



# CET4001B Big Data Technologies

School of Computer Engineering and Technology

# **CET4001B Big Data Technologies**

**Teaching Scheme**

**Credits: 03 + 01**

**Theory: 3 Hrs / Week**

**Practical: 2Hrs/Week**

## **Course Objectives:**

- Understand the various aspects and life cycle of Big Data
- Learn the concepts of NoSQL for Big Data
- Design an application for distributed systems on Big Data.
- To understand and analyse different storage technologies required for Big Data
- To explore the technological foundations of Big Data Analytics
- To understand the role of various visualization techniques and explore the various Big Data visualization tools.

## **Course Outcomes:**

- Recognize the characteristics of Big Data
- Ability to demonstrate information retrieval of Big Data
- Analyse the HADOOP and Map Reduce technologies associated with big data
- Perform analytics to learn the usage of distributed processing framework
- To investigate the impact of different visualizations for real world applications

# Course outline

CCA (30 marks) + LCA (30 marks) + End term exam (40 marks) = Total (100 marks)

CCA = **Midterm** (15 marks Unit 1 & 2) + **Active learning** (Online in campus MCQ exam 10 marks Unit 3 & 4) + **Case study** (5 marks Unit 5)

LCA = **Continuous lab assignment** 10 marks + **Mini project** 10 marks + **End sem lab exam** 10 marks

# Syllabus

## Unit- I : Introduction to Big Data

- **What is Big Data**
- Overview of big data analytics
- Traditional database systems vs. big data systems
- Different V's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle

# Motivation For BIG DATA

## 1. Huge volume of data:

Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns due to different applications like twitter, Facebook, Instagram.

## 2. Complexity of data types and structures:

Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings also digital traces being left on the web and other digital repositories for subsequent analysis

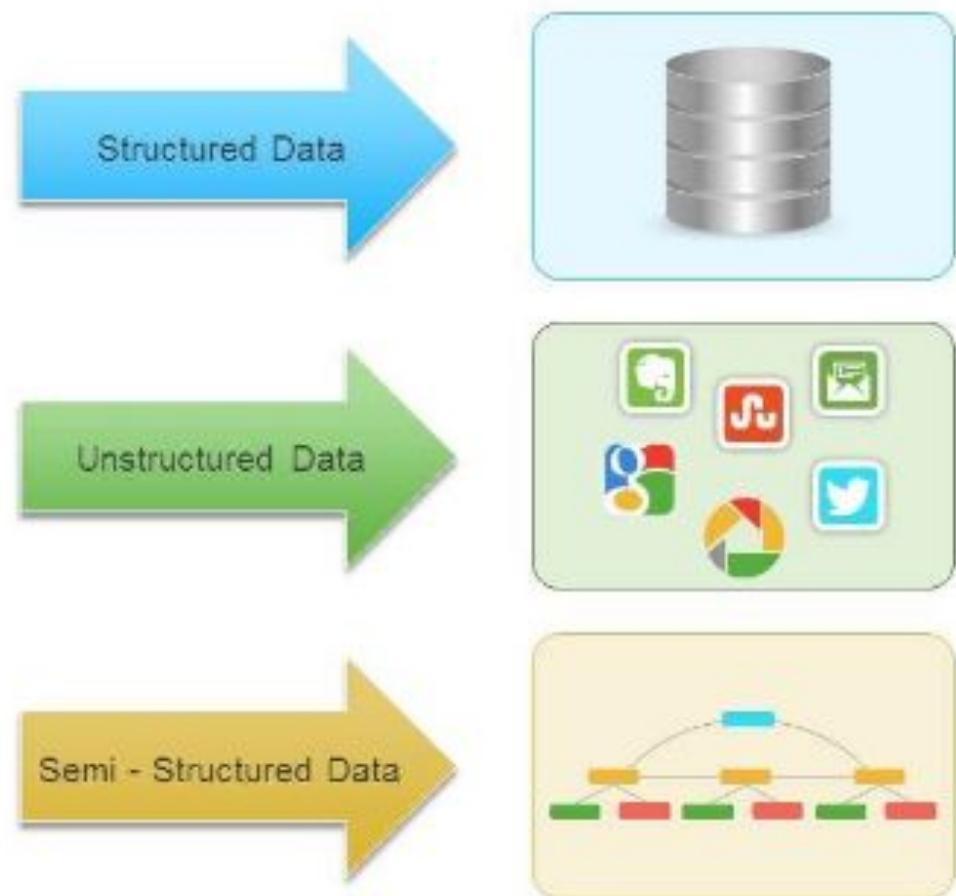
# Motivation for BIG data Contd..

3. High Speed of new data creation and growth:

Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

4. Distributed computing environments and Massively Parallel Processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data

# Motivation for BIG data Contd..



- **Structured data** is data that has been organized in a way that makes it easy to understand and use.
- It is typically stored in a database, spreadsheet, or other structured format.
- Structured data is often used in applications that need to process large amounts of data quickly and efficiently.
- **Semi-structured data** is a type of data that has some structure, but not as much as structured data.
- It is often used to store data that is not easily categorized or classified, such as text documents, XML files, and HTML pages.

# Unstructured data types



Text files and documents



Server, website  
and application  
logs



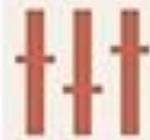
Sensor data



Images



Video files



Audio files



Emails

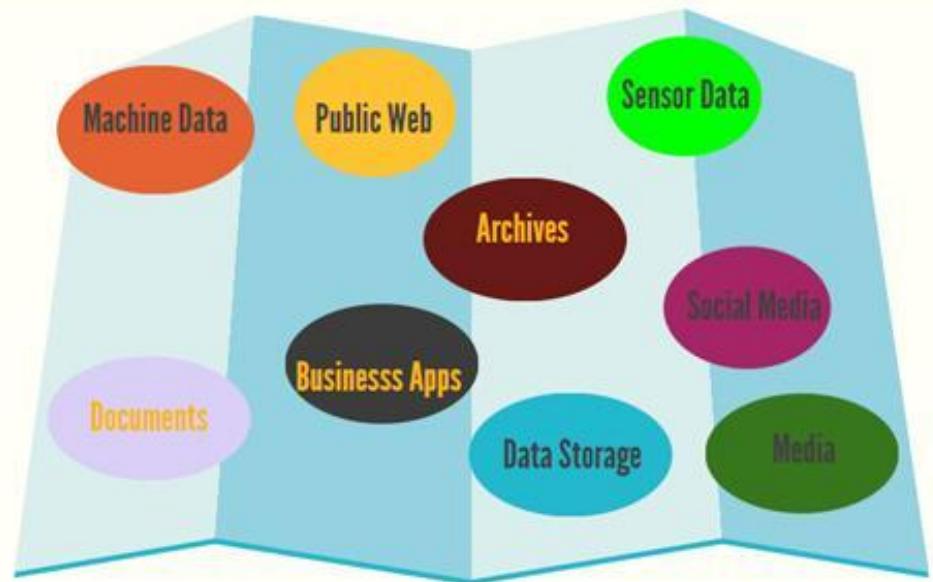


Social media  
data

# Big Data Sources

---

## BIG DATA SOURCES



# Big Data Sources

Contd..

Data is created constantly, and at an ever-increasing rate:

## Sources of Big Data:

1. Mobile phones, social media, imaging technologies -all these and more create new data, and that must be stored somewhere for some purpose
2. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time

# Examples of big data

---

Photos and video footage uploaded to the World Wide Web.

Video surveillance, such as the thousands of video cameras spread across a city .

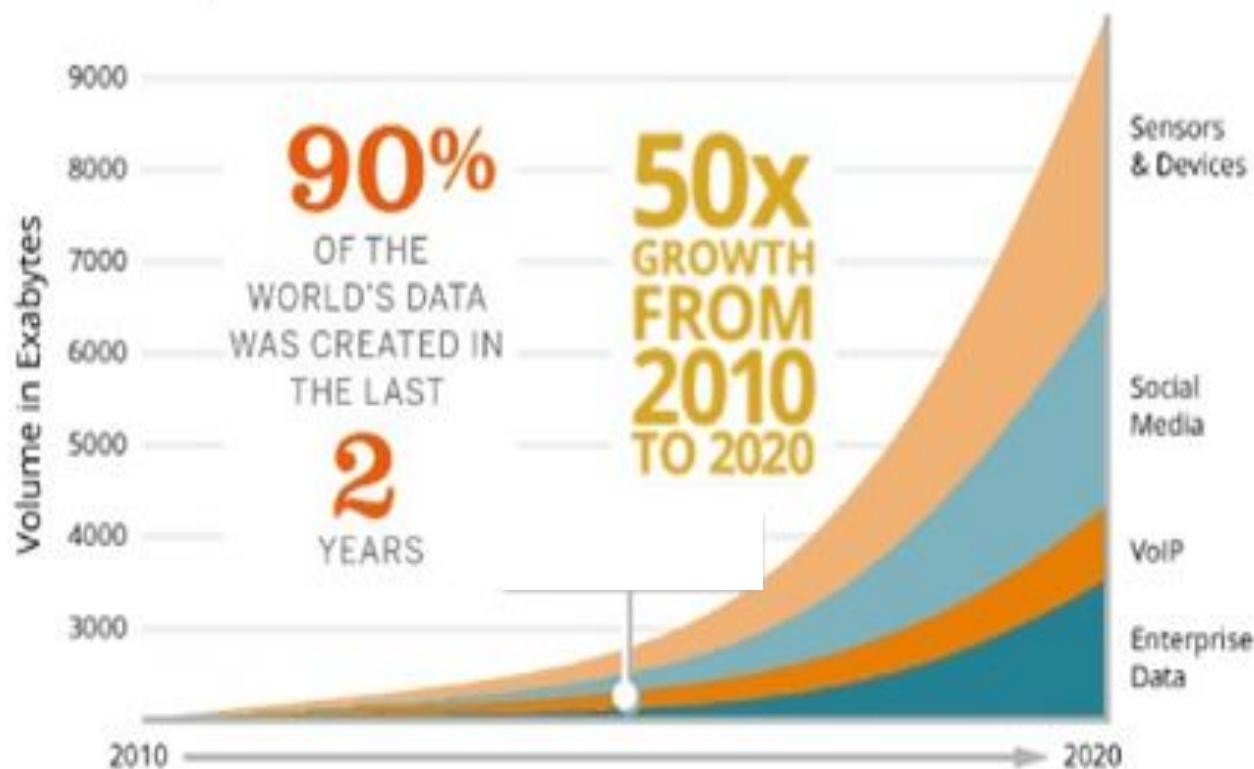
Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones

Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures

# Statistics of big data

**BIG IN GROWTH, TOO.**

1 exabyte (EB) = 1,000,000,000,000,000,000 bytes



# Unit I- Introduction to Big Data

- What is Big Data
- **Overview of Big Data Analytics**
- Traditional database systems vs big data systems
- Different v's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle



# DEFINITION OF BIG DATA

---

Not a single definition.....

- Big data is high volume, high velocity, high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. ---*Doug Laney, Gartner, 2012.*
- Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

# Example: Healthcare application



# Overview of Big Data Analytics

---

17

- **Big data analytics** is the often-complex process of examining large and varied data sets, or **big data**, to uncover information -- such as **hidden patterns**, unknown **correlations**, market trends and customer preferences -- that can help organizations make informed business decisions.

