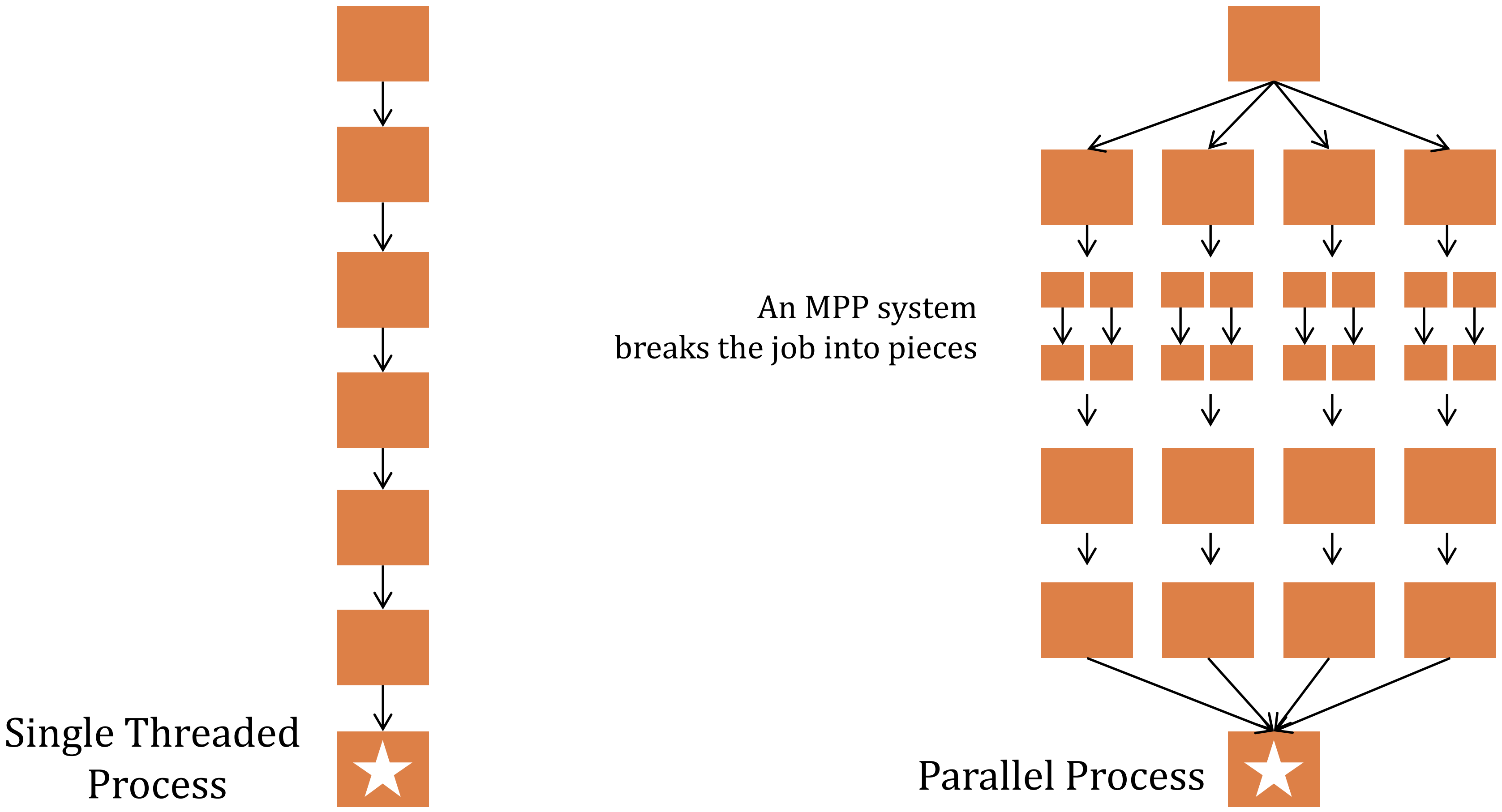
A MPP database spreads data out across multiple sets of CPU and disk space. Think logically about dozens or hundreds of personal computers each holding a small piece of a large set of data. This allows much faster query execution, since many independent smaller queries are running simultaneously instead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



OLTP vs. MPP vs. Hadoop

OLTP	MPP
Examples: Oracle, DB2, SQL Server etc.	Examples: Netezza, Teradata, Vertica etc.
It needs to read data from disk to memory before start processing, so very fast in memory calculation.	Takes the processing as close possible to the data, so less data movement
It is good for smaller OLTP (transaction) operations. It also maintains very high level of data integrity.	It is good for batch processing. Some of the MPP (Netezza, Vertica) overlooks integrity like enforcing unique key for the sake of batch performance.

