CSE 424

Big Data

Introduction to Big data

Slides 1

Instructor: Asst. Prof. Dr. Hüseyin ABACI

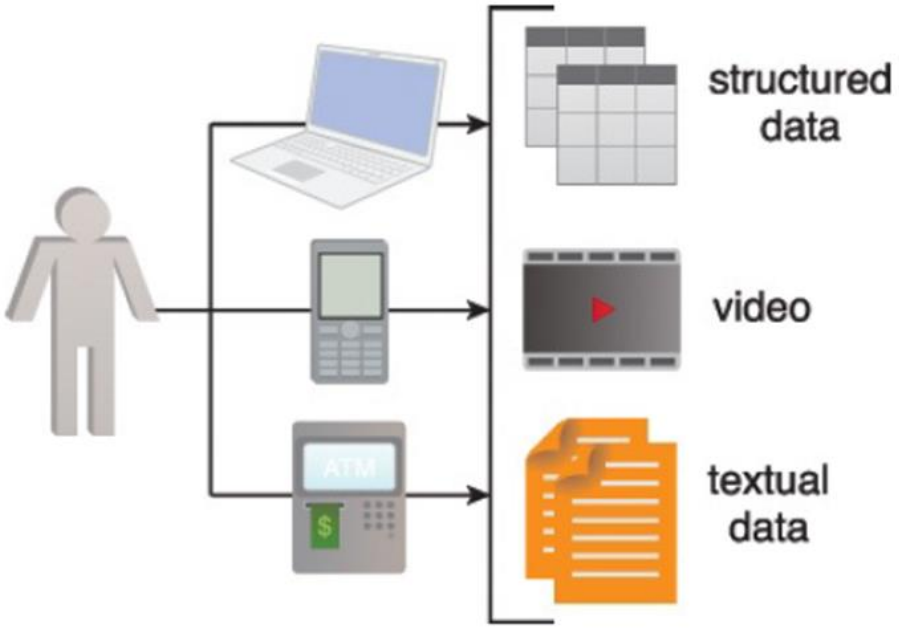
Outline

- Understanding Big Data
- Big Data Characteristics
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value
- Concepts and Terminology
  - Datasets, Data Analysis
  - Data Analytics
- Different Types of Data
  - Structured Data
  - Unstructured Data
  - Semi-structured Data
  - Metadata
- Big Data Analytics Lifecycle
  - Business Case Evaluation
  - Data Identification
  - Data Acquisition and Filtering
  - Data Extraction
  - Data Validation and Cleansing
  - Data Aggregation and Representation
  - Data Analysis
  - Data Visualization
  - Utilization of Analysis Results

Thomas Erl, Wajid Khattak, and Paul Buhler, Big Data Fundamentals Concepts: Drivers & Techniques, 2016
Bahga, Arshdeep, and Vijay Madisetti. Big data science & analytics: A hands-on approach. VPT, 2016.

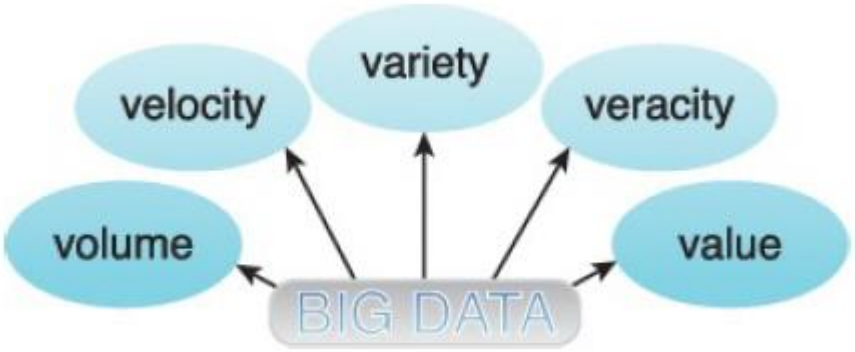
Understanding Big Data

- Big Data is a field dedicated to the analysis, processing, and storage of large collections of data that frequently originate from disparate sources.
- Big Data solutions and practices are typically required when traditional data analysis, processing and storage technologies and techniques are insufficient.
- Specifically, Big Data addresses distinct requirements, such as the combining of multiple unrelated datasets, processing of large amounts of unstructured data and harvesting of hidden information in a time-sensitive manner.



