CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

APEX INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Introduction to Data Science (21CST-292)

Faculty: Dr. Jitender Kaushal (E14621)

Associate Professor

Lecture -

DISCOVER . LEARN . EMPOWER

Big Data Analytics Life Cycle

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

COURSE OUTCOMES

On completion of this course, the students shall be able to:-

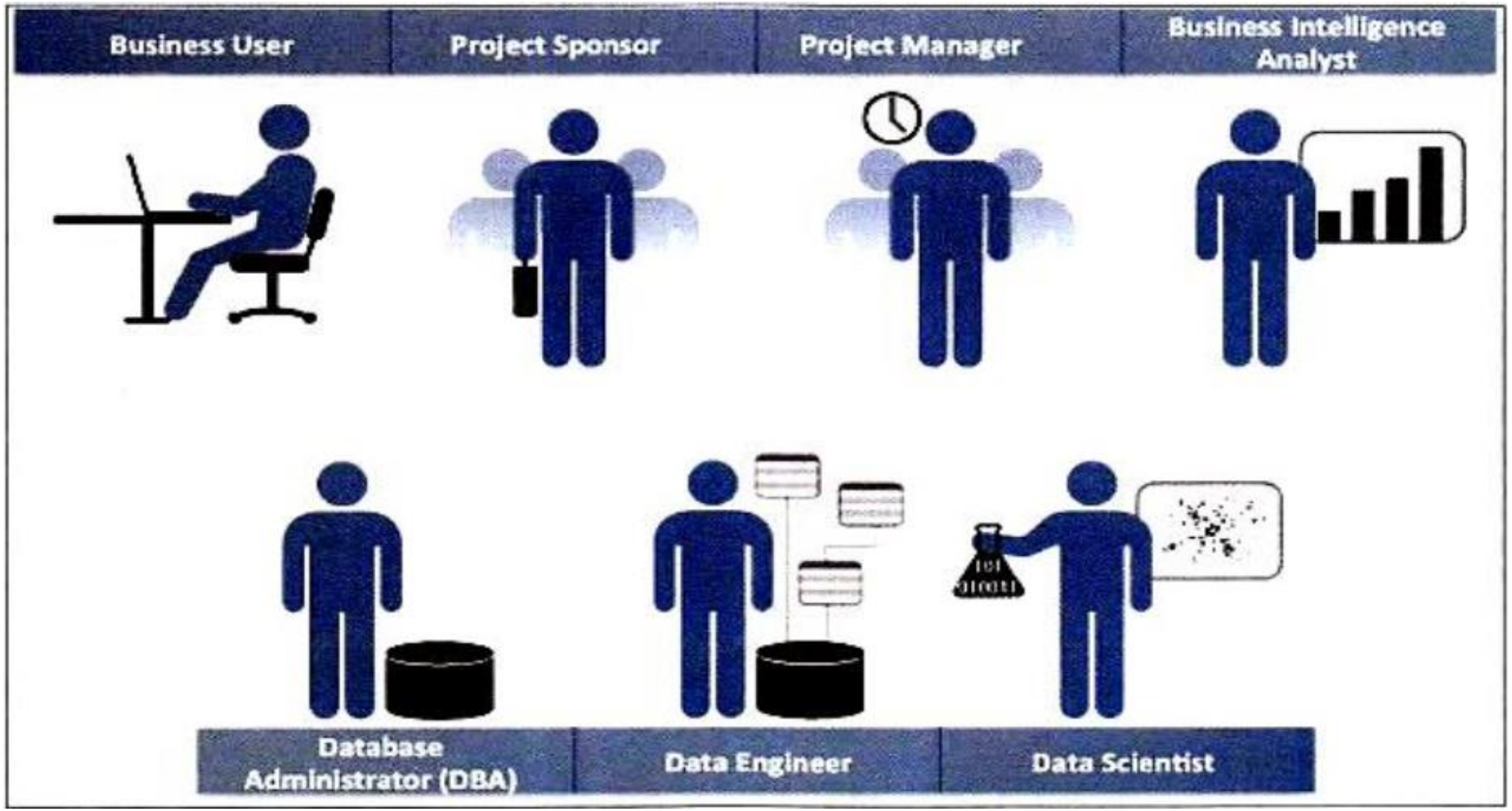
CO3	To learn and understand Data Life Cycle, Data Preparation.

Data Analytics Lifecycle

- Big Data analysis **differs** from traditional data analysis primarily **due to the volume, velocity, value, veracity and variety** of characteristics of the data being processed.
- To address the distinct requirements for analyzing Big Data, **a step-by-step methodology** is needed to organize the activities and tasks involved with **acquiring, processing, analyzing and repurposing data**.

Data Analytics Lifecycle (cont..)

From a Big Data **adoption and planning** perspective, it is important that in **addition to the lifecycle**, consideration be **made for issues** of **training, education, tooling, and staffing** of a data analytics team.