# Big Data Analytics..

- Big data analytics is a form of advanced analytics, which involves complex applications .
- Following are the type of analytics:



- *Descriptive analytics* answers the question of what happened.
- *Diagnostic Analytics*, historical data can be measured against other data to answer the question of why something happened.
- *Predictive analytics* tells what is likely to happen.
- *Prescriptive analytics* is to literally prescribe what action to take to eliminate a future problem or take full advantage of a promising trend.
- *Decisive*: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

# Descriptive Analytics

- Steps:
  - Identify the attributes, then assess/evaluate the attributes
  - Estimate the magnitude to correlate the relative contribution of each attribute to the final solution
  - Accumulate more instances of data from the data sources
  - If possible, perform the steps of evaluation, classification and categorization quickly
  - Yield a measure of adaptability within the OODA (Observe, Orient, Decide, Act) loop
- At some threshold, crossover into diagnostic and predictive analytics

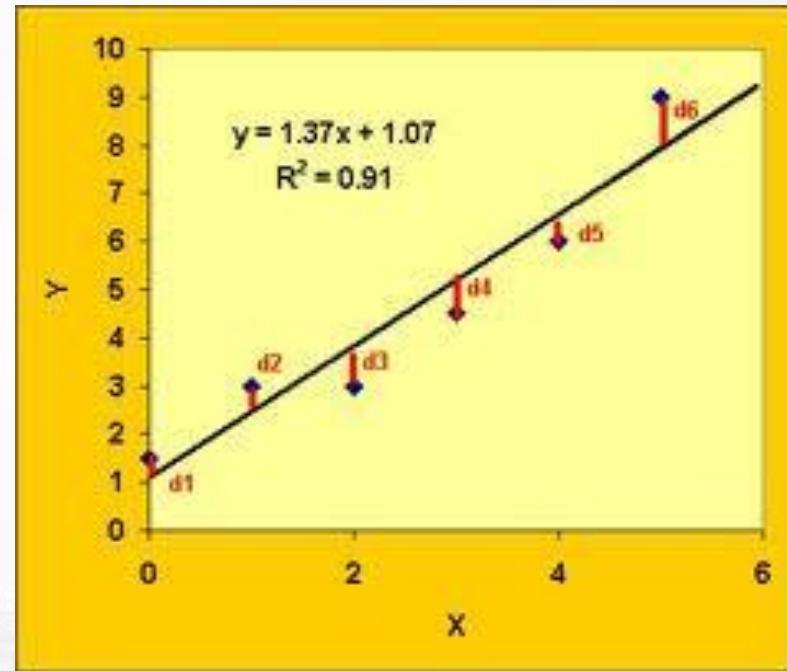


"Data don't make any sense,  
we will have to resort to statistics."



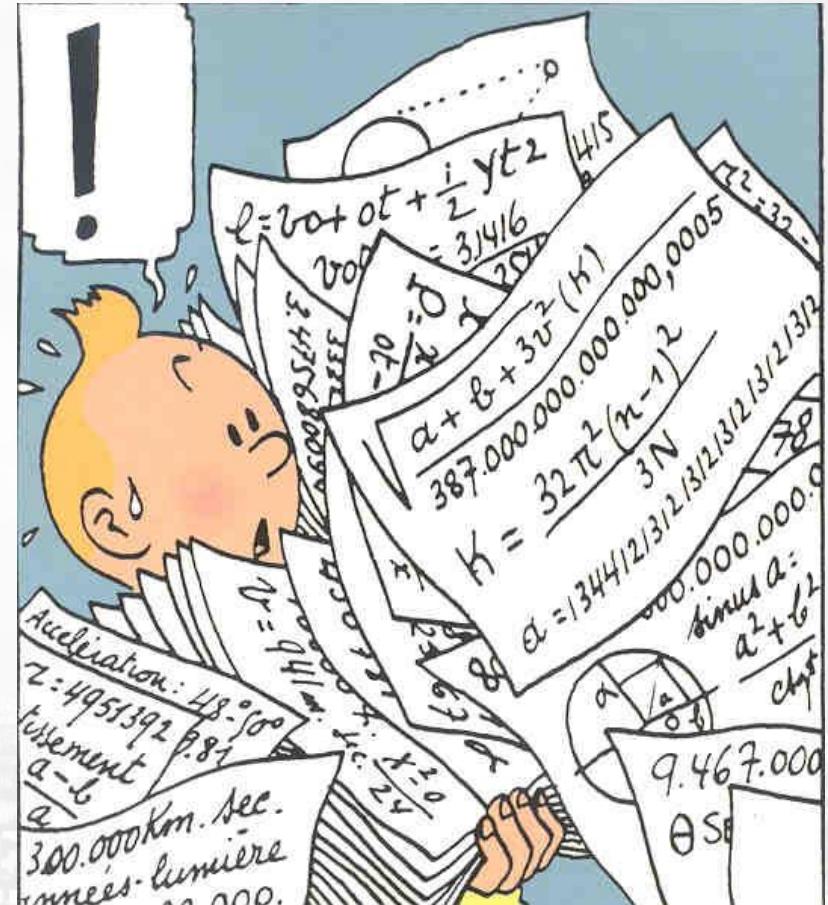
# Diagnostic Analytics

- **Steps:**
  - Begin with descriptive analytics
  - Extract patterns from large data quantities via data mining
  - Correlate data types for explanation of near-term behavior – past and present
  - Estimate linear/non-linear behavior not easily identifiable through other approaches.
- Example: by classifying past insurance claims, estimate the number of future claims to flag for investigation with a high probability of being fraudulent.



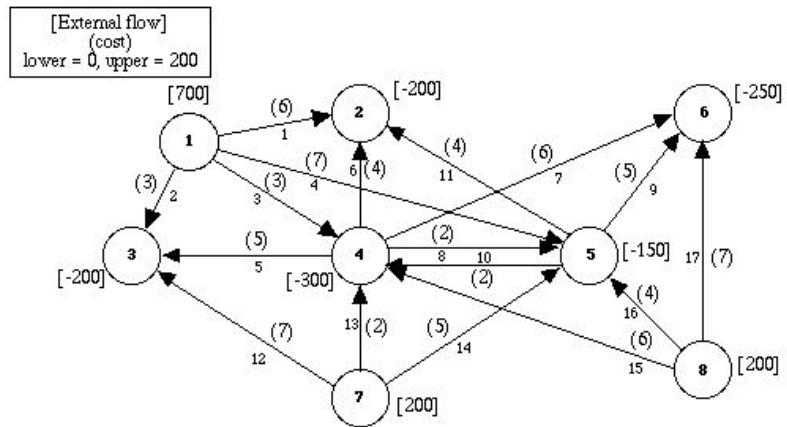
# Predictive Analytics

- Steps:
  - Begin with descriptive AND diagnostic analytics
  - Choose the right data based on domain knowledge and relationships among variables
  - Choose the right techniques to yield insight into possible outcomes
  - Determine the likelihood of possible outcomes given initial boundary conditions
  - Remember! Data driven analytics is non-linear; do NOT treat like an engineering project



# Prescriptive Analytics

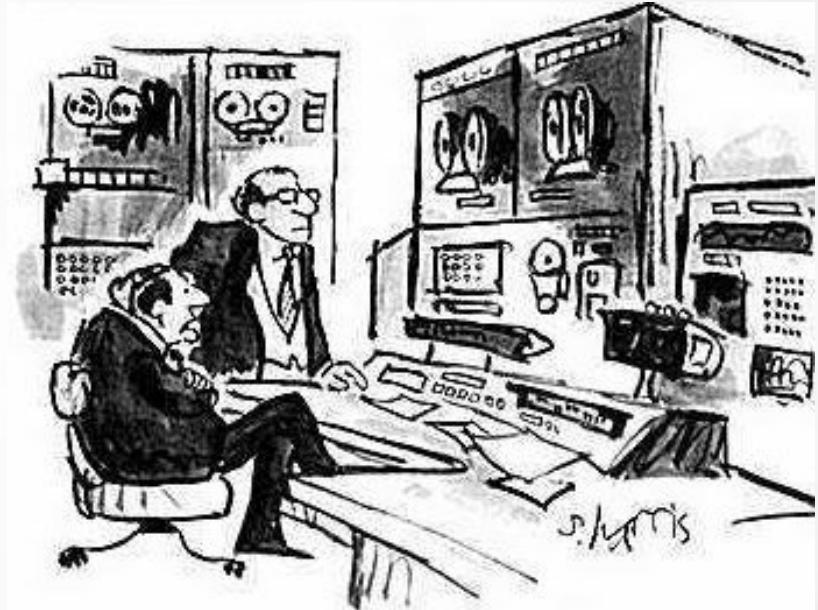
- Steps:
  - Begin with predictive analytics
  - Determine what should occur and how to make it so
  - Determine the mitigating factors that lead to desirable/undesirable outcomes
  - “What-if” analysis with local or global optimization
    - Find the best set of prices and advertising frequency to maximize revenue
    - The right set of business moves to make to achieve that goal



# Decisive Analytics

- **Steps:**

- Given a set of decision alternatives, choose the one course of action to do from possibly many
- But, it may not be the optimal one.
- Visualize alternatives – whole or partial subset
- Perform exploratory analysis – what-if and why
  - How do I get to there from here?
  - How did I get here from there?

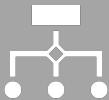


"What it comes down to is this thing is capable of telling us a lot more than we really want to know."

# *What-if analysis*



## **What-if analysis:**



Process of calculating backward to find out an input by providing a specific output.



Works in opposite fashion of formulae



What-if analysis helps to find out what input will result in a specific output.

# *What-if analysis (contd)*



Example of formula:



$$y = a * (x^2)$$



Problem = given the values for input variables  $a$  and  $x$ , we can compute value of output variable  $y$

# *What-if analysis (contd)*

Example of what-if analysis:

Suppose a student plans to score an average of 80 in semester exam. She scored 82, 70, 83 and 76 in the subjects English, Mathematics, Computer Science and Mechanics respectively.

Statistics exam is due to happen shortly, we want to calculate the marks she needs to score in Statistics to achieve an average of 80 in the semester.

- *What-if analysis(contd)*

## *Before and after scenarios*

Subject	Score	Subject	Score
English	82	English	82
Mathematics	70	Mathematics	70
Computer Science	83	Computer Science	83
Mechanics	76	Mechanics	76
Statistics		Statistics	89
Average	77.75	Average	80

# Big data Analytics

- Big data can **deliver value** in almost any area of business or society:



Report on Big Data in Big Companies

Big Data Technologies

# Key Roles For A Successful Analytics Project

---

## Key Roles for a Successful Analytics Project

- **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

# 1. Business User

Someone who understands the domain area and usually benefits from the results.

This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized. [ put into operation / use ]

Usually a business analyst or subject matter expert in the project domain fulfills this role.

## 2. Project Sponsor

Responsible for the commencement of the project.  
[origin/ source]

Provides the stimulus and requirements for the project and defines the core business problem.  
[impulse/stimulus]

This person sets the priorities for the project and clarifies the desired outputs.