Big Data Characteristics

- This section explores the five Big Data characteristics that can be used to help differentiate data categorized as "Big" from other forms of data.
- The five Big Data traits shown in Figure below are commonly referred to as the Five Vs:



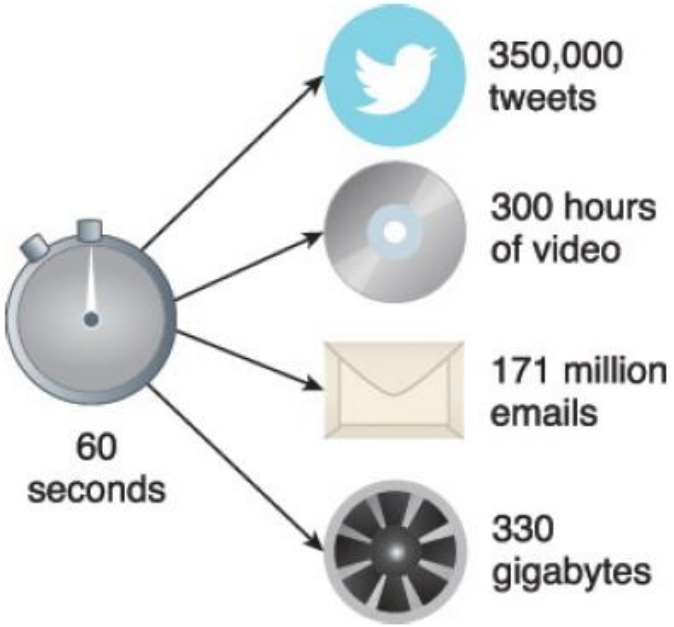
The Five Vs of Big Data.

Big Data Characteristics - Volume

- Organizations and users world-wide create over 2.5 EBs of data a day.
- The anticipated volume of data that is processed by Big Data solutions is substantial and ever-growing.
- Typical data sources that are responsible for generating high data volumes can include:
  - Online transactions, such as point-of-sale and banking
  - Scientific and research experiments, such as the Large Hadron Collider and Atacama Large Millimeter/Submillimeter Array telescope
  - Network traffic traces
  - Sensors, such as GPS sensors, RFIDs, smart meters and telematics
  - Social media, such as Facebook and Twitter

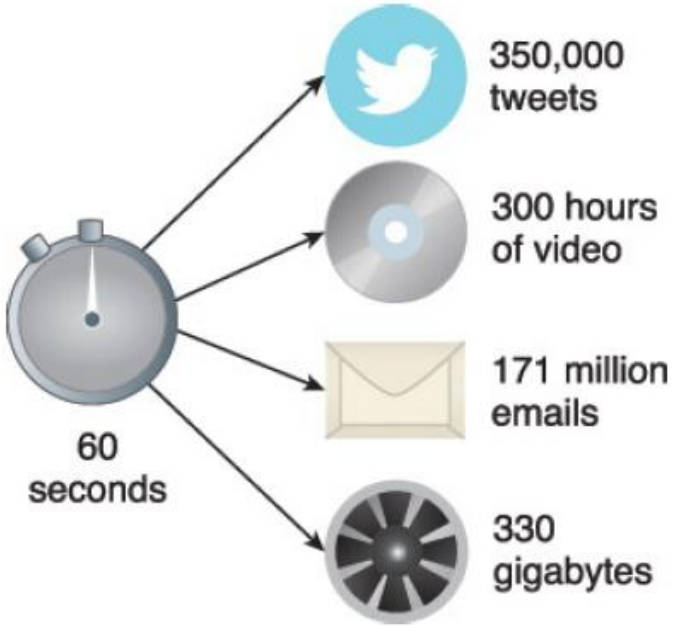
Data velocity is put into perspective when considering that the following data volume can easily be generated in a given minute: 350,000 tweets, 300 hours of video footage uploaded to YouTube, 171 million emails, 330 GBs of sensor data from a jet engine, Amazon receives 4300 new visitors Instagram users like nearly 1.73 million photos, Apple users download nearly 51,000 apps.

An **exabyte** (EB) > **petabyte** (PB) > **terabyte** (TB) > **gigabyte** (GB) > **megabyte** (MB) > **kilobyte** (KB) > **byte** (B).