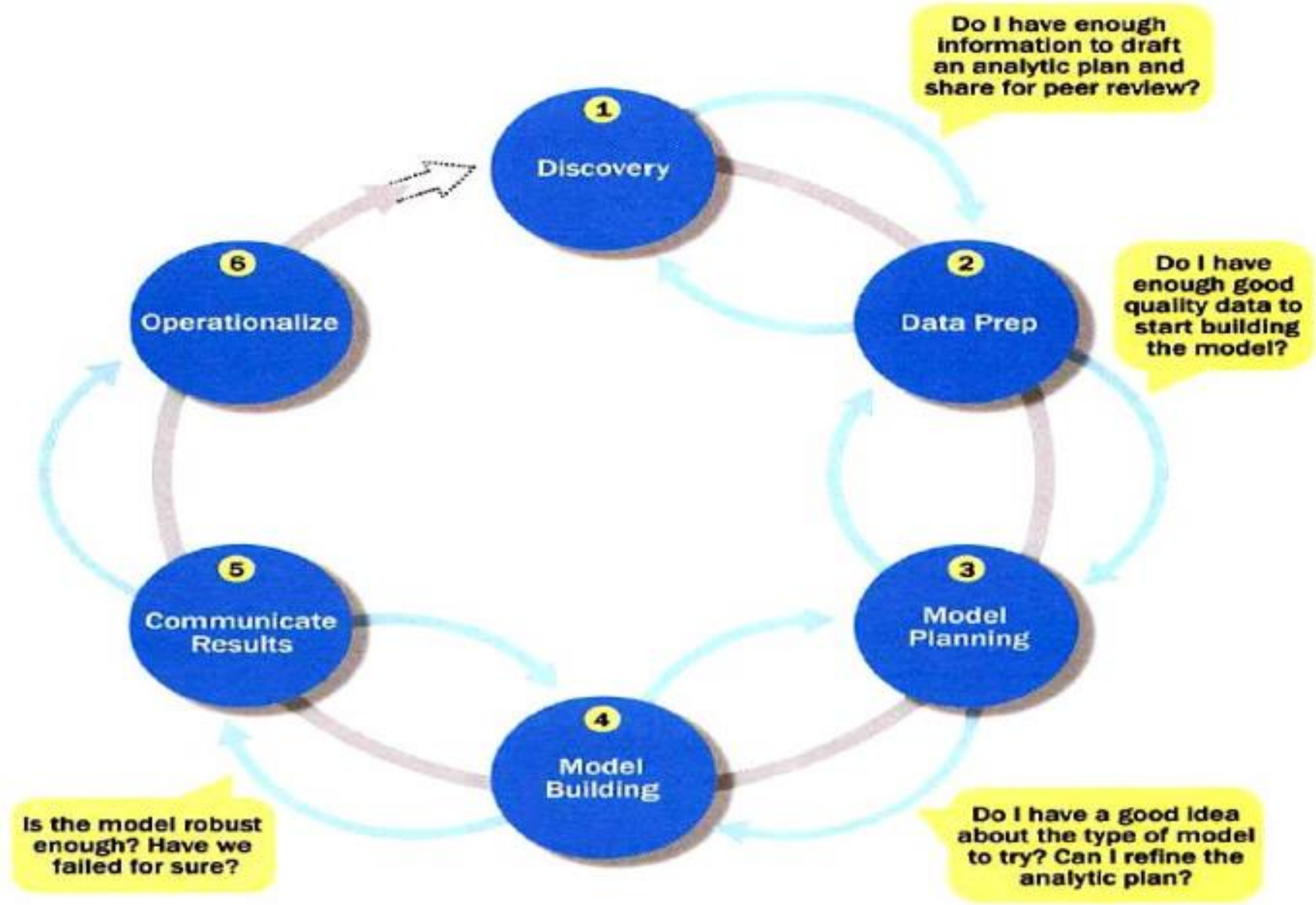
Key Roles for a Successful Analytics Project

- **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

Data Analytics Lifecycle (cont..)

- The data analytic lifecycle is **designed for Big Data problems and data science projects**
- The cycle is **iterative to represent a real project**
- Work can **return to earlier phases** as new information is uncovered

Data Analytics Lifecycle-Abstract view



Phase 1: Discovery

In this phase,

- The data science team must **learn and investigate the problem**,
- Develop **context and understanding**, and
- Learn about the **data sources needed** and **available** for the project.
- In addition, the **team formulates initial hypotheses** that can later be tested with data.

Phase 1: Discovery (cont.)

- The team should perform **five main** activities during this step of the discovery phase:
- **Identify data sources:** Make a list of data sources the team may need to test the initial hypotheses outlined in this phase.
- *Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.*
- **Capture aggregate data sources:** This is for previewing the data and providing high-level understanding.
- *It enables the team to gain a quick overview of the data and perform further exploration on specific areas.*
- **Review the raw data:** Begin understanding the interdependencies among the data attributes.
- *Become familiar with the content of the data, its quality, and its limitations.*

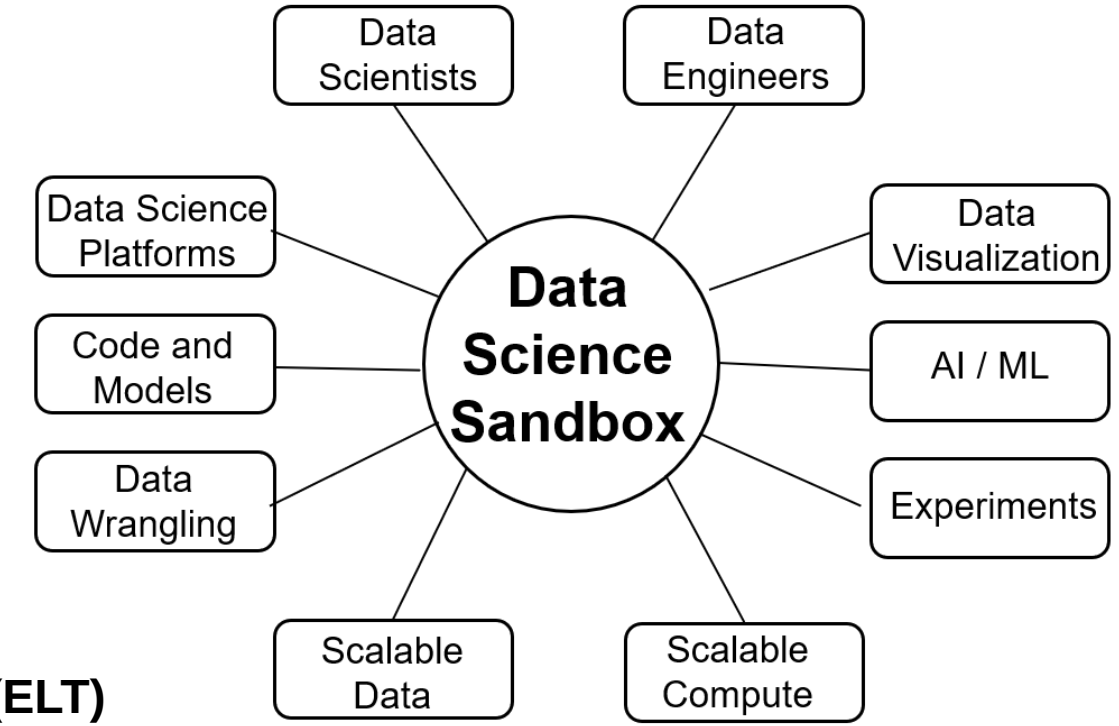
Phase 1: Discovery (cont.)