MPP	Hadoop
Stores data in a matured internal structure. So data loading and data processing is efficient.	There are no such structured architecture for data stored on Hadoop. So, accessing and loading data is not as efficient as conventional MPP systems.
It support only relational models.	Support virtually any kind of data.
However the main objective of MPP and Hadoop is same, process data parallely near storage.	

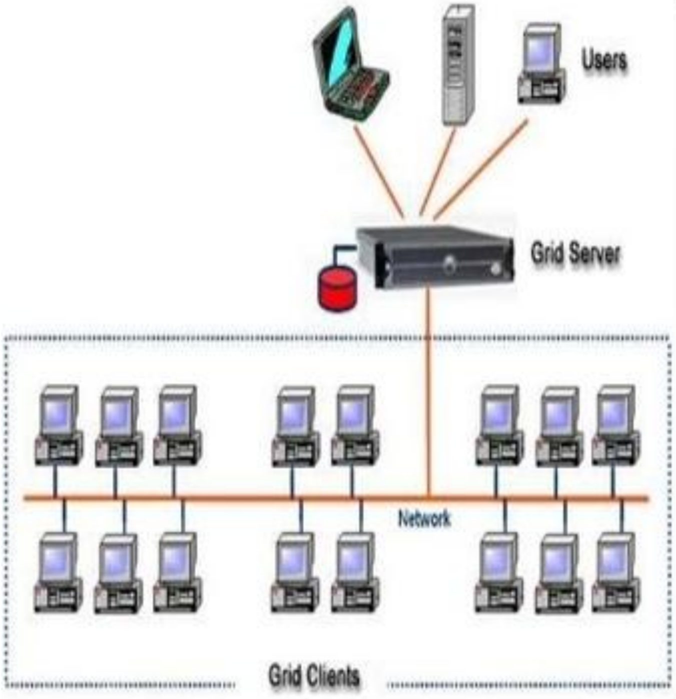
How to choose what?

- OLTP Databases (Oracle,DB2, MySQL, MS SQL, Exadata):
  - Transaction based application
  - Smaller DWH
- MPP (Netezza, Teradata, Vertica):
  - Bigger Data warehouse (may be having tables with size more than 4-5 TB)
  - Needs no or little pre-processing
  - Needs faster batch processing speed
  - In database analytics
- Only Hadoop:
  - All data as heavily unstructured (documents, audio, video etc)
  - Need to process in batch

Grid Computing

- ❑ Grid Computing can be defined as a network of computers working together to perform a task that would rather be difficult for a single machine.
- ❑ The task that they work on may include analysing huge datasets or simulating situations which require high computing power.
- ❑ Computers on the network contribute resources like processing power and storage capacity to the network.
- ❑ Grid Computing is a subset of distributed computing, where a virtual super computer comprises of machines on a network connected by some bus, mostly Ethernet or sometimes the Internet.
- ❑ It can also be seen as a form of parallel computing where instead of many CPU cores on a single machine, it contains multiple cores spread across various locations.

How Grid Computing works?



In general, a grid computing system requires:

- **At least one computer, usually a server, which handles all the administrative duties for the System**
- **A network of computers running special grid computing network software.**
- **A collection of computer software called middleware**

Cloud Computing

- ❑ It is a internet-based computing and relies on sharing computing resources on-demand rather than having local PCs and other devices.
- ❑ It is the delivery of on-demand computing services - from applications to storage and processing power over the internet and on a pay-as-you-go basis.
- ❑ It uses high-capacity networks, low-cost computers, and storage devices and adopts hardware virtualization, service-oriented architecture, and utility computing.
- ❑ Rather than owning their own computing infrastructure or data centers, companies can rent access to anything from applications to storage from a cloud service provider and can scale up and scale down as per their computing demands.
- ❑ There are 3 types of cloud environment named public cloud, private cloud and hybrid cloud.

