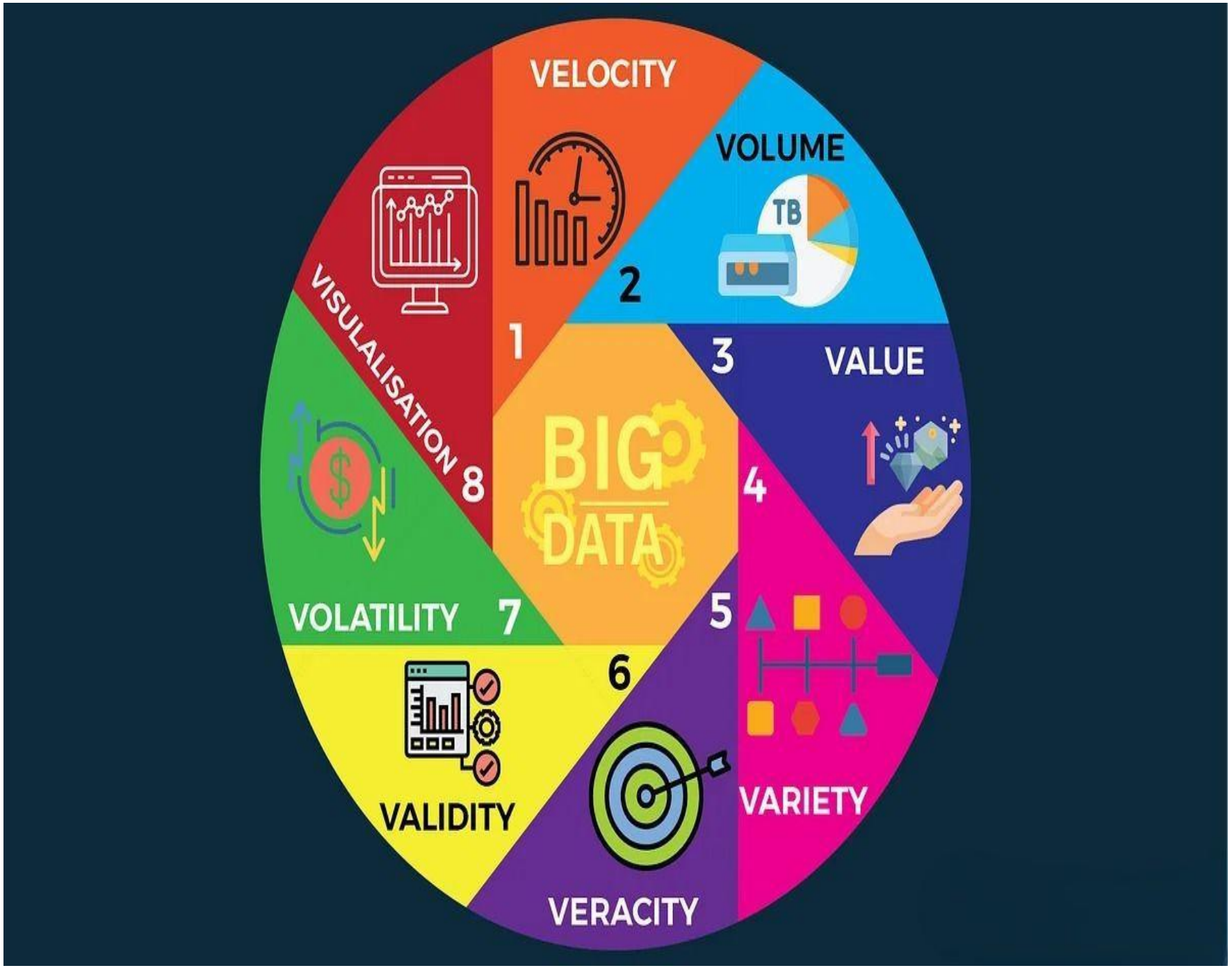
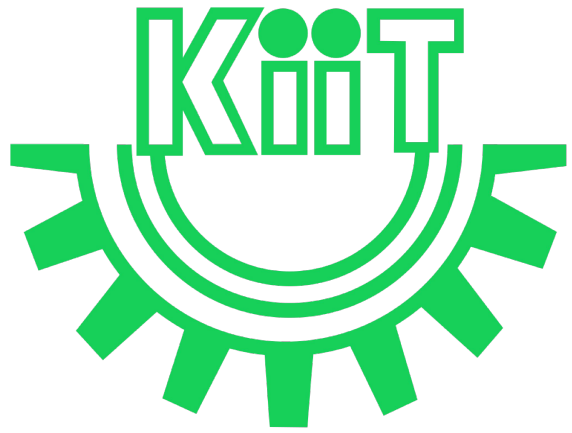


CS 3032: Big Data

Lec-4

In this Discussion . . .