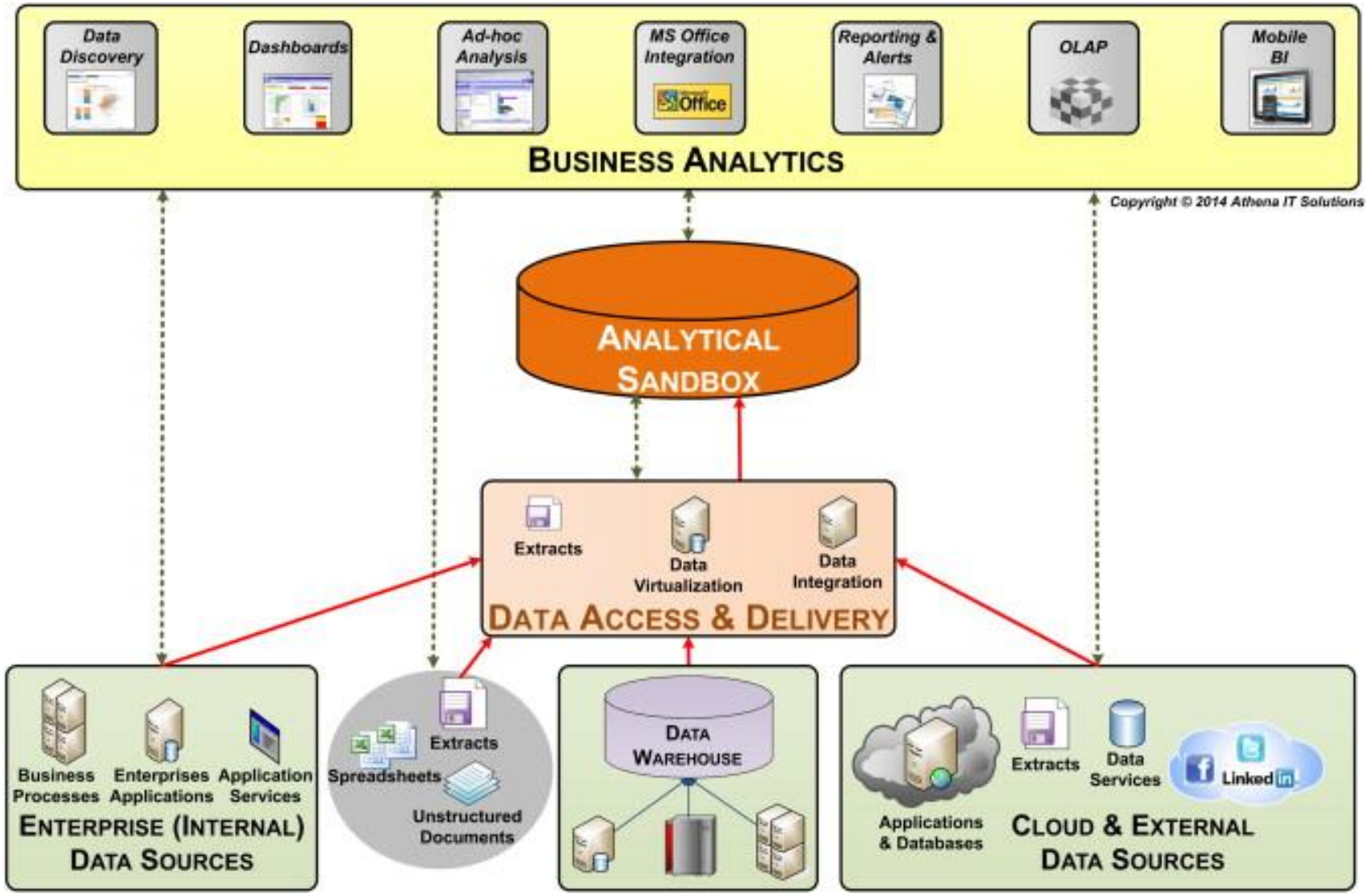
- **Evaluate the data structures and tools needed:** The data type and structure dictate which tools the team can use to analyze the data.
- **Scope the sort of data infrastructure needed for this type of problem:** In addition to the tools needed, the **data influences the kind of infrastructure that's required**, such as **disk storage** and **network capacity**.
- Unlike many traditional stage-gate processes, in which the team can advance only when specific criteria are met, the **Data Analytics Lifecycle is intended to accommodate more ambiguity**.
- For each phase of the process, it is **recommended to pass certain checkpoints** as a way of gauging **whether the team is ready to move to the next phase** of the Data Analytics Lifecycle.

Phase 2: Data preparation

This phase includes

- Steps to explore, Preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic **sandbox (workspace)**, in which the team can work with data and perform analytics for the duration of the project.
- The team needs to execute **Extract, Load, and Transform (ELT) or extract, transform and load (ETL)** to get data into the sandbox.
- In ETL, users perform processes to **extract data from a datastore, perform data transformations, and load the data back into the datastore.**
- The ELT and ETL are sometimes abbreviated as ETLT. Data should be transformed in the ETLT process so the team can work with it and analyze it.