Public Cloud

- ❑ It is the most common type of cloud computing deployment.
- ❑ The cloud resources (like servers and storage) are owned and operated by a third-party cloud service provider and delivered over the internet.
- ❑ With a public cloud, all hardware, software and other supporting infrastructure are owned and managed by the cloud provider.
- ❑ In a public cloud, the same hardware, storage and network devices are shared with other organizations or cloud “tenants,” and the adopter access services and manage account using a web browser.
- ❑ Public cloud deployments are frequently used to provide web-based email, online office applications, storage and testing and development environments.
- ❑ Advantages of public clouds are lower costs, no maintenance, high reliability etc.

Private Cloud

- A private cloud consists of cloud computing resources used exclusively by one business or organization.
- The private cloud can be physically located at your organization's on-site datacenter or it can be hosted by a third-party service provider.
- The services and infrastructure are always maintained on a private network and the hardware and software are dedicated solely to the organisation.
- It is often used by government agencies, financial institutions, any other mid- to large-size organizations with business-critical operations seeking enhanced control over their environment.
- Advantages of private clouds are more flexibility, more control, and more scalability etc.

Hybrid Cloud

- A hybrid cloud combines on-premises infrastructure or a private cloud with a public cloud.
- It allow data and apps to move between the two environments.
- Many organizations choose a hybrid cloud approach due to business imperatives such as meeting regulatory and data sovereignty requirements, taking full advantage of on-premises technology investment or addressing low latency issues.
- A hybrid cloud platform gives organizations many advantages—such as greater flexibility, more deployment options, security, compliance and getting more value from their existing infrastructure.
- When computing and processing demand fluctuates, hybrid cloud computing gives businesses the ability to seamlessly scale up their on-premises infrastructure to the public cloud to handle any overflow—without giving third-party datacenters access to the entirety of their data.

Fault Tolerance

- Fault tolerance refers to the ability of a system (computer, network, cloud cluster, etc.) to continue operating without interruption when one or more of its components fail.
- The objective of creating a fault-tolerant system is to prevent disruptions arising from a single point of failure, ensuring the high availability and business continuity of mission-critical applications or systems.
- Fault-tolerant systems use backup components that automatically take the place of failed components, ensuring no loss of service. These include:
  - **Hardware systems** that are backed up by identical or equivalent systems. For example, a server can be made fault tolerant by using an identical server running in parallel, with all operations mirrored to the backup server.
  - **Software systems** that are backed up by other software instances. For example, a database with customer information can be continuously replicated to another machine. If the primary database goes down, operations can be automatically redirected to the second database.
  - **Power sources** that are made fault tolerant using alternative sources. For example, many organizations have power generators that can take over in case main line electricity fails.

Analytic Processes and Tools

Self-Study from the book

Points to cover

- ❑ Spreadsheets and Analytics Tool
- ❑ Analytics Engine
- ❑ CRM and Online Marketing Solutions

Analysis vs. Reporting

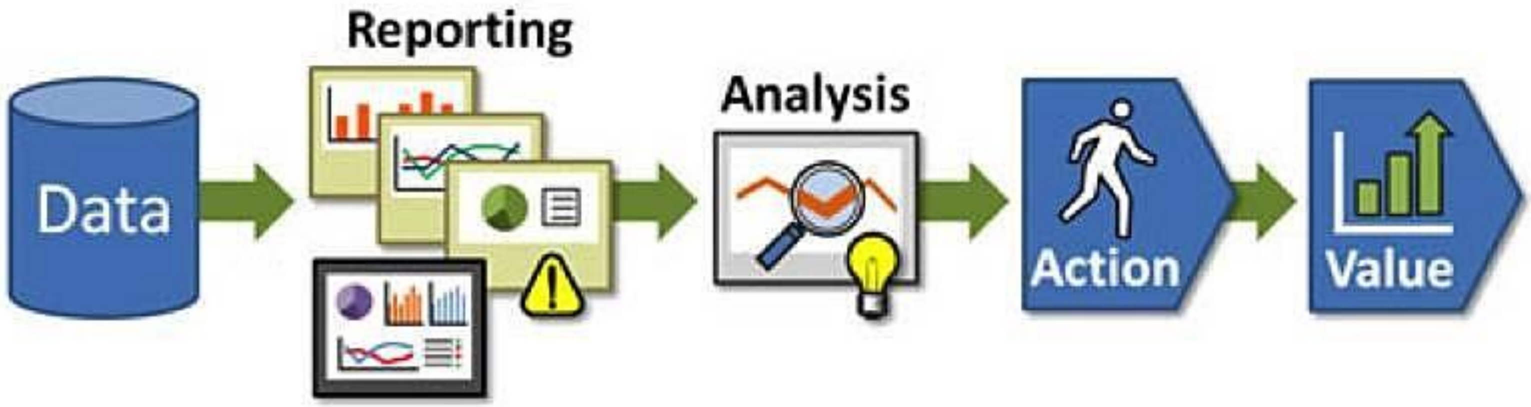
Reporting: The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