- Big Data analytics lifecycle



Big Data Analytics Life Cycle

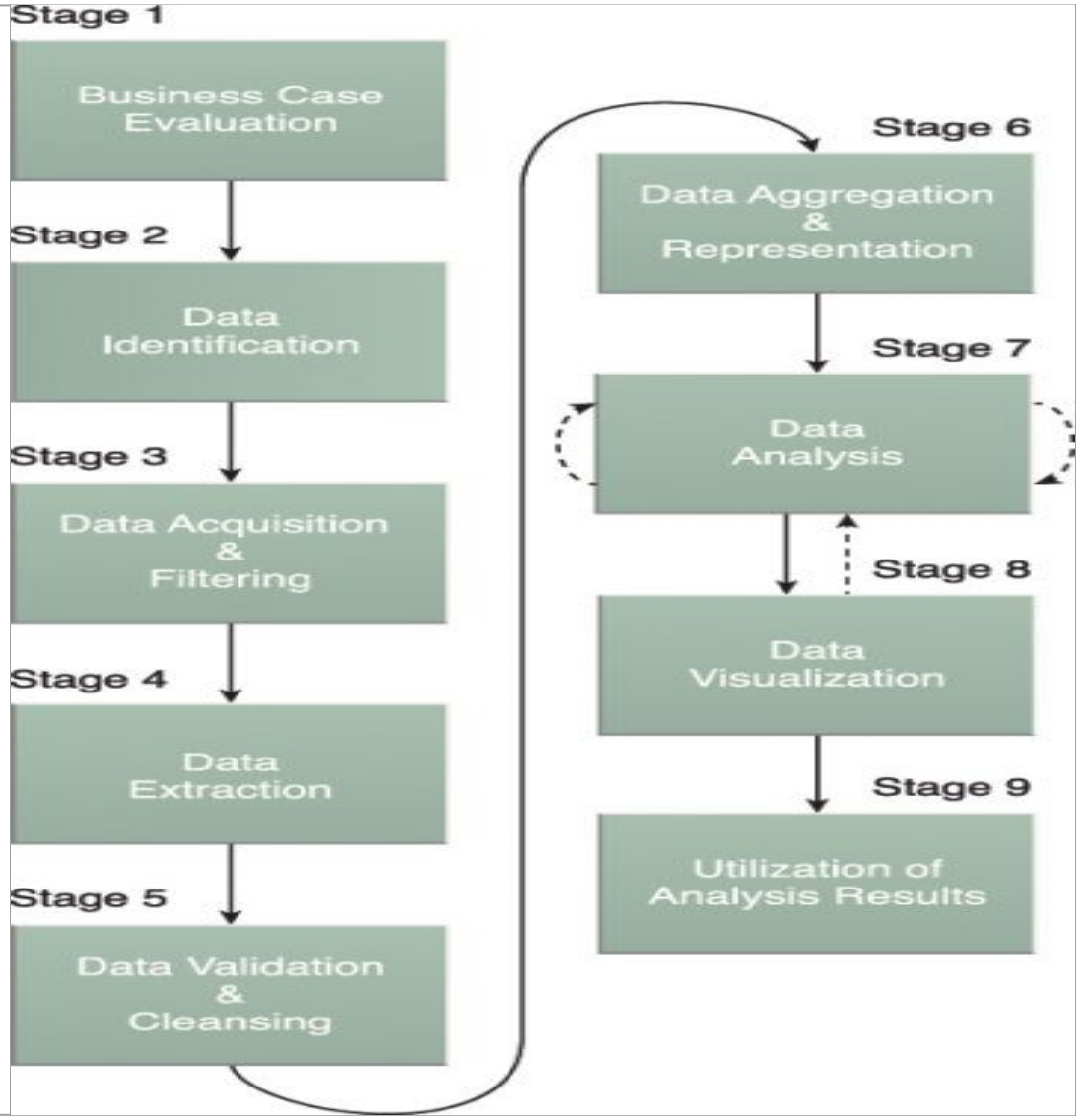
- Big Data analysis differs from traditional data analysis primarily due to the elements of Big Data (or V's of Big Data) like volume, velocity and variety characteristics of the data being processed.
- To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.

Big Data Analytics Life Cycle

- 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.
- The Big Data analytics lifecycle can be divided into the following **nine stages** presented in the next slide.

Big Data Analytics Life Cycle

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

- Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.
- The **Business Case Evaluation stage** requires **that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.**

1. Business Case Evaluation (Contd.)

- An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle.
- The further identification of Key Performance Indicators (KPI) during this stage can help determine assessment criteria and guidance for the evaluation of the analytic results.
- If KPIs are not readily available, efforts should be made to make the goals of the analysis project SMART, which stands for specific, measurable, attainable, relevant and timely.