- Data sandboxes help teams make more informed decisions by giving them access to valuable insights in large datasets.
- A data sandbox is a place where you can test and experiment with data.
- The **IBM Netezza 1000** is an example of a data sandbox platform which is a stand-alone analytic data mart.
- An example of a logical partition in an enterprise data warehouse, which also serves as a data sandbox platform, is the IBM Smart Analytics System.

Phase 2: Data preparation

Rules for Analytics Sandbox

- When developing the analytic sandbox, **collect all kinds of data** there, as team members need access to high volumes and varieties of data for a Big Data analytics project.
- This can include everything from **summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs,** depending on the kind of analysis, the team plans to undertake.
- A good rule is **to plan for the sandbox to be at least 5–10 times the size of the original datasets**, partly because copies of the data may be created that serve as specific tables or data stores for specific kinds of analysis in the project.

Phase 2: Data preparation

Performing ETLT

- As part of the ETLT step, **it is advisable to make an inventory of the data and compare the data currently available with datasets** the team needs.
- Performing this sort of gap analysis provides a framework for understanding which datasets the team can take advantage of today and where the team needs to initiate projects for data collection or access to new datasets currently unavailable.
- A component of this subphase involves extracting data from the available sources and determining data connections for raw data, **online transaction processing (OLTP) databases, online analytical processing (OLAP) cubes**, or other data feeds.
- **Data conditioning** refers to the process of **cleaning data, normalizing datasets**, and **performing transformations** on the data.

Common Tools for the Data Preparation Phase

Several tools are commonly used for this phase:

Hadoop can perform massively parallel ingest and custom analysis for web traffic analysis, GPS location analytics, and combining of massive unstructured data feeds from multiple sources.

Alpine Miner provides a graphical user interface (GUI) for creating analytic workflows, including data manipulations and a series of analytic events such as staged data-mining techniques (for example, first select the top 100 customers, and then run descriptive statistics and clustering).

OpenRefine (formerly called Google Refine) is a free, open source, powerful tool for working with messy data. A GUI-based tool for performing data transformations, and it's one of the most robust free tools currently available.

Similar to OpenRefine, **Data Wrangler** is an interactive tool for **data cleaning and transformation**. Wrangler was developed at Stanford University and can be used to perform many transformations on a given dataset.

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.
- During this phase that the team refers to the **hypotheses** developed in Phase 1, when they first became acquainted with the data and understanding the business problems or domain area.

Common Tools for the Model Planning Phase

Here are several of the more common ones:

- **R** has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code. In addition, it has the ability to interface with databases via an ODBC connection and execute statistical tests.
- **SQL Analysis** services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- **SAS/ ACCESS** provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, and OLE DB. SAS itself is generally used on file extracts, but with SAS/ ACCESS, users can connect to relational databases (such as Oracle or Teradata).

Phase 4: Model Building

- In this phase the data science team needs **to develop data sets for training, testing, and production** purposes. These data sets enable the **data scientist to develop the analytical model and train it** ("training data"), while holding aside some of the data ("holdout data" or "test data") for testing the model.
- the team develops datasets 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 **sufficient for running the models**, or if it will **need a more robust environment** for executing models and workflows (for example, fast hardware and parallel processing, if applicable).
- **Free or Open Source tools:** Rand PL/R, Octave, WEKA, Python
- **Commercial Tools:** Matlab, STATISTICA.

Phase 5: Communicate Results

- In **Phase 5,** After executing the model, the team needs to compare the outcomes of the modeling to the criteria established for success and failure.

- The team considers how best to articulate the findings and outcomes to the various team members and stakeholders, taking into account warning, assumptions, and any limitations of the results.

- The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize

- In the final **phase 6, Operationalize)**, the team communicates the benefits of the project 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 of users.
- This approach enables the team to learn about the performance and related constraints of the model in a production environment on a small scale and make adjustments before a full deployment.
- 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.

Common Tools for the Model Building Phase

Free or Open Source tools:

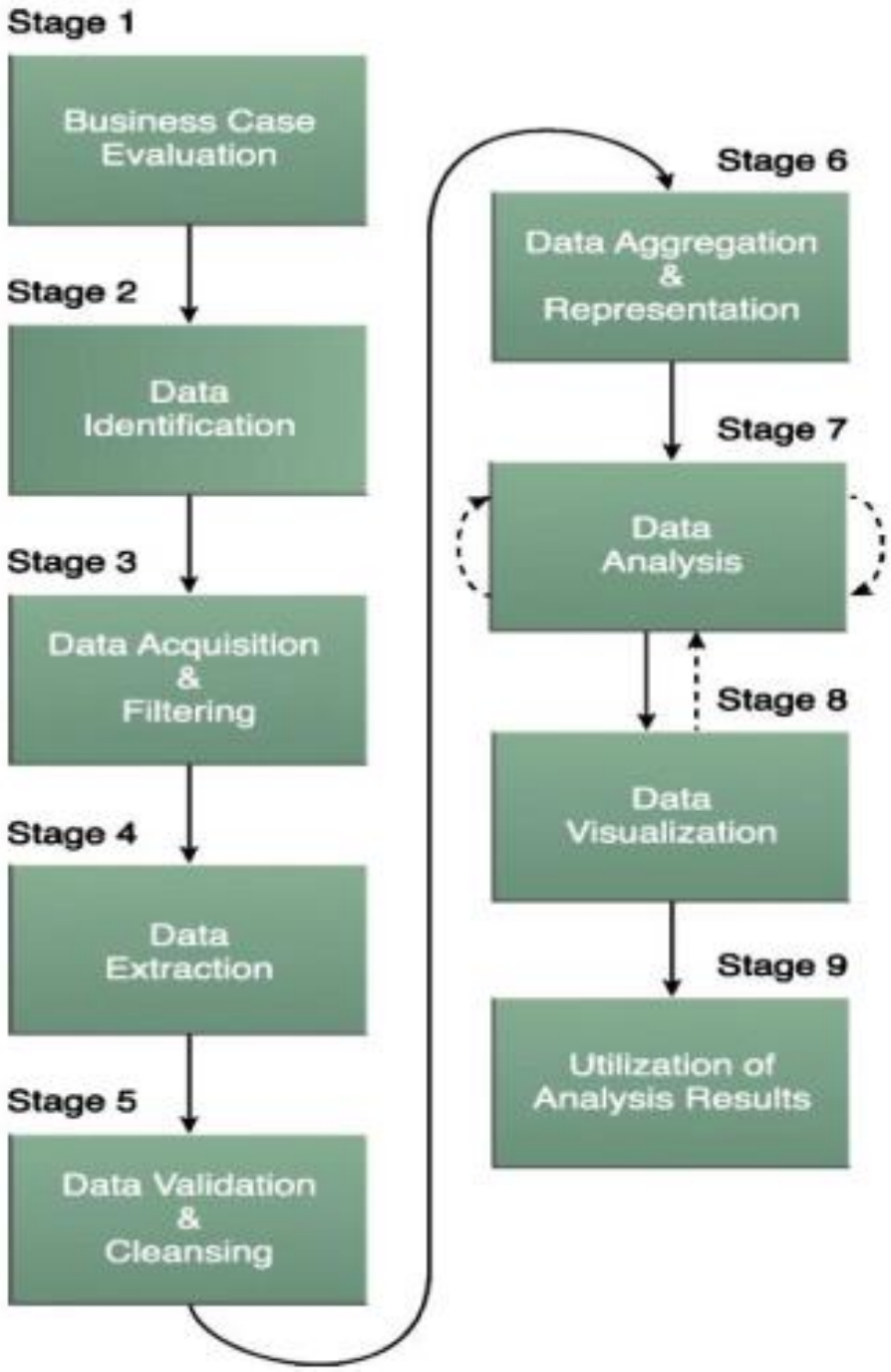
- **R and PL/ R** was described earlier in the model planning phase, and PL/ R is a procedural language for PostgreSQL with R. Using this approach means that R commands can be executed in database.
- **Octave** , a free software programming language for computational modeling, has some of the functionality of Matlab. Because it is freely available, Octave is used in major universities when teaching machine learning.
- **WEKA** is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
- **Python** is a programming language that provides toolkits for machine learning and analysis, such as scikit-learn, numpy, scipy, pandas, and related data visualization using matplotlib.
- **SQL** in-database implementations, such as MADlib, provide an alterative to in memory desktop analytical tools.
- **MADlib** provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL or Greenplum.

Data Analytics Lifecycle

Overview

The Big Data analytics lifecycle can be divided into the following nine stages, as shown in;

1. **Business Case Evaluation**
2. **Data Identification**
3. **Data Acquisition & Filtering**
4. **Data Extraction**
5. **Data Validation & Cleansing**
6. **Data Aggregation & Representation**
7. **Data Analysis**
8. **Data Visualization**
9. **Utilization of Analysis Results**



1. Business Case Evaluation

- Before any Big Data project can be started, it needs to be clear what the **business objectives and results** of the data analysis should be.
- 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.
- Once an overall business problem is defined, the **problem is converted into an analytical problem**.

2. Data Identification

- The Data Identification stage **determines the origin of data**. Before data can be analysed, it is important to know what the sources of the data will be.
- Especially **if data is procured from external suppliers**, it is necessary to clearly identify what the **original source of the data is and how reliable** (frequently referred to as the veracity of the data) the dataset is.
- The **second stage of the Big Data Lifecycle is very important** because if the input data is unreliable, the output data will also definitely be unreliable.
- Identifying a **wider variety of data sources** may **increase the probability of finding hidden patterns and correlations**.

