The objectives :

- to get introduced to the basic concepts in Data analytics.
- to understand the basic processing, storage and programming models for Data Analytics.
- the different phases of Data Analytics Project.

So is the Digital Generation who wants ‘their demand’ there, at their gadget.

Thus its competition all around where

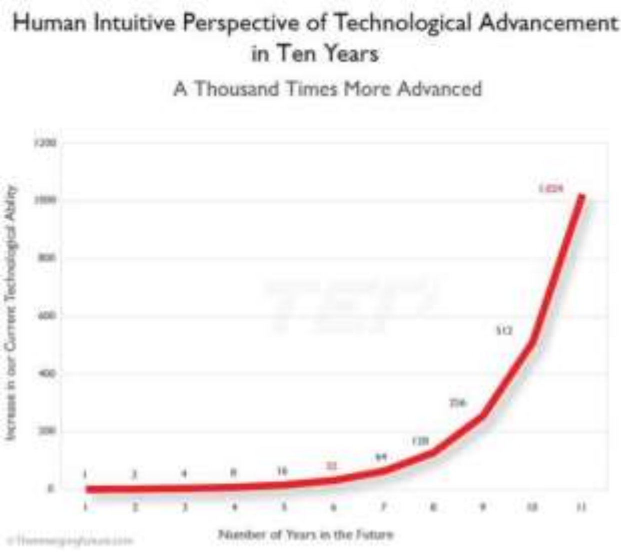
‘Fiction is demanded to become reality’

and so every next hour its an,

‘astounding Technology coming up’

‘Technologies (hardware, software, com

grow in an exponential order’…

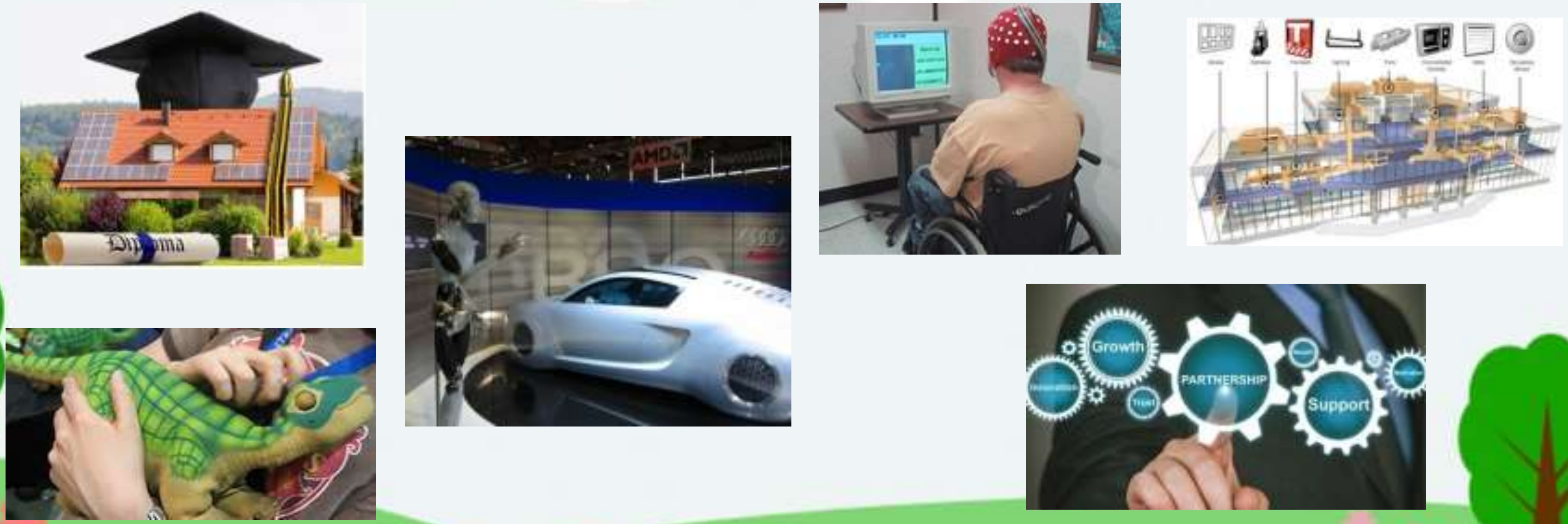


Speed of technology growth [6]

Towards 'on demand' services are today's technologies:

1. **Smart Technologies:** smart home, smart buildings, smart environments, smart cities, smart industries, smart automobiles, smart country, smart world, smart earth……

using techniques and technologies conceptualized as 'Internet of things' and 'Analytics'

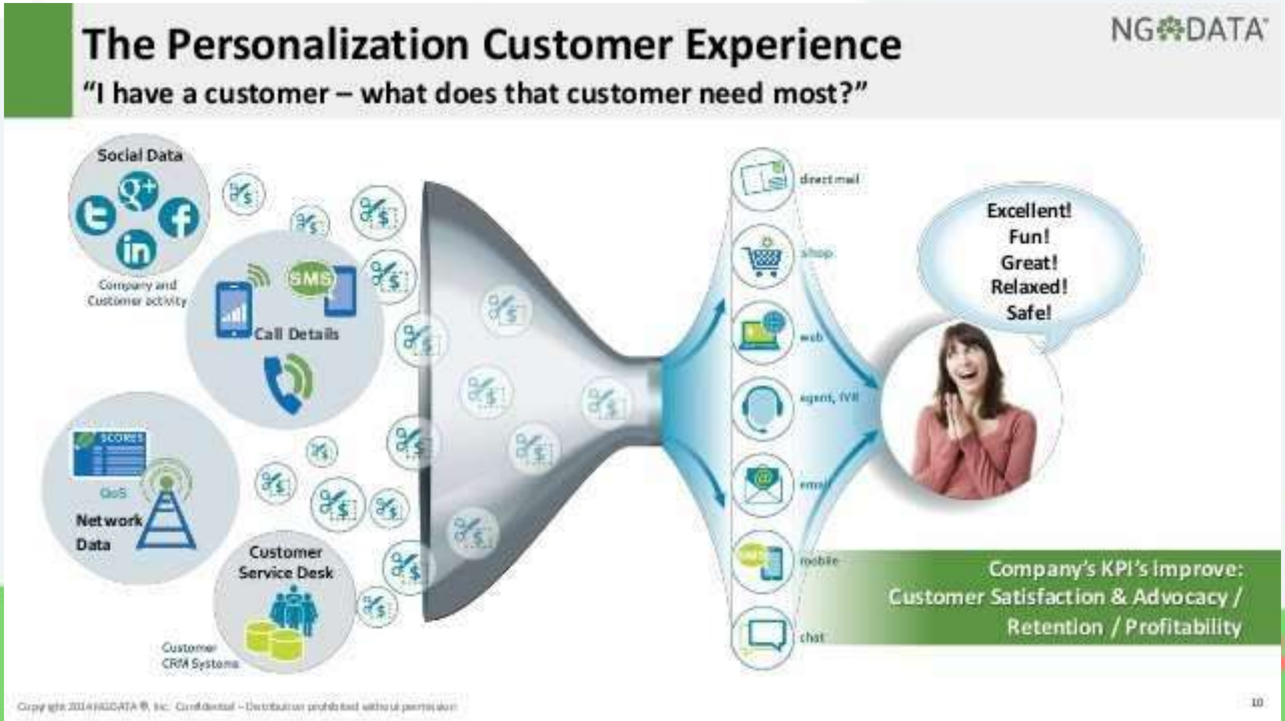


Smart technologies[8]

2. Analytics Technologies: for making perfect performance decisions on all sectors, business, medicine, social networks, bio-informatics, geo-informatics, sports, media, politics, vision framing, environment etc.

where information mound in associated massive data is obtained using techniques and technologies conceptualized as **‘Data Analytics’.**

Even they are **smart and personalized** decision making technologies. **(Analytics compliments Smartness.)**



The Personalization Customer Experience

“I have a customer – what does that customer need most?”

NGDATA
Social Data
Company and Customer activity
SMS
Call Details
SCORES
QoS
Network Data
Customer Service Desk
Customer CRM Systems
direct mail
shop
web
agent, IVR
email
mobile
chat
Excellent! Fun! Great! Relaxed! Safe!
Company’s KPI’s improve: Customer Satisfaction & Advocacy / Retention / Profitability
Copyright 2014 NGDATA ®, Inc. Confidential – Distribution prohibited without permission
10

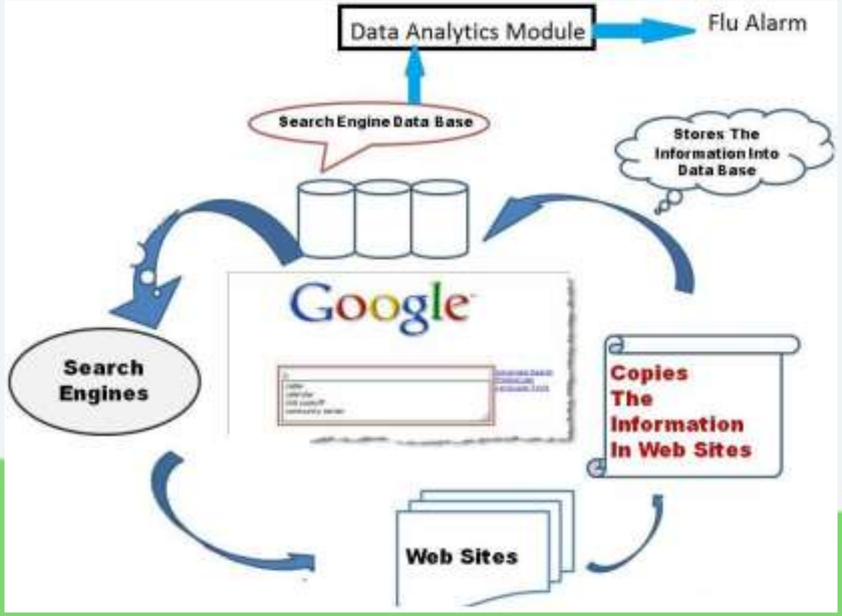
Personalized Marketing[9]

ON DATA ANALYTICS

> **Hiding within those mounds of data is knowledge that could change the life of a patient or change the world.**
>
> ***Atul Butte, Stanford***

Google,

once could conclude with its big data expertise, to raise warning for flu in the U.S. by analyzing the queries having a **'flu theme'** well before the conventional public health services [10].



Google's Prediction[5]

Well, that's Big Data Analytics.

