Chapter 1:Introduction to Big Data

Om Prakash Mahato

Assistant Director

Nepal Telecommunications Authority

What is Big Data ?

1.1 Big Data Overview

- Big data is a large and complex collection of data sets which may not be correlated and which is impossible to process using traditional data processing applications.
- - Big Data implies the method of extracting value from data, where the size of data is unpredictable.
- - Big Data comes into existence because it is really a smart idea to use simple algorithm in large data to achieve higher efficiency and accuracy in analysis.
- - The world has very huge amount of data but they are not properly analyzed into the required information. To generate appropriate information from the large size of data we have, Big data acts as an important tool in today's era.
- - The amount of data sets arises from various sources such as web, e-commerce, social network, bank transactions, sensor technology, mobile computing and so on.

Characteristics of Big Data

1**. Volume:**
- Volume indicates the quantity of generated and stored data.
- The size of data determines the potential value and insight.
- It also determines whether it is considered to be big data or not.
- The volume of data in the world is increasing exponentially .

2. Variety:
- Variety indicates the different types and nature of data.
- All the data present in big data analysis may not be of same type.
- Even a single application may be generating variety of data.
- This increases complexity in big data analysis and knowledge extraction.
- The data may be web data, relational data, XML, structured data, streaming data, graph data and so on.
- For efficient extraction of information or patterns, all these variety of data must be linked together and analyzed together.

3. Velocity:
- Velocity indicates the speed at which data is generated and processed to meet the demands.
- The data obtained is of dynamic nature, so they must be analyzed very fast to provide efficient and effective knowledge.

Challenges in Big Data

1.Big data consists of **huge amount of data sets**. The main challenge evolve in identifying the appropriate data from such mass of data and determining how to make best use of the relevant data.

2. Even though the data and analysis method are determined, there is struggle in finding the appropriate and **skilled manpower** capable of working with both new technology and data analysis for relevant business insight.

3. The variety of data types and formats may generate hindrance in the data analysis as it is very difficult to connect variety of data points for a single insight. **Data integration** is the important aspect of effective big data analysis, but it is also one of the major challenge that prevails.

4. The technology landscape in the data world is **evolving very fast**. So, efficient handling of the technology along with **adaptation to cope with technology** is must for big data analysis.

5. The organizational structure for big data project management should be apart from other project management task because this field is very much different and needs a **strong and motivational project management team**.

6. During the big data analysis, the data analysts do not get full benefits from the data they have due to the **security concerns about data protection**.

7. **The technology infrastructure** necessary to work with big data is **very expensive**.

1.2 Background of Data Analytics

- ***Big Data Analytics***
- - Big Data analytics is the process of analyzing the huge volume of variety of data so as to discover the hidden patterns and other essential information that can be used for decision making process.

- The traditional data analytics tools can not be used for big data due to its unstructured data format.

- The most commonly used technologies in big data analytics are NoSQL databases, Hadoop and Map Reduce.etc.

- It focuses on efficient use of a simple model applied to large volume of data.

Big Data Analyst

- Data analyst is the person who is responsible to collect the relevant data, analyze it using big data analytics tools and use the outcome for gaining competitive advantages for the organizational decision making.
- - The various skills required by the data analyst are as follows:
  1. Analytical skill
  2. Communication skill
  3. Critical thinking
  4. Attention to detail
  5. Mathematical skill

Objective of big data analytics

- Big Analytics supporting the following objectives for working with Big Data Analytics:

1. Avoid sampling / aggregation;

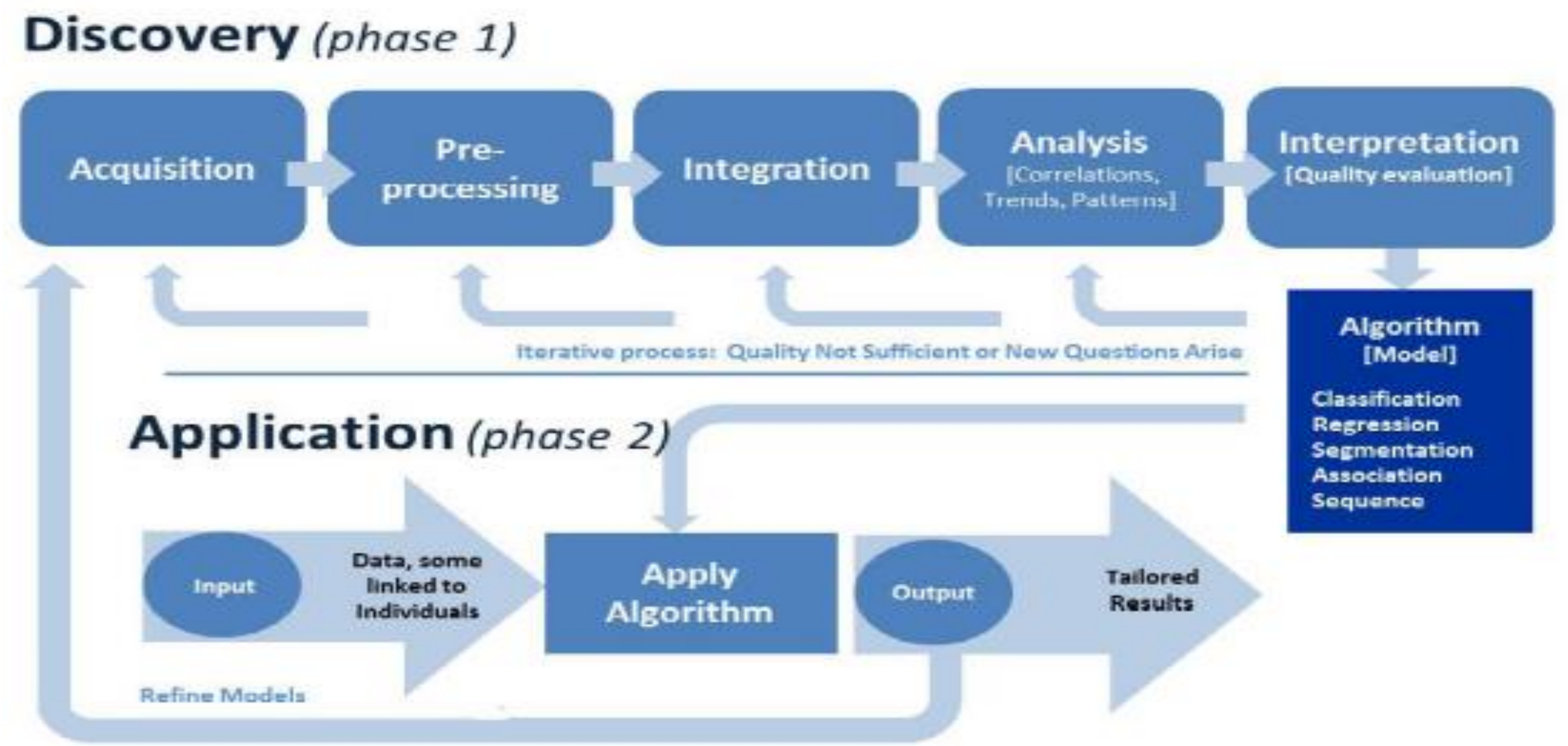
2. Reduce data movement and replication;

Data redundancy is the key issue for data analysis in big data environments. Three main reasons for data redundancy are: (1) addition of nodes, (2) expansion of datasets, and (3) data replication

3. Bring the analytics as close as possible to the data.

4. Optimize computation speed.

Process of Data Analytics



Process of Data Analytics

Phase 1 - Discovery Phase

- Discovery phase is the phase in data analytics process in which knowledge are gathered from the available data by analyzing the data to discover the hidden pattern.
- This phase consists of following processes:

a) Acquisition
b) Pre-processing
c) Integration
d) Analysis
e) Interpretation
f) Algorithmic model

- Acquisition is the process in which the data necessary for the concerned analytical problems are collected or gathered. The data can be gathered through internal or external sources.
- Pre-processing is the process that converts the collected data into the standard format that can be easily used in further processes.