# 3. Project Manager

- Ensures that key milestones and objectives are met on time and at the expected quality. [ a significant stage/event]

# 4. Business Intelligence Analyst

- Provides **business domain expertise** based on :
  - **A deep understanding of the data,**
  - **Key Performance Indicators (KPIs)**
  - **key metrics**
  - **Business intelligence from a reporting perspective**
- Business intelligence analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

# KPIs vs Key metrics

- **KPIs** are measurable values that show you how effective you are at achieving business objectives.
- **Metrics** are different in that they simply track the status of a specific business process.
- Thus **KPIs** track whether you hit business **objectives/targets**, and **metrics** track **processes**

# KPI

---

- Example of KPI
- Target of teams was to increase sales revenue by 20% this year end (2021)

**Team A**  
increase in sales = 21%

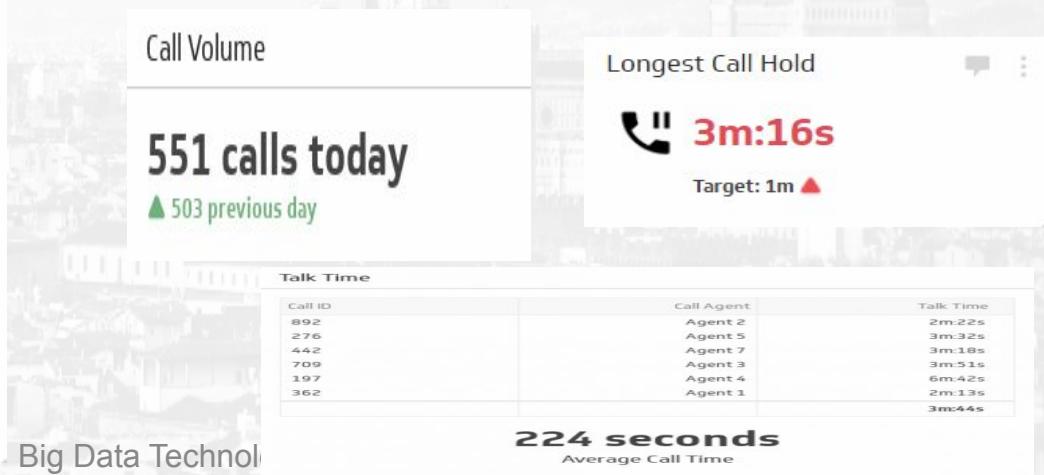
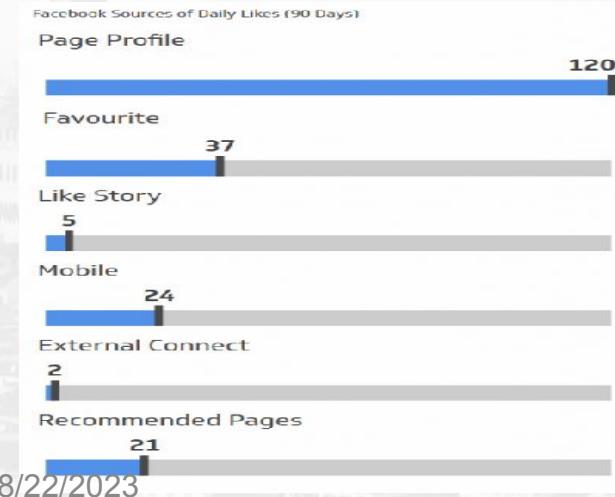


**Team B**  
increase in sales = 18%

# Example for KPI and Metrics

## • Best Social Media Marketing Metrics

- Likes
- Engagement
- Followers growth
- Traffic conversions
- Social interactions
- Social sentiment
- Social visitor goals
- Social shares
- Web visitors from social channel
- Social visitors conversion rates



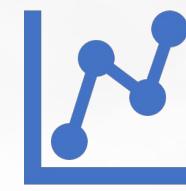
# Business intelligence



**Business Intelligence (BI)** refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information.



The purpose of Business Intelligence is to support better business decision making.



Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

## 5. Database Administrator (DBA)

Provisions and configures the database environment to support the analytics needs of the working team.

- These responsibilities may include

providing access to key databases or tables and

ensuring the appropriate security levels are in place related to the data repositories.

## 6. Data Engineer

Leverages deep technical skills to assist with tuning Query Language queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.

While the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics.

# *Analytics Sandbox*

An Analytics Sandbox is a separate environment that is part of the architecture, used by multiple users and is maintained with the support of IT.

- **Key Characteristics**

The environment is controlled by the analyst

- Allows them to install and use the data tools of their choice
- Allows them to manage the scheduling and processing of the data

Enables analysts to explore and experiment with internal and external data

Can hold and process large amounts of data efficiently from many different data sources –

- *big data (unstructured), transactional data (structured), web data, social media data, documents etc.*

## 7. Data Scientist

The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

Provides subject matter expertise for:

- analytical techniques,
- data modeling,
- applying valid analytical techniques to given business problems.

Ensures overall analytics objectives are met.

Designs and executes analytical methods and approaches with the data available to the project.

# Roles contd...

- Each role plays a critical part in a successful analytics project.
- Although seven roles are listed, fewer or more people can accomplish the work depending on
  - the scope of the project,
  - organizational structure and
  - the skills of the participants.

# Syllabus

## **Introduction to Big Data:**

- What is Big Data
- Overview of big data analytics
- **Traditional database systems vs. Big Data Systems**
- 7 v's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle



Traditional database systems  
vs  
Big Data systems

#	Dimension	Traditional	Big Data
1	Data Source	Mainly internal	Both inside and outside organization including traditional data warehouse
2	Data Structure	Pre-defined structure	Unstructured in nature
3	Data Relationship	By default, stable and interrelationship	Unknown relationship
4	Data Location	Centralized	Physically highly distributed
5	Data Analysis	After the complete build	Intermediate analysis, as you go
6	Data Reporting	Mostly canned with limited & pre-defined interaction paths	Reporting in all possible direction across the data in real time mode
7	Cost Factor	Specialized high end hardware and software	Inexpensive commodity boxes in cluster mode
8	CAP Theorem	Consistency – Top priority	Availability – Top priority

**CAP Theorem:** **Consistency:** All replicas of a piece of data have the same value at the same time. **Availability:** Every request receives a response, without guarantee that it contains the most recent write. **Partition tolerance:** The system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes.

# Analytics Difference

## Traditional Analytics (BI)

## vs

## Big Data Analytics

### Focus on

- Descriptive analytics
- Diagnosis analytics

- Predictive analytics
- Data Science

### Data Sets

- Limited data sets
- Cleansed data
- Simple models

- Large scale data sets
- More types of data
- Raw data
- Complex data models

### Supports

**Causation:** what happened, and why?

**Correlation:** new insight  
More accurate answers

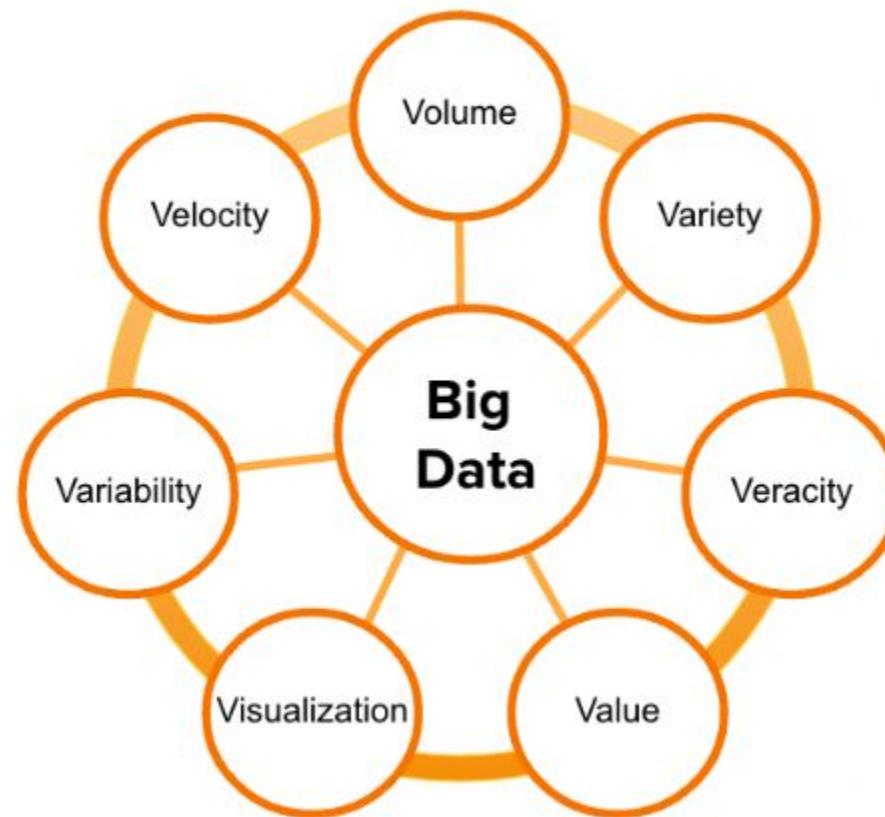
# Syllabus

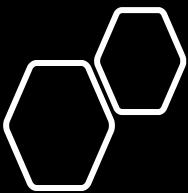
## **Introduction to Big Data:**

- What is Big Data
- Overview of big data analytics
- Traditional database systems vs. Big Data Systems
- **Different V's of big data**
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle



# Characteristics of Big Data





# 1. Volume:

Big data first and foremost has to be “**big**,” and size in this case is measured as volume.

## Example:

From clinical data associated with lab tests and physician visits, to the administrative data surrounding payments, this well of information is already expanding.

When that data is coupled with greater use of precision medicine, there will be a big data explosion in health care, especially as genomic and environmental data become more ubiquitous.

# 2. Velocity:

Velocity in the context of big data refers to two related concepts familiar to anyone in healthcare: the rapidly increasing speed at which new data is being created by technological advances, and the corresponding need for that data to be digested and analyzed in near real-time.

Example:

As more and more medical devices are designed to monitor patients and collect data, there is great demand to be able to analyze that data and then to transmit it back to clinicians and others.

This “internet of things” of healthcare will only lead to increasing velocity of big data in healthcare.

### 3. Variety:

With increasing volume and velocity comes increasing variety. This third “V” describes just what you’d think: the huge diversity of data types that healthcare organizations see every day.

- Example: Electronic health records and medical devices.

Each one might collect a different kind of data, which in turn might be interpreted differently by different physicians—or made available to a specialist but not a primary care provider.

- **Challenges:**

Standardizing and distributing all of that information so that everyone involved is on the same page.

# 4. Variability

Variability is different from variety.

A coffee shop may offer six different blends of coffee, but if you get the same blend every day and it tastes different every day, that is variability.

The same is true of data. If the meaning constantly changes, it can significantly impact your data homogenization.

# 5. Veracity

- Veracity refers to the level of trustiness or messiness of data, and if higher the trustiness of the data, then lower the messiness and vice versa.
- Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights.
- It also refers to the assurance of **quality/ integrity/ credibility/ accuracy** of the data.
- Veracity and Value both together define the data quality, which can provide great insights to data scientists..
- Veracity ensures the data is accurate, which requires processes to keep the insufficient data from accumulating in your systems.
- The simplest example is when contacts enter your marketing automation system with false names and inaccurate contact information.

# 6. Visualization

Visualization is critical in today's world.

Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.

## 7. Value

- big data must have value.

That is, if you're going to invest in the infrastructure required to collect and interpret data on a system-wide scale, it's important to ensure that the insights that are generated are based on accurate data and lead to measurable improvements at the end of the day.

Organizations might use the same tools and technologies for gathering and analyzing the data they have available, but how they then put that data to work is ultimately up to them.

After addressing volume, velocity, variety, variability, veracity, and visualization — which takes a lot of time, effort, and resources —, you want to be sure your organization is getting value from the data.

# Overview of 'V' s of Big Data

## The 7 Vs OF BIG DATA

Just having Big Data is of no use unless we can turn it into value

The speed at which the data is generated

The different types of data

The trustworthiness of the data in terms of accuracy

Value

Velocity

Variety

Veracity

Volume

Variability

Visualisation

The size of the data

The data whose meaning is constantly changing

The data in a manner that's readable and accessible

# Syllabus

## **Introduction to Big Data:**

- What is Big Data
- Overview of big data analytics
- Traditional database systems vs big data systems
- 5 v's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle

# Importance of Big Data

- Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including:
  - New revenue opportunities
  - More effective marketing
  - Better customer service
  - Improved operational efficiency
  - Competitive advantages over rivals



# Importance of Big Data

Cont....

## *1. It helps companies to better understand and serve customers:*

- Examples include the recommendations made by Amazon or Netflix., Coca-Cola( Customer Acquisition and Retention)

## *2. It allows companies to optimize their processes:*

- Faster and Better Decision Making
- Example
  - UOB Bank from Singapore use Big Data for Risk Management
  - Uber is able to predict demand, dynamically price journeys and send the closest driver to the customers

# Importance of Big Data

Cont....