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**.

Goal of Analysis and Reporting



Reporting uses data to track the performance of your business, while an analysis uses data to answer strategic questions about your business. Though they are distinct, reporting and analysis rely on each other. Reporting sheds light on what questions to ask, and an analysis attempts to answer those questions.

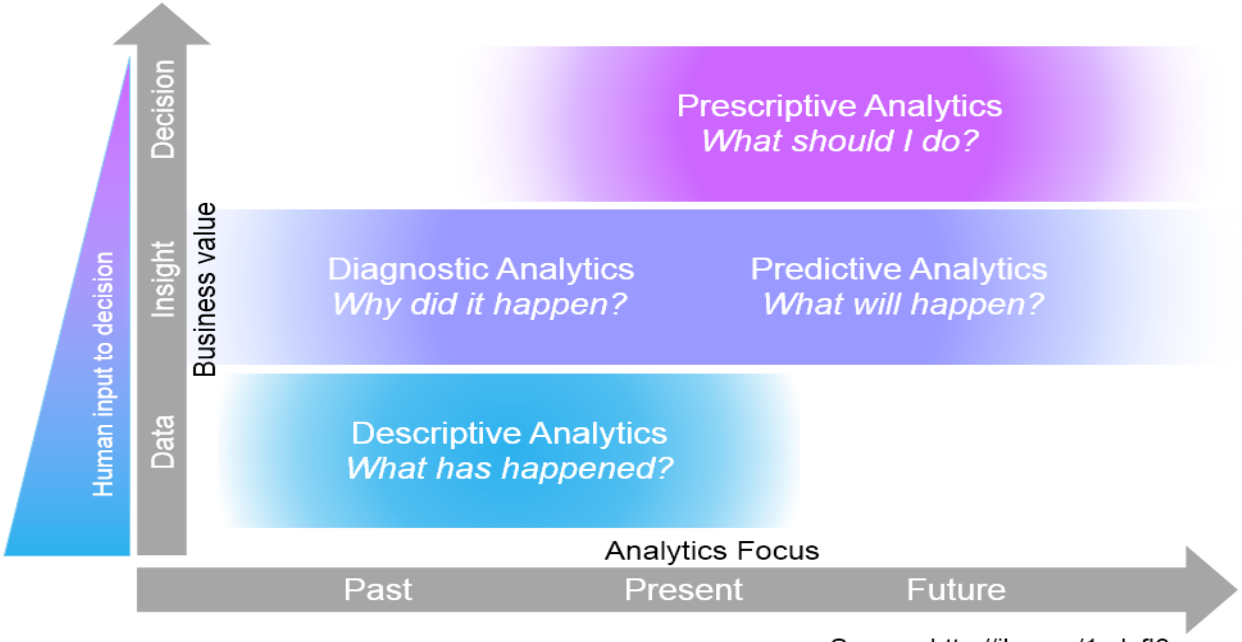
Simply put,

- ❑ Data Reporting Reveals The Right Questions.
- ❑ Data Analysis Helps Find Answers.

