Name: A. Shiva Surya Saran
Roll No.: BE-B 20

Data Analytics

Assignment 1

Unit 1: Introduction and Life Cycle

1. Explain Big data with Data Structures and Big data Repositories.

Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time. In short, such data is so large and complex that none of the traditional data management tools can store it or process it efficiently.

Big Data' could be found in three forms:

- Structured
  Structured is one of the types of big data and by structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc., will be present in an organized manner.

- Unstructured
  Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

- Semi-structured
  Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database) yet contains vital information or tags that segregate individual elements within the data. Thus, we come to the end of types of data.

Big Data Repositories are:

i. **Healthcare**
   Big Data has already started to create a huge difference in the healthcare sector. With the help of predictive analytics, medical professionals and HCPs are now able to provide personalized healthcare services to individual patients. Apart from that, fitness wearables, telemedicine, remote monitoring – all powered by Big Data and AI – are helping change lives for the better.

ii. **Banking**
The banking sector relies on Big Data for fraud detection. Big Data tools can efficiently detect fraudulent acts in real-time such as misuse of credit/debit cards, archival of inspection tracks, faulty alteration in customer stats, etc.

iii. **Manufacturing**
The most significant benefit of Big Data in manufacturing is improving the supply strategies and product quality. It helps create a transparent infrastructure, thereby, predicting uncertainties and in competencies that can affect the business adversely.

iv. **IT**
One of the largest users of Big Data, IT companies around the world are using Big Data to optimize their functioning, enhance employee productivity, and minimize risks in business operations.

v. **Transportation**
Big Data Analytics holds immense value for the transportation industry. In countries across the world, both private and government-run transportation companies use Big Data technologies to optimize route planning, control traffic, manage road congestion, and improve services. Additionally, transportation services even use Big Data to revenue management, drive technological innovation, enhance logistics, and of course, to gain the upper hand in the market.

2. What is the difference between BI and Data Science

FACTOR	**DATA SCIENCE**	**BUSINESS INTELLIGENCE**
Concept	It is a field that uses mathematics, statistics, and various other tools to discover the hidden patterns in the data.	It is basically a set of technologies, applications and processes that are used by the enterprises for business data analysis.
Focus	It focuses on the future.	It focuses the past and present.
Data	It deals with both structured as well as unstructured data.	It mainly deals only with structured data.
Flexibility	Data science is much more flexible as data sources can be added as per requirement.	It is less flexible as in case of business intelligence data sources need to be pre-planned.
Method	It makes the use of scientific method.	It makes the use of analytic method.
Complexity	It has a higher complexity in comparison to business intelligence.	It is much simpler when compared to data science.
Expertise	Its expertise is data scientist.	Its expertise is business user.
Questions	It deals with the questions what will happen and what if.	It deals with the question what happened.

Tools	Its tools are SAS, BigML, MATLAB, Excel etc.	Its tools are InsightSquared Sales Analytics, Klipfolio, ThoughtSpot, Cyfe, TIBCO Spotfire etc.

3. Explain Emerging Big Data Ecosystem and new approach with diagram.

There are four main groups of players

- Data devices
  Games, smartphones, computers, etc.
- Data collectors
  Phone and TV companies, Internet, Gov't, etc.
- Data aggregators – make sense of data
  Websites, credit bureaus, media archives, etc.
- Data users and buyers
  Banks, law enforcement, marketers, employers, etc.

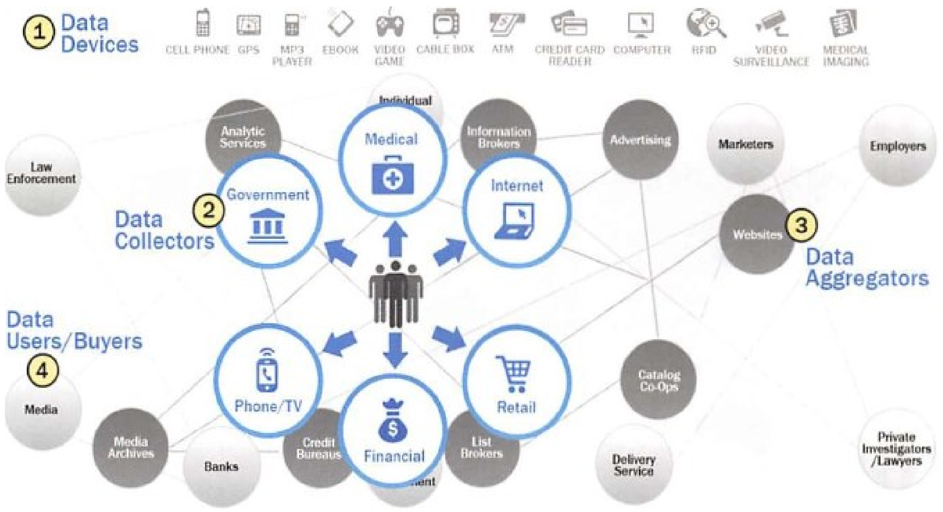


Fig. Emerging Big Data ecosystem

4. Explain Current Analytical Architecture

Analytics architecture refers to the systems, protocols, and technology used to collect, store, and analyse data. It also focuses on multiple layers, starting with data warehouse architecture, which defines how users in an organization can access and interact with data. Storage is a key aspect of creating a reliable analytics process, as it will establish both how your data is organized, who can access it, and how quickly it can be referenced.