1. Business Case Evaluation (Contd.)

- Based on business requirements that are documented in the business case, it can be determined whether the business problems being addressed are really Big Data problems.
- In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

1. Business Case Evaluation (Contd.)

- Also Note that another outcome of this stage is the determination of the underlying budget required to carry out the analysis project.
- Any required purchase, such as tools, hardware and training, must be understood in advance so that the anticipated investment can be weighed against the expected benefits of achieving the goals.
- Initial iterations of the Big Data analytics lifecycle will require more up-front investment of Big Data technologies, products and training compared to later iterations where these earlier investments can be repeatedly leveraged.

2. Data Identification

- The **Data Identification stage** is dedicated to **identifying the datasets required for the analysis project and their sources.**
- Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations.
- For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.

2. Data Identification (Contd.)

- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.
- In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched against a predefined dataset specification.

2. Data Identification (Contd.)

- In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled.
- Some forms of external data may be embedded within blogs or other types of content-based web sites, in which case they may need to be harvested via automated tools.

3. Data Acquisition & Filtering

- During the **Data Acquisition and Filtering stage**, **the data is gathered from all of the data sources that were identified during the previous stage.**
- *The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.*

3. Data Acquisition & Filtering (Contd.)

- Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter.
- 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.

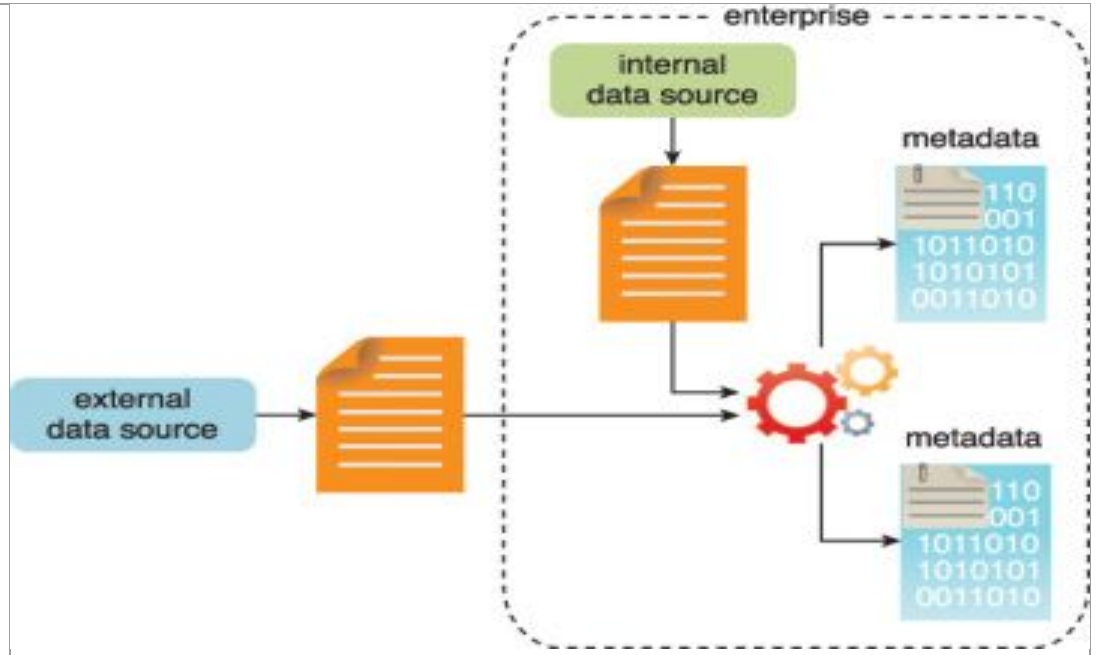
3. Data Acquisition & Filtering (Contd.)

- 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.
- Therefore, it is advisable to store a verbatim copy of the original dataset before proceeding with the filtering. To minimize the required storage space, the verbatim copy can be compressed.

3. Data Acquisition & Filtering (Contd.)

- Both internal and external data needs to be persisted once it gets generated or enters the enterprise boundary.
- For batch analytics, this data is persisted to disk prior to analysis. In the case of real-time analytics, the data is analyzed first and then persisted to disk.

- Metadata can be added via automation to data from both internal and external data sources to improve the classification and querying.