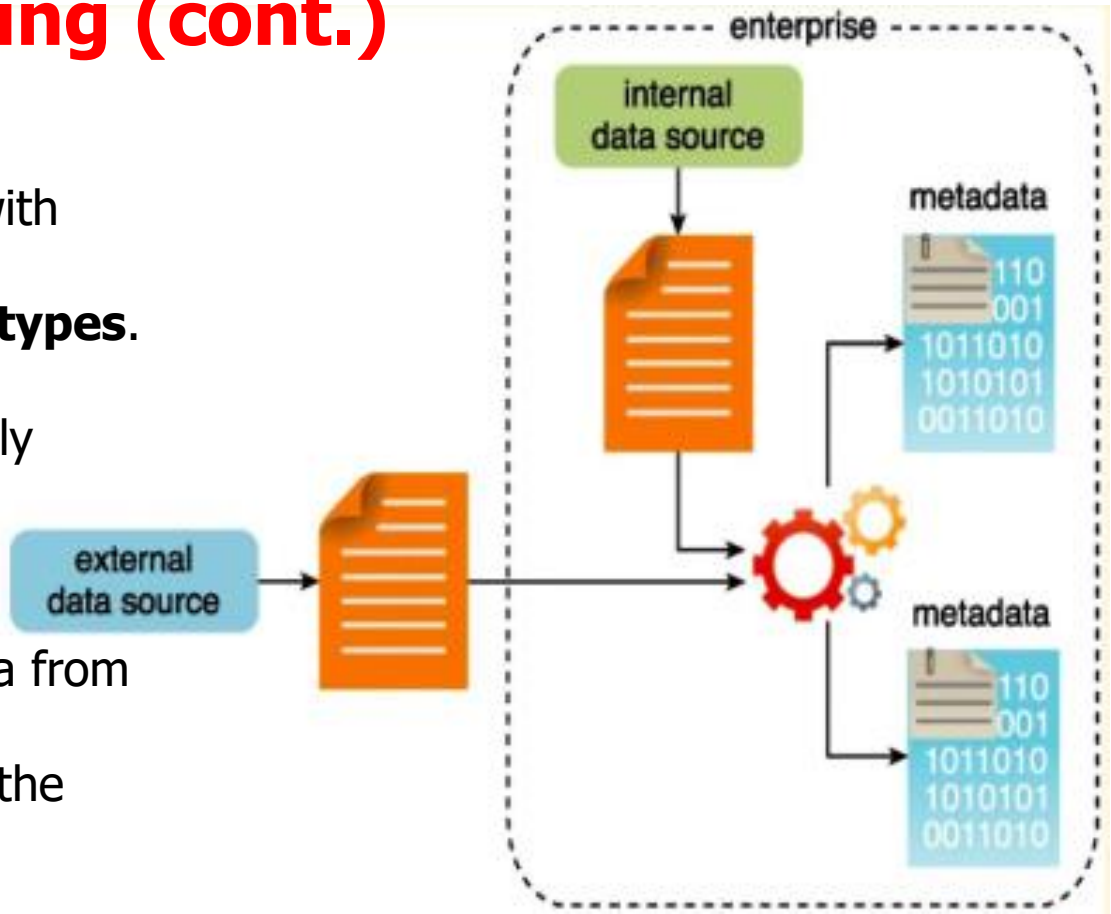
3. Data Acquisition and Filtering

- The Data Acquisition and Filtering Phase **builds upon the previous stage** of the Big Data Lifecycle.
- In this stage, the data is gathered from different sources, both from within the company and outside of the company.
- After the acquisition, a **first step of filtering is conducted to filter out corrupt data**.
- Additionally, data that is **not necessary for the analysis will be filtered out as well**.
- The **filtering step will be applied on each data source individually**, so before the data is aggregated into the data warehouse.
- In many cases, especially where **external, unstructured data is concerned, some or most of the acquired data may be irrelevant (noise)** and can be **discarded** as part of the filtering process.

3. Data Acquisition and Filtering (cont.)

- Data classified as "corrupt" can include records with **missing or nonsensical values or invalid data types**.

Data that is filtered out for one analysis may possibly be valuable for a different type of analysis.



- **Metadata can be added** via automation to data from both internal and external data sources to improve the classification and querying.
- Examples of appended **metadata include dataset size and structure, source information, date and time of creation or collection and language-specific information**.

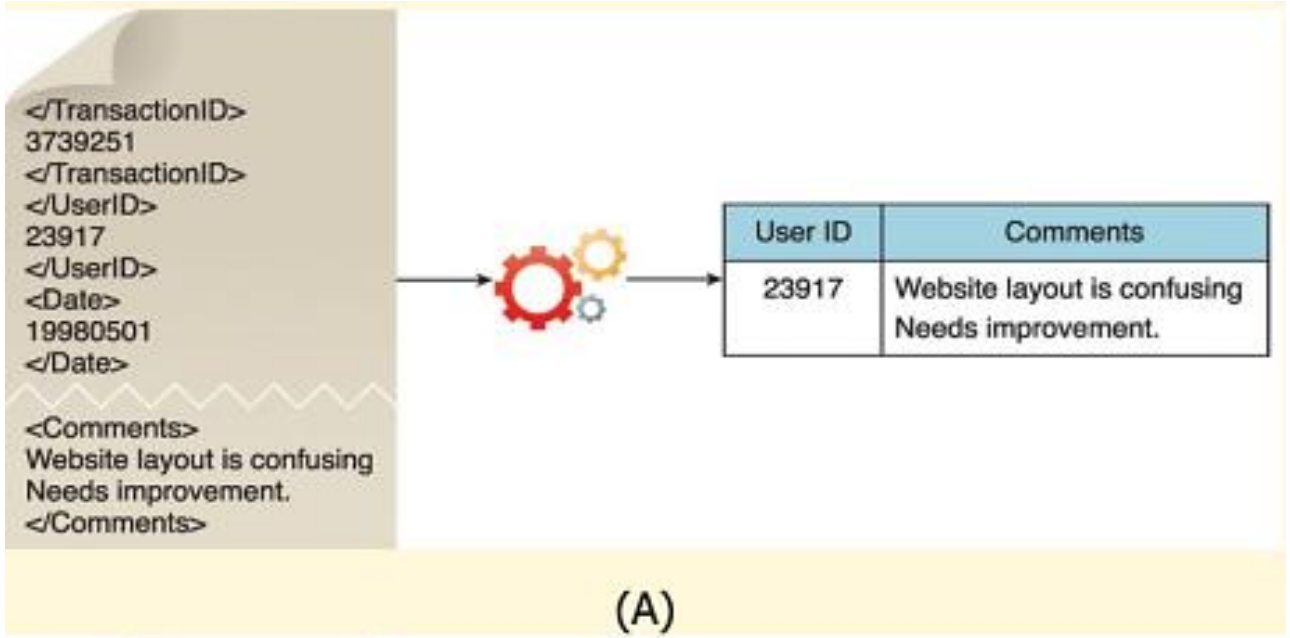
4. Data Extraction

- Some of the data identified in the **two previous stages may be incompatible** with the Big Data tool that will perform the actual analysis.
- In order to deal with this problem, the Data Extraction stage is dedicated **to extracting different data formats from data sets** (e.g. the data source) and **transforming these into a format the Big Data tool** is able to process and analyse.
- The complexity of the transformation and the extent in which is necessary to transform data is greatly dependent on the Big Data tool that has been selected.
- The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

4. Data Extraction (cont.)

(A). Illustrates the extraction of comments and a user ID embedded Within an XML document without the need for further transformation.

(B). Demonstrates the extraction of the latitude And longitude coordinates of a user from a single JSON field.