## *3. It improves our health care:*

- Government agencies can now predict flu outbreaks and track them in real time and pharmaceutical companies are able to use big data analytics to fast-track drug development.

## *4. It helps us to improve security:*

- Government and law enforcement agencies use big data to foil terrorist attacks and detect cyber crime.

## *5. It allows sport stars to boost their performance:*

- Sensors in balls, GPS trackers on their clothes allow athletes to analyze and improve upon what they do.

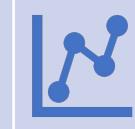
## *6. Cost Reduction:*

- Big Data Technologies like Hadoop and Cloud based analytics bring sufficient cost advantages when it come to storing large data

# Real world challenges



Exploiting the opportunities that Big Data presents requires new data architectures, including analytic sandboxes, new ways of working, and people with new skill sets.



These drivers are causing organizations to set up analytic sandboxes and build Data Science teams.

# Real world Challenges contd..

## **1. Dealing with data Growth**

- The most obvious challenge associated with big data is simply storing and analyzing all that information.



## **2. Recruiting and retaining big data talent**

- In order to develop, manage and run applications that generate insights, organizations need professionals with big data skills.
- Potential pitfalls of big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced data scientists and data engineers to fill the gaps.

# Active Learning

## State Business Goals

- Decreasing expenses through operational cost efficiencies
- Creating new avenues for innovation and disruption
- Accelerating the speed with which new capabilities and services are deployed
- Launching new product and service offerings

# Real world Challenges

contd..

## Generating insights in a timely manner

- Business goals can be achieved if data scientists can extract insights from Big Data and can act upon on those quickly.
- Although some organizations are fortunate to have data scientists (most may not be), there is a growing talent gap that makes finding and hiring data scientists in a **timely manner difficult**

# Real world Challenges

contd..

## 4. Integrating disparate data sources

- The variety associated with big data leads to challenges in data integration.
- Big data comes from a lot of different places — enterprise applications, social media streams, email systems, employee-created documents, etc. Combining all that data and reconciling it so that it can be used to create reports can be incredibly difficult.

## 5. Validating data

- Often organizations are getting similar pieces of data from different systems, and the data in those different systems doesn't always agree.
- For example, the ecommerce system may show daily sales at a certain level while the enterprise resource planning (ERP) system has a slightly different number.

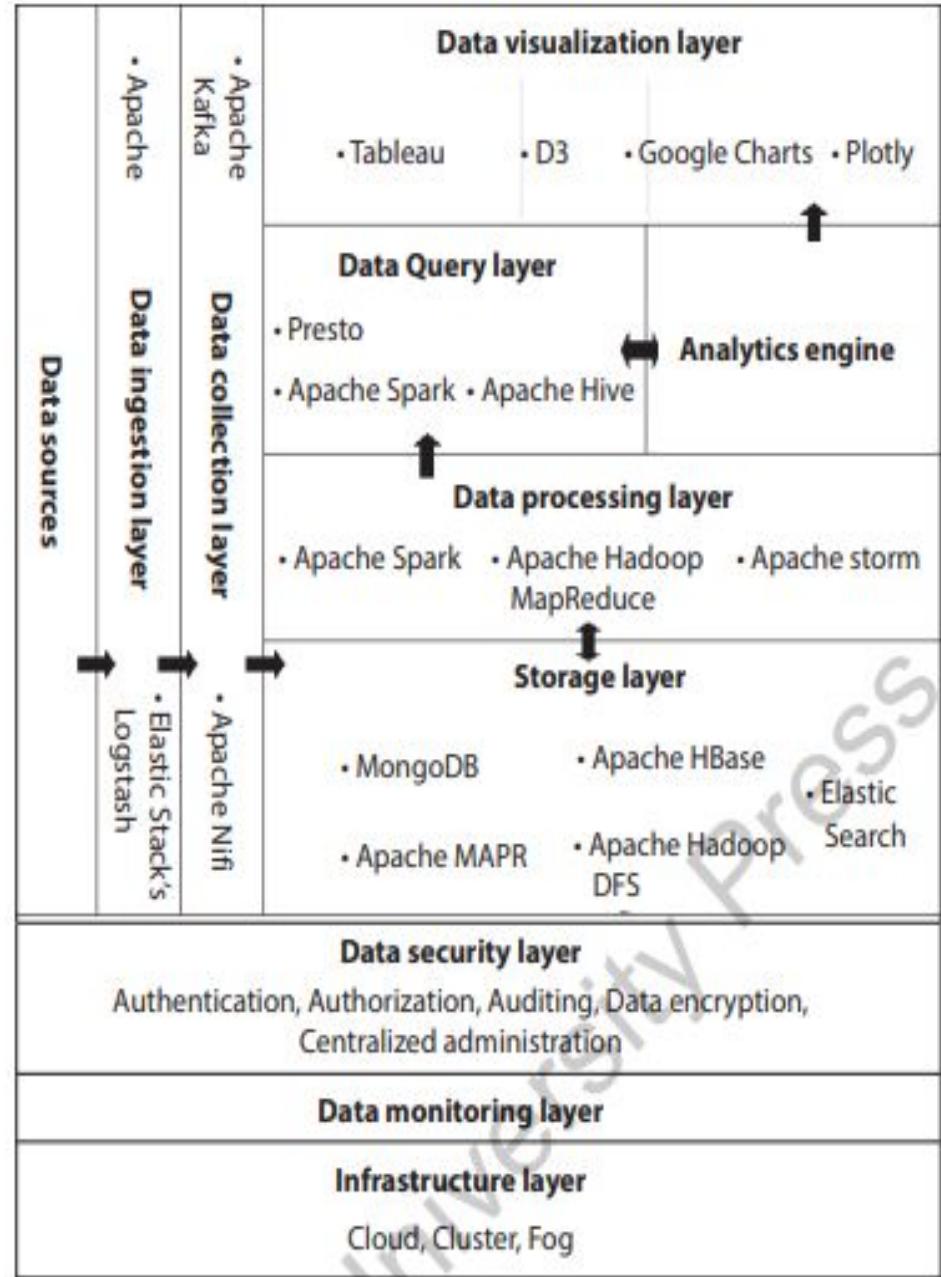
# Syllabus

## Introduction to Big Data:

- What is Big Data
- Overview of big data analytics
- Traditional database systems vs big data systems
- 5 v's of big data
- Importance of big data and real challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle



# Architecture of Big Data Systems



# Architecture of Big data Systems

## 4 Core Layers of Big Data Systems

### Architecture:

- Data Storage layer
- Data Processing layer
- Data Query layer
- Data Visualization layer

## Traditional Data Systems:

- Physical layer
- Logical layer
- View layer

# Architecture of Big data Systems (Cont....)

## 1. Data Storage layer:

- Necessity to handle heterogeneity using different data stores
- Polyglot persistence: Approach to identify an effective data store for a particular data
- To store large amount of unstructured data , Hadoop Distributed File System (HDFS) can be used.
- For Object based storage Simple Storage System(S3) can be used
- Functionality of this layer is handled by 2 sublayers
  - Physical layer- Handles large volume of heterogeneous real-time data
  - Data layer- Maintains data blocks and the global namespace to access data
    - It also maintains tools to organize, access and retrieve heterogeneous data

# Architecture of Big data Systems

(contd...)

## 2. Data Processing layer:

Data collected in the storage layer is processed in this layer in batch or real-time mode

- Batch Processing is used for offline Analytics
  - E.g. Hadoop is a batch processing system with Map-Reduce programming technique
- Real-time processing is used for online analytics
  - E.g. Apache storm processes streaming data in real time to make the decision
  - Spark is time-efficient, in-memory data processing engine that can execute streaming, machine learning or SQL-workloads
- Along with MapReduce, Spark it also supports tools for statistical modelling, machine Learning

# Architecture of Big data Systems

## (Cont....)

### 3. Data query layer:

- This layer aims at obtaining data values or valuable insights from the processing layer
- Hive: used by data analysts to query, summarize, explore and analyze unstructured data to obtain actionable business insights
- Analytics Engine- It extends the functionality of the data processing layer with domain specific tools for decision making
- Tools in this layer performs descriptive, predictive, diagnostic analytics



# Architecture of Big data Systems

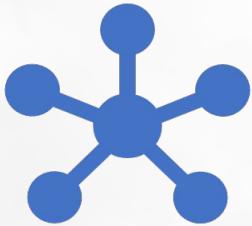
(Cont....)

## 4.Data Visualization layer:

- This layer presents the value of the data in a presentable , understandable formats
- It makes use of Dashboards, Graphs and tables tools for visualization
- E.g. Google Chart-
  - It is a JavaScript based charting library meant to enhance web applications by adding interactive charting capability.
  - Google Charts provides wide variety of charts. For example, line charts, spline charts, area charts, bar charts, pie charts and so on.
- E.g. D3-
  - It is programming tool for visualization
  - User must be knowledgeable on Java Script to visualize the collected data effectively

# Architecture of Big data Systems

(Cont....)



Following layers offer common services to the core layers also called as *service layers*.



## 1. Data Ingestion layer:

This layer determines the value of information extracted

Data coming from different sources is prioritized, validated, categorized and routed to the destination for effective storage and access

Data may be ingested in batches periodically or in real time

E.g. Sqoop-

- supports bulk data transfer between Hadoop and structured stores such as ORACLE, MYSQL

E.g. Elastic Logstash-

- aggregates data from multiple sources and routes it to Elastic Search Engine

# Architecture of Big data Systems

(Cont....)

## 2. Data Collector layer:

- This layer transports data from ingestion layer to the rest of the data pipeline
- E.g. Kafka
  - It is a message oriented middleware used for data collection
  - It collaborates with Storm, Hbase, Spark for real time analysis of data

## 3. Data Security layer:

This layer provides authentication, Authorization, audit, data encryption and central administration for big data systems

E.g. Knox in Hadoop stack, Kerberos, HDFS encryption

# Architecture of Big data Systems (Cont....)

## 4. Data Monitoring layer:

- It includes tools for monitoring the performance at infrastructure, framework analytics engine, data store and application levels

## 5. Infrastructure layer:

- This layer provides the hardware to host various big data frameworks in cloud infrastructure that is highly scalable and preferable



# Syllabus

## Introduction to Big Data:

- What is Big Data
- Overview of big data analytics
- Traditional database systems vs big data systems
- 5 v's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications
- Data analytics life cycle

# Syllabus

## Introduction to Big Data:

- What is Big Data
- Overview of big data analytics
- Traditional database systems vs big data systems
- 5 v's of big data
- Importance of big data and real world challenges
- Architecture of big data systems
- Big data applications 
- Data analytics life cycle

- Big Data Analytics Life Cycle

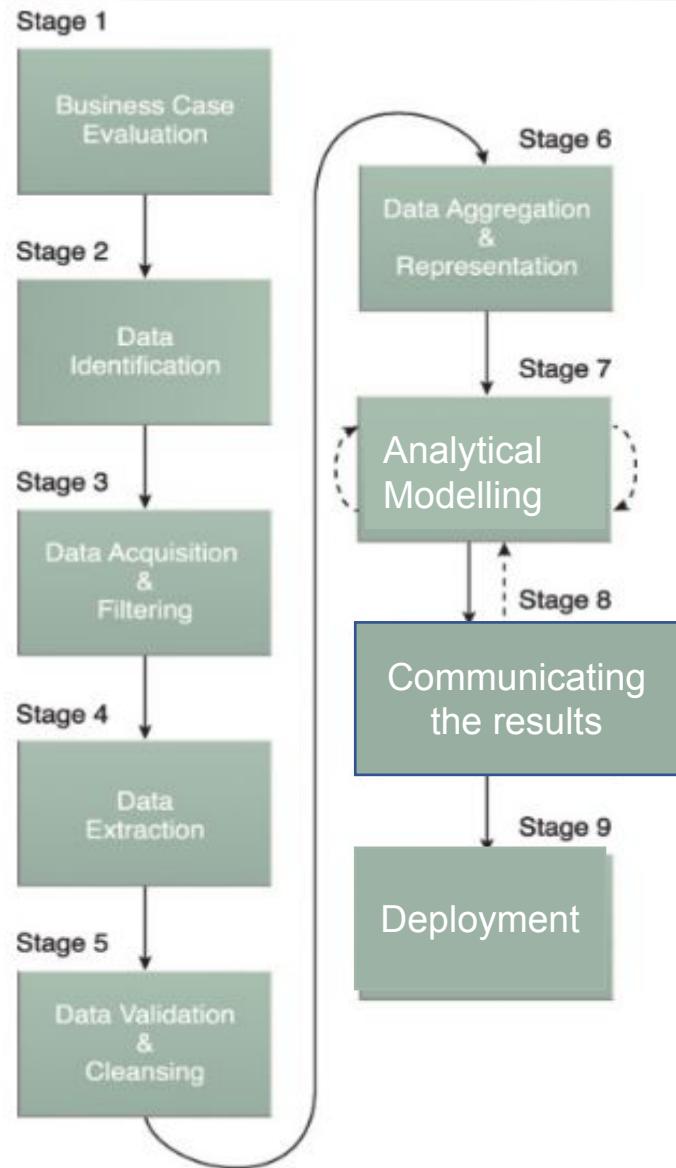
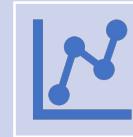


Figure 3.6 The nine stages of the Big Data analytics lifecycle.