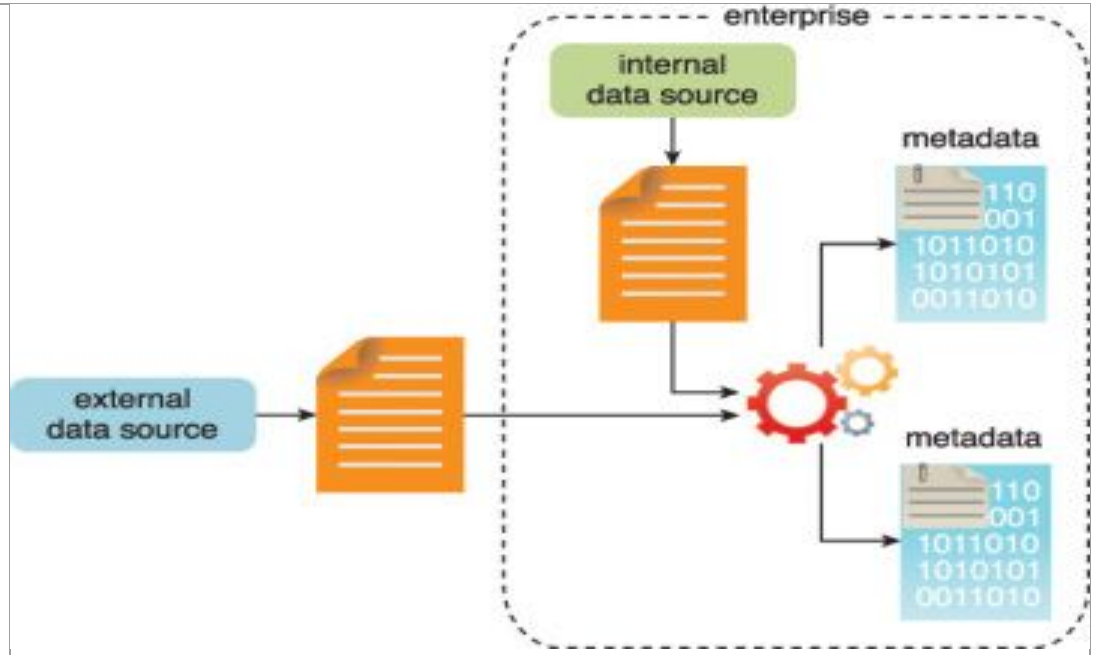


Metadata is added to data from internal and external sources.

3. Data Acquisition & Filtering (Contd.)

- Both internal and external data needs to be persisted once it gets generated or enters the enterprise boundary.
- For batch analytics, this data is persisted to disk prior to analysis. In the case of real-time analytics, the data is analyzed first and then persisted to disk.

- Examples of appended metadata include dataset size and structure, source information, date and time of creation or collection and language-specific information.

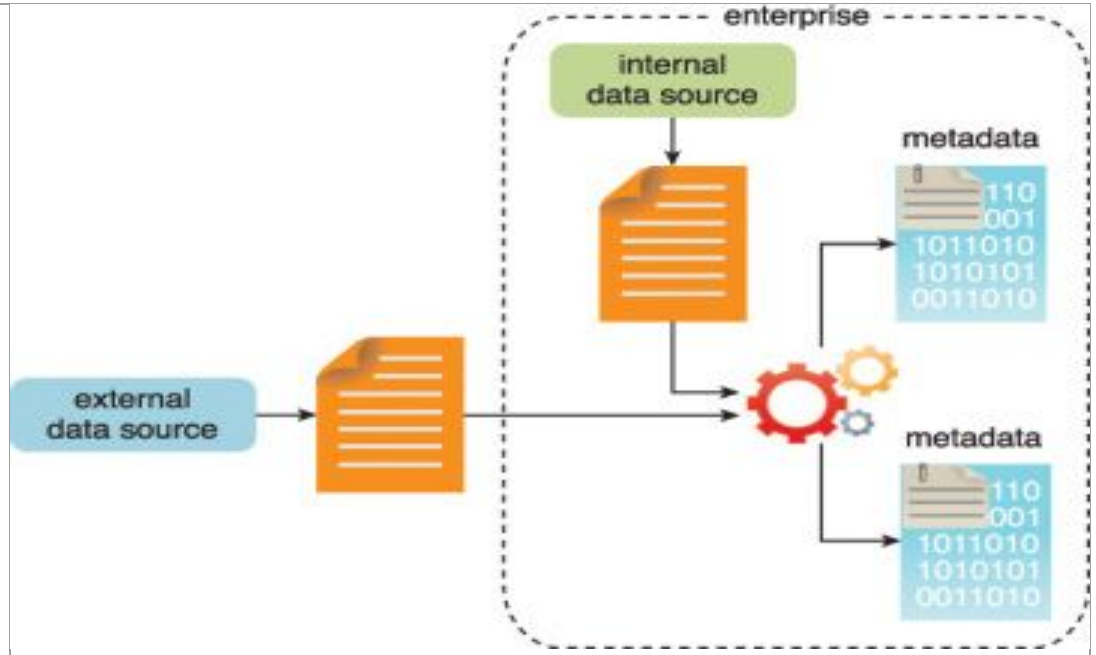


Metadata is added to data from internal and external sources.

3. Data Acquisition & Filtering (Contd.)

- Both internal and external data needs to be persisted once it gets generated or enters the enterprise boundary.
- For batch analytics, this data is persisted to disk prior to analysis. In the case of real-time analytics, the data is analyzed first and then persisted to disk.

- It is vital that metadata be machine-readable and passed forward along subsequent analysis stages.
- This helps maintain data provenance throughout the Big Data analytics lifecycle, which helps to establish and preserve data accuracy and quality.



Metadata is added to data from internal and external sources.

4. Data Extraction

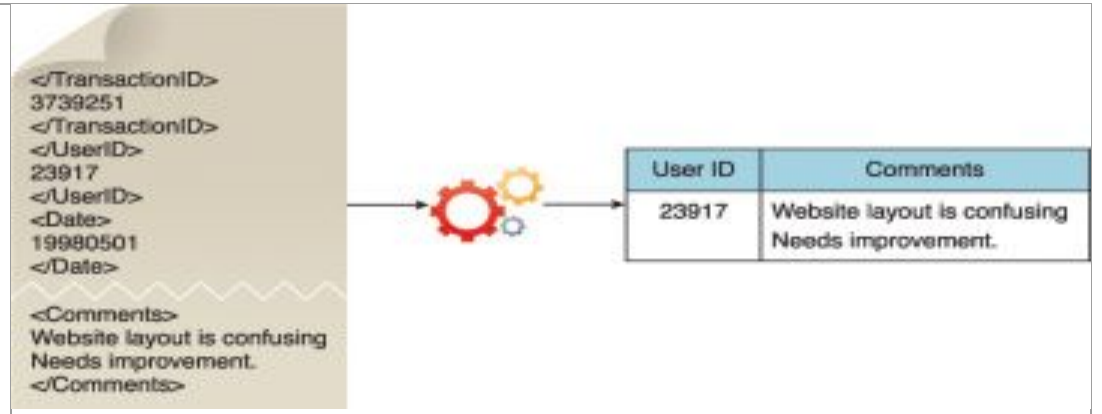
- Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution.
- The need to address disparate types of data is more likely with data from external sources.
- The **Data Extraction lifecycle stage**, **is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis**.

4. Data Extraction (Contd.)

- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution.
- For example, extracting the required fields from delimited textual data, such as with web-server log files, may not be necessary if the underlying Big Data solution can already directly process those files.
- Similarly, extracting text for text analytics, which requires scans of whole documents, is simplified if the underlying Big Data solution can directly read the document in its native format.

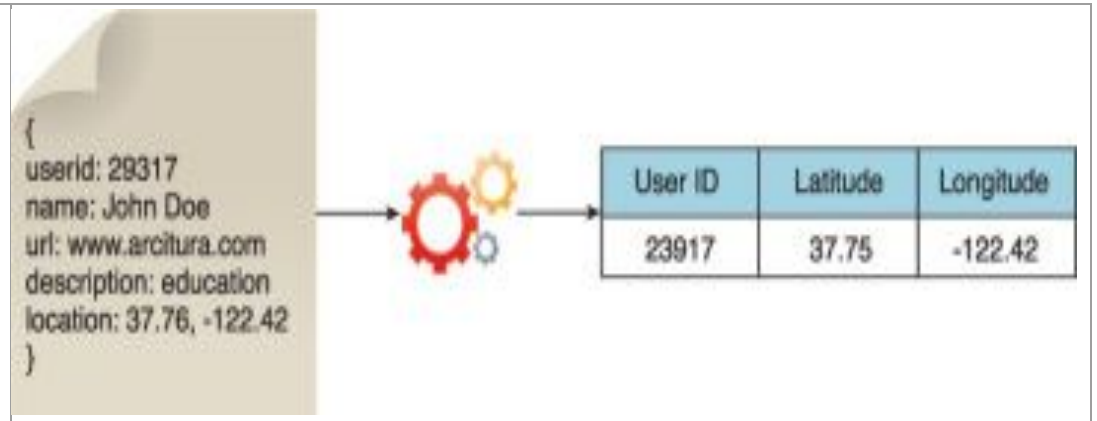
4. Data Extraction (Contd.)

- Extraction of comments and a user ID embedded within an XML document without the need for further transformation.



Comments and user IDs are extracted from an XML document.

- Extraction of the latitude and longitude coordinates of a user from a single JSON field.



The user ID and coordinates of a user are extracted from a single JSON field.

4. Data Extraction (Contd.)

- Further transformation is needed in order to separate the data into two separate fields as required by the Big Data solution.

5. Data Validation & Cleaning

- Invalid data can skew and falsify analysis results.
- Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analysis can be unstructured without any indication of validity.
- Its complexity can further make it difficult to arrive at a set of suitable validation constraints.

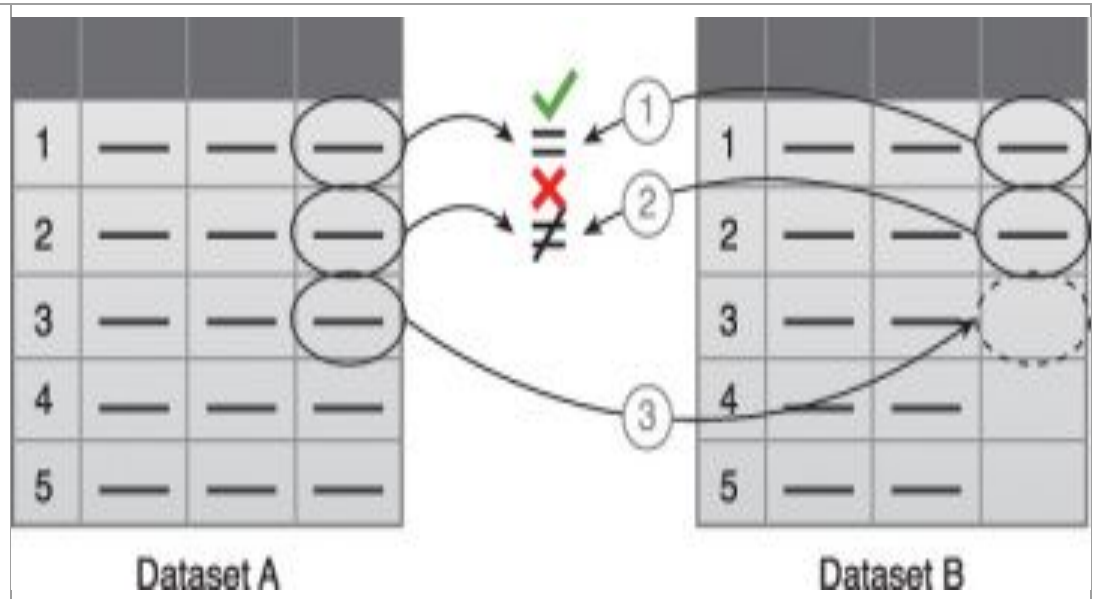
5. Data Validation & Cleaning (Contd.)

- **The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.**
- Big Data solutions often receive redundant data across different datasets.
- This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.

5. Data Validation & Cleaning (Contd.)

- For instance:

- 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.

5. Data Validation & Cleaning (Contd.)

- For batch analytics, data validation and cleansing can be achieved via an offline ETL operation.
- For real-time analytics, a more complex in-memory system is required to validate and cleanse the data as it arrives from the source.
- Provenance can play an important role in determining the accuracy and quality of questionable data.

5. Data Validation & Cleaning (Contd.)

Data that appears to be invalid may still be valuable in that it may possess hidden patterns and trends.

The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.

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 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.

7. Data Analysis

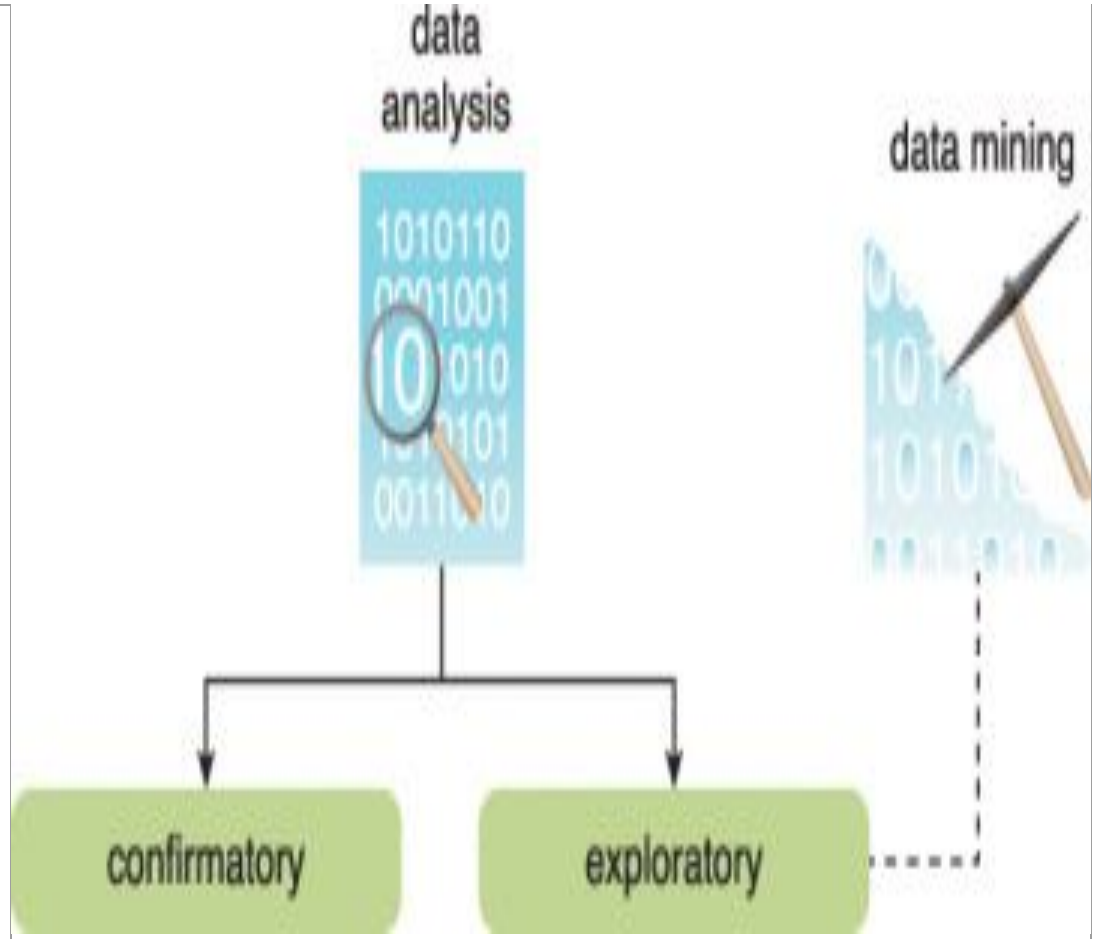
- The **Data Analysis stage** is dedicated to **carrying out the actual analysis task, which typically involves one or more types of analytics.**
- This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
- The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

7. Data Analysis (Contd.)

- 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 (Contd.)

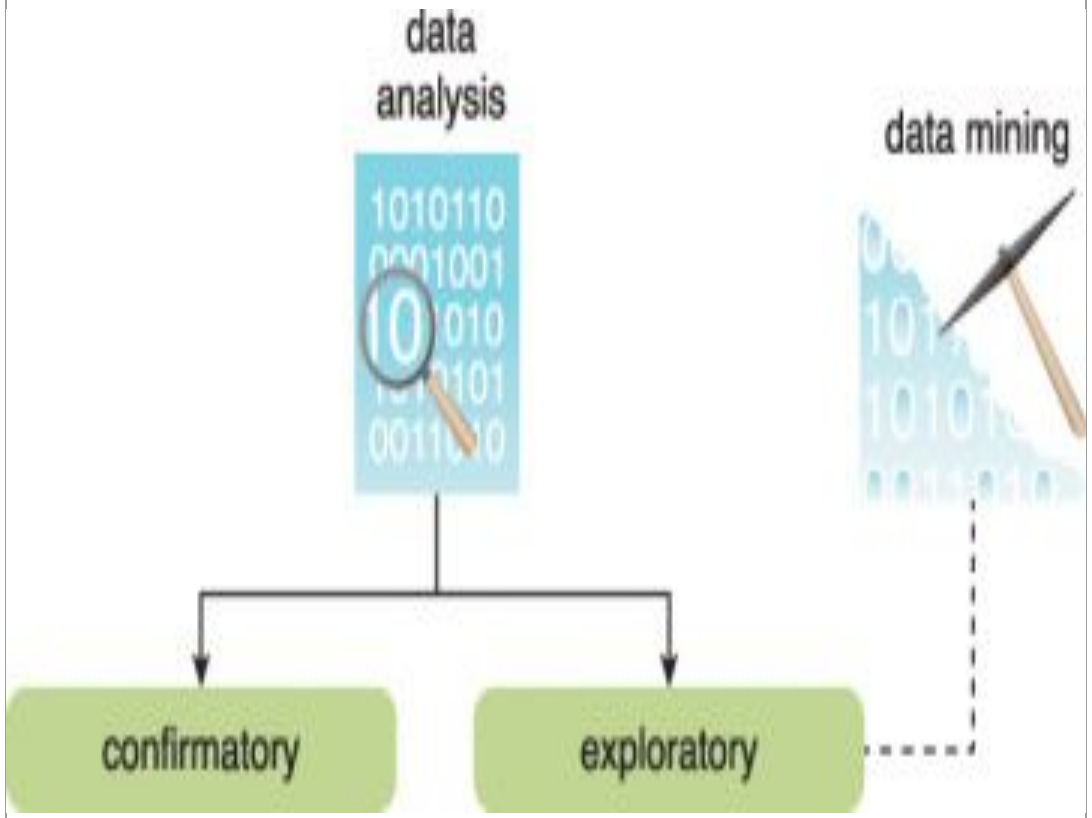
- **Confirmatory data analysis** is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand.
- The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions.
- Data sampling techniques are typically used.
- Unexpected findings or anomalies are usually ignored since a predetermined cause was assumed.



Data analysis can be carried out as confirmatory or exploratory analysis.

7. Data Analysis (Contd.)

- **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.
- Although it may not provide definitive answers, this method provides a general direction that can facilitate the discovery of patterns or anomalies.



Data analysis can be carried out as confirmatory or exploratory analysis.