Phase-1 Conti..

- Integration is the process in which the relevant data are integrated from different sources and the redundant data are eliminated.
- - Analysis is the process in which the relationships among the data items are searched so as to yield some useful information. This process is used to discover hidden patterns from the data.
- - Interpretation is the process in which the results of the analysis are examined and the quality of the results are determined. It provides the information whether or not the analysis results can be used for decision making.
- - After all these steps, if the result is trustworthy, then an algorithmic model is generated that incorporates all the knowledge discovered in analysis phase, which can be used further with new set of data to check for the outcome.

Phase 2 - Application

- Application phase is the phase in which the algorithmic model generated as a result of data analytics is used in the real domain.

- With the help of the model, the organization acts based on the outcome of the model.

- The algorithm is applied to some sort of input from the organization so as to gain the outcome based on the discovered knowledge or pattern. This helps in proper formulation of strategy by an organization.

1.3 Role of Distributed System in Big Data

Distributed System

- - Distributed system is the system that is composed of autonomous computers, connected through a network and middleware for coordinating and sharing the resources, in such a way that the whole system looks as a single integrating computer facility for the user.

Use of Distributed System to solve Big Data problems

- - Big data consists of massive amount of data which can not be stored in a single computer or node.

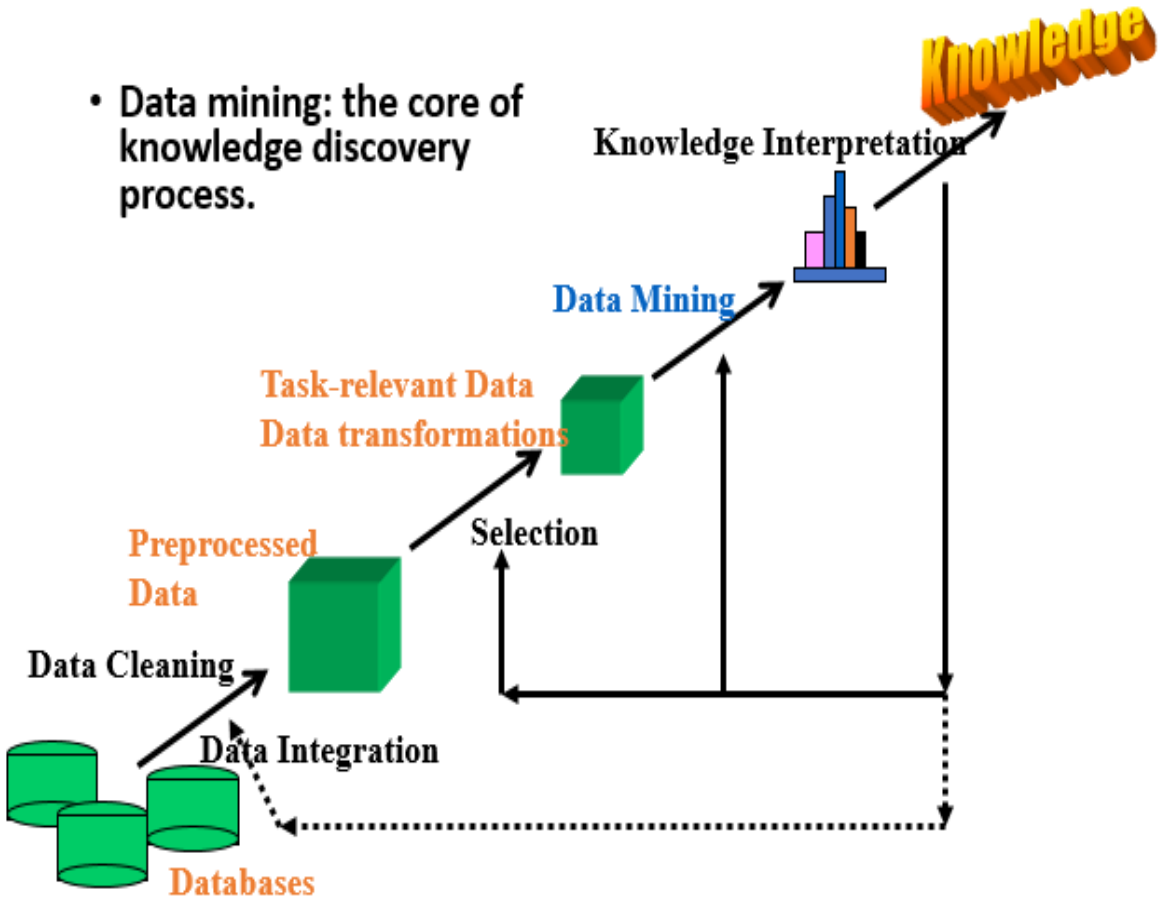
- So, there is necessity for big data to be distributed across multiple nodes.