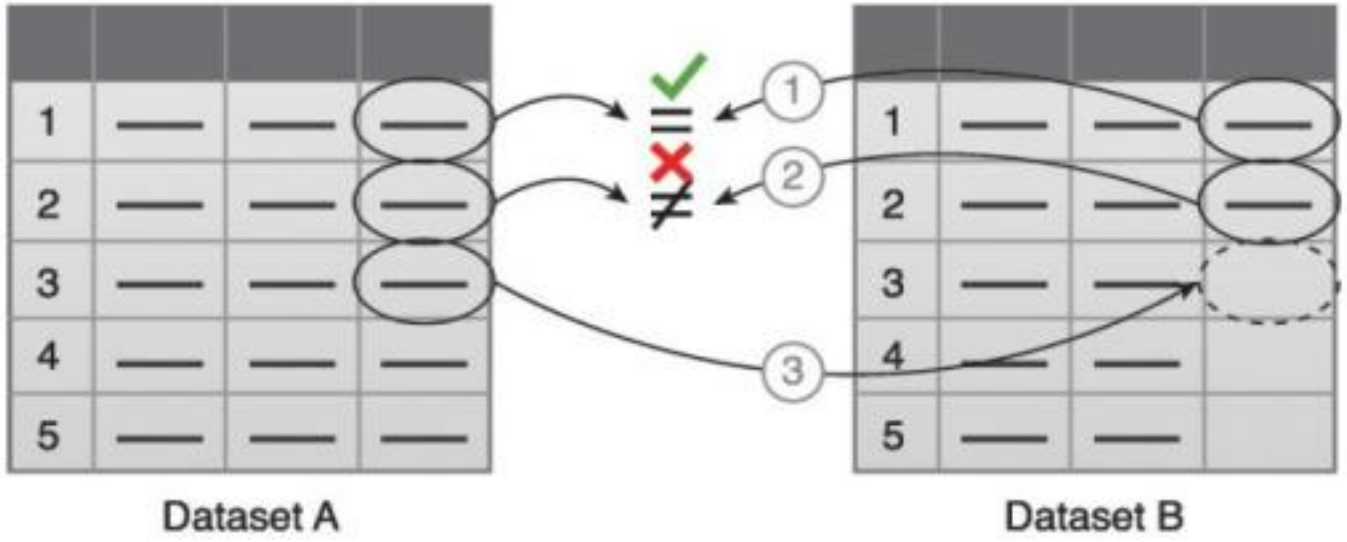


5. Data Validation and Cleansing

- **Data** that is **invalid leads to invalid results**. In order to ensure only the appropriate data is analysed, the **Data Validation and Cleansing stage of the Big Data Lifecycle is required**.
- During this stage, data is validated against **a set of predetermined conditions and rules** in order to ensure the data is not corrupt.
- An **example** of a validation rule would be to exclude all persons that are older than 100 years old since it is very unlikely that data about these persons would be correct due to physical constraints.
- The Data Validation and Cleansing stage is **dedicated to** often establishing complex **validation rules and removing any known invalid data.**

5. Data Validation and Cleansing (cont.)

- **For example**, as illustrated in Fig. 1, the first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A. If a value is missing, it is inserted from Dataset A.



- Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

6. Data Aggregation and Representation

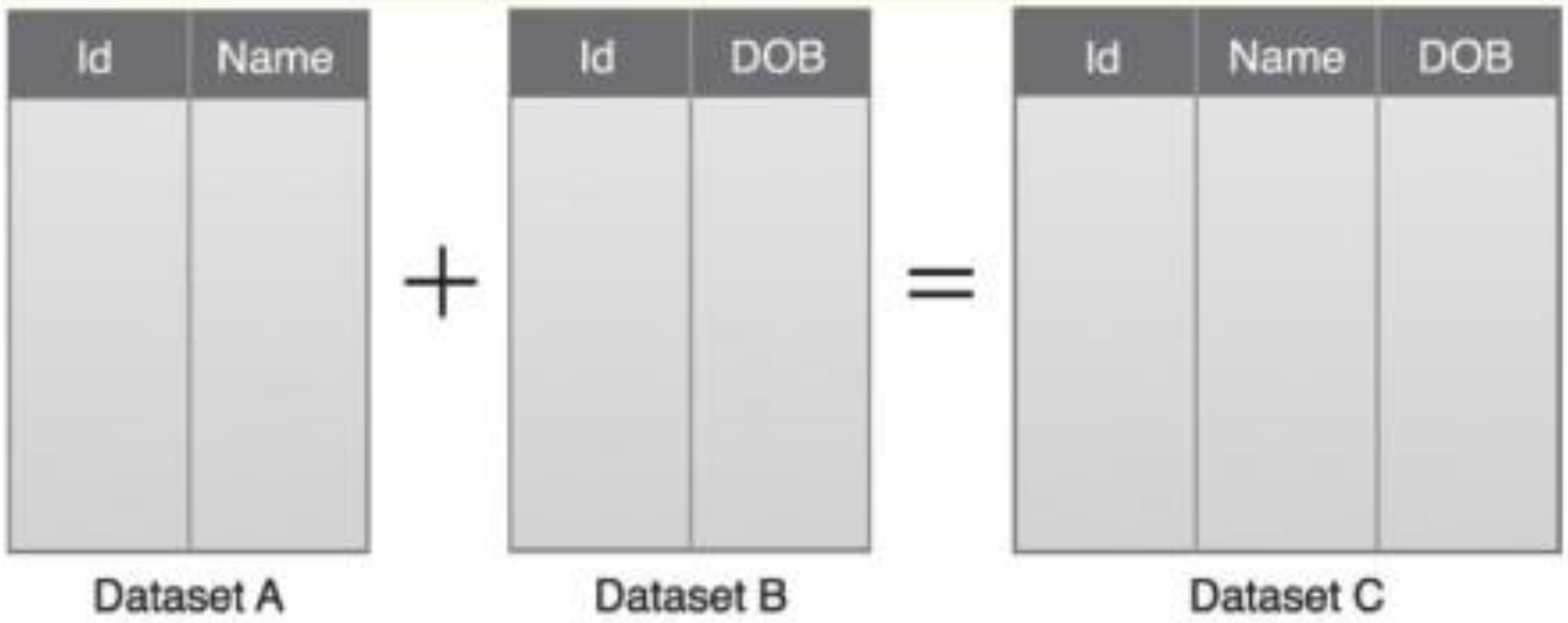
- Data may be **spread across multiple datasets**, requiring that dataset be joined together to conduct the actual analysis.
- In order **to ensure that only the correct data** will be analysed **in the next stage**, it might be **necessary to integrate multiple datasets**.
- The Data Aggregation and Representation stage is **dedicated to integrate multiple datasets to arrive at a unified view**.
- Additionally, **data aggregation will greatly speed up the analysis process** of the Big Data tool, because the tool will not be required to join different tables from different datasets, greatly speeding up the process.