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:

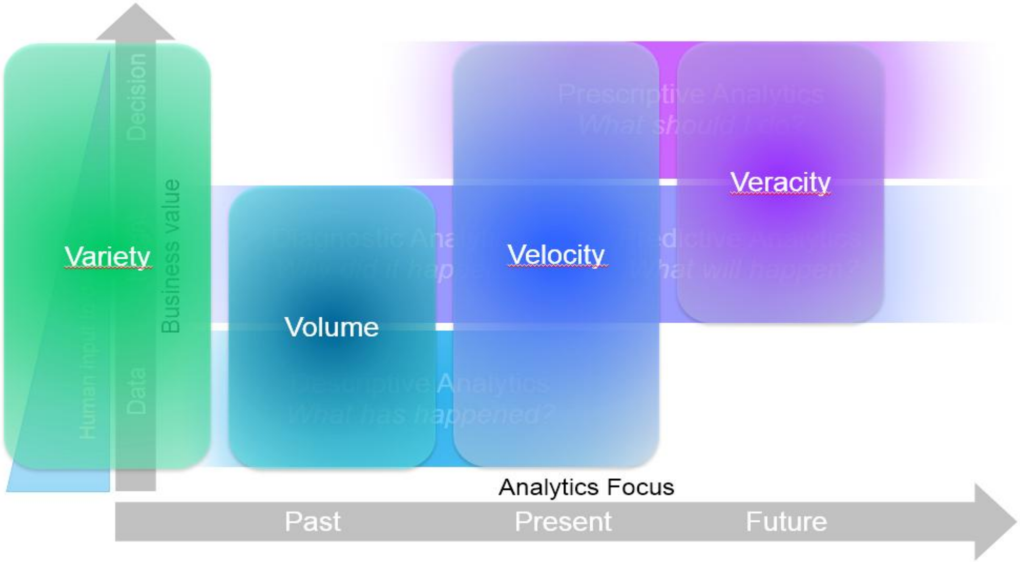


Source: http://ibm.co/1gJyfl3

Types of Analytics

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



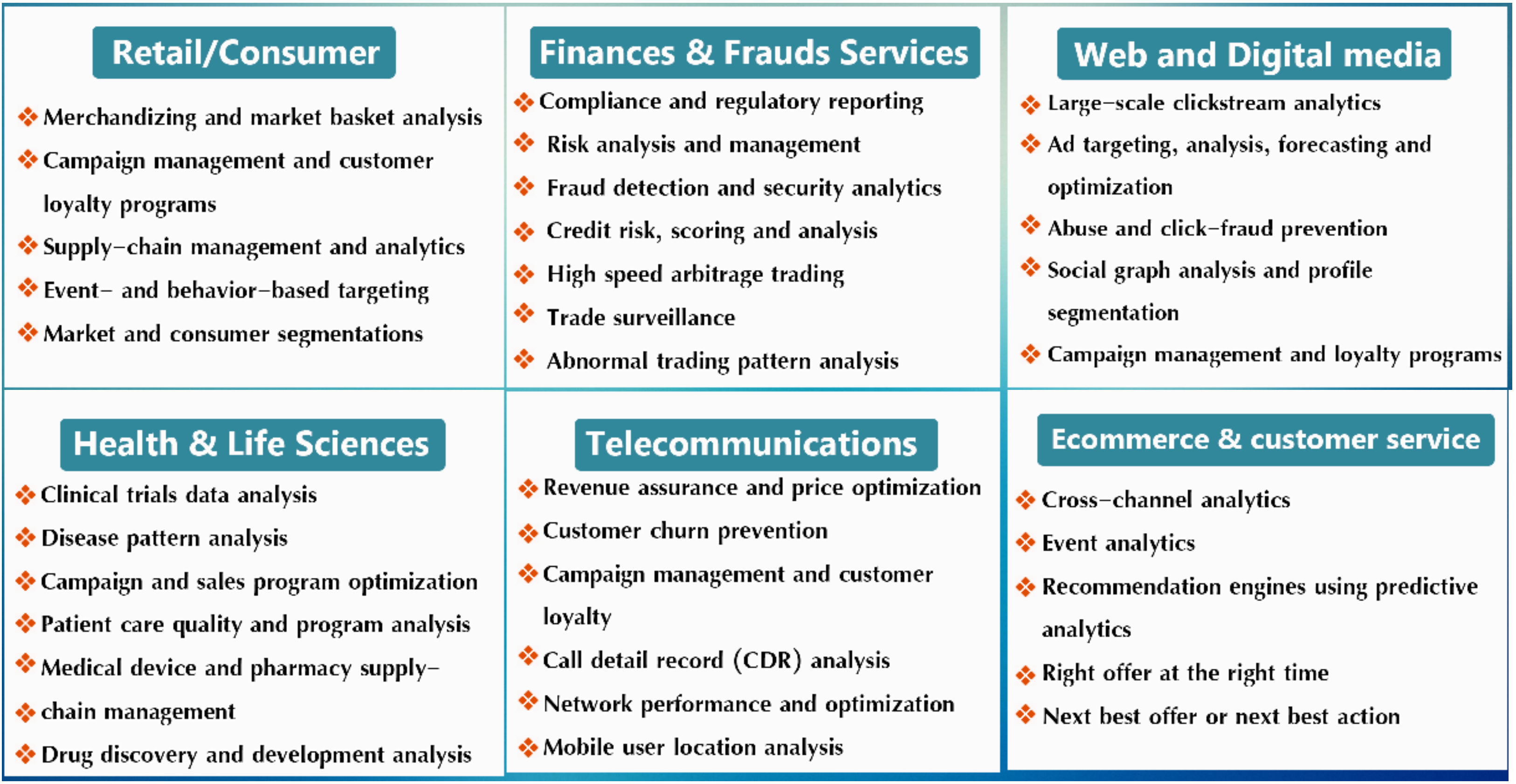