- Distributed system helps to solve big data problems without the requirement of a single resource capable to handle it. This makes the big data analytics cost efficient and performance improvement.

1.4 Role of Data Scientist

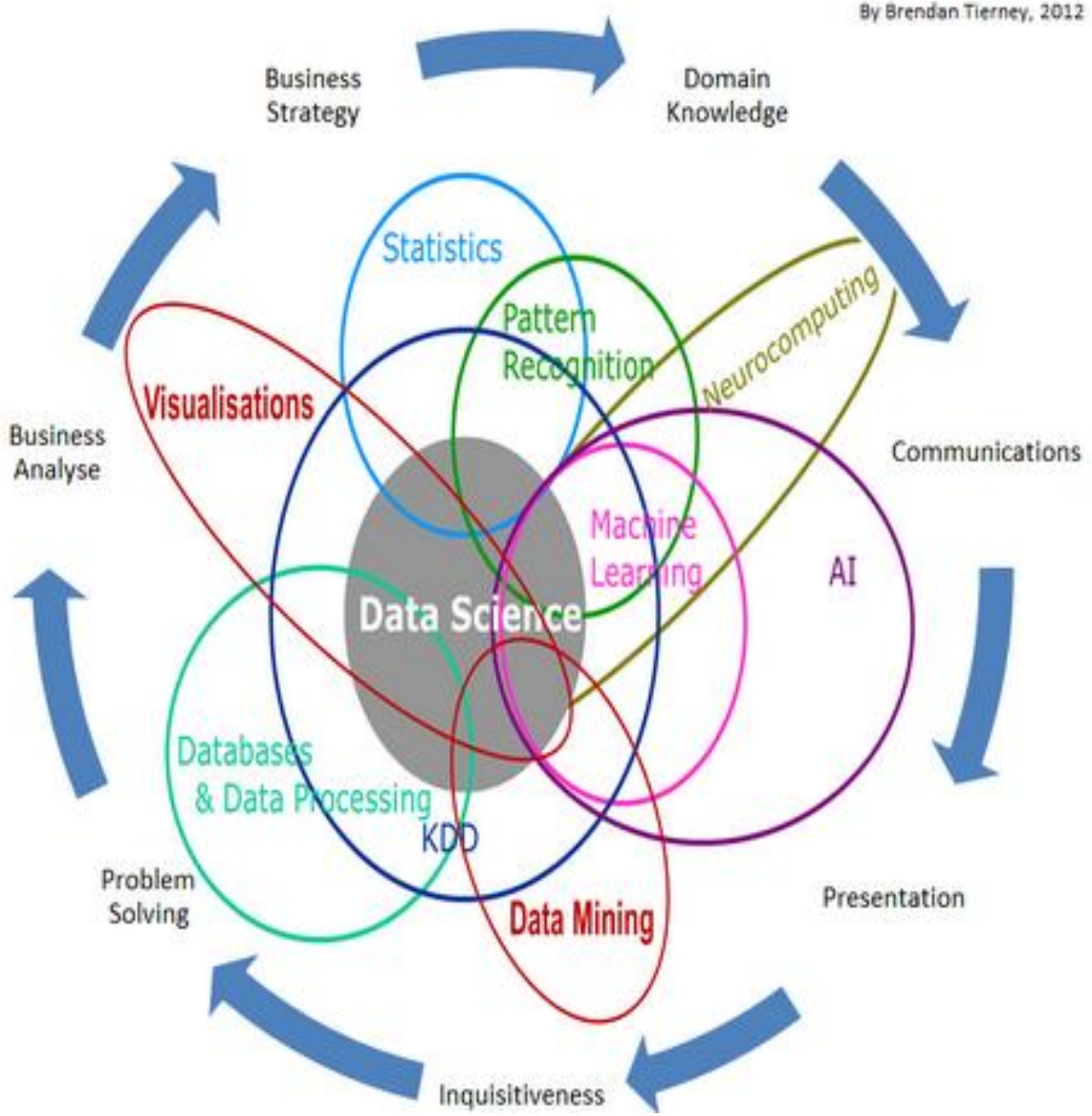
Knowledge Discovery Process

- Data mining: the core of knowledge discovery process.



Data Science Is Multidisciplinary

By Brendan Tierney, 2012



1.4 Role of Data Scientist Conti..

Data Scientist

Data science is the field that deals with scientific methods, processes and systems so as to extract knowledge from the data.

- It requires techniques from different fields like mathematics, statistics, computer science and so on.

- Data scientist is the person who apply powerful tools and advanced statistical modeling techniques to make discoveries about business processes and problems through the curious problem solving mechanism applied on the available data sets.

- Data scientist has the ability to handle the raw data using the latest technologies, can perform the analysis and can present the acquired knowledge in understandable manner.

Roles of Data Scientist

1. Use the big data technologies to make new discoveries from the data.
2. Develop structure from the large volume of unstructured data and generate possibility for analysis.
3. Identify the data sets from multiple sources, integrate them and eliminate the unnecessary data present.
4. Communicate the results and suggest implications for new business direction.
5. Display information in clear and understandable way.

Skills Needed for Data Scientist