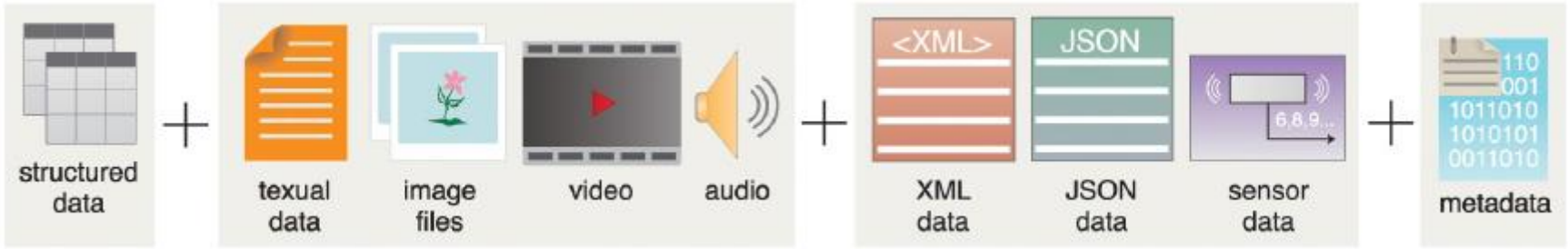
Big Data Characteristics - Velocity

- In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time.
- From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter.
- Coping with the fast inflow of data requires the enterprise to design highly elastic and available data processing solutions and corresponding data storage capabilities.



Big Data Characteristics - Variety

- Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.
- Figure below provides a visual representation of data variety, which includes structured data in the form of financial transactions, semi-structured data in the form of emails and unstructured data in the form of images.



structured un-structured semi-structured data about data

Big Data Characteristics - Veracity

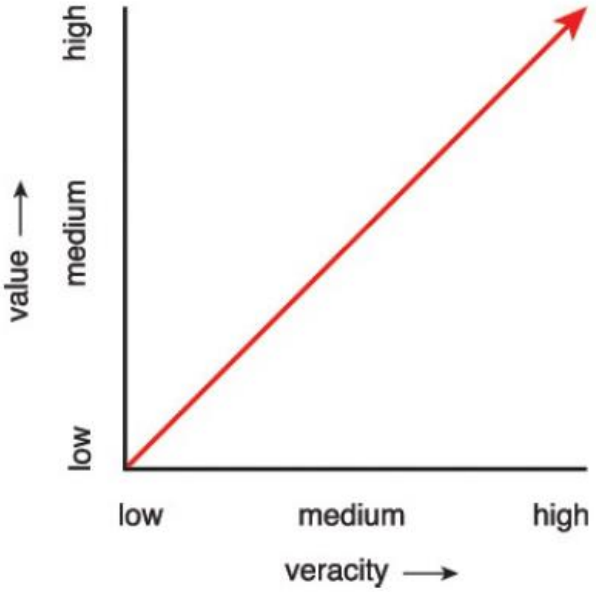
- Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise.

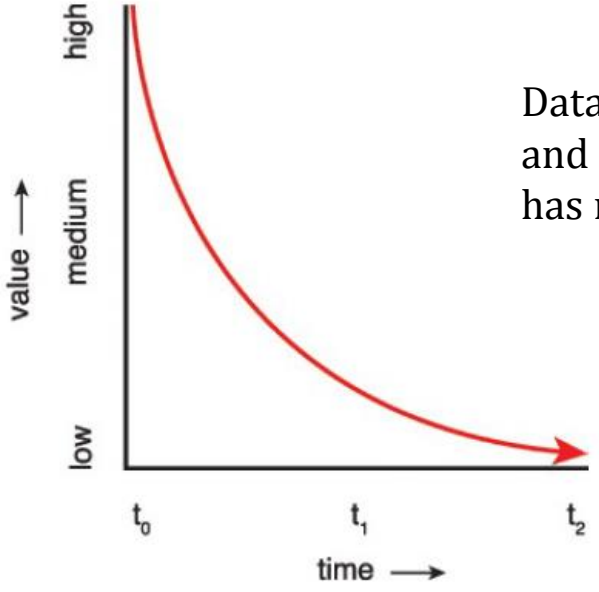
- In relation to veracity, data can be part of the signal or noise of a dataset. Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information.

- Data that is acquired in a controlled manner, for example via online customer registrations, usually contains less noise than data acquired via uncontrolled sources, such as blog postings.