Source: http://ibm.co/1gJyfl3

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.

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

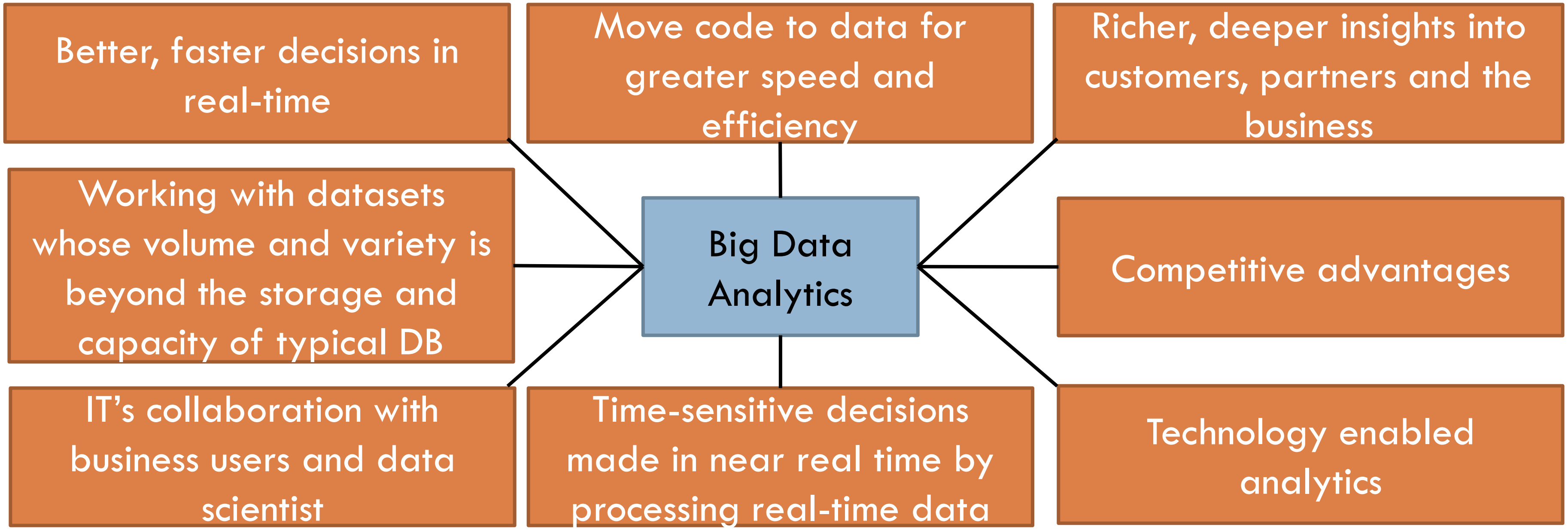
Telecommunications

- 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 ?