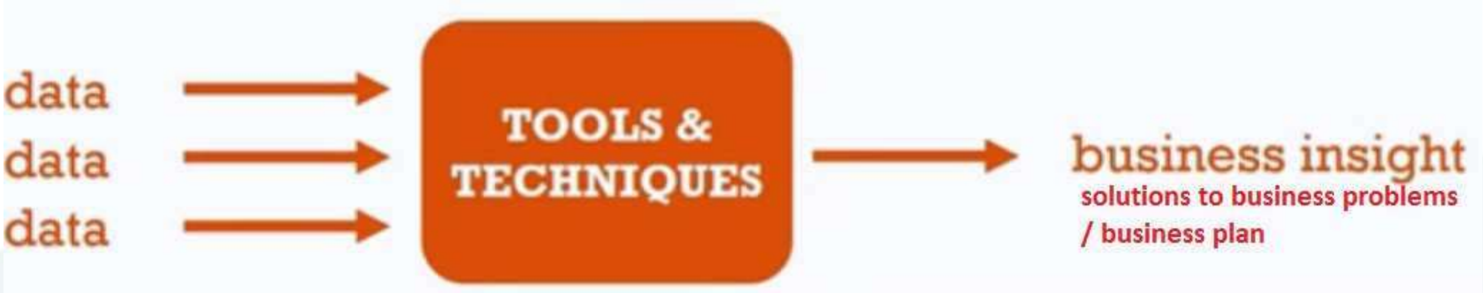
Data Analytics- Definition

Data when suitably filtered and analyzed along with other Sources related Data and a suitable Analytics applied can provide valuable information to organizations, industries, business, medicine, health, disaster management etc. in the form of prediction, recommendation, decision and the like [12].

BIG DATA ANALYTICS for applications which depend upon the value of BIG DATA

Business Analytics

Business Analytics involves business planning / making business insights/ arriving at solutions for business problems using the information and statistics from relevant/ associated data sources by applying different tools and techniques.



Business Analytics [4]

- The **data associated** with an analytics problem can be from social networks (Facebook, Twitter), relevant databases, spreadsheets. These data are **identified, gathered and organized**. Then it is subjected to analysis using tools and techniques.

- The **tools and techniques** can be statistical models or machine learning concepts etc. The tools and techniques involved are for descriptive analytics, predictive analytics, discovery analytics and/or prescriptive analytics. These analytics are for generating the statistics and other information that shall eventually lead to relevant solutions.

Applications of Business Analytics

Personalized Marketing [6]

I like red plates!
I like green plates!

Many shopping companies use Big Data Analytics for **personalized marketing** to make their customers happy.
They have **Recommendation engines** for the same.

Mobile Advertising [6]

The Big Data analytic engine of a shopping company knows the personalized needs of its customers from **shopping history**. When offers come up on the products of their interest in a particular place where the customer is around, he/she gets informed over their mobile phones.

The Big Data source associated with **customer's geographical position** is also used here.

Data Analytics-Advantages in Manufacturing Industry [10]

Big Data Analytics will always improve the functioning of any associated organization or business.

For example, In manufacturing Unit, Data Analytics can improve the following processes:

- **Procurement**- to find efficient and cost-effective suppliers.
- **Product Development-** make innovative design based on demand.

- **Manufacturing**-to find problems that can come up in the machines quality of the product.
- **Distrbution**- to enhance supply and increase inventory based on de r, holidays, economy etc.
- **Marketing**- understanding customer behaviors for personalized marketing.
- **Prices management-** Manage prices based on related co

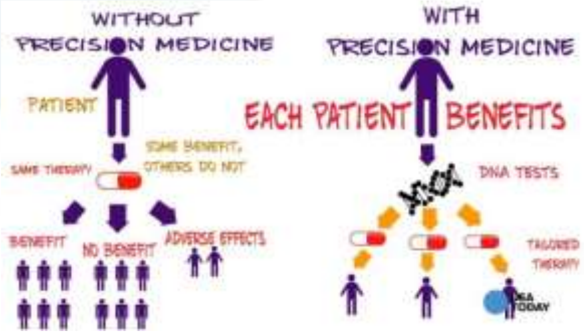
Medical Analytics: [12], [18]

Is used for

i) Precision Medicine- It involves **analyzing** his/ her genetic data, environment, day to day activities,

to **predict** the health problems so that **prescriptions** can be followed to prevent diseases.

In case a disease is diagnosed, it provides **personalized treatment** targeting a individual to provide the correct composition of medicine in the correct doze.



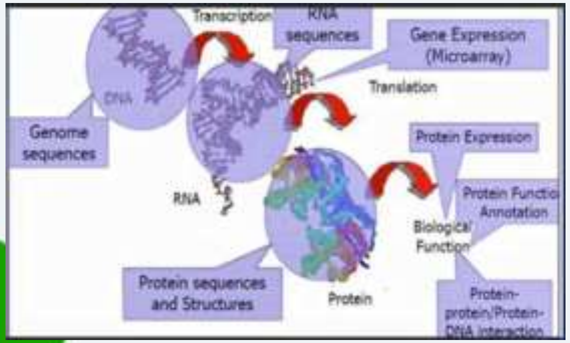
Precision Medicine-[19]

Data sources to be analyzed:[12]

1. Sensor data- from the digital bio-medical equipments, Fitness devices.

1. Organizational data- data from bio technology database maintained by public organizations like the **National Centre for Biotechnology Information,** the knowledge databases-**Gene ontology, Unified Medical Language System**.

3. People Data- self reported data from Health apps which record various human body measurements like blood sugar, pressure, heart rate, oxygen saturation level etc. or from social networks help identify actual changes that has happened some time in the body.


Transcription
RNA sequences
Gene Expression (Microarray)
DNA
Translation
Genome sequences
Protein Expression
RNA
Protein Function Annotation
Biological Function
Protein sequences and Structures
Protein
Protein-protein/Protein-DNA interaction

National Center for Biotechnology Information-[12]

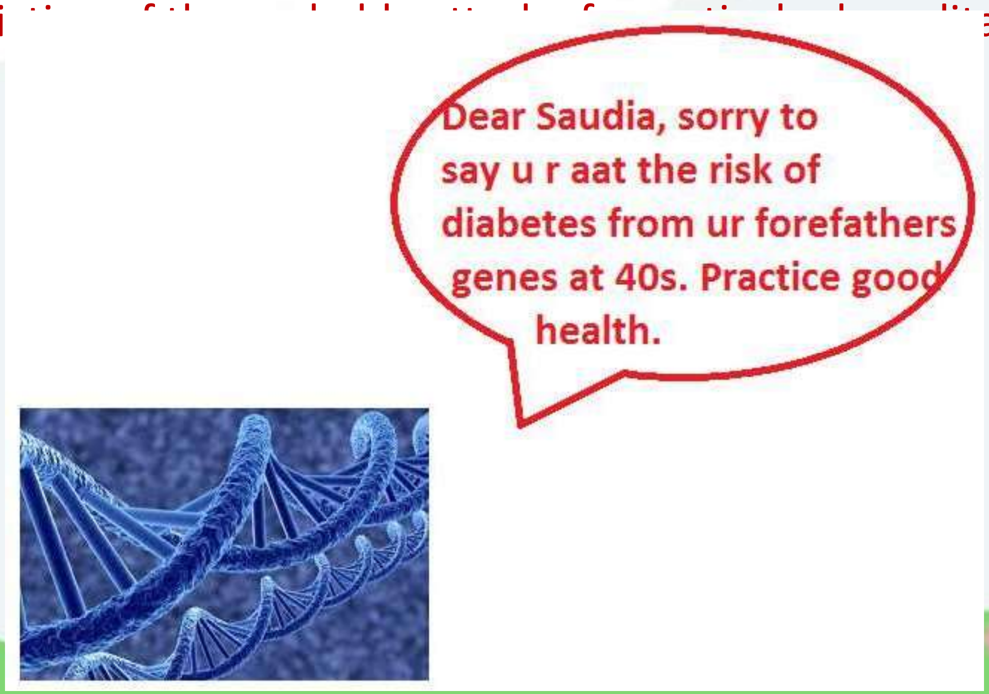
These data sources when integrated can increase the life time and well being of a human being.

Data Analytics on Bio-informatics data [12], [15]

The very **huge Genomic data** are analyzed for **personalized treatment**, **personalized/ precision medicines, better health profiles** of many genetic diseases like Diabetes, Arthritis, breast Cancer, Heart Diseases etc by analyzing all the big data that constitutes the components of disease like genomic data, metabolites, tissues, ecosystems etc.

Millions of genomes of the order of many exabytes are to sequenced for the purpose.

In fact, predi ary disease is under proposal.


Dear Saudia, sorry to say u r aat the risk of diabetes from ur forefathers genes at 40s. Practice good health.

Smart Data Analytics[12]

The very **huge Smart data** from the sensor networks of Smart projects viz… smart cities, smart homes etc. are analyzed for pollution control, security by preventing from thefts, homicide, energy conservation, traffic maintanance, disaster management and many more.

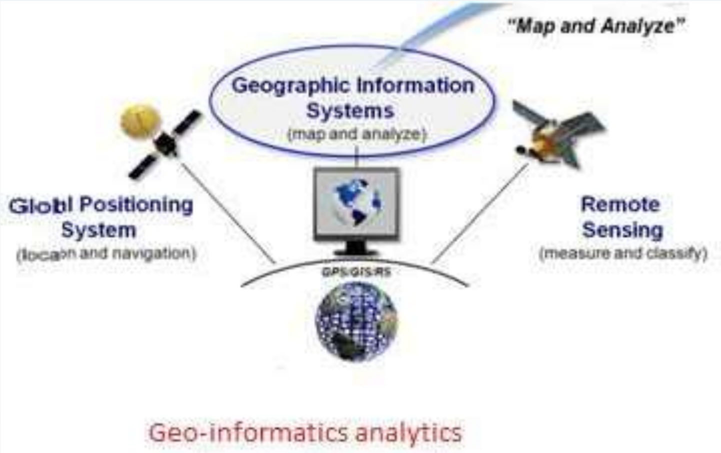
Smart city signals to Sensor network [13]

Data Analytics on Spatial Data [14]

The very **huge Spatial data** from **GPS, Radar, Lidar, Aerial data** are used for identifying, visualizing and analyzing patterns of an area with specific condition or characteristic for:

- **Tracking movements of vehicles between destinations,**
- **Public Safety,**
- **Emergency management,**
- **Climate analysis etc.**

Eg: **Economic analysis is done** based on the different attributes in the vehicle movement patterns: taxi id, distance travelled, fare etc...



Data Analytics on Geo-spatial Data [15]

Big Data Analytics-Possibilities in Financial Services [9]

In **Finance,** Data Analytics can improve the following aspects:

- **Credit Scoring**- to find people with highest credit worthiness [10].

Credit Score [10]

Fraud detection [11]

- **Fraud Detection-** to predict fraudulent transactions and customers in the financial services industry and to formulate strategies to prevent or minimize damages from it [11].

- **Claims analysis**-to find positive and negative factors that affe , [13].

Fraud detection [12]

Exercise:

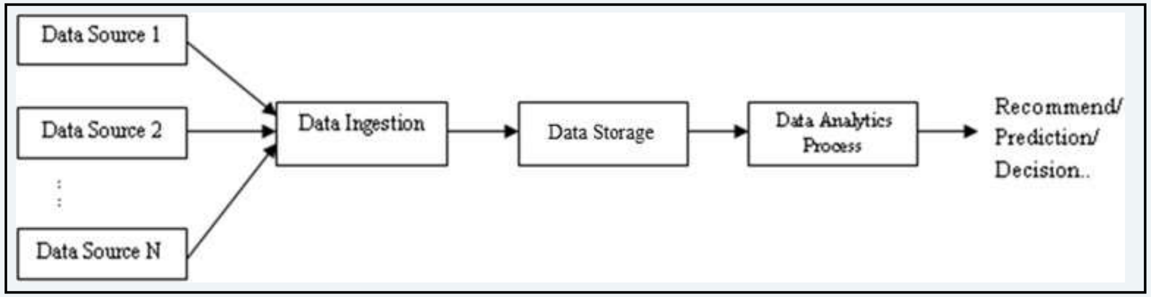
Find out any five business problems that needs analytics.

In this manner, different organizations, fields: Transportation, Health, Education, Government, Travel, Healthcare, Telecom Consumer Goods Industry etc. can get benefitted in all their respective aspects through Big Data Analytics.

Explore…..

Thus, from all those applications, **the general flow of Data Analytics** is as follows:

1. Identifying the associated Big Data Centers for the problem.
2. Filtering the required data.
3. Make the analytics using relevant concepts like: Predictive Analysis/ Sentiment Analysis/ Natural Language Processing/ Machine Learning/ Image Processing .
4. Make prediction/ decision/ suitable recommendation.



Big Data Analytics Process

Examples: Twitter   Flume   HDFS   Hive, R..   Output Message

STAGES OF BIG DATA BUSINESS ANALYTICS [9, 14-17]

The different stages of business analytics are:

1. **Descriptive analytics:**

Here the information that is present in the data is obtained and summarized. It is primarily involved in **finding all the statistics** that describes the data.

Eg: How many buyers bought A.C. in the month of December previous years?

2. Diagnostic/ Discovery Analytics:

This stage involves finding out the reason for the statistics determined in analytics stage. Otherwise it involves, **why that statistics** have happened?

Eg: Why there is an increase/ decrease in the sales of A.C.in the month of December?

3. Predictive Analytics:

This stage involves predicting the possible future events based on the information obtained from the Descriptive and/or Discovery Analytics stage risks involved can be identified.

Eg: **What shall be** the sales improvement next year **(making insights for futur**

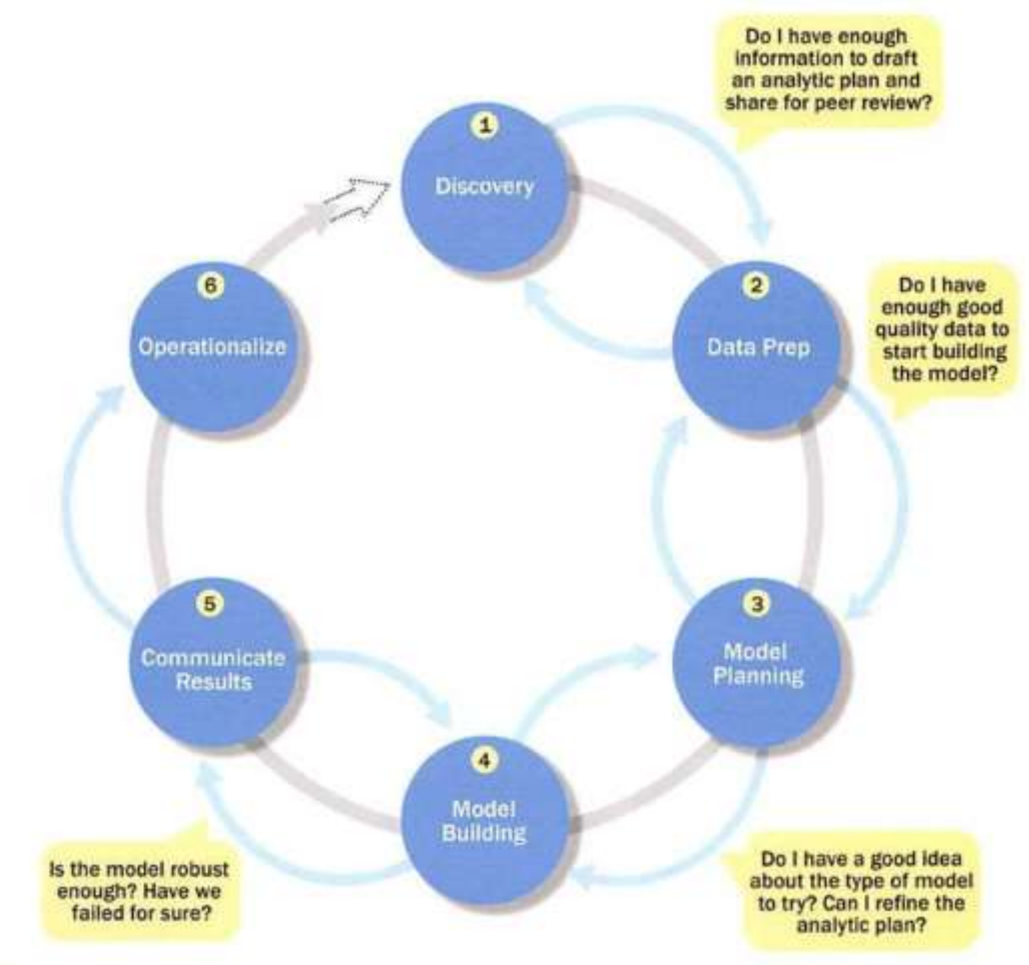
4. **Prescriptive Analytics:**

It involves **planning actions or making decisions** to improve the Business based on the predictive

Eg: **How much amount of material** should be procured to increase the production?

LIFE CYCLE OF BIG DATA BUSINESS ANALYTICS PROJECT [9], [12]

The life cycle has a circular movement. An iterative movement happens between two phases until sufficient information is obtained by the project team to move to the next phase. The life cycle of Business Analytics involves the stages as shown in the figure below.


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

Data Analytics Methodology [9]

Life Cycle of Big data Business Analytics Project [9], [18] contd…

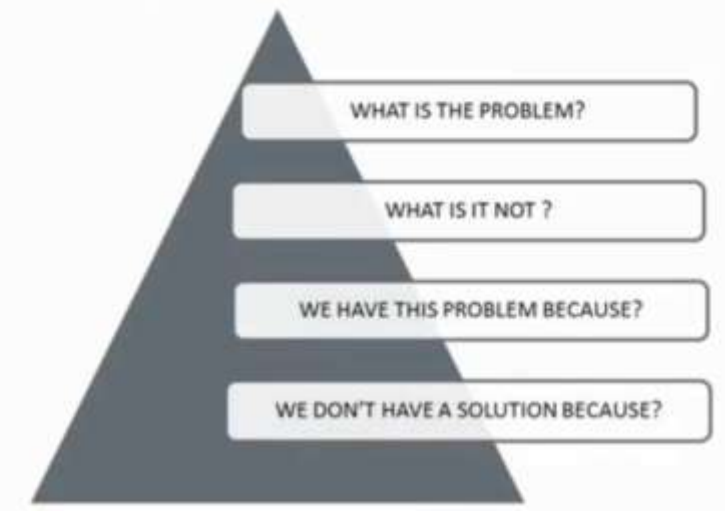
The life cycle of Business Analytics involves the following phases:Discovery, Data Preparation, Model Planning, Model Building and Communicate Results.

1. **Phase 1: Discovery**

a) The most important activity in this phase is **Problem Definition** which involves

i) Stating the problem generally. Then **gathering initial hypotheses** from stakeholders and domain experts to have better perspective about what the problem is.

ii) Analyzing all the literature associated with the problem and answer the following questions: What is the problem?- State the problem generally.
What is it not? – problems that are not caused Why the problem?- find the cause of the problem
Why no solution? – to find out why the problem is not solved.
The answer to the question shall solve the problem.


WHAT IS THE PROBLEM?
WHAT IS IT NOT ?
WE HAVE THIS PROBLEM BECAUSE?
WE DON'T HAVE A SOLUTION BECAUSE?

Problem Definition [9]

Examples of Business analytics problems:

1. T fix the best price for a particular product.

1. To find the best market for a product.

1. To find the best time in a year to procure stock.

4. To make best design modifications for a product.

b) Learning the business **domain**

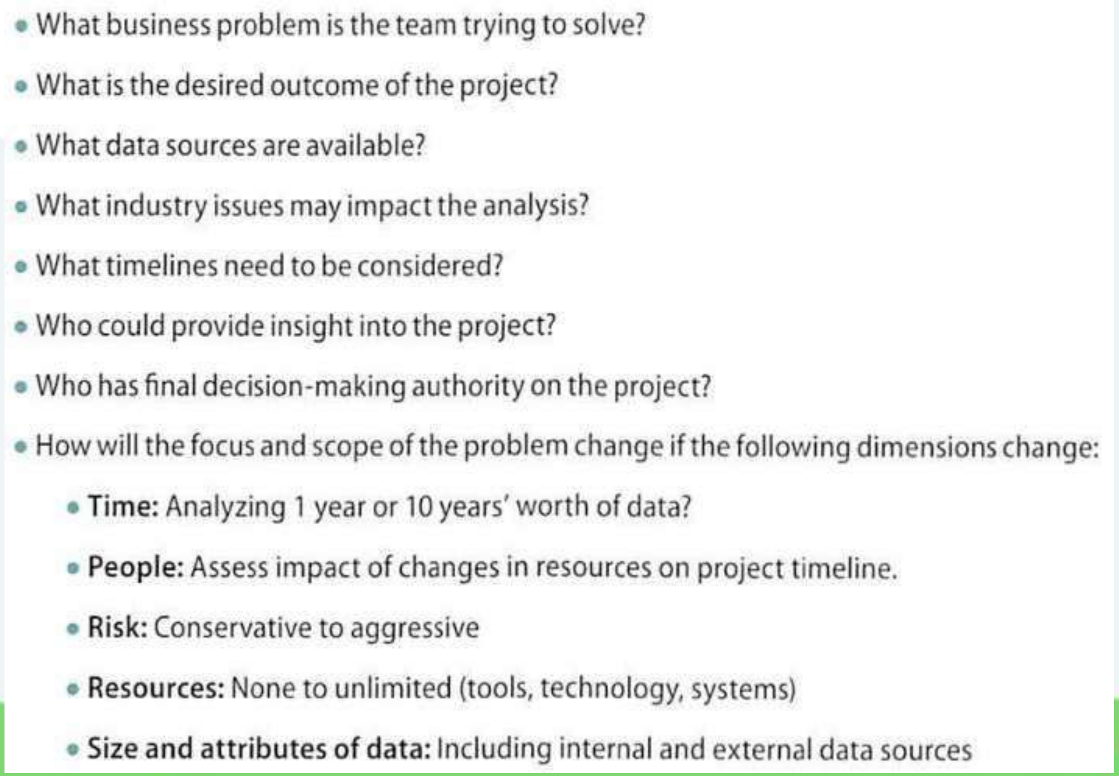
It involves

i) finding the required **domain or business knowledge** which are to be acquired for planning and building models in the later Phase 3 and Phase 4. This assessment is needed to find the resources needed for the analytics project.

ii) identifying the **key stakeholders** who will include anyone who will benefit from the project and who knows more about the kind of requirements.

iii) discussing with **the project sponsors** to get an idea of the requirements and the potential working solution.

Common questions to interview a project sponsor:



- What business problem is the team trying to solve?
- What is the desired outcome of the project?
- What data sources are available?
- What industry issues may impact the analysis?
- What timelines need to be considered?
- Who could provide insight into the project?
- Who has final decision-making authority on the project?
- How will the focus and scope of the problem change if the following dimensions change:
    - **Time:** Analyzing 1 year or 10 years' worth of data?
    - **People:** Assess impact of changes in resources on project timeline.
    - **Risk:** Conservative to aggressive
    - **Resources:** None to unlimited (tools, technology, systems)
    - **Size and attributes of data:** Including internal and external data sources

c) Assessing the **resources** for project viz., tools, technology, people, data are also assessed during the phase I.

It involves finding

i) the tools and techniques necessary for **building** the project.

ii) the tools and techniques needed to **operationalize** the project.

iii) the tools and techniques required in **future for the long term success** of the project.

iv) if the **skills needed are present** in the organization or if it should be created.

v) if **data** available is enough/ should it be brought from somewhere/ should the available data be transformed to carry out the project. Also the volume, type and time span of the data for testing initial hypotheses.

2. Phase 2: Data Preparation

The iterative stage is an important stage of the project lifecycle and takes more than 50% of the project's time.

a) Preparing a analytical **Sandbox (Workspace)**

It involves

i) Creating a similar and complete **synthetic dataspace** to do the analytics. (The analytics is not done on the actual data associated with the problem).

The size of the sandbox must be **5-10 times** the size of act

ii) Collecting **all kinds and types of data** associated with the

data,

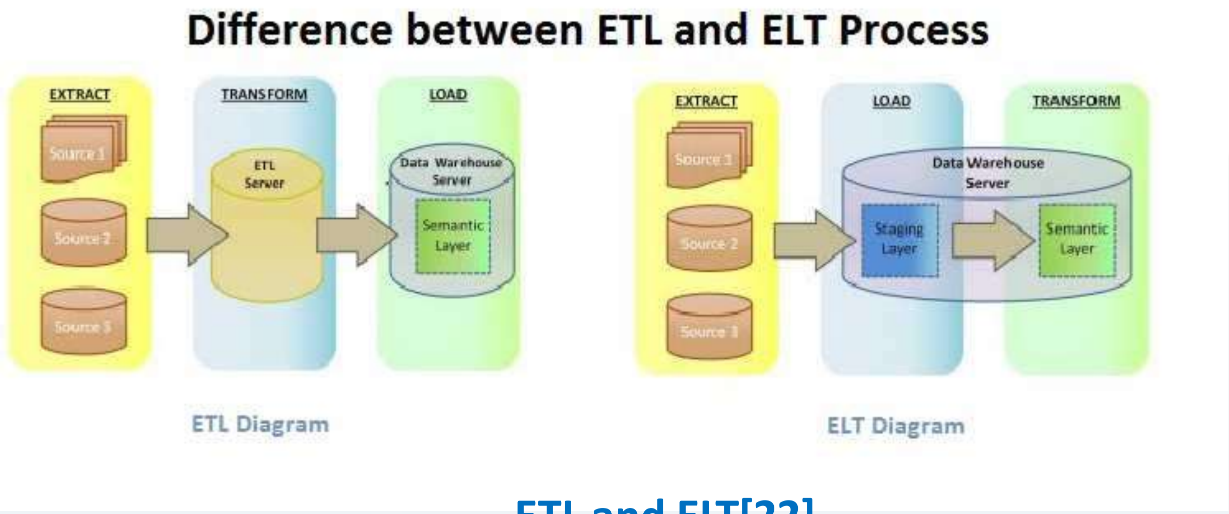
structured data, text data, call logs, web logs etc..

Sandbox[21]

b) The stage involves data extracting, loading and transforming (ELT) processes for analytics.

It involves

i) **extracting data** in the raw form, **loading** into the datastore and the **transforming** the data in to a new state or leave it in the original state (The last stage is done by the analysts). In analytical sandbox, the traditional **ETL is not followed** because transformation may filter fine nuances in the data which may be needed for certain analytics as is the case with **fraudulent transaction** detection.



ETL and ELT[22]

Large sized data from different datastores can be parallelized to move into sandbox using the tools: HADOOP.

Many websites and social networks provide APIs to access data for a project. Eg: **Twitter API** helps download millions of Tweets. Twitter data is public.

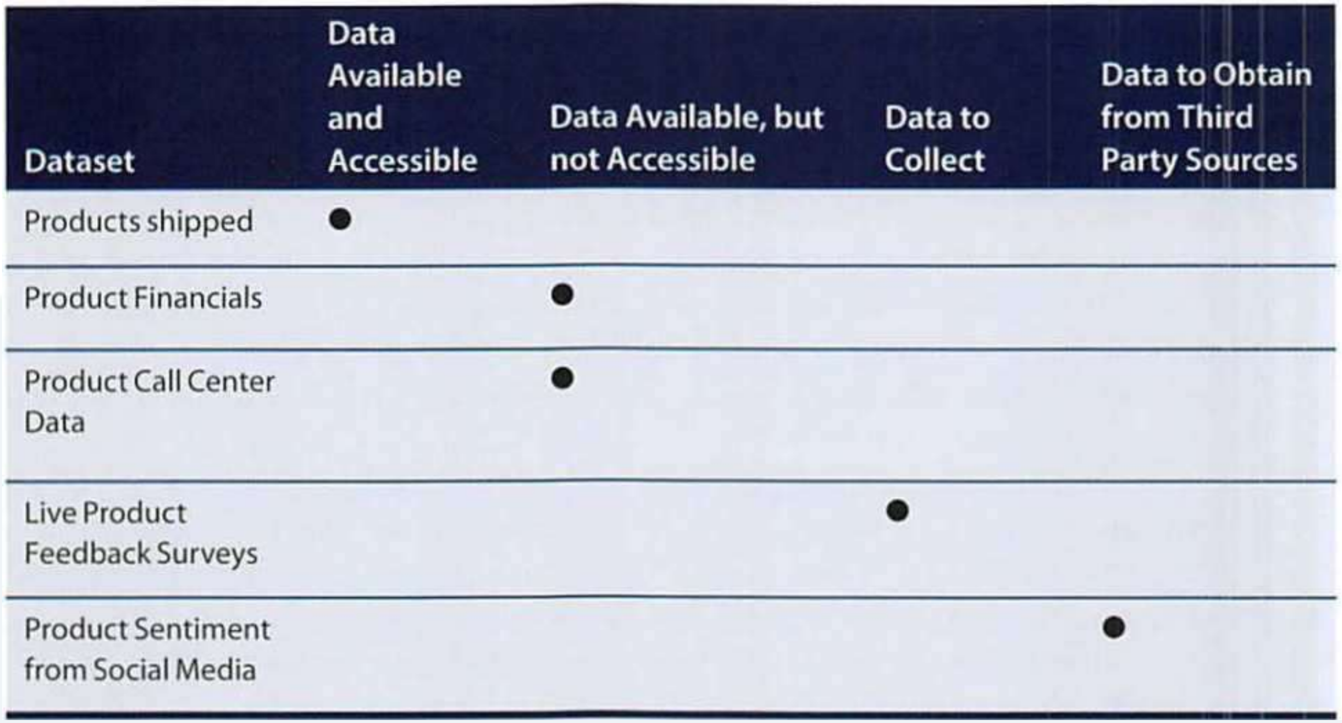
c) Developing familiarity with the data.

It involves

i) understanding the data in detail.

ii) visualizing the data to identify the trends, outliers, and relationships between data variables.

iii) finding the datasets that are needed but not available and triggering ways to collect new data through open APIs or shares or by purchase.



Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Sample Data Set for a Project [12]

d) Conditioning the data.

It involves

i) cleaning the data.

ii) normalizing the data: checking the consistency of data, checking the consistency of data types, reviewing the content of data columns, evidences of errors (breakage in data feed).

iii) it involves analyzing which aspects of a particular database will be useful for later analysis stages.

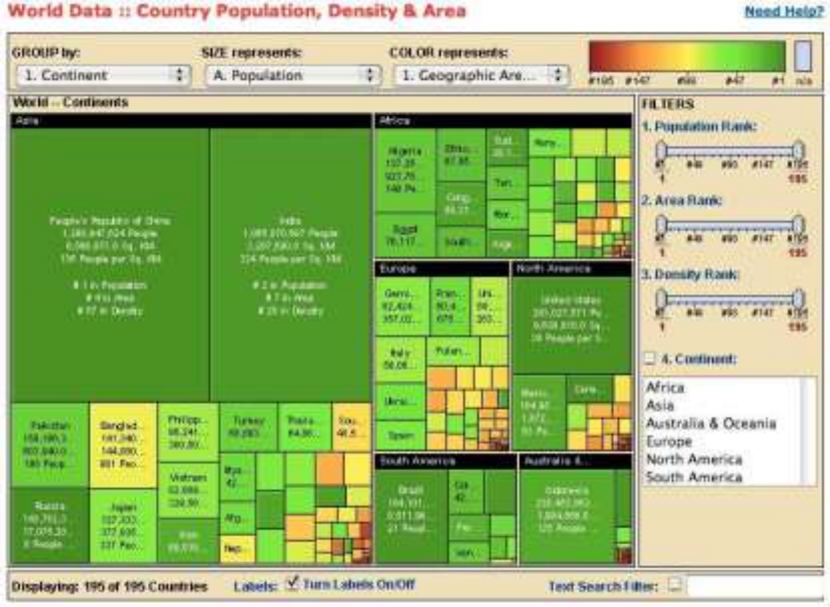
iv) it is normally done by data owners, DBA or data engineers unde4 the supervision of the Data Scientist.

e) Survey and Visualize the data.

It involves

i) examining if the data has any unexpected values or dirty data.

ii) The visualization approach is : overview first, zoom and filter and then details on demand [23].



World Data :: Country Population, Density & Area

GROUP by: 1. Continent

SIZE represents: A. Population

COLOR represents: 1. Geographic Are...

Displaying: 195 of 195 Countries

Labels: ☑ Turn Labels On/Off

Text Search Filter: □

A visualization tool [24]

Tools for Data Preparation

i) **Hadoop**- for parallel ingestion of massive data from multiple sources.

ii)**Alpine Miner-** for creating analytic workflows (eg: first selecting first 10 customers and the doing descriptive analytics then clustering)

iii)**Open Refine-** robust GUI based open source for data transformation.

iv) **Data Wrangler-** for data cleaning and data transformation.

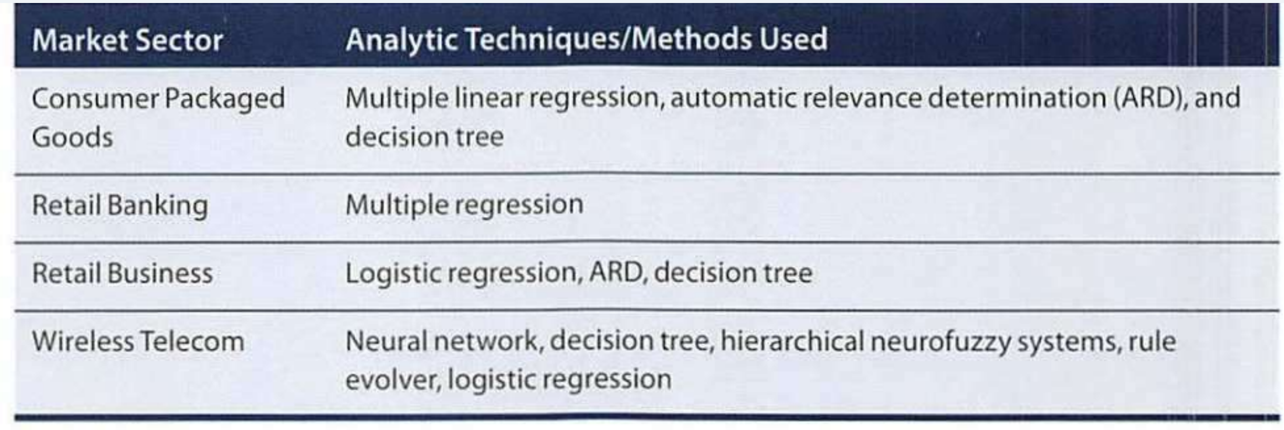
3. Phase 3: Model Planning

a) The data science team identifies the candidate models to do data clustering, classifying and for finding the relationships between data corresponding to the initial hypotheses for project objectives.

It involves finding

i) if a single model or a series of analytical models required. Eg models: association rules, logistic regression

ii) capture the most essential predictors and variables. Eg: If the team plans regression analysis, identify candidate predictors and outcome variables of the model.



Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression

Research on Model Planning [12]

Tools for Model Planning

i) **R**- has 5000 packages for data analysis and garphical representation. It can interface with databases with ODBC for parallel ingestion of massive data from multiple sources.

ii) **SQL Analysis Services-** for database analytics of data mining functions, aggregations and basic predictive models.

iii) **SAS/ACCESS** can be used to connect to relational databases, (Oracle and Teradata), enterprise applications (SAP and Salesforce.com),

.

4. Phase 4: Model Building [12]

a) The data science team develops datasets for training, testing and production purposes.

It involves

i) using the training set of data for conducting the initial experiments and using the test data sets for validating an approach.

ii) checking at the end if the model accounts for all the data and if it has robust predictive power.

Also all the assumptions made in the modeling process must be regarded and the following questions are to be answered:

1. Does the model appear valid and accurate for the test data?
2. Does the model output make sense to the domain experts?
3. Does the model avoid intolerable mistakes?
4. Are more data needed?
5. Will the model chosen support run-time requirements?
6. Is a Different form of model required for meeting the problem?

Tools for Model Building

Commercial tools:

i) **SAS Enterprise Miner**- allows users to run predictive and descriptive models based on enterprise data.

ii)**SPSS Modeler-** offers methods to explore and analyse data through a GUI.

iii) **Matlab-** for a variety of data analytics algorithms and data exploration.

iv) **STATISTICA, MATHEMATICA-** data mining and analytics tools.

Open Source tools:

i) **R**- R commands can be executed in database.

ii) **Octave-** has functionality of Matlab and is suitable for machine learning.

iii) **WEKA-** a free data mining software.

iv) **Python-**programming language for machine learning and analysis.

v) **SQL-**for database implementations.

5. Phase 5: Communicate Results[12]

a) The data science team considers how best to articulate thhe findings and outcomes of the various team members and stakeholders, taking into account caveats, assumptions and limitations.

Communicating Results[25]

It involves finding

i) if the team succeeded in objectives.

ii) how the results are to be communicated to the team members and stakeholders. Often all the results are recorded and most significant ones are shared with the stakeholders.

6. Phase 6: Operationalize [12]

a) The data science team communicates the benefits to the team more broadly and sets up a pilot project to deploy the work in a controlled way before broadening the work to a full enterprise or ecosystem.

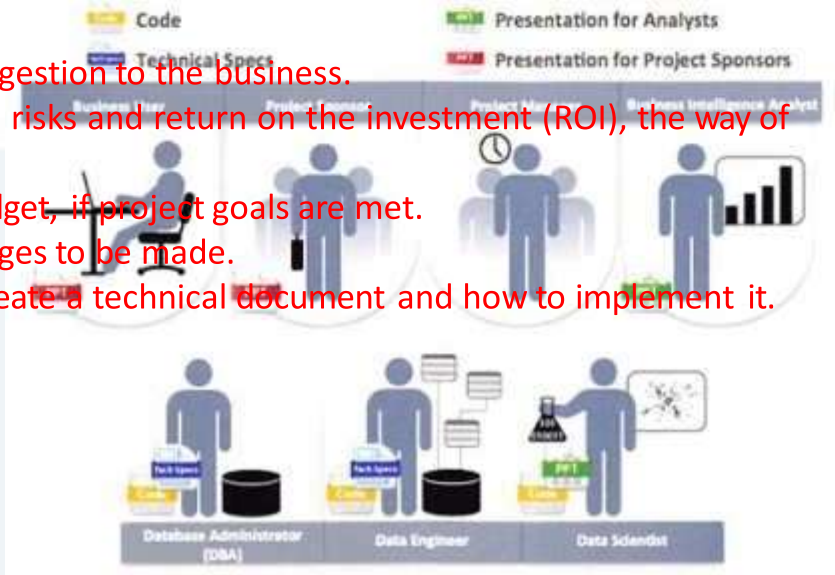
It involves

i) learning the performance and related constraints of the model on a small scale and make all the necessary adjustments before making the final full deployment.

ii) performing ongoing monitoring of model accuracy and if accuracy degrades, retrain the model.

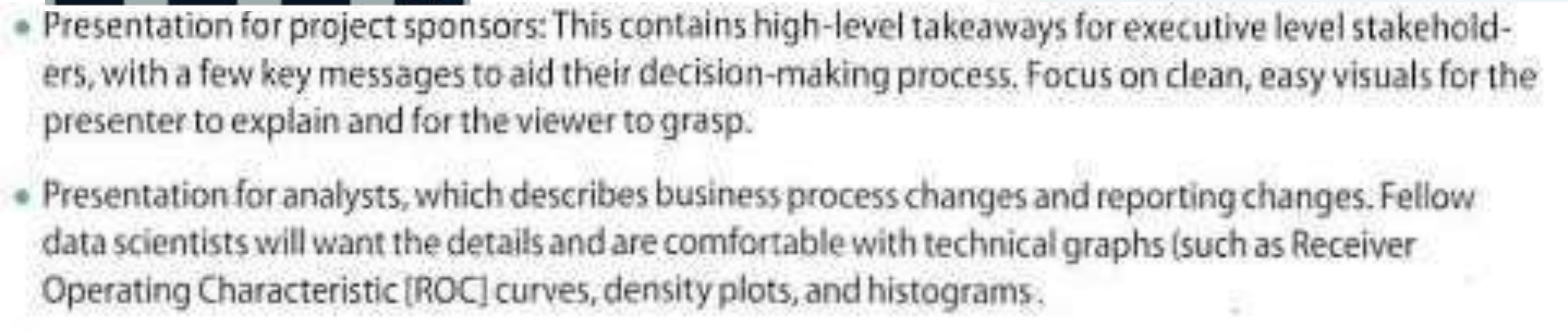
Key outputs of the project from the main stake holders:

1. **Business User:** benefits of the business and implication/ suggestion to the business.
2. **Project Sponsor:** impact of the project on the business, the risks and return on the investment (ROI), the way of putting the project into the enterprise.
3. **Project Manager:** if project is completed within time and budget, if project goals are met.
4. **Business Intelligence Analyst:** reports and dashboards changes to be made.
5. **Data Engineer and DBA:** share the code from the project, create a technical document and how to implement it.
6. **Data Scientist:** share the code, explain the project to other stakeholders.


Code
Technical Specs
Presentation for Analysts
Presentation for Project Sponsors
Business User
Project Sponsor
Project Manager
Business Intelligence Analyst
Database Administrator (DBA)
Data Engineer
Data Scientist

Key outputs from a successful analytics project [12]

So the key deliverables of the project for the main stake holders are:



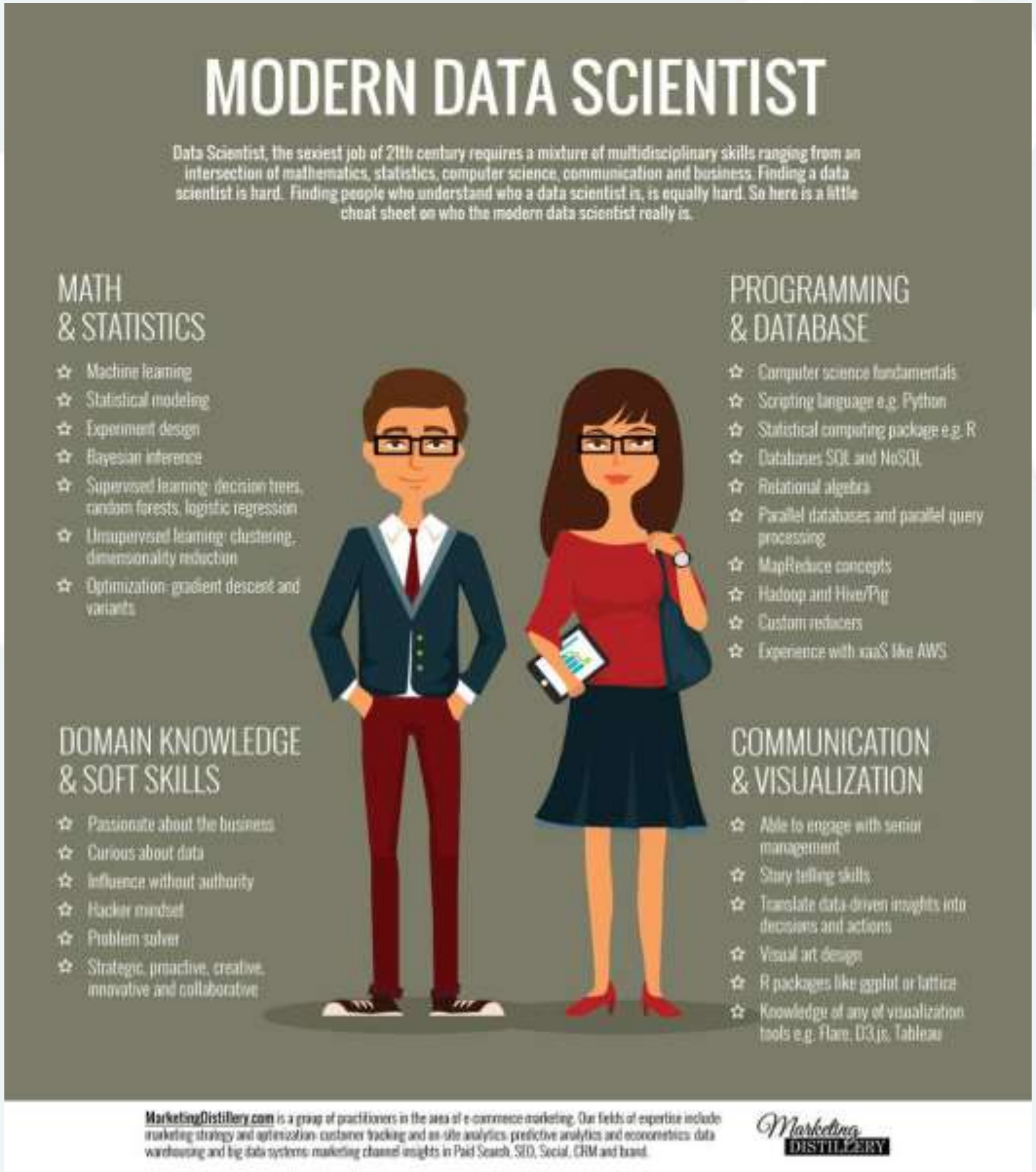
- Presentation for project sponsors: This contains high-level takeaways for executive level stakeholders, with a few key messages to aid their decision-making process. Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.
- Presentation for analysts, which describes business process changes and reporting changes. Fellow data scientists will want the details and are comfortable with technical graphs (such as Receiver Operating Characteristic [ROC] curves, density plots, and histograms .
- Code for technical people.
- Technical specifications of implementing the code.

ROLE OF DATA SCIENTISTS [9].

A data scientist should be a person who

- looks at data and makes out the trends in it.
- does descriptive, discovery, predictive and prescriptive analytics on the data i.e., finds out the hidden story in the data, makes insights and takes suitable actions/ decisions.
- works with application developers to find the suitable data for analysis.
- make the plan for doing analytics for specific results.
- makes effective data mining architecture.
- makes reports.

Who is a data scientist?



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

Becoming a Data Scientist [19]

KEY ROLES FOR A SUCCESSFUL ANALYTICS PROJECT [9].

1. Business User:

- Understands the domain area
- Benefits from the results
- Knows the values of the results.

A subject matter expert fulfills the role.

2. Project Sponsor:

- Funds the project
- defines the core business problem
- Finds the values of the results.
- Sets priorities for the desired project.

2. Project Manager:

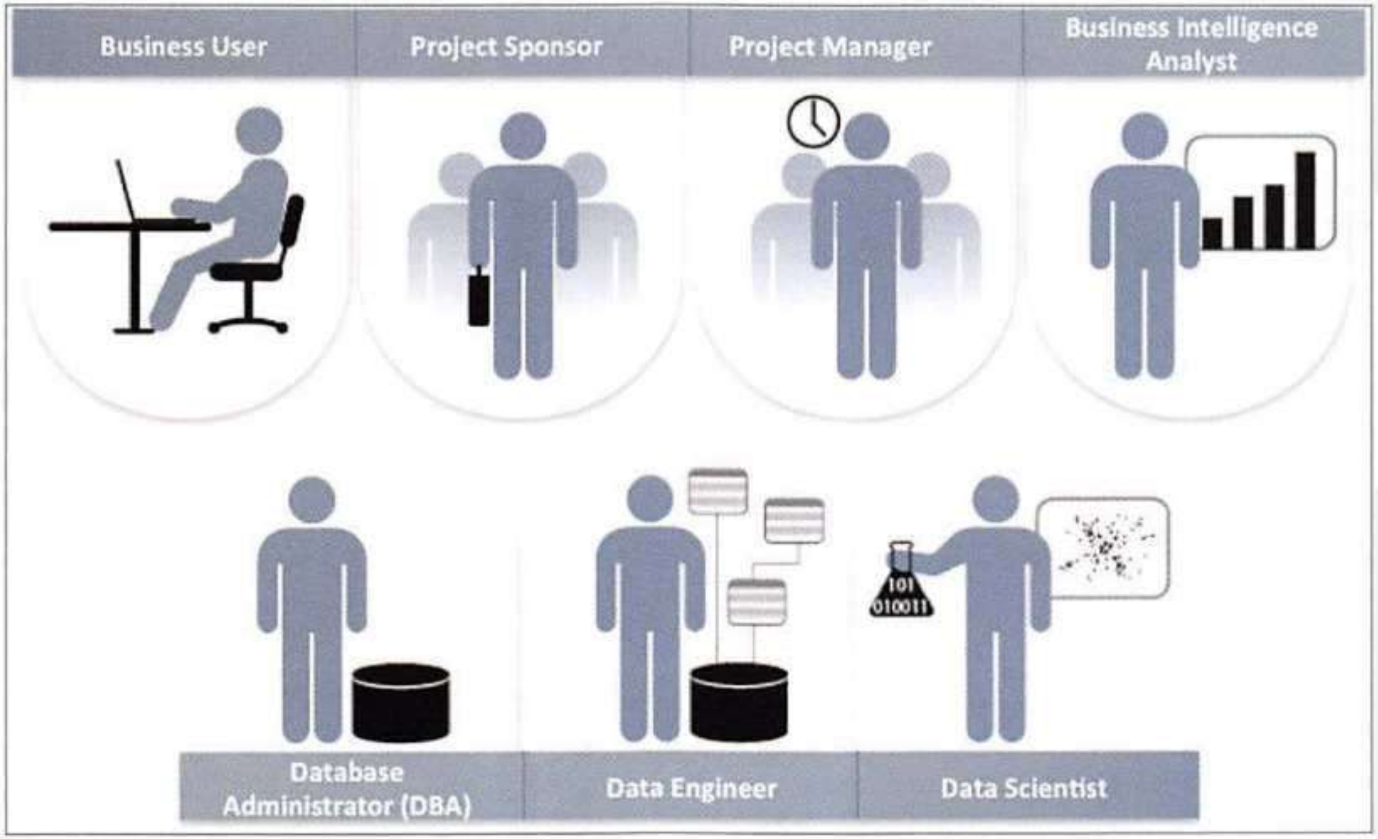
- Ensures that the milestones are met on time.

2. Business Intelligence Manager:

- Reports an understanding of the data, Key performance indicators, Key metrics.

2. Database Administrator:

- Configures database environment to support analytics.



Key Roles for a successful Analytics Project [9].

6. Data Engineer:

- Leverages technical skills to assist tuning SQL queries.

6. Data Scientist:

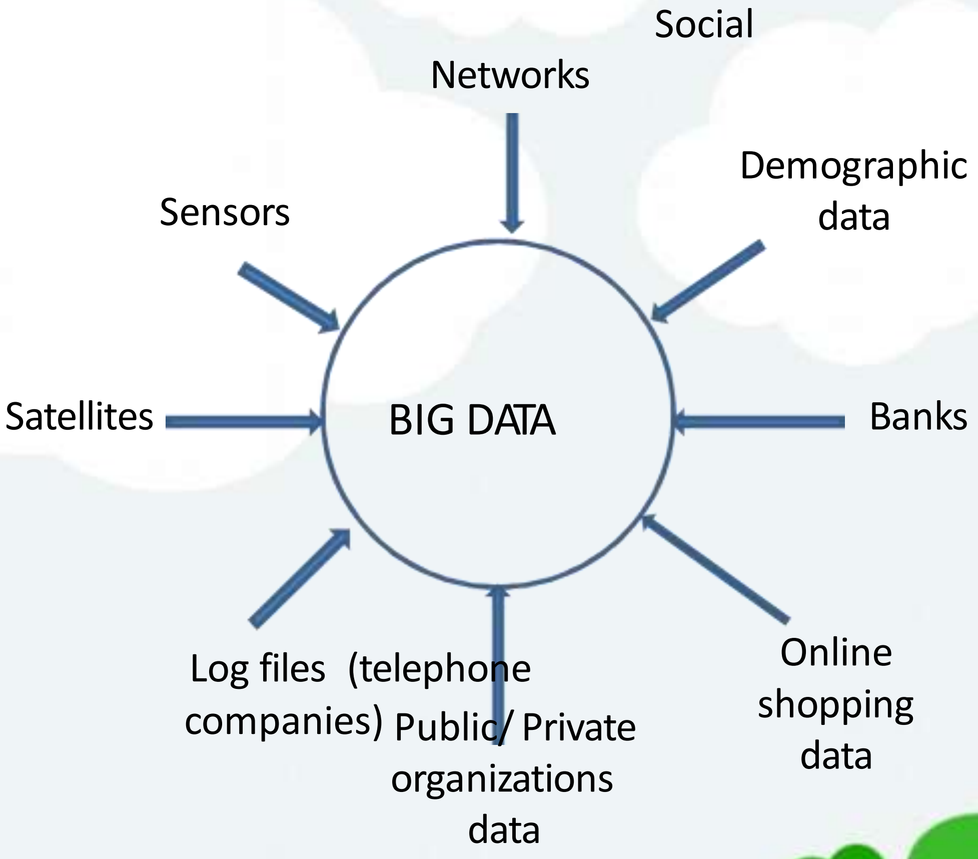
- Provides subject expertise for analytical techniques, data modeling for business problems.

BIG DATA

As the name implies, Big Data is huge data. In fact, very huge of the order of Petabyte (1PB=1000TB) or exabyte (1EB=1000PB)

Examples:

1. **Tweets from Twitter**-1000s/ second.
2. **Facebook**- 1000 of comments, 293,000 statuses, 1,36,000 photos uploaded/minute.
3. **Walmart, a departmental store chain**-one million customer transactions/ hour



Social Networks

Demographic data

Sensors

Satellites

BIG DATA

Banks

Log files (telephone companies)

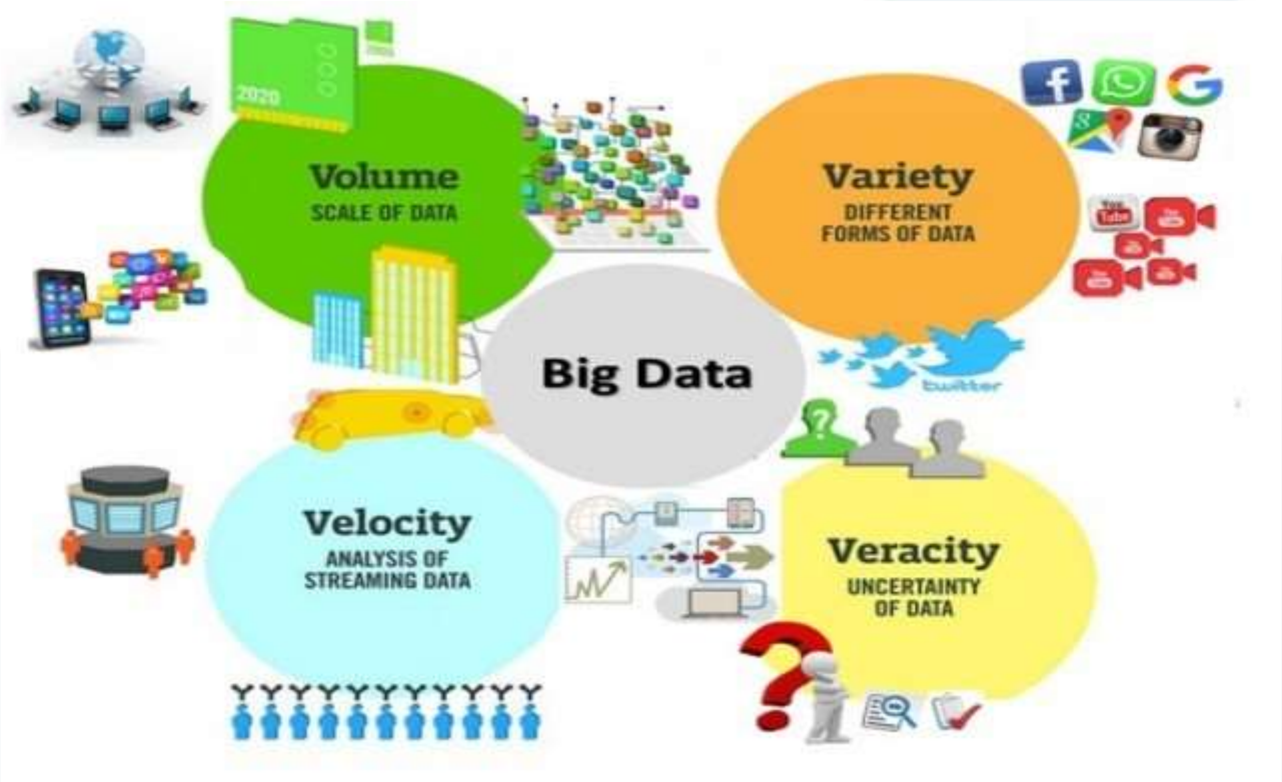
Public/ Private organizations data

Online shopping data

Big data Sources

2. BIG DATA Features

From the examples of Big Data mentioned in the previous slide, Big Data can be characterized to have the following features:



Four Vs of Big Data [2]

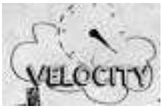
The big data growth is associated with the following four ‘Vs’.

1. **Volume:** It corresponds to the amount of data associated by an organization or individuals. The data amount can be of the order of petabytes, exabytes, zettabytes (1000 EB) .

   By 2020 , it is estimated that online transactions shall be 450 billion / day.

2. **Velocity:** It corresponds to the rate at which data is generated, captured, shared and processed in/ from Big Data sources. The data flow from such sources is continuous and more fast and in large quantity like a water fall.

   Case: Ebay analyzes 5 million transactions/ day to analyze frauds from PayPal.



3. **Variety:** Big Data is generated from different types of sources: internal/ external and are of different formats : structured/ unstructured/ semi-structured like records, text, images, video, audio, cookies, JSON etc.

4. **Veracity:** The data is messy in nature: unclear/ inconsistent/ incorrect mostly. It’s a challenge to find the right piece of data which is correct and consistent for analysis/ processing.

Big data Types and Sources

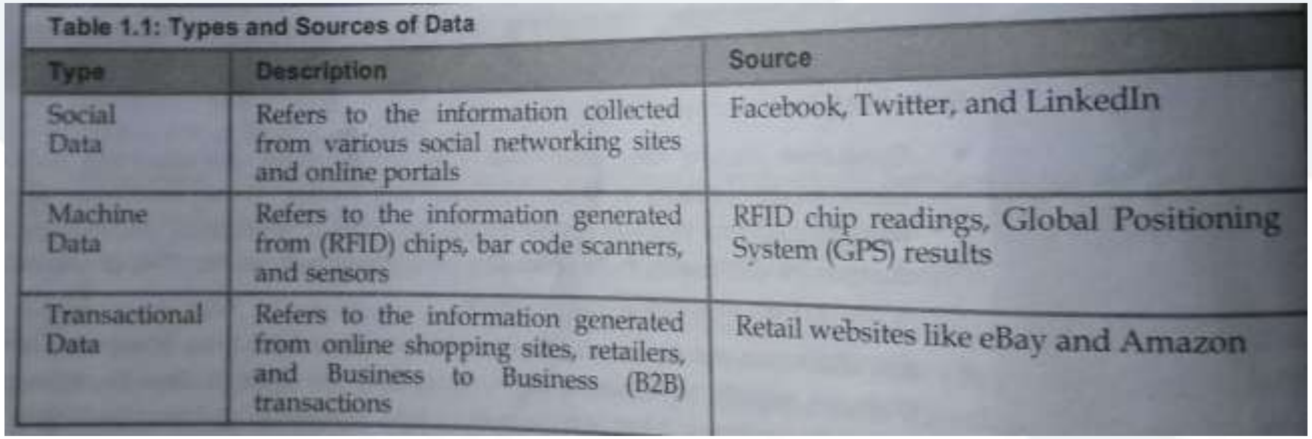


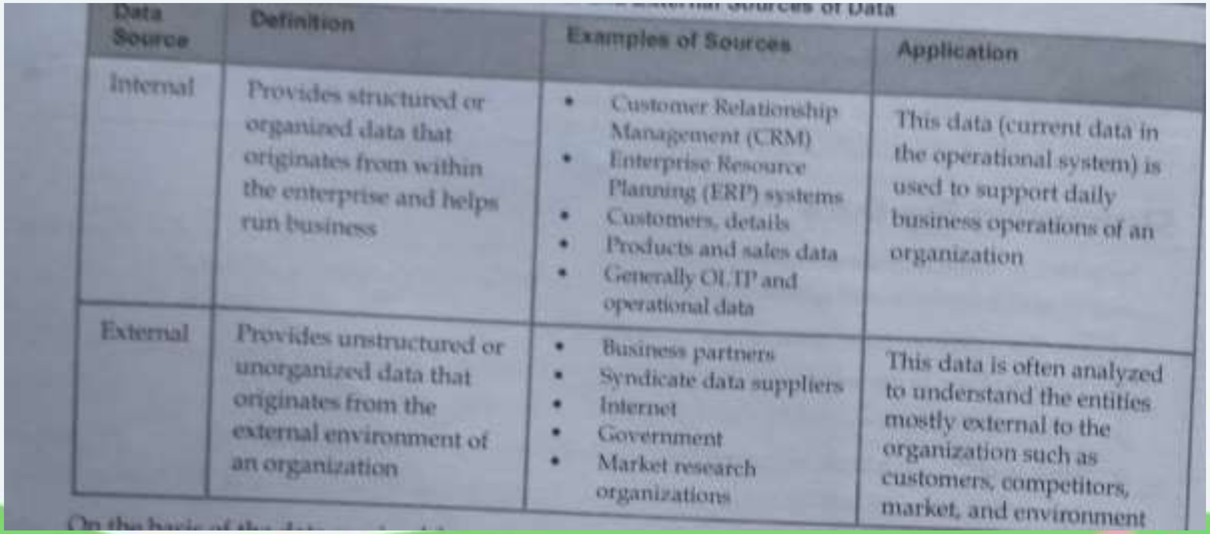
Table 1.1: Types and Sources of Data

Type	Description	Source
Social Data	Refers to the information collected from various social networking sites and online portals	Facebook, Twitter, and LinkedIn
Machine Data	Refers to the information generated from (RFID) chips, bar code scanners, and sensors	RFID chip readings, Global Positioning System (GPS) results
Transactional Data	Refers to the information generated from online shopping sites, retailers, and Business to Business (B2B) transactions	Retail websites like eBay and Amazon

Big data Types [10]

Types of Big Data fall under the broad categories: Internal and External data.

- Internal data are organizational or enterprise data.
- External data are social data.

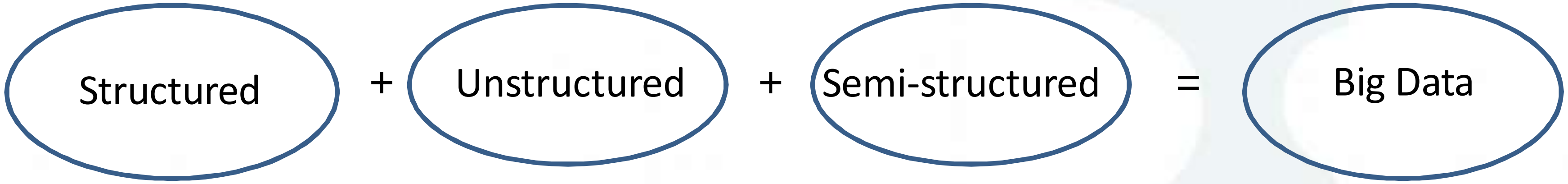


Data Source	Definition	Examples of Sources	Application
Internal	Provides structured or organized data that originates from within the enterprise and helps run business	• Customer Relationship Management (CRM) • Enterprise Resource Planning (ERP) systems • Customers, details • Products and sales data • Generally OLTP and operational data	This data (current data in the operational system) is used to support daily business operations of an organization
External	Provides unstructured or unorganized data that originates from the external environment of an organization	• Business partners • Syndicate data suppliers • Internet • Government • Market research organizations	This data is often analyzed to understand the entities mostly external to the organization such as customers, competitors, market, and environment

Big data Types, Sources and Examples [10]

The Internal and External categories of data also comprises of

i. Structured data ii.Unstructured data iii.Semi-structured data



Structured data-repeated data pattern with pre-defined number of fields. Easy to be queried, sorted and processed. Eg: Relational databases, Files

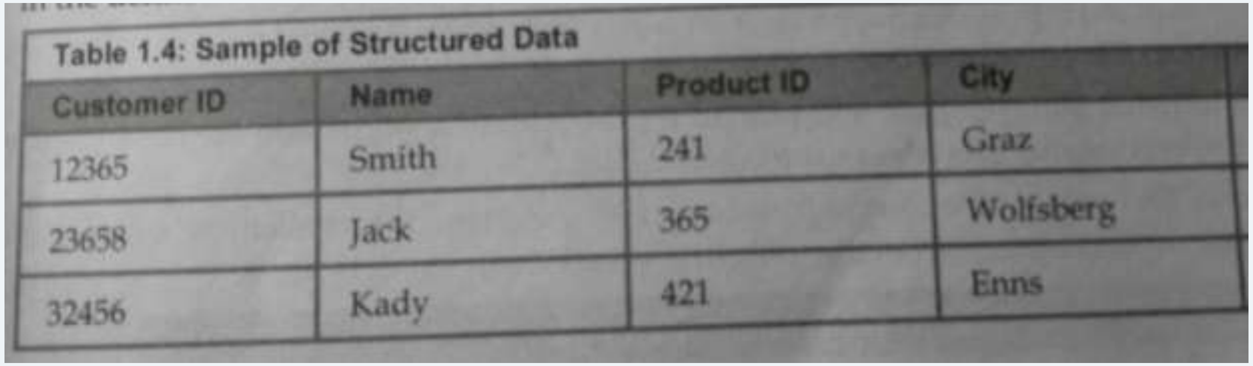


Table 1.4: Sample of Structured Data

Customer ID	Name	Product ID	City
12365	Smith	241	Graz
23658	Jack	365	Wolfsberg
32456	Kady	421	Enns

Structured Data [10]

Semi-Structured Data:

The data does not follow proper structure of data models as in relational databases.

The data is stored inconsistently in rows and columns.

Sources:

- File systems such as Web data in the form of Cookies.
- Data exchange Format such as JavaScript Object Notation (JSON) Data.
- Sensor data from RFID, infrared, wireless technology, GPS

Example:

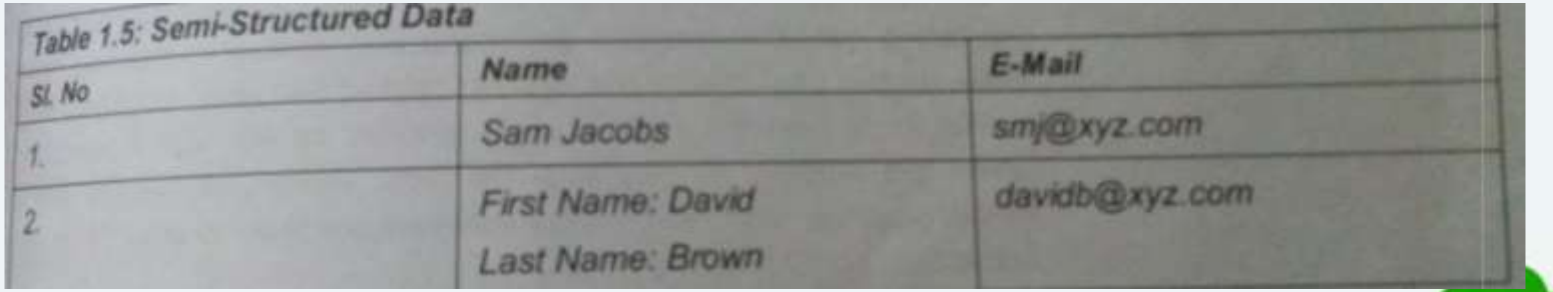


Table 1.5: Semi-Structured Data

Sl. No	Name	E-Mail
1.	Sam Jacobs	smj@xyz.com
2.	First Name: David Last Name: Brown	davidb@xyz.com

Semi-Structured Data [10]

Unstructured Data:

The data can or cannot have repeating patterns. The data consists of :

- metadata: Additional information related to data.
- inconsistent data: data from the files, social media, websites, satellites etc.
- data of different formats: data such as emails, text, audio, video or images.

The data sources include:

- **Text both internal and external to an organization**: Documents, logs, survey results, emails etc.
- **Social media:** Data from social networking platforms, including You Tube, Facebook, Twitter, LinkedIn and Flickr.
- **Mobile data:** Text messages, location information.

The challenges with big data sources include:

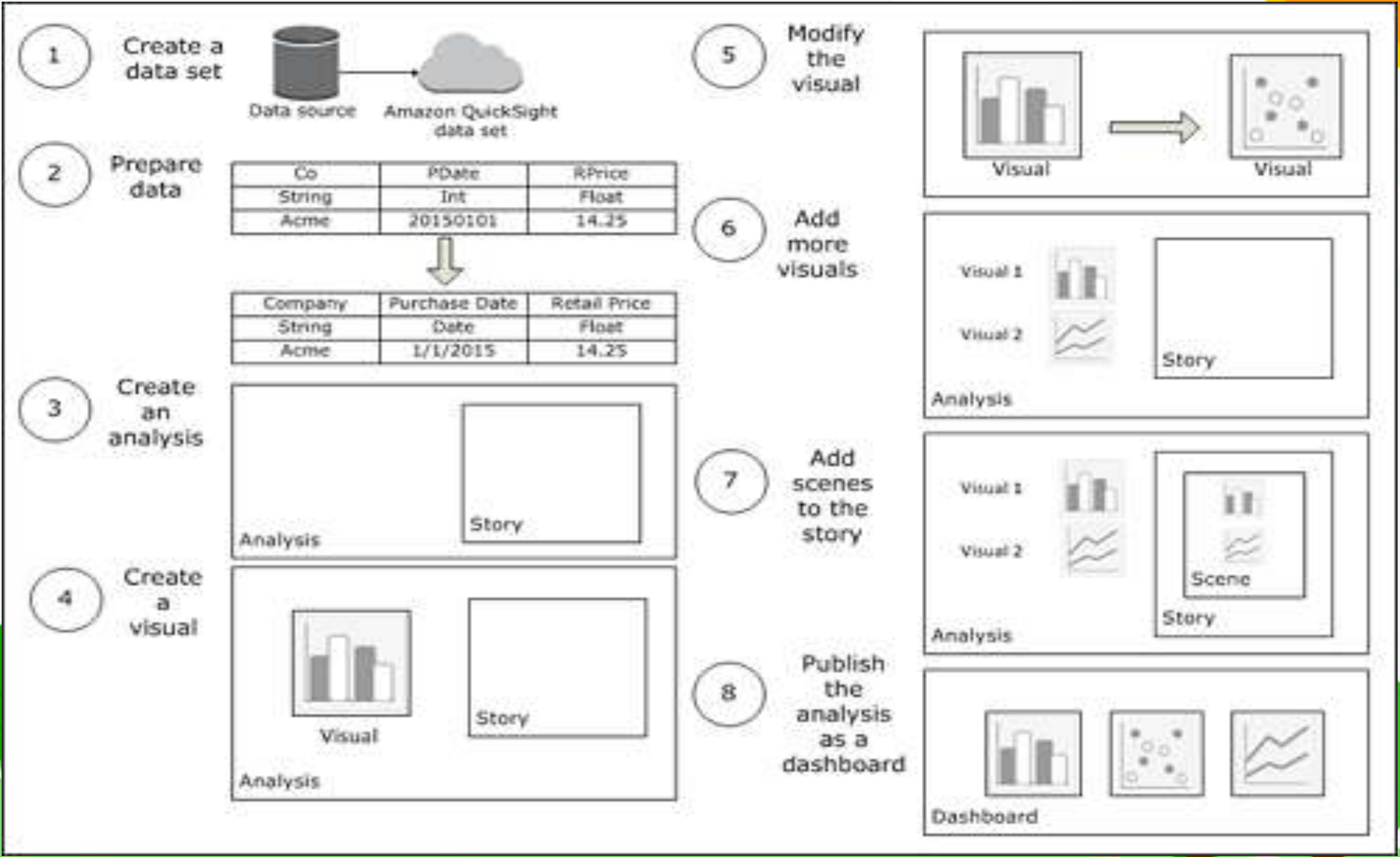
- Identifying the data.
- Sorting, Organizing and storing the data of different formats.
- Relating unstructured data and querying it for logical conclusion

UnStructured Data [17]

ASSIGNMENT

Write in detail the data sources involved and the predictions that can be made by the following prediction systems

- Any Biomedical Analytics System,
- Mobile Advertising,
- Sentimental Analysis,
- Disaster Management,
- Recommendation Engines,
- Smart Cities



Co	PDate	RPrice
String	Int	Float
Acme	20150101	14.25

Company	Purchase Date	Retail Price
String	Date	Float
Acme	1/1/2015	14.25