Analytics architecture helps you not just store your data but plan the optimal flow for data from capture to analysis. Understanding these steps can give you a better idea of your hardware and logistics needs and clue you in on the best tools to use.

One important use for analytics architecture in your organization is the design and construction of your preferred data storage and access mechanism. Many companies prefer a more structured approach, using traditional data warehouses or data mart models to keep data more organized and easily sorted for access later.

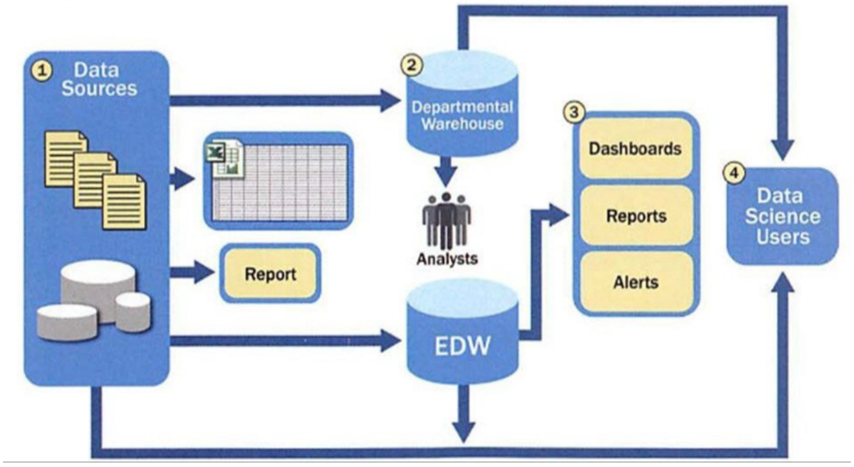


Fig. Typical analytic architecture

For data sources to be loaded into the data warehouse, there is need that the data should be well understood, normalized with the suitable data type definitions and in structured format.

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 structure as the main EDW and allow users to do some level of more in-depth analysis.

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

At the end of this workflow, analysts get data provisioned for their downstream analytics. 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.

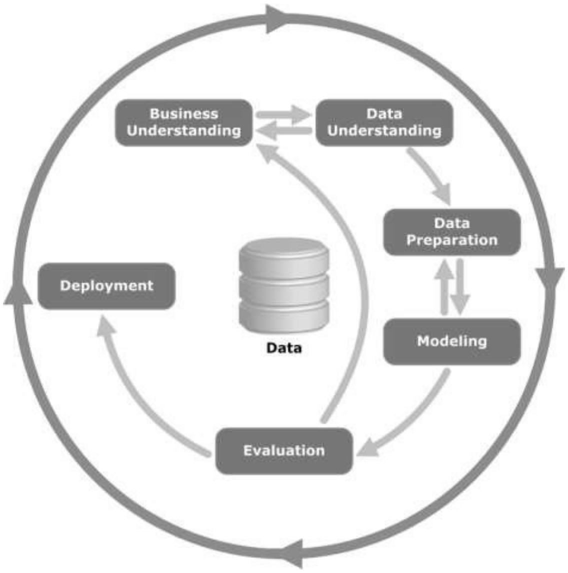
Because new data sources slowly accumulate in the EDW due to the rigorous validation and data structuring process, data is slow to move into the EDW, and the data schema is slow to change.

- High-value data is hard to reach and leverage, and predictive analytics and data mining activities are last in line for data.
- Data moves in batches from EDW to local analytical tools. This workflow means that data scientists are limited to performing in-memory analytics (such as with R, SAS, SPSS, or Excel), which will restrict the size of the datasets they can use.
- Data Science projects will remain isolated and ad hoc, rather than centrally managed.

The current Data Warehousing solutions continue offering reporting and BI services to support management and mission-critical operations.

5. Explain Data Analytic Life Cycle with diagram

Data science projects differ from most traditional Business Intelligence projects and many data analysis projects in that data science projects are more exploratory in nature. The Data Analytics Lifecycle is designed specifically for Big Data problems and data science projects. The lifecycle has six phases.



- **Business Understanding** − This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition. A preliminary plan is designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

- **Data Understanding** − The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation** − The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modelling tools.
- **Modelling** − In this phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, it is often required to step back to the data preparation phase.
- **Evaluation** − At this stage in the project, you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to evaluate the model thoroughly and review the steps executed to construct the model, to be certain it properly achieves the business objectives.

  A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
- **Deployment** − Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process.