6. Data Aggregation and Representation (cont.)

- Performing this stage can become complicated because of differences in:

• Data Structure – Although the data format may be the same, the data model may be different.
• Semantics – A value that is labeled differently in two different datasets may mean the same thing, for example "surname" and "last name."

- Whether data aggregation is required or not, **it is important to understand that the same data can be stored in many different forms. A simple example of data aggregation where two datasets a reaggregated together using the Id field.**

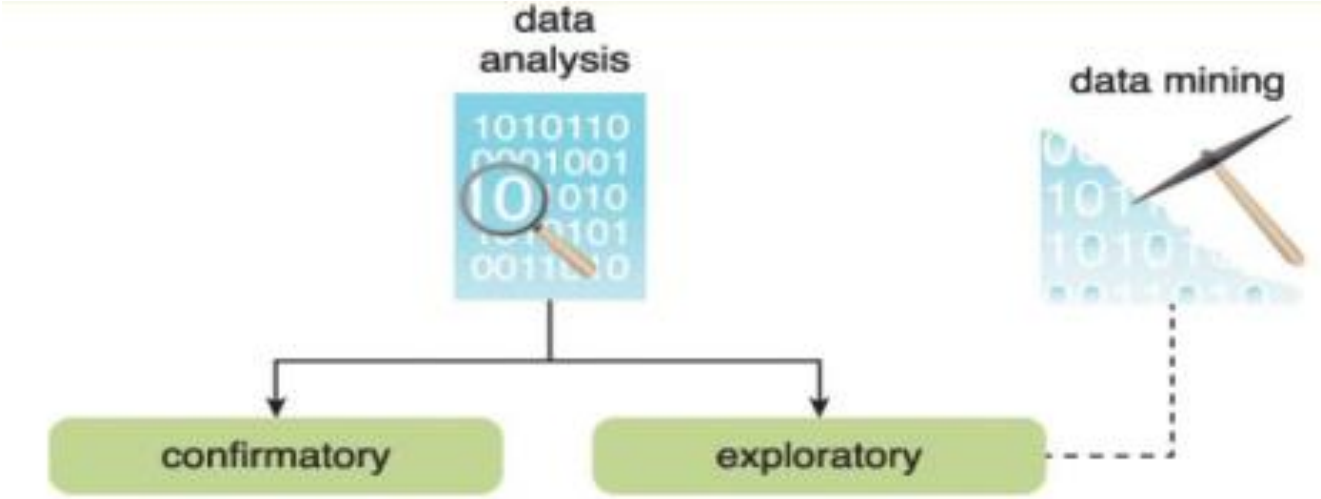


7. Data Analysis

- The Data Analysis stage of the Big Data Lifecycle stage is dedicated **to carrying out the actual analysis task**.
- It runs the **code or algorithm** that makes the calculations that will **lead to the actual result**.
- Data Analysis can be simple or really complex, depending on the required analysis type.
- In this stage the **'actual value'** of the Big Data project will be **generated**. **If all previous stages** have been **executed carefully**, the **results will be factual and correct**.
- Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison.
- On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a **statistical or mathematical model** to depict relationships between variables.

7. Data Analysis (cont.)

- Data analysis can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining, as shown in Figure,



➢ **Confirmatory** data analysis is **a deductive approach** where the cause of the **phenomenon being investigated** is proposed beforehand. The **proposed cause or assumption is called a hypothesis**.

➢ **Exploratory** data analysis is **an inductive approach** that is closely **associated with data mining**. **No hypothesis or predetermined assumptions are generated**. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon.

8. Data Visualization

- The ability to analyse massive amounts of data and find useful insight is one thing; communicating the results in a way that **everybody can understand** is something completely different.
- The Data visualization stage is dedicated to using **data visualization techniques** and **tools to graphically communicate the analysis results for effective interpretation** by business users. Frequently this requires **plotting data points in charts, graphs or maps**.
- The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated.

8. Data Visualization (cont.)

8. Data Visualization (cont.)

9. Utilization of Analysis Results

- **After the data analysis** has been performed and the **result** have been presented, the final step of the Big Data Lifecycle is **to use the results in practice**.
- The Utilization of Analysis results is **dedicated to determining how and where the processed data can be further utilized** to leverage the result of the Big Data Project.
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results **to produce "models" that encapsulate new insights and understandings** about the nature of the patterns and relationships that exist within the data that was analyzed.
- A model may look like a mathematical equation or a set of rules. Models can be **used to improve business process logic and application system logic**, and they can **form the basis of a new system or software program**.

Difference between data analysis and data utilization

Data Analysis

Purpose

To obtain information you want from data

Extract the information you need and graph it

Analyze

Data Utilization

Purpose

To use data to help business activities

Develop strategies based on data

Take action

Data

Validate results

THANK YOU