Big Data Analytics

Better, faster decisions in real-time

Move code to data for greater speed and efficiency

Richer, deeper insights into customers, partners and the business

Working with datasets whose volume and variety is beyond the storage and capacity of typical DB

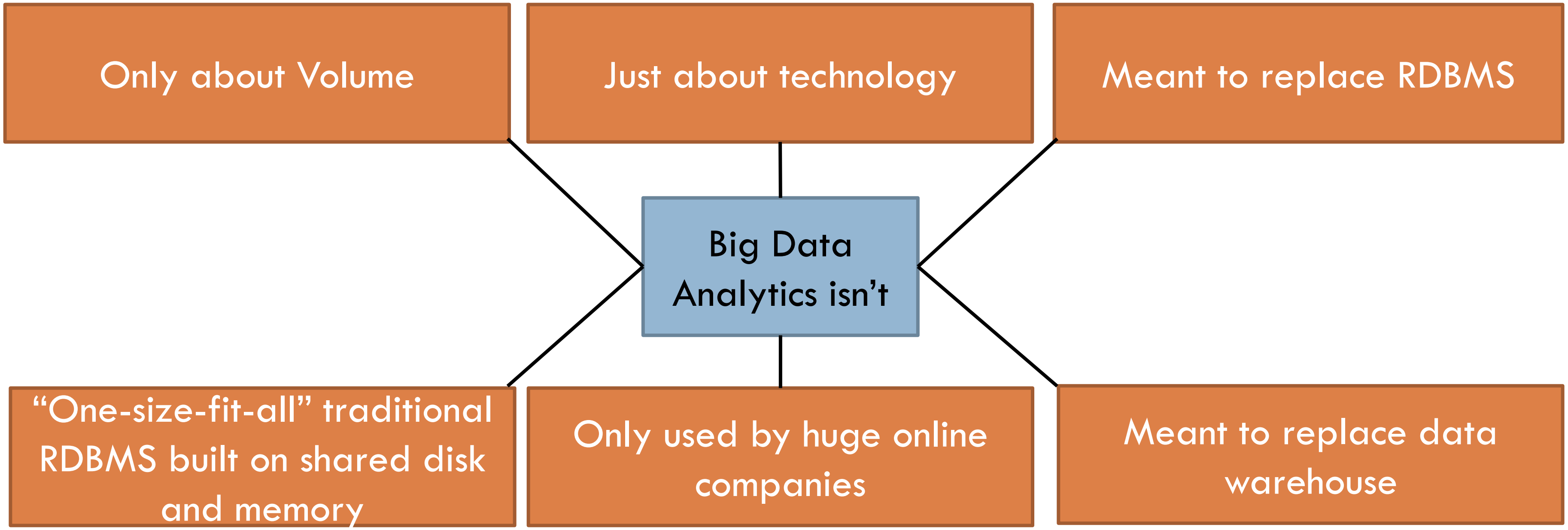
Competitive advantages

IT's collaboration with business users and data scientist

Time-sensitive decisions made in near real time by processing real-time data

Technology enabled analytics

What is Big Data Analytics isn't?



Big Data Analytics isn't

Only about Volume

Just about technology

Meant to replace RDBMS

"One-size-fit-all" traditional RDBMS built on shared disk and memory

Only used by huge online companies

Meant to replace data warehouse

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

Summary

Detailed Lessons

Introduction to Data, Big Data Characteristics, Types of Big Data, Challenges of Traditional, Systems, Web Data, Evolution of Analytic Scalability, OLTP, MPP, Grid Computing, Cloud Computing, Fault Tolerance, Analytic Processes and Tools, Analysis Versus Reporting, Statistical Concepts, Types of Analytics.

How was the journey?

THANK YOU!

Appendix

- **Data Mining:** Data mining is the process of looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/previously unknown relationships amongst the data. It is a multi-disciplinary skill that uses machine learning, statistics, AI and database technology.
- **Natural Language Processing (NLP):** NLP gives the machines the ability to read, understand and derive meaning from human languages.
- **Text Analytics (TA):** TA is the process of extracting meaning out of text. For example, this can be analyzing text written by customers in a customer survey, with the focus on finding common themes and trends. The idea is to be able to examine the customer feedback to inform the business on taking strategic action, in order to improve customer experience.
- **Noisy text analytics:** It is a process of information extraction whose goal is to automatically extract structured or semi-structured information from noisy unstructured text data.

Appendix cont…

Example of Data Volumes

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies