Assignment 1 DA data analysis

Computer Engineering (Savitribai Phule Pune University)

Assignment 1

1. **Explain Big Data & its Characteristics?**

Data is created constantly, and at an ever-increasing rate. Mobile phones, social media, imaging technologies to determine a medical diagnosis-all these and more create new data, and that must be stored somewhere for some purpose. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time. Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information. These challenges of the data deluge present the opportunity to transform business, government, science, and everyday life.

Three attributes stand out as defining Big Data characteristics:
• Huge volume of data: Rather than thousands or millions of rows, Big Data can be billions of rows and
millions of columns.

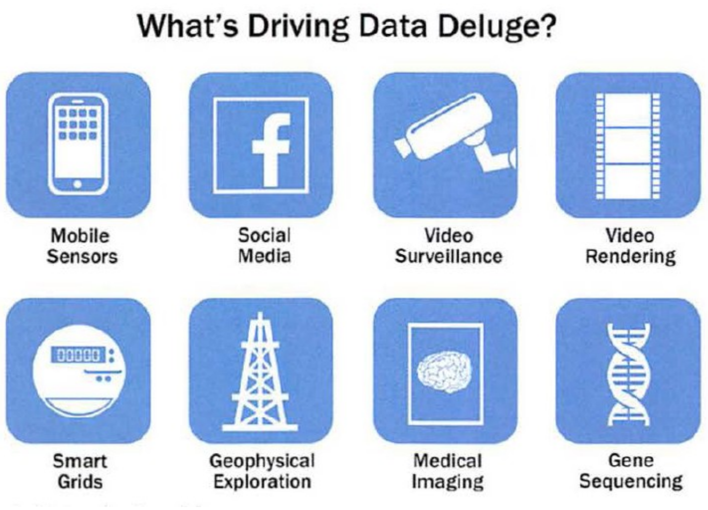
• Complexity of data t ypes and structures: Big Data reflects the variety of new data sources, formats,
and structures, including digital traces being left on the web and other digital repositories for subsequent
analysis.

• Speed of new data creation and growth: Big Data can describe high velocity data, with rapid data
ingestion and near real time analysis.

Although the volume of Big Data tends to attract the most attention, generally the variety and velocity
of the data provide a more apt definition of Big Data. (Big Data is sometimes described as having 3 Vs:
volume, variety, and velocity.) Due to its size or structure, Big Data cannot be efficiently analyzed using only traditional databases or methods.

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.

McKinsey's definition of Big Data implies that organizations will need new data architectures and analytic sandboxes, new tools, new analytical methods, and an integration of multiple skills into the new role of the data scientist



The rate of data creation is accelerating, driven by many of the items in Figure 1
For example, in 2012 Facebook users posted 700 status updates per second worldwide, which can be leveraged to deduce latent interests or political views of users and show relevant ads. For instance, an update in which a woman changes her relationship status from "single" to "engaged" would trigger ads on bridal dresses, wedding planning, or name-changing services

Facebook can also construct social graphs to analyze which users are connected to each other as an interconnected network. In March 2013, Facebook released a new featu re called "Graph Search," enabling users and developers to search social graphs for people with similar interests, hobbies, and shared locations

Data 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. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze

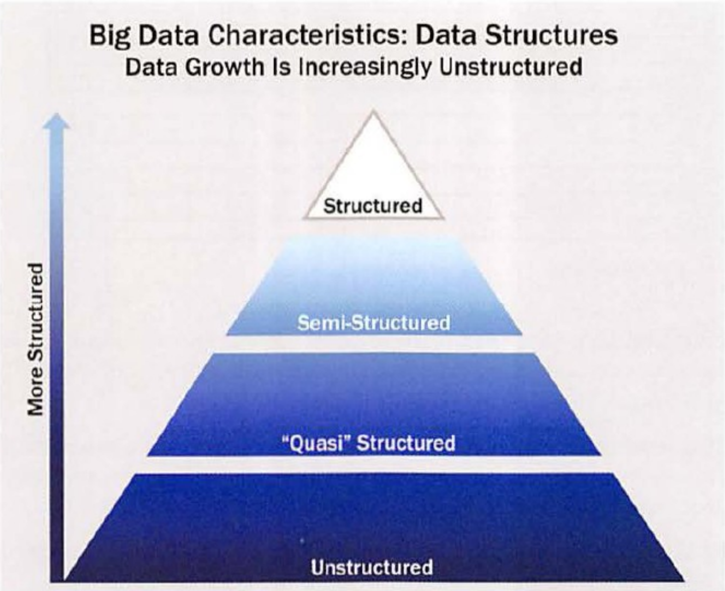

Big Data Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

Figure 1-3 shows four types of data structures, with 80-90% of future data growth coming from nonstructural data types.

Here are examples of how each of the four main types of data structures may look.
o **Structured data**: Data containing a defined data type, format, and structure (that is, transaction data,
online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets)

Semi-structured data: Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML
schema).
o **Quasi-structured data**: Textual data with erratic data formats that can be formatted with effort,
tools, and time (for instance, web clickstream data that may contain inconsistencies in data values
and formats).
o **Unstructured data**: Data that has no inherent structure, which may include text documents, PDFs,
images, and video.

2. Explain BI Versus Data Science

The four business drivers shown in Table 1-2 require a variety of analytical techniques to address them properly. Although much is written generally about analytics, it is important to distinguish between Bland Data Science

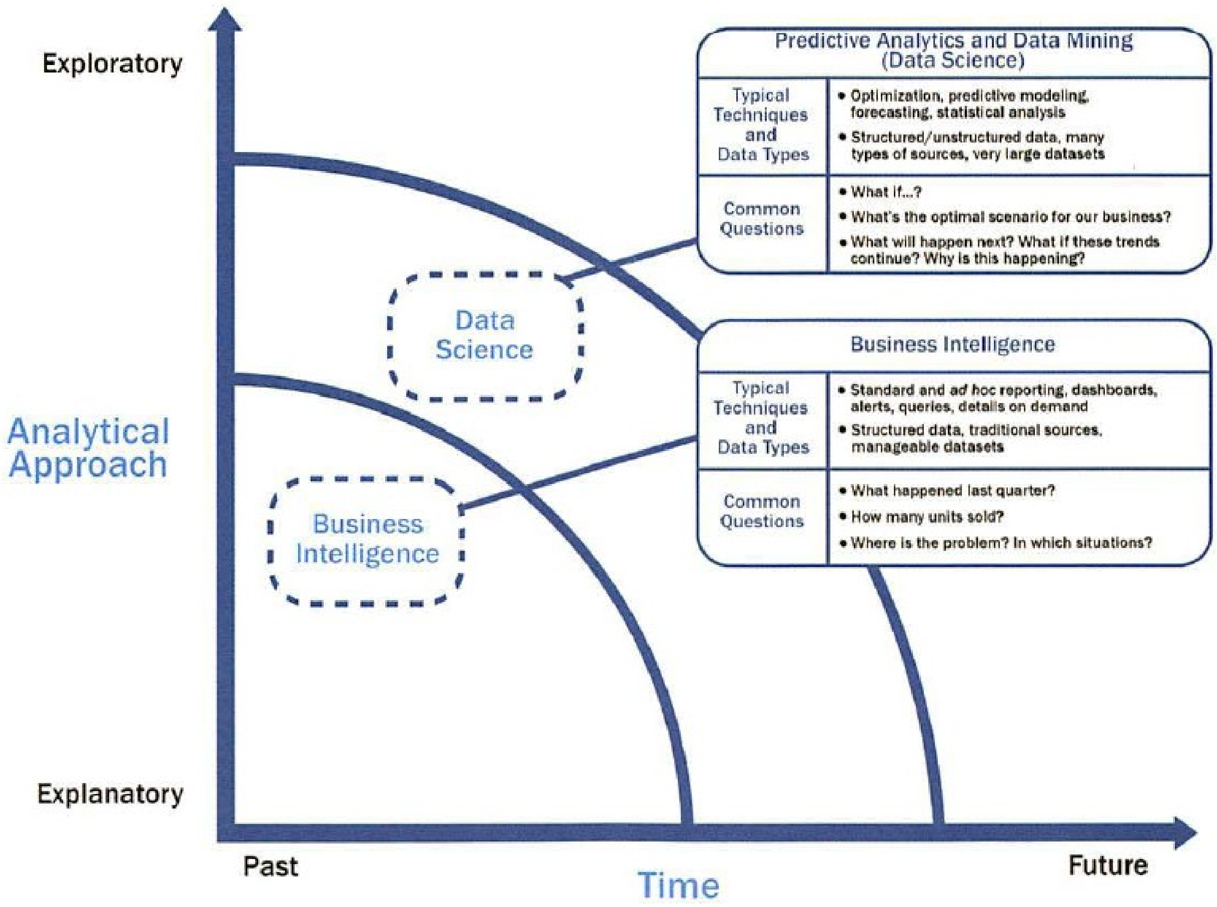
One way to evaluate the type of analysis being performed is to examine the time horizon and the kind of analytical approaches being used. Bl tends to provide reports, dashboards, and queries on business questions for the current period or in the past

Bl systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year.

BI provides hindsight and some insight and generally answers questions related to "when" and "where" events occurred

By comparison, Data Science tends to use disaggregated data in a more forward-looking, exploratory
way, focusing on analyzing the present and enabling informed decisions about the future.

Where BI problems tend to require highly structured data organized in rows and columns for accurate reporting, Data Science projects tend to use many types of data sources, including large or unconventional
datasets. Depending on an organization's goals, it may choose to embark on a 81 project if it is doing reporting, creating dashboards, or performing simple visualizations, or it may choose Data Science projects if it needs to do a more sophisticated analysis with disaggregated or varied datasets.



Predictive Analytics and Data Mining (Data Science)	
Typical Techniques and Data Types	• Optimization, predictive modeling, forecasting, statistical analysis • Structured/unstructured data, many types of sources, very large datasets
Common Questions	• What if...? • What's the optimal scenario for our business? • What will happen next? What if these trends continue? Why is this happening?

Business Intelligence	
Typical Techniques and Data Types	• Standard and *ad hoc* reporting, dashboards, alerts, queries, details on demand • Structured data, traditional sources, manageable datasets
Common Questions	• What happened last quarter? • How many units sold? • Where is the problem? In which situations?

3. Explain Current Analytical Architecture

Data Science projects need workspaces that are purpose-built for experimenting with data, with flexible and agile data architectures. Most organizations still have data warehouses that provide excellent support for traditional reporting and simple data analysis activities but unfortunately have a more difficult time supporting more robust analyses.

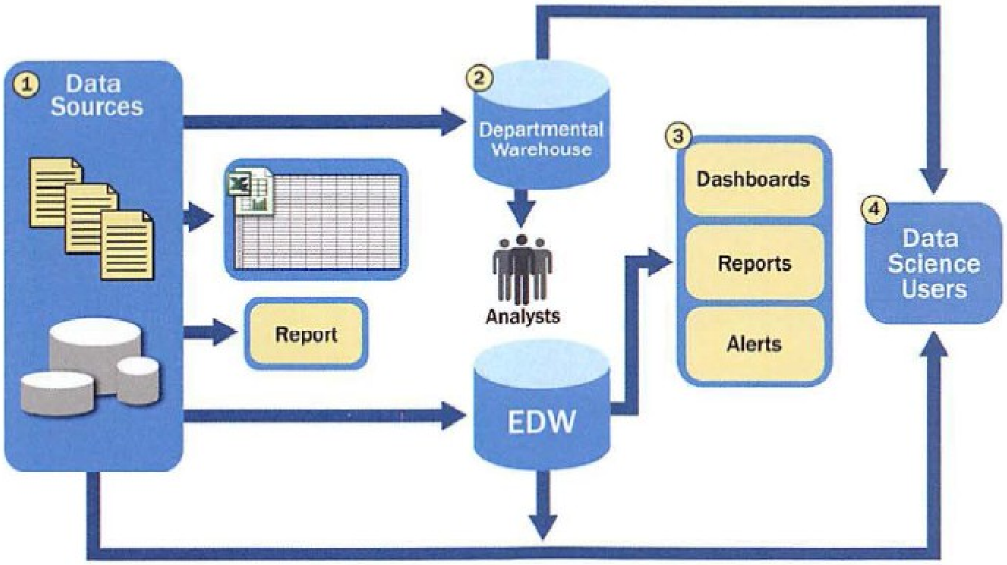


Figure shows a typical data architecture and several of the challenges it presents to data scientists and others trying to do advanced analytics

1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions. Although this kind of centralization enables security, backup, and fai lover of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.

2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts may not have the same constraints for security and structu re as the main EDW and allow users to do some level of more in-depth analysis. However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

3. Once in the data warehouse, data is read by additional applications across the enterprise for Bl and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

4. At the end of this workflow, analysts get data provisioned for their downstream analytics. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools. Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset. Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis-and any insights on the quality of the data or anomalies-rarely are fed back into the main data repository.

Because new data sources slowly accumulate in the EDW due to the rigorous validation and data structuring process, data is slow to move into the EDW, and the data schema is slow to change.

Departmental data warehouses may have been originally designed for a specific purpose and set of business needs, but over time evolved to house more and more data, some of which may be forced into existing schemas to enable Bland the creation of OLAP cubes for analysis and reporting

Although the EDW achieves the objective of reporting and sometimes the creation of dashboards, EDWs generally limit the ability of analysts to iterate on the data in a separate nonproduction environment where they can conduct in-depth analytics or perform analysis on unstructured data.

4. Explain Big Data Ecosystem and a New Approach to Analytics

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging. As this new digital economy continues to
evol ve, the market sees the introduction of data vendors and data cleaners that use crowds ourcing (such as Mechanical Turk and Ga laxyZoo) to test the outcomes of machine learning techniques

As the new ecosystem takes shape, there are four main groups of players within this interconnected
web

Data devices [shown in the (1) section of Figure] and the "Sensornet" gat her data from multiple locations and continuously generate new data about th is data. For each gigabyte of new data created,
an additional petabyte of data is created about that data. [2)

• Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location. This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in ca rs can give drivers real-time updates and offer alternative routes to avoid traffic delays.

• Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of prod ucts purchased together. Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certa in types of retail promotions.

- Data collectors [the blue ovals, identified as (2) within Figure 1-1 1] include sample entities that col lect data from the device and users.

- Data resul ts from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is will ing to pay for premium TV content

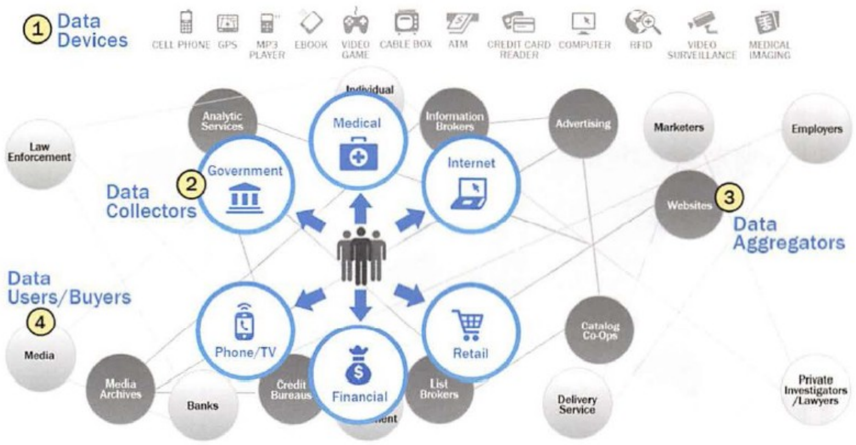
- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips

- Data aggregators (the dark gray ovals in Figure 1-11, marked as (3)) make sense of the data collected from the various entities from the "SensorNet" or the "Internet ofThings." These organizations compile data from the devices and usage patterns collected by government agencies, retail stores, and websites. ln turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

- Data users and buyers are denoted by (4) in Figure 1-11. These groups directly benefit from the data collected and aggregated by others within the data value chain.

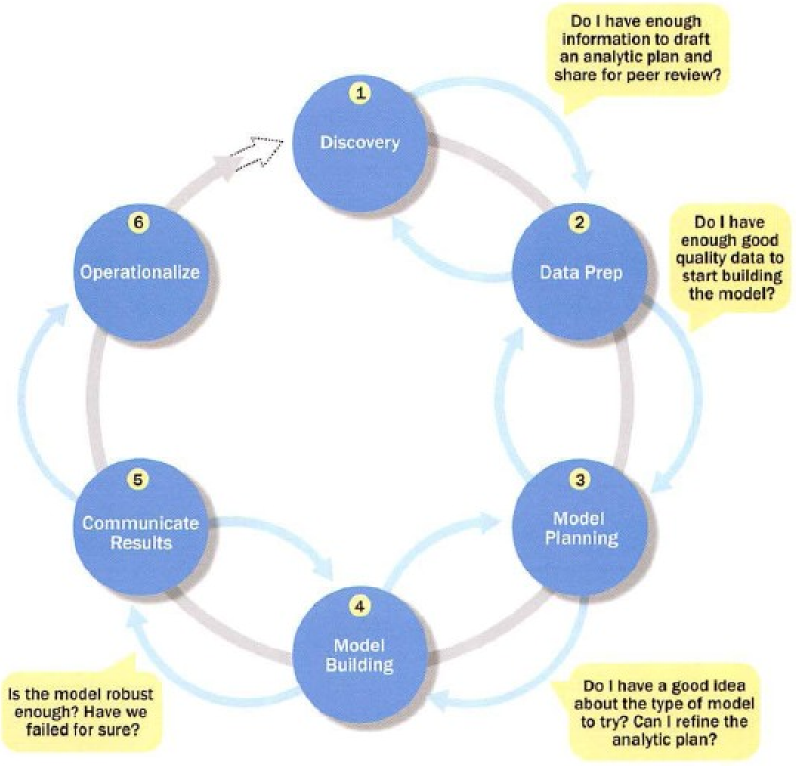
- Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit. To provide input for this analysis, retail banks may purchase data from a data aggregator. This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bil ls on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodeling projects. Obtaining data from these various sources and aggregators will enable a more targeted marketing campaign, which would have been more challenging before Big Data due to the lack of information or high-performing technologies.

- Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analyzing related blogs and online comments. Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects fi rst and how it moves, based on which geographic areas are tweeting about it or discussing it via social media.



5. Explain in detail Data Analytic Life Cycle

The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six phases, and project work can occur in several phases at once. For most phases in the lifecycle, the movement can be either forward or backward


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

Here is a brief overview of the main phases of the Data Analytics Lifecycle:

- Phase 1- Discovery: In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn. The team assesses the resources available to support the project in terms of people, technology, time, and data. Important activities in this phase include framing the business problem as an analytics challenge that can be addressed in subsequent phases and formulating initial hypotheses (IHs) to test and begin learning the data.

- Phase 2- Data preparation: Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project. The team needs to execute ext ract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. The ELT and ETL are sometimes abbreviated as ETLT. Data should be t ransformed in the ETLT process so the team can work with it and analyze it. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data

Phase 3-Model planning: Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

- Phase 4-Model building: In Phase 4, the team develops data sets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase. The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and work flows (for example, fast hardware and parallel processing, if applicable).

- Phase 5-Communicate results: In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

- Phase 6-0perationalize: In Phase 6, the team delivers final reports, briefings, code, and technical documents. In addition, the team may run a pilot project to implement the models in a production environment.