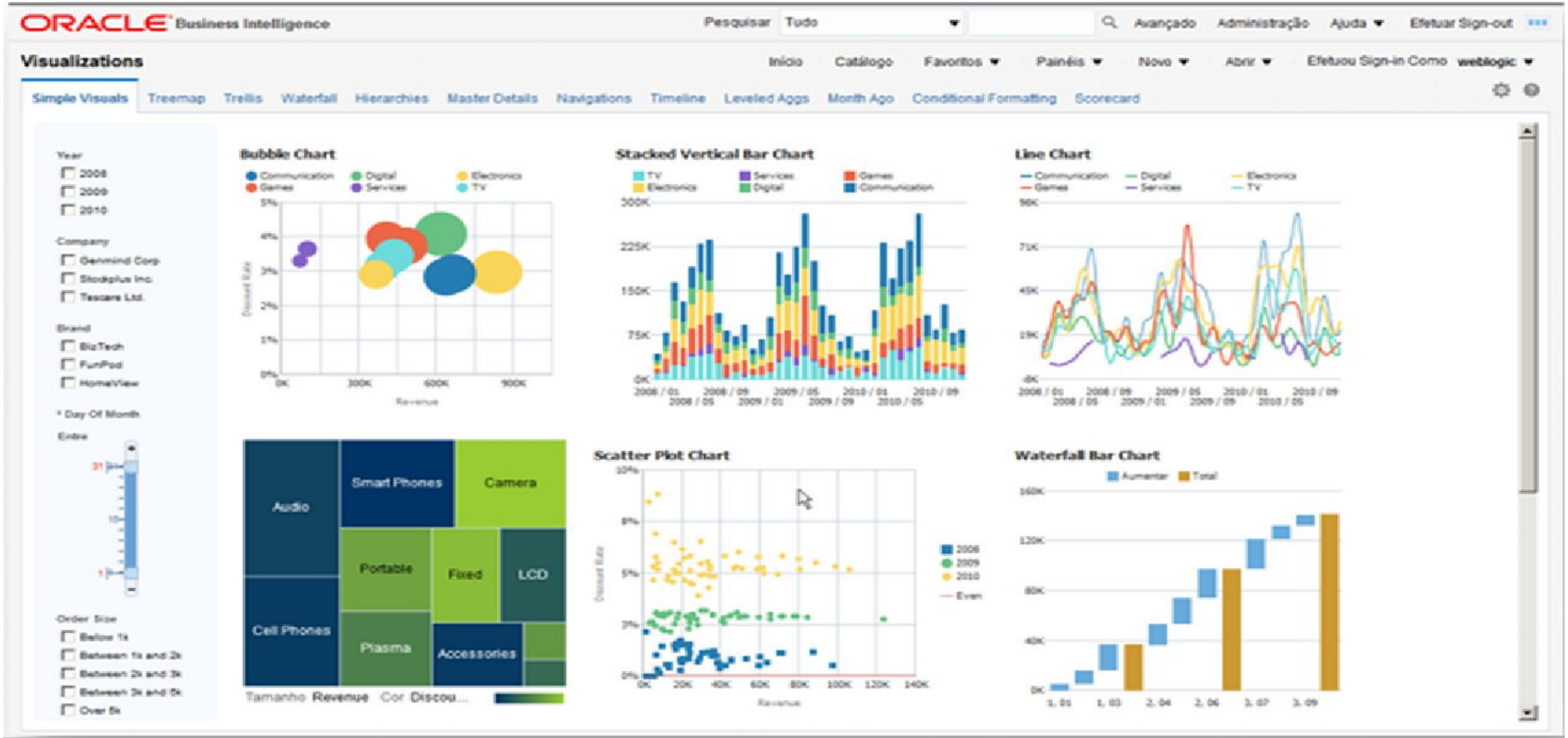
8. Data Visualization

- The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.
- 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.**

8. Data Visualization (Contd.)

- **Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback, as indicated by the dashed line leading from stage 8 back to stage 7.**
- 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 (Contd.)



9. Utilization of Analysis Results

- Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportunities to utilize the analysis results.
- The **Utilization of Analysis Results stage**, is dedicated to **determining how and where processed analysis data can be further leveraged.**

9. Utilization of Analysis Results (Contd.)

- 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.

9. Utilization of Analysis Results (Contd.)

- Common areas that are explored during this stage include the following:

Input for Enterprise Systems	Business Process Optimization	Alerts
● The data analysis results may be automatically or manually fed directly into enterprise systems to enhance and optimize their behaviors and performance. ● For example, an online store can be fed processed customer-related analysis results that may impact how it generates product recommendations. ● New models may be used to improve the programming logic within existing enterprise systems or may form the basis of new systems.	● The identified patterns, correlations and anomalies discovered during the data analysis are used to refine business processes. ● An example is consolidating transportation routes as part of a supply chain process. Models may also lead to opportunities to improve business process logic.	● Data analysis results can be used as input for existing alerts or may form the basis of new alerts. ● For example, alerts may be created to inform users via email or SMS text about an event that requires them to take corrective action

Cloud Computing & Big Data

References

1. https://www.informit.com/articles/article.aspx?p=2473128&seqNum=11
2.