# 1. Business Case Evaluation



It must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.



An evaluation of a Big Data analytics business case helps decision-makers to understand the business resources which helps business challenges to tackle including KPIs .



The outcome of this stage is the understand budget (h/w, s/w) required to carry out the analysis project.

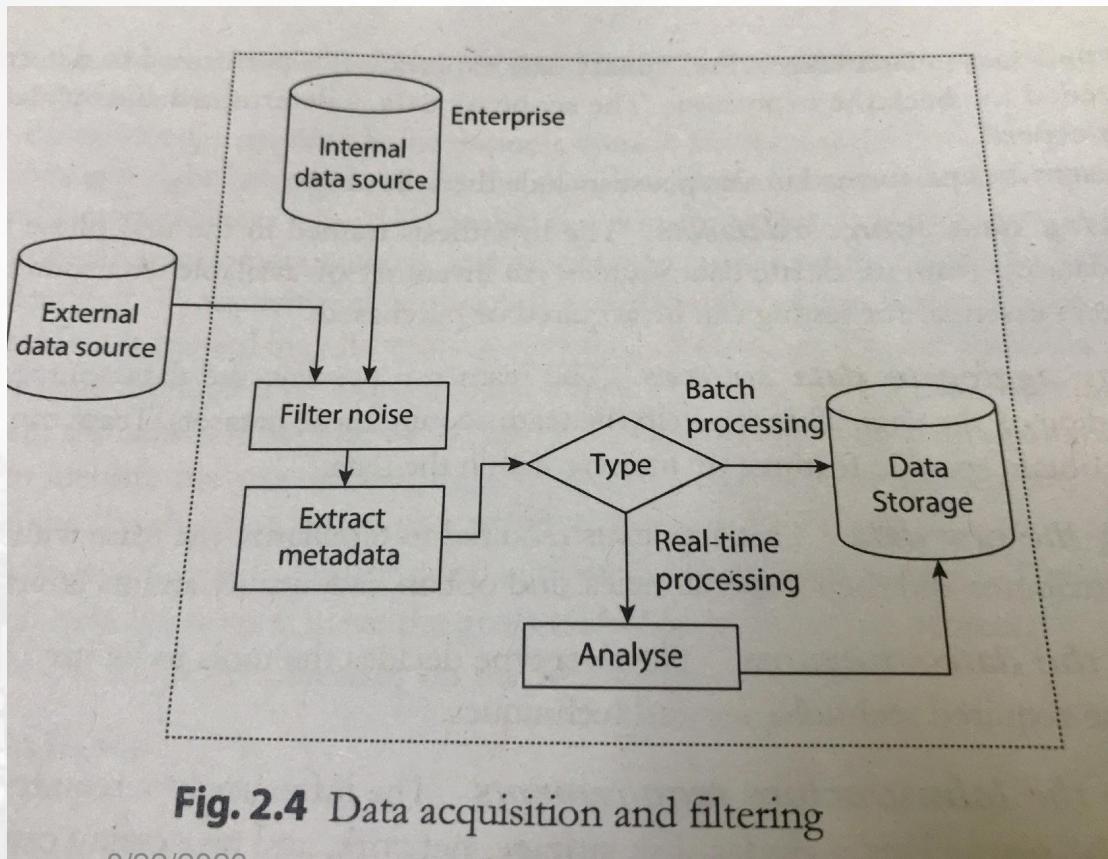


Initial iterations of the Big Data analytics lifecycle will require more up-front investment of Big Data technologies, products and training compared to later iterations

## 2. Data Identification

- Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations.
- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be categorized into 2 types
  - Internal datasets: such as data marts and operational systems, are typically compiled and matched against a pre-defined dataset specification.
  - External datasets: publicly available datasets, content-based web sites, blogs.
- Review the raw data.
- Evaluating the data structures.
- Decide on the infrastructure requirements.

# 3. Data Acquisition and Filtering

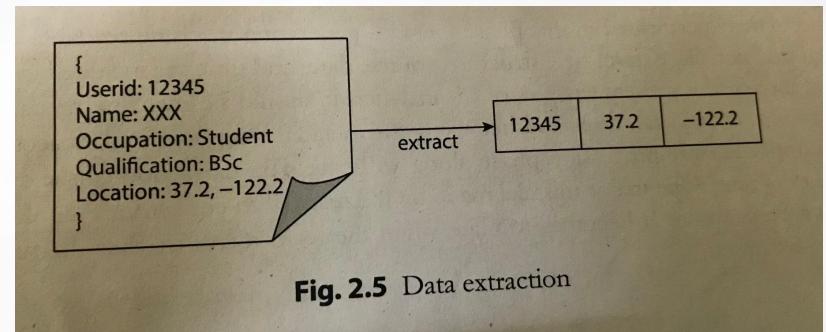


**Fig. 2.4** Data acquisition and filtering

- The data is gathered from all of the data sources that were identified during the previous stage.

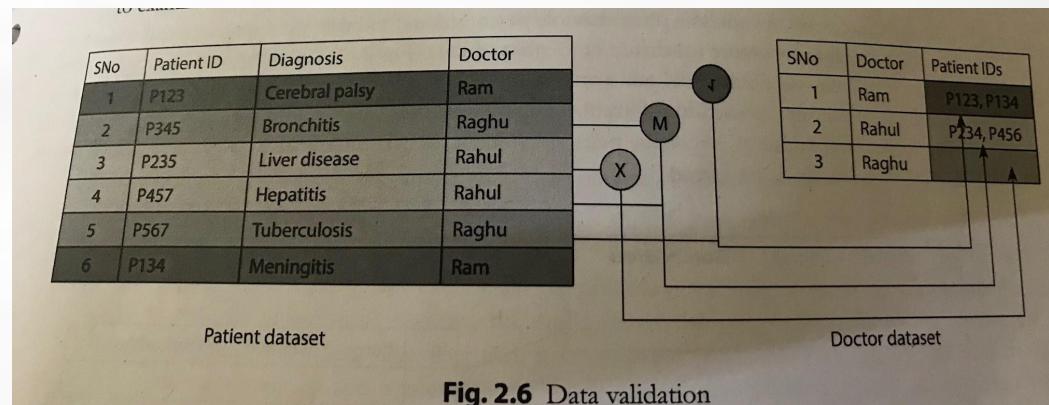
# 4. Data Extraction

- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution.
- E.g., extracting the required fields from delimited textual data, such as with webserver log files.
- Similarly, extracting text for text analytics, which requires scans of whole documents, is simplified if the underlying Big Data solution can directly read the document in its native format.



# 5.Data Validation and Cleansing

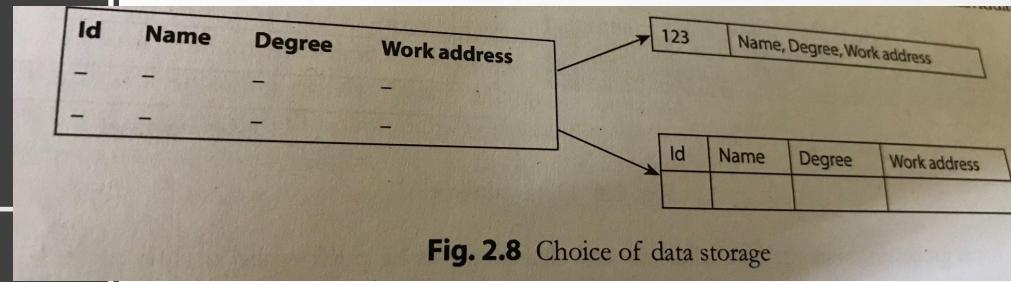
- Examining the cleanliness of the data
- Checking for consistency of data by identifying missing and inconsistent values.
- Assessing the consistency of the data types by checking if values suit the data type.
- Reviewing the contents of the data columns for relevant and consistent values
- Looking for validity of incoming data by checking for extreme or incorrect values.



**Fig. 2.6** Data validation

# 6. Data Aggregation and Representation

- The objective of this phase is to integrate multiple datasets to arrive at unified view.
- The tools for data indigestion, filtering ,extraction, validation ,cleansing and aggregation are Hadoop, open refiner, Alpine miner, Data Wrangler.



**Fig. 2.8** Choice of data storage

Id	Name	Home Address

+

Id	Name	degree	Home Address	Work Address

=

Id	Name	degree	Work Address

**Fig. 2.7** Data aggregation

## 7. Analytical Modelling

- The data analysis helps to decide the hypothesis which can be used to know the data.
- Analytical modelling includes two sub-phases
  1. Model Planning
    1. Data Exploration
      - Helps to clean the data to gain data quality.
    2. Model Selection
      - Commonly used tools are R, SQL Analysis services, SAS/ Access for RDBMS
  2. Model Building
    - Develop analytical model that fits on the training data, evaluated against test data which is fitted after several iterations.

## 8. Communicating the results

Record all the findings and then select the most significant ones and share with the other stakeholders.

The team made recommendations for future work or improvements to existing processes.

## 9. Deployment

This phase deals with deploying the analytical models in a production environment.