Big Data Characteristics - Value

- Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.
- Value is also dependent on how long data processing takes because analytics results have a shelf-life; for example, a 20 minute delayed stock quote has little to no value for making a trade compared to a quote that is 20 milliseconds old.
- As demonstrated, value and time are inversely related. The longer it takes for data to be turned into meaningful information, the less value it has for a business. Stale results inhibit the quality and speed of informed decision-making.





Data that has high veracity and can be analyzed quickly has more value to a business.

Big Data Characteristics – Value (Cont.)

- Apart from veracity and time, value is also impacted by the following lifecycle-related concerns:
  - How well has the data been stored?
  - Were valuable attributes of the data removed during data cleansing?
  - Are the right types of questions being asked during data analysis?

Concepts and Terminology

- As a starting point, several fundamental concepts and terms need to be defined and understood.

- Datasets

- Data Analysis

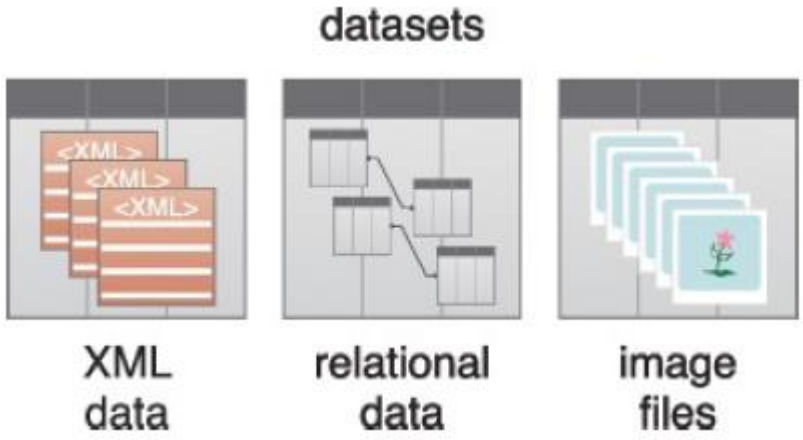
- Data Analytics

  There are four general categories of analytics that are distinguished by the results they produce:

  - descriptive analytics
  - diagnostic analytics
  - predictive analytics
  - prescriptive analytics

Concepts and Terminology - Datasets

- Collections or groups of related data are generally referred to as datasets. Each group or dataset member (datum) shares the same set of attributes or properties as others in the same dataset. Some examples of datasets are:
  - tweets stored in a flat file
  - a collection of image files in a directory
  - an extract of rows from a database table stored in a CSV formatted file
  - historical weather observations that are stored as XML files



Datasets can be found in many different formats.

Datasets – Primary Types

- Human-generated (social media etc.) and machine-generated (sensors etc.) data can come from a variety of sources and be represented in various formats or types.

- The primary types of data are:
  - structured data
  - unstructured data
  - semi-structured data

- These data types refer to the internal organization of data and are sometimes called data formats.

Datasets – Structured Data

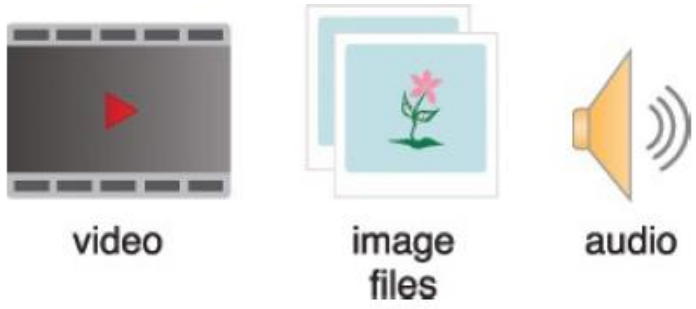
- Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database.
- Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems.

- Examples of this type of data include
  - banking transactions,
  - invoices,
  - customer records.

Structured data stored in a tabular form.

Datasets – Unstructured Data

- Data that does not conform to a data model or data schema is known as unstructured data.
  - It is estimated that unstructured data makes up 80% of the data within any given enterprise. Unstructured data has a faster growth rate than structured data.
- This form of data is either textual or binary and often conveyed via files that are self-contained and non-relational.
  - A **text file** may contain the contents of various tweets or blog postings.
  - **Binary files** are often media files that contain image, audio or video data.
- Unstructured data cannot be directly processed or queried using SQL.
  - If it is required to be stored within a relational database, it is stored in a table as a Binary Large Object (BLOB).
  - Alternatively, a Not-only SQL (NoSQL) database is a non-relational database that can be used to store unstructured data alongside structured data.