1.Knowledge of basic tools like statistical programming languages (R and Python) and database.
2. Basic Statistics
3. Machine Learning
4. Calculus and Linear Algebra
5. Data munging skills (Conversion of data to necessary format)
6. Data Visualization and Communication Skills
7. Software Engineering
8. Problem Solving
9. Business Skills

Data Scientist vs Data Analyst

- Data scientist is responsible to formulate the questions for business progress and proceed in solving them. Data analysts is responsible to pursue a solution for the provided questions.
- Data scientist is expected to convert the data and analysis into a business scenario. Data analyst is expect just to analyze the data to find patterns.

1.5 Current Trend in Big Data Analytics

1. Big data analytics services are available on the cloud. This makes possibility for faster and easier analysis and processing.

2. Hadoop is being general purpose and enterprise wide solutions for big data.

3. Emergence of big data lakes. It removes the necessity for designing the data set before entering the data. It allows all the data to be collected in the repository or data lake.

4. The value of data is being more. All the business organizations are enrolled to get valuable insights from their data for their progress and growth.

5. Use of NoSQL database helps to store the unstructured data that is the most prevalent data in the world today.

6. Use of in memory analytics allows speeding of the analytic processes.

7. Variety is becoming the single driving factor for big data investments.

8. The rise of metadata catalogs helps the users to discover and understand relevant data which are worth analyzing.