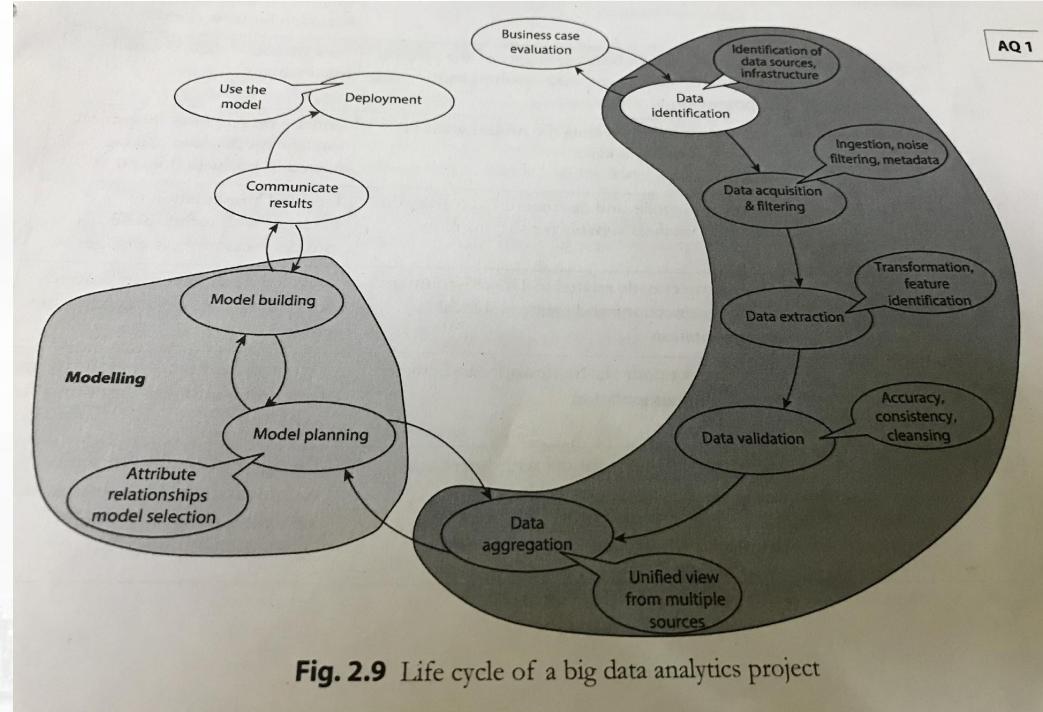
The output of these models can also be used to prescribe some actions such as:

Optimizing business process

Creating alerts

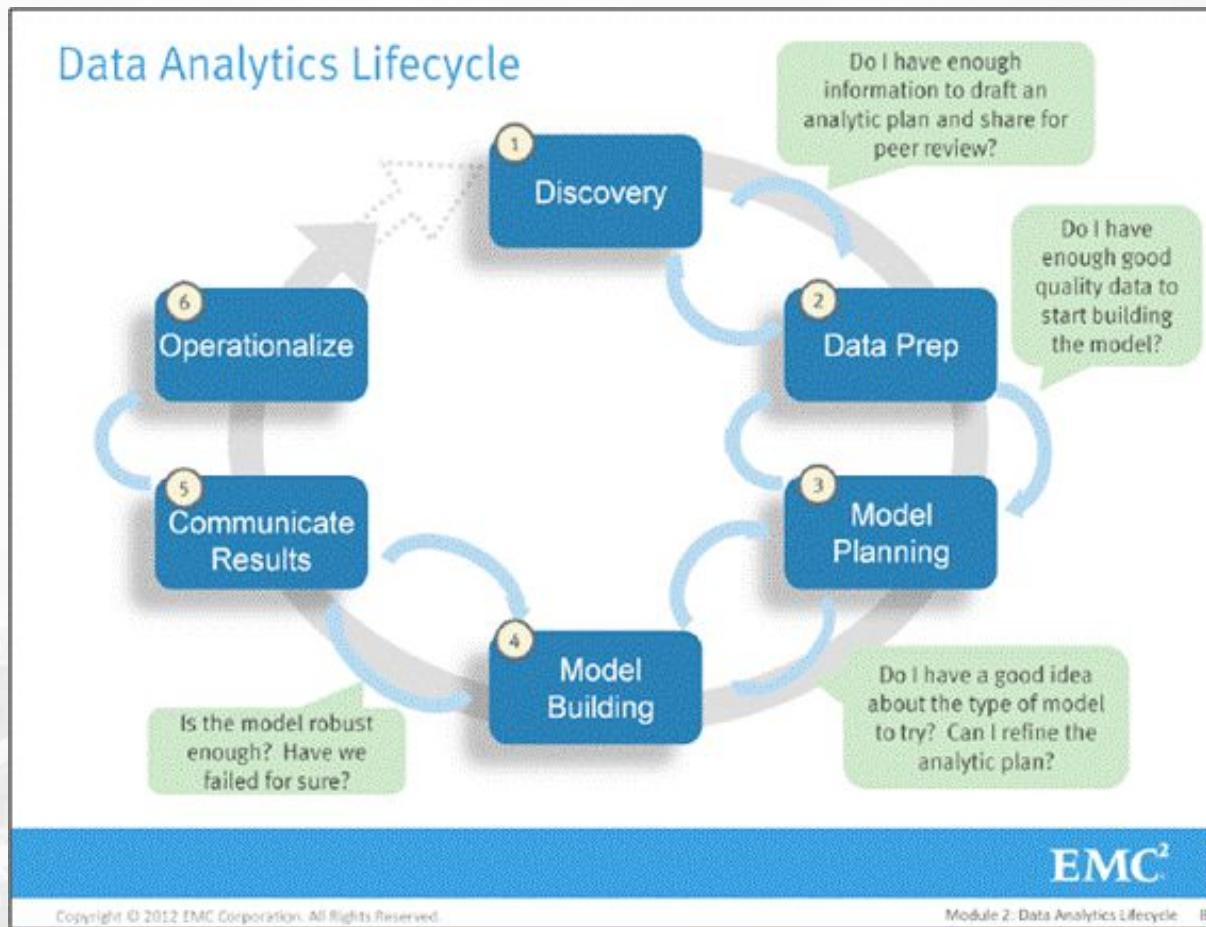
Extending the functionality of enterprise systems

# Summary: Life Cycle of BDA Project



**Fig. 2.9** Life cycle of a big data analytics project

# Main phases of Data Analytics Lifecycle



# Data Analytics Life Cycle



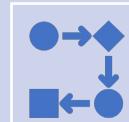
The Data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion.



The lifecycle draws from established methods in the realm of data analytics and decision science.



This synthesis was developed after gathering input from data scientists and consulting established approaches that provided input on pieces of the process.



Traditional projects follows the process centric approach(WATERFALL/ SPIRAL) to develop the project.



SDLC can *not be applied directly* for the data analytics projects as it is *data centric* projects.



We have to follow CRISP-DM approach for data oriented projects.

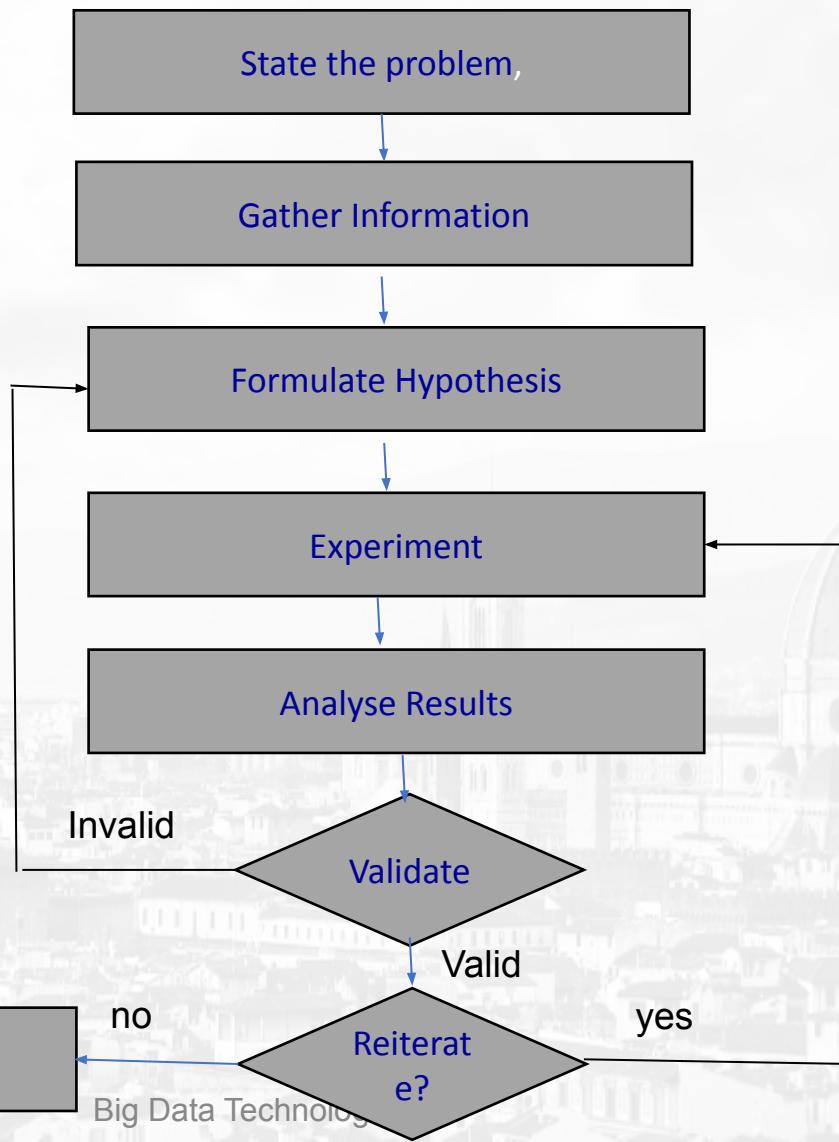
# Life Cycle of Data Centric Projects

Techniques  
Include:

- Scientific Method
- Cross Industry Standard Process for Data Mining(CRISP-DM)
- Sample, Explore, Modify, Model, Access (SEMMA)
- DELTA Framework
- Applied Information Economics(AIE) Approach
- Magnetic, Agile, Deep Analytic skills

# Life Cycle of Data Centric Projects

## 1. Scientific Method

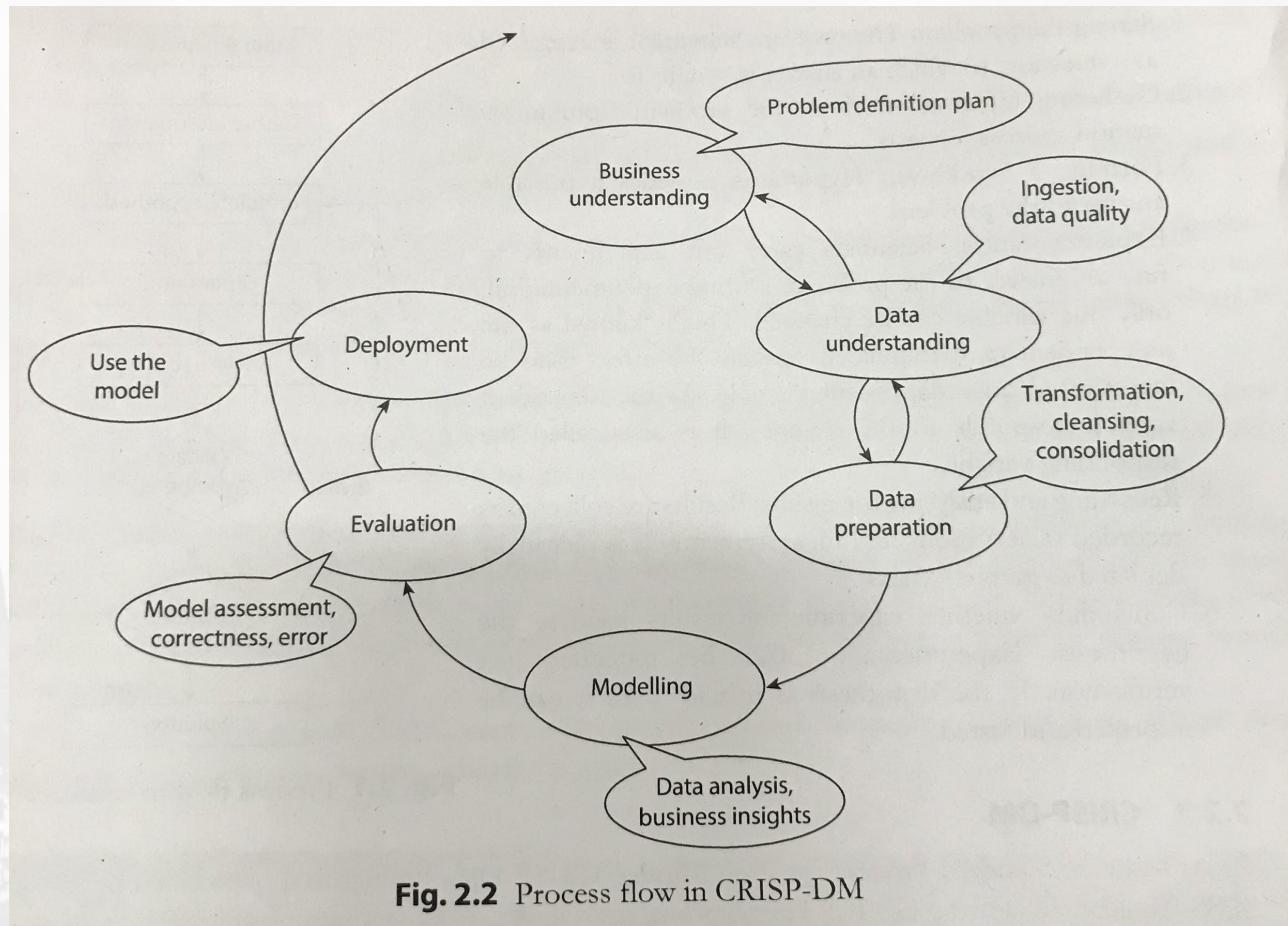


# Life Cycle of Data Centric Projects

2. CRISP-DM Phases include
  1. Business Understanding
  2. Data Understanding
  3. Data Preparation
  4. Modelling
  5. Evaluation
  6. Deployment



# Process Flow in CRISP



**Fig. 2.2** Process flow in CRISP-DM

# 3. SEMMA Methodology

SEMMA is another methodology developed by SAS for data mining modeling.