Video, image and audio files are all types of unstructured data.

Datasets – Semi-structured Data

- Semi-structured data has a defined level of structure and consistency, but is not relational in nature. Instead, semi-structured data is hierarchical or graph-based. This kind of data is commonly stored in files that contain text.
- Due to the textual nature of this data and its conformance to some level of structure, it is more easily processed than unstructured data.

XML, JSON and sensor data are semi-structured.

- Examples of common sources of semi-structured data include
  - spreadsheets,
  - RSS feeds
  - sensor data.

Datasets – Metadata

- Metadata provides information about a dataset's characteristics and structure. This type of data is mostly machine-generated and can be appended to data.

- The tracking of **metadata is crucial** to Big Data processing, storage and analysis because it provides information about the pedigree of the data, particularly when processing semi-structured and unstructured data.

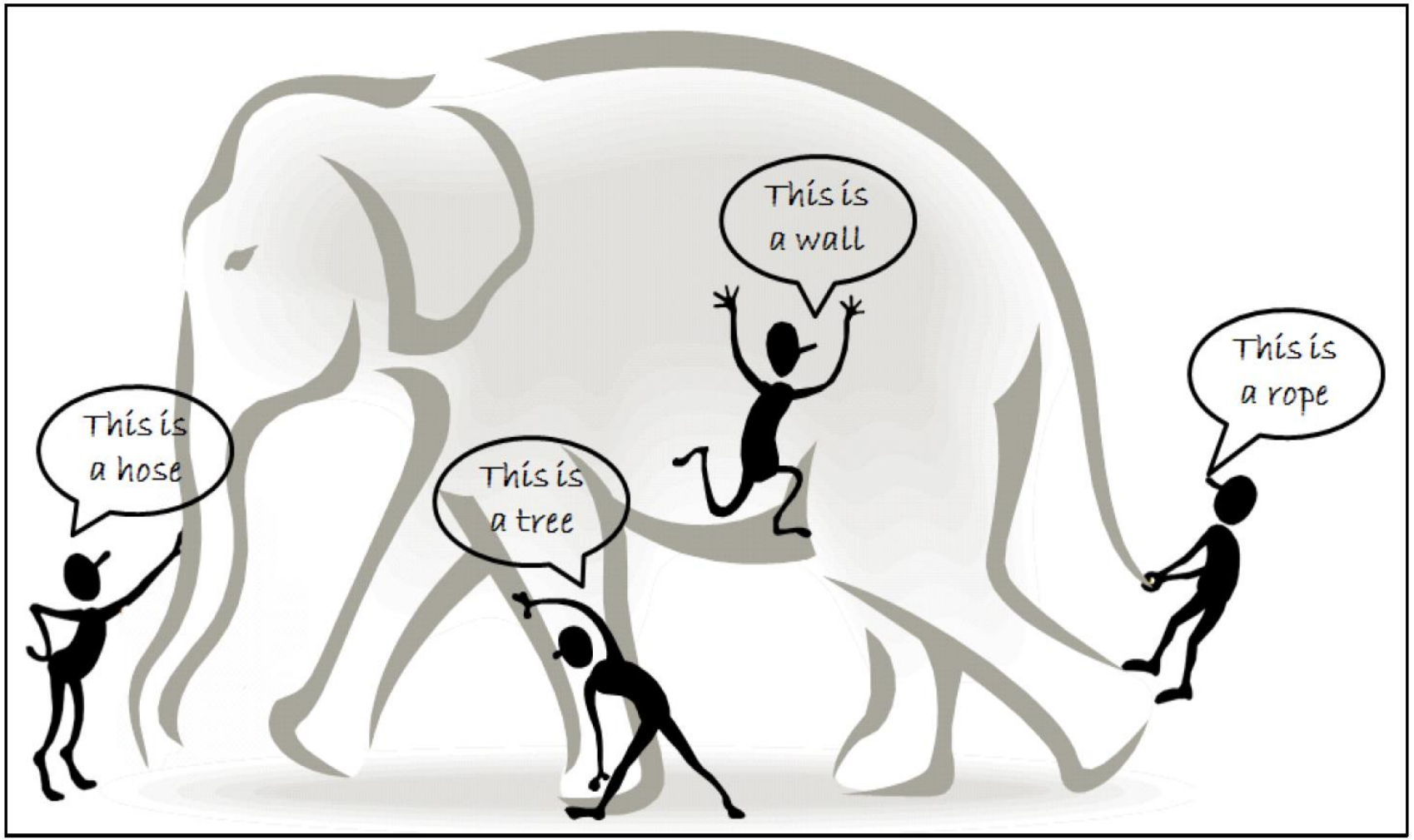
- Examples of metadata include:
  - XML tags providing the author and creation date of a document
  - attributes providing the file size and resolution of a digital photograph
  - language-specific information