It stands for **S**ample, **E**xplore, **M**odify, **M**odel, and **A**sse. Here is a brief description of its stages –

**Sample** – The process starts with data sampling, e.g., selecting the dataset for modeling.

- The dataset should be large enough to contain sufficient information to retrieve, yet small enough to be used efficiently.
- This phase also deals with data partitioning.

**Explore** – This phase covers the understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities, with the help of data visualization.

# SEMMA Methodology

cont....

**Modify** – The Modify phase contains methods to select, create and transform variables in preparation for data modeling.

**Model** – In the Model phase, the focus is on applying various modeling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.

**Assess** – The evaluation of the modeling results shows the reliability and usefulness of the created models.

## Difference between CRISM–DM and SEMMA

- SEMMA focuses on the modeling aspect, whereas CRISP-DM gives more importance to stages of the cycle prior to modeling such as understanding the business problem to be solved, understanding and preprocessing the data to be used as input,
- E.g machine learning algorithms.

# 4. DELTA FRAMEWORK/ model

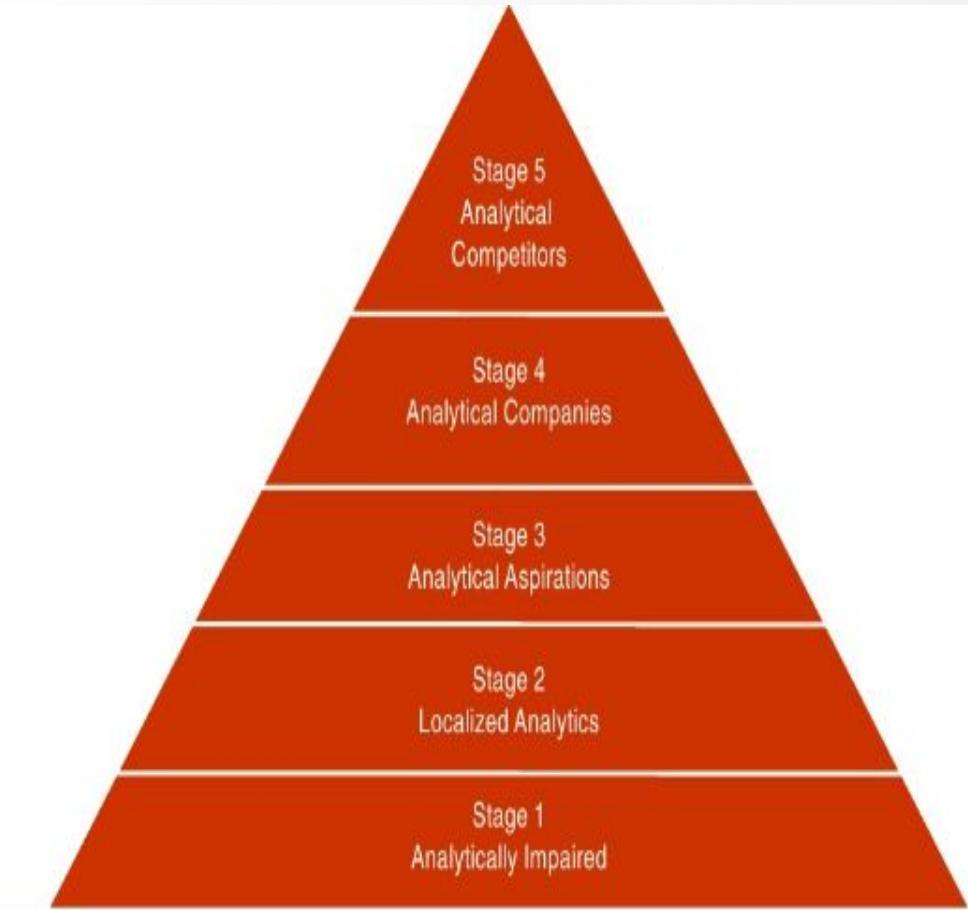
- Suggested by Davenport.



Analytical DELTA pieces

8/22/2023

- Companies can implement the Delta elements into 5 stages



Big Data Technologies

Thomas H. Davenport – Analytics at Work

# Mapping DELTA elements with organizational analytical maturity

Success Factor	Stage 1 Analytically Impaired	Moving to:			
		Stage 2 Localized Analytics	Stage 3 Analytical Aspirations	Stage 4 Analytical Companies	Stage 5 Analytical Competitors
<b>Data</b>	Inconsistent, poor quality, poorly organized	Data useable, but in functional or process silos	Organization beginning to create centralized data repository	Integrated, accurate, common data in central warehouse	Relentless search for new data and metrics
<b>Enterprise</b>	n/a	Islands of data, technology, and expertise	Early stages of an enterprise-wide approach	Key data, technology and analysts are centralized or networked	All key analytical resources centrally managed
<b>Leadership</b>	No awareness or interest	Only at the function or process level	Leaders beginning to recognize importance of analytics	Leadership support for analytical competence	Strong leadership passion for analytical competition
<b>Targets</b>	n/a	Multiple disconnected targets that may not be strategically important	Analytical efforts coalescing behind a small set of targets	Analytical activity centered on a few key domains	Analytics support the firm's distinctive capability and strategy
<b>Analysts</b>	Few skills, and these attached to specific functions	Isolated pockets of analysts with no communication	Influx of analysts in key target areas	Highly capable analysts in central or networked organization	World-class professional analysts and attention to analytical amateurs

## 5. Applied Information Economics Approach(AIE)

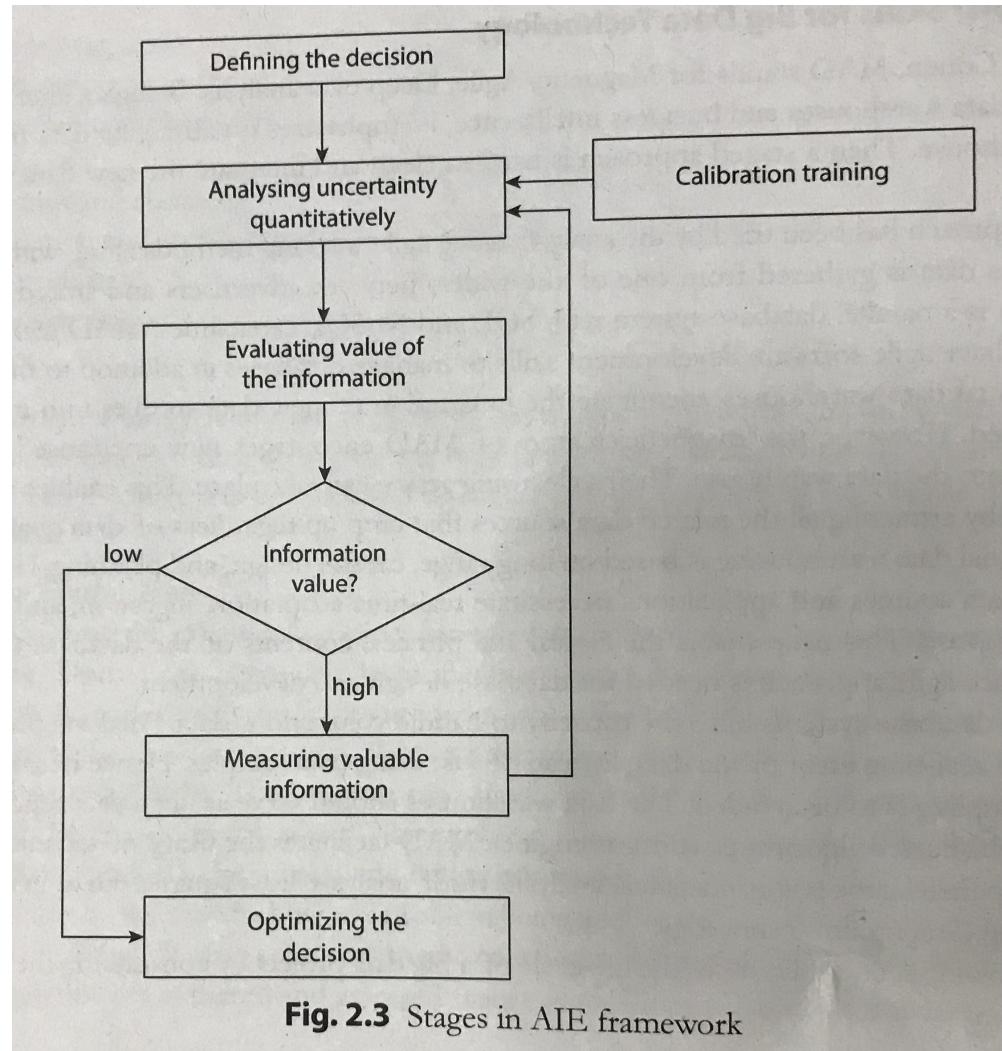
- It is a proven method for measuring intangibles, optimizing decisions and avoiding catastrophe.