Concepts and Terminology - Data Analysis

- Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal of data analysis is to support better decision making.
- A simple data analysis example is the analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the daily temperature.
- The results of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information. Carrying out data analysis helps establish patterns and relationships among the data being analyzed.

Concepts and Terminology - Data Analysis

- **The blind men and the giant elephant:** the localized (limited) view of each blind man leads to a biased conclusion.



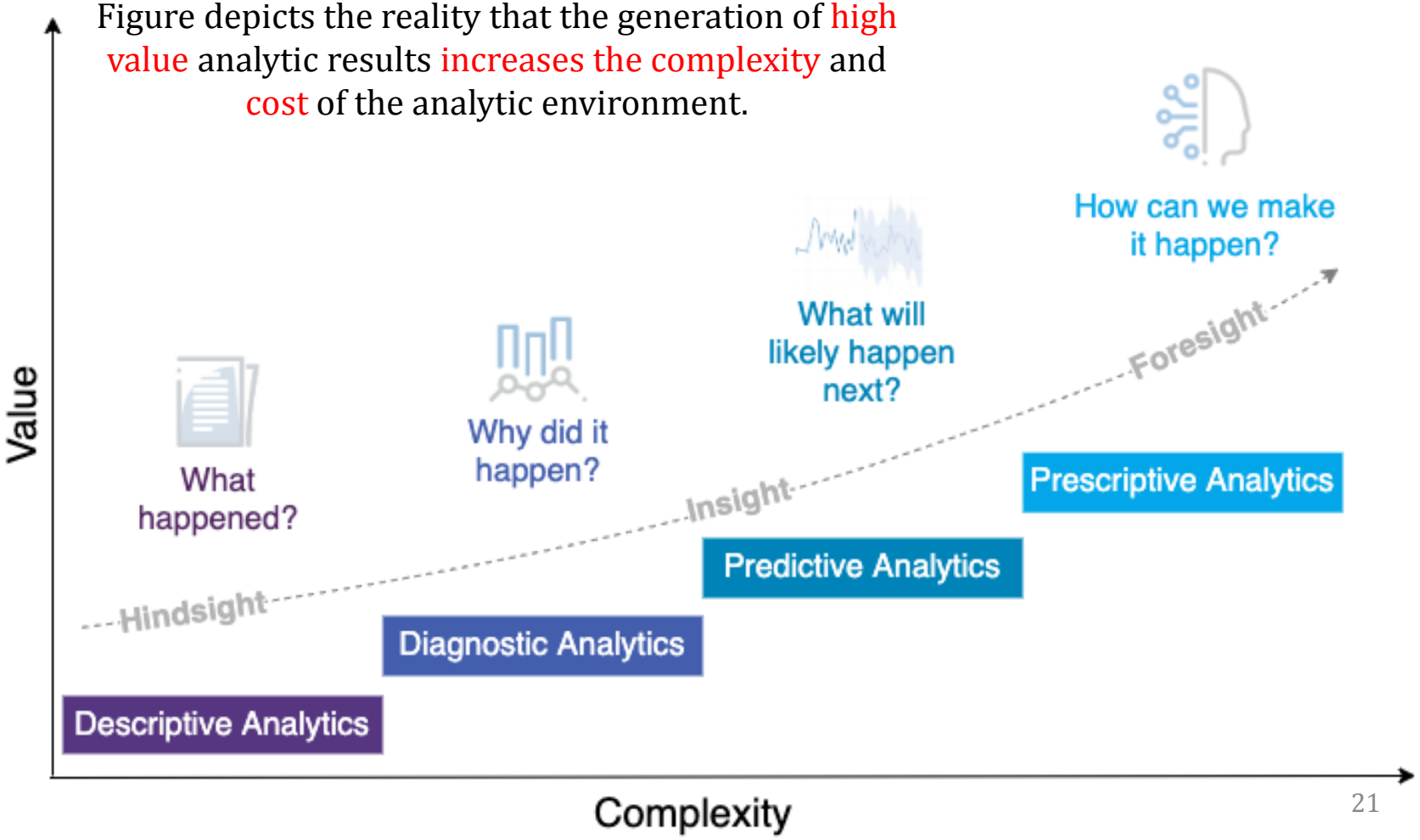
Concepts and Terminology - Data Analytics