Applied Information Economics (AIE) is:

1. The practical application of scientific and mathematical methods to quantify the value of management choices - regardless of how difficult the measurement challenge appears to be.
2. The optimization of the decision by optimizing the information gathering process itself – the highest payoff measurements are identified by computing the economic value of information.
3. The emphasis on using forecasting methods that have been scientifically tested to measurably reduce error of expert estimates

- "Quantifying the risk and comparing its risk/return with other investments sets AIE apart from other methodologies. It can substantially assist in financially justifying a project-- especially projects that promise significant intangible benefits." *The Gartner Group*
- "AIE represents a rigorous, quantitative approach to improving IT investment decision making....this investment will return multiples by enabling much better decision making. Giga recommends that IT executives learn more about AIE and begin to adopt its tools and methodologies, especially for large IT projects." *Giga Information Group*

# AIE Cont...



**Fig. 2.3** Stages in AIE framework



## Big Data Applications

# Big Data Applications

## Sports Domain

- To understand and study player movement
- E.g. Nike uses big data for eco-friendly product design

## Sentiment Analysis

- To understand changing customer interest, identify potential customer
- E.g. Delta Airlines

## Behavioral Analysis

- To understand customer behavior
- E.g. Amazon's product recommendations

## Healthcare

# Big Data Applications

## Customer Segmentation

- It is the grouping of similar users on their purchases and recommending suitable items for them based on personal or group interest.
  - e.g. Pandora provides music recommendation based on static profile, related songs, user interest, location.
  - Netflix uses collaborative filtering algo. to recommend the movies.
  - Amazon

## Prediction

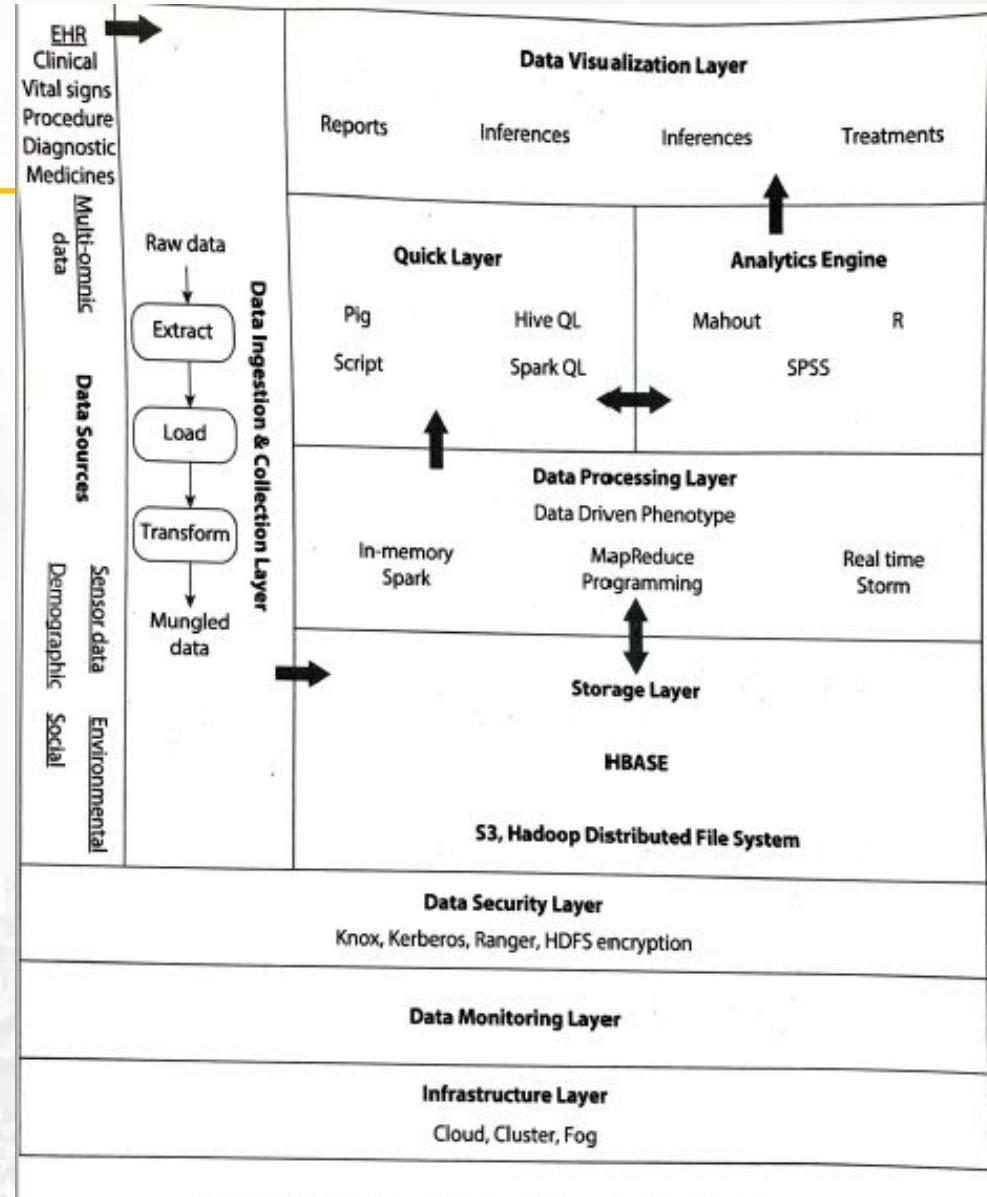
- It is the outcome done on historical information.

## Fraud Detection

- To detect prevent and eliminate internal and external frauds.
- Unusual usage pattern of a debit and credit cards can alert a bank of stolen card.

## Personalized Healthcare

# Big Data Architecture for personalized Healthcare.



**Fig. 1.5** Big data architecture for personalized healthcare

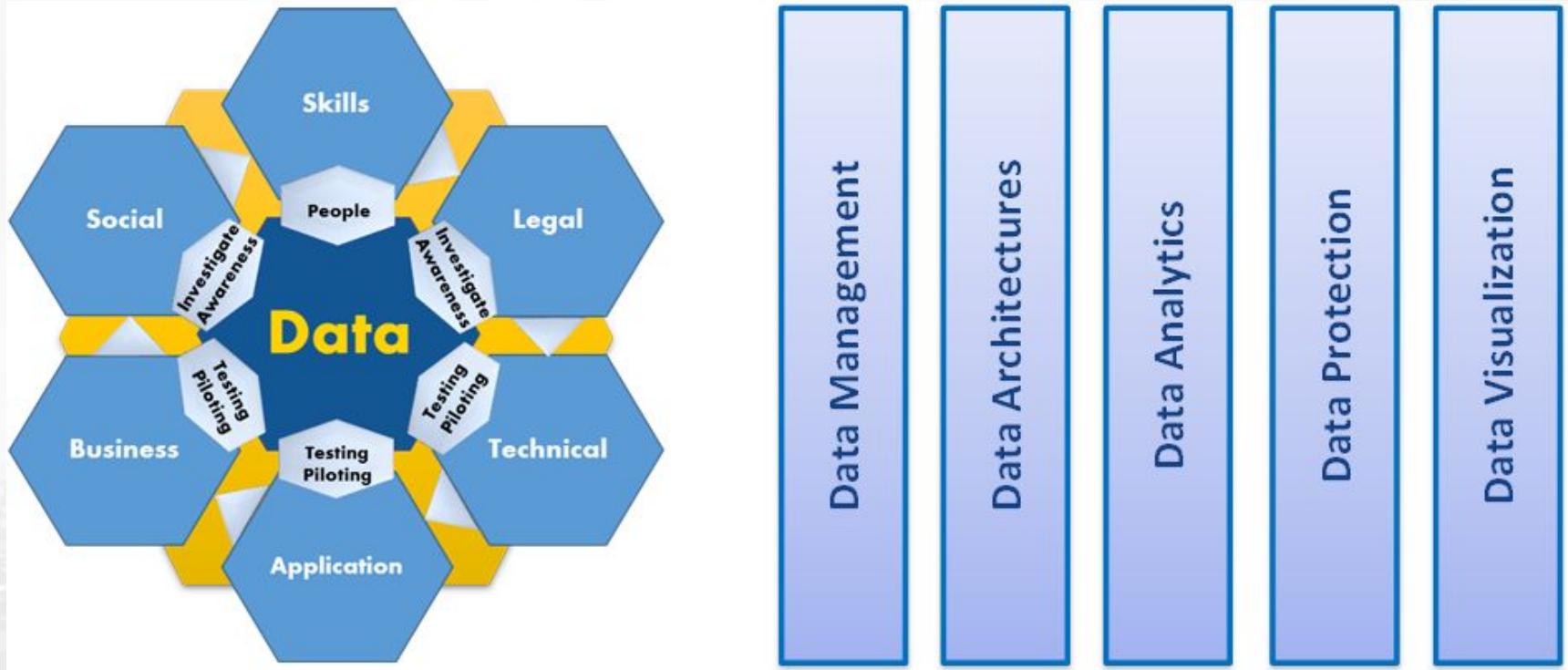
## Personal Health Care Cont....

The data processing layer extracts the Big Data Driven phenotype.

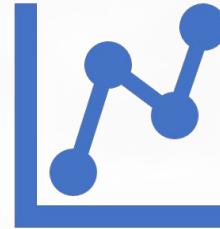
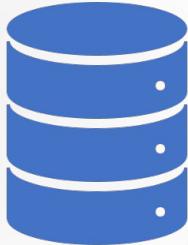
The analytic layer uses the following:

- *Descriptive analytics* to evaluate various statistics and visualize them using charts.
- *Diagnostics analytics* using survival analysis and regression to correlate survival rate of patients with heart failure.
- *Predictive analytics* using classification , clustering and inferential analysis to predict survival rate for a new patient.
- *Prescriptive analytics* for treatment plan and decision support.

# Multiple dimensions of Big Data



# DATA



**Availability of data and the access to data sources.**

**There is a broad range of data types and data sources:**

- structured and unstructured data
- multi-lingual data sources
- data generated from machines and sensors
- data-at-rest
- data-in-motion.

# DATA contd...

**Value is generated by:**

- acquiring data,
  - combining data from different sources
  - providing access to it while ensuring data integrity and preserving privacy.
- 
- **Value is added by**
  - Pre-processing,
  - Validating,
  - Analyzing
  - Augmenting
  - Ensuring data integrity and accuracy



# 1. Skills

Ensuring the availability of highly and rightly skilled people who have an excellent grasp of the best practices and technologies for delivering Big Data Value within applications and solutions.



There will be the need for data scientists and engineers who have expertise in :

analytics

statistics

machine  
learning

data mining

data  
management

## 2. Legal:

- The increased importance of data will intensify the debate on
  - data ownership and usage,
  - data protection and privacy,
  - security,
  - liability,
  - cybercrime,
  - Intellectual Property Rights (IPR) and
  - impact of insolvencies on data rights.

# 3. Technical

Key aspects including

real-time  
analytics,

low latency  
and scalable  
data  
processing,

new and rich  
user  
interfaces,

data  
interaction  
and

linking data,  
information  
and content



All have to be advanced to open up new opportunities and to sustain or develop competitive advantages.

# 4. Application



Business and market ready applications need to be a core target to allow activities to have market impact.



Novel applications and solutions must be developed and validated based on technologies and concepts in ecosystems.

# 5. Business



A more efficient use of Big Data and understanding data as an economic asset carries great potential for the economy and society.

The setup of Big Data Value ecosystems and the development of appropriate business models on top of a strong Big Data Value ecosystem must be supported in order to generate the desired positive impact on economy and employment



# 6. Social

Big Data will provide solutions for major societal challenges, such as

The improved efficiency in healthcare information processing  
or

Reduced CO2 emissions through climate impact analysis.



In parallel it is critical for an accelerated adoption of Big Data to increase awareness on the benefits and the Value that Big Data can create for business, the public sector, and the citizen

# What is Business Intelligence?

- Collecting and refining information from many sources (internal and external)
- Analyzing and presenting the information in useful ways (dashboards, visualizations)
- So that people can make better decisions
- That help build and retain competitive advantage.

# BI

- Tools and techniques to turn data into meaningful information.
  - Process: Methods used by the organization to turn data into knowledge.
  - Product: Information that allows businesses to make decisions.

# BI Applications

- Customer Analytics
- Human Capital Productivity Analysis
- Business Productivity Analytics
- Sales Channel Analytics
- Supply Chain Analytics
- Behavior Analytics

# References

- G. Sudha Sadhasivam, Thirumahal Rajkumar. Big Data Analytics. Oxford University Press ( Chapter 1, Chapter 2)
- Kevin Roebuck. Storing and Managing Big Data - NoSQL, HADOOP and More, Emereopty Limited, ISBN: 1743045743, 9781743045749
- David Dietrich, Barry Hiller. Data Science and Big Data Analytics, 6<sup>th</sup> edition, EMC education services, Wiley publications, 2015, ISBN0-07-120413-X
- <https://www.blue-granite.com/blog/advantages-of-the-analytics-sandbox-for-data-lakes>
- [https://www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209 \[image\]](https://www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209)
- [https://informationcatalyst.com \[image\]](https://informationcatalyst.com)
- [https://www.slideshare.net/hktripathy/lecture2-big-data-life-cycle \[image\]](https://www.slideshare.net/hktripathy/lecture2-big-data-life-cycle)