- Data analytics is a **broader term** that encompasses data analysis.
- In Big Data environments, data analytics has developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analysing large volumes of data from different sources.
- The Big Data analytics lifecycle generally involves identifying, preparing and analysing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches.

Concepts and Terminology - Data Analytics

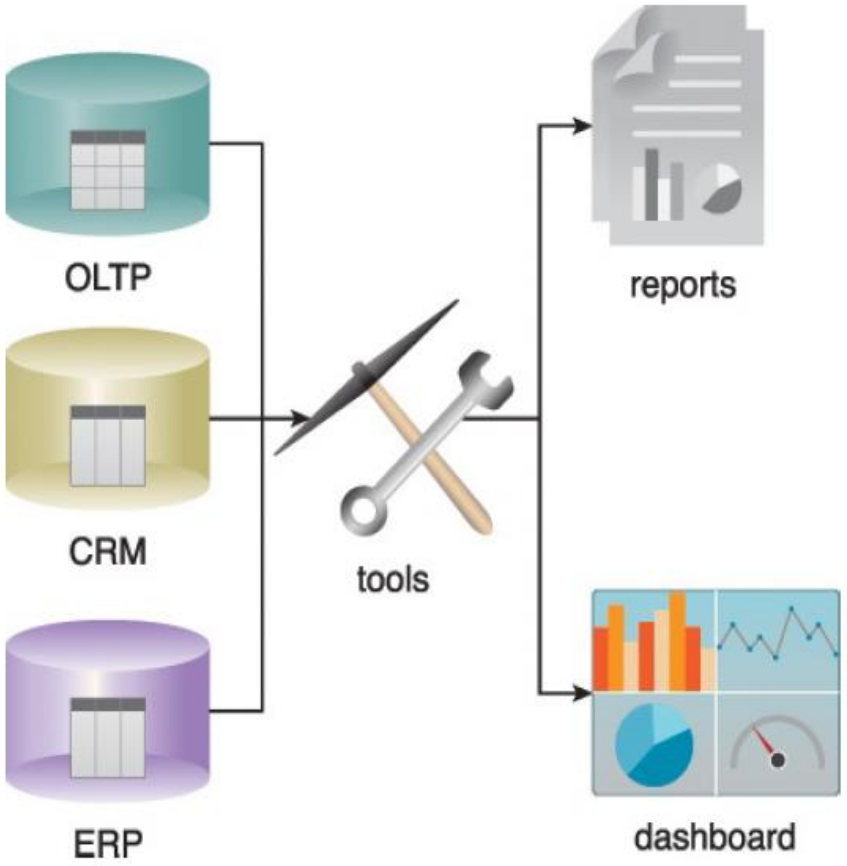
- The different analytics types leverage different techniques and analysis algorithms. This implies that there may be varying data, storage and processing requirements to facilitate the delivery of multiple types of analytic results.

Figure depicts the reality that the generation of high value analytic results increases the complexity and cost of the analytic environment.



Concepts and Terminology - Descriptive Analytics
What has happened?

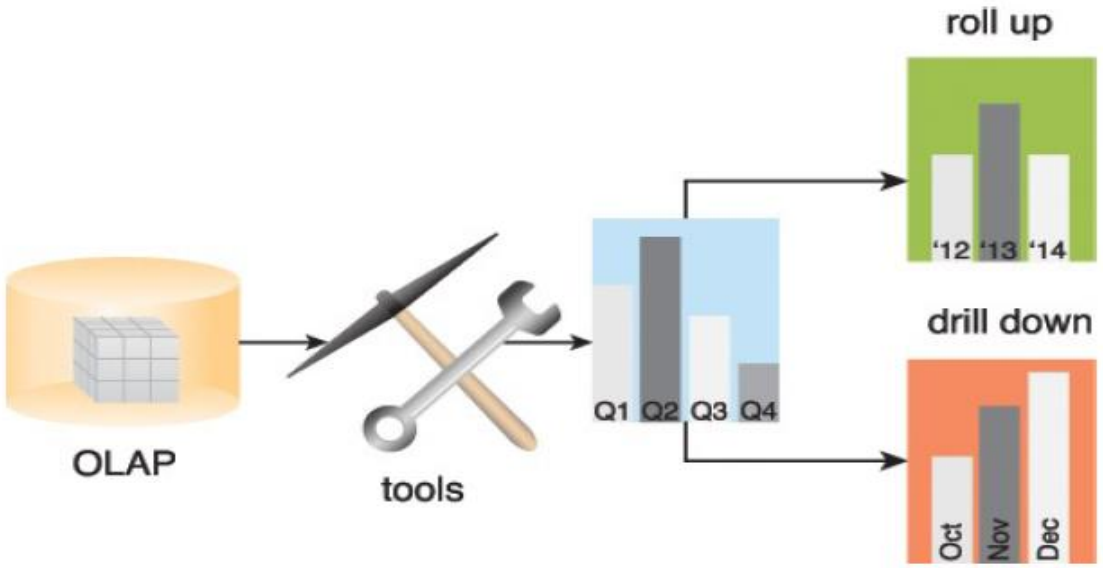
- Descriptive analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.
  - What was the sales volume over the past 12 months?
  - What is the number of support calls received as categorized by severity and geographic location?
  - What is the monthly commission earned by each sales agent?
- The reports are generally static in nature and display historical data that is presented in the form of data grids or charts. Queries are executed on operational data stores from within an enterprise, for example a Customer Relationship Management system (CRM) or Enterprise Resource Planning (ERP) system.
- It is estimated that 80% of generated analytics results are descriptive in nature.



Descriptive analytics are often carried out via ad-hoc reporting or dashboards.

Concepts and Terminology - Diagnostic Analytics

Why did it happen?

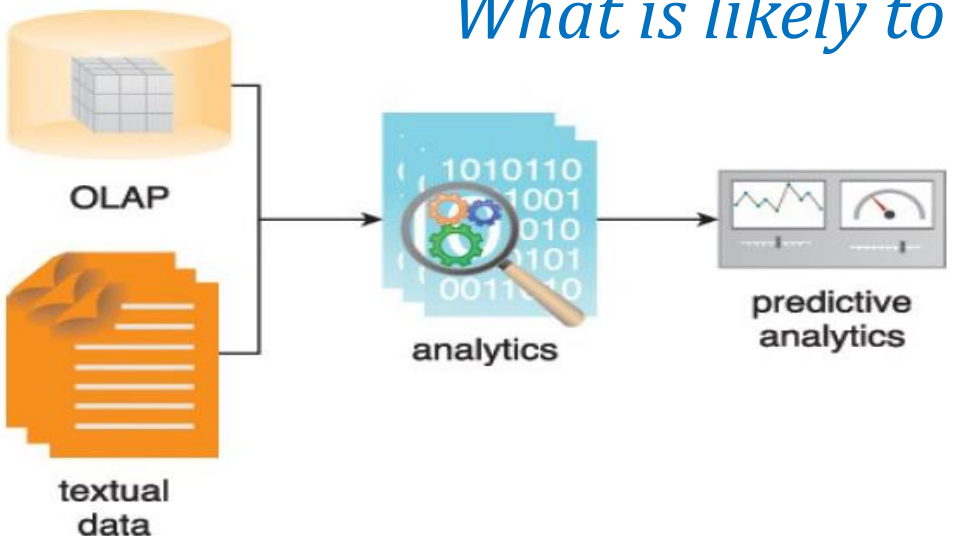


Diagnostic analytics usually require collecting data from multiple sources and storing it in a structure that lends itself to performing drill-down and roll-up analysis

- Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
  - Why were Q2 sales less than Q1 sales?
  - Why have there been more support calls originating from the Eastern region than from the Western region?
  - Why was there an increase in patient re-admission rates over the past three months?
- Diagnostic analytics provide more value than descriptive analytics but require a more advanced skillset.

Concepts and Terminology - Predictive Analytics

What is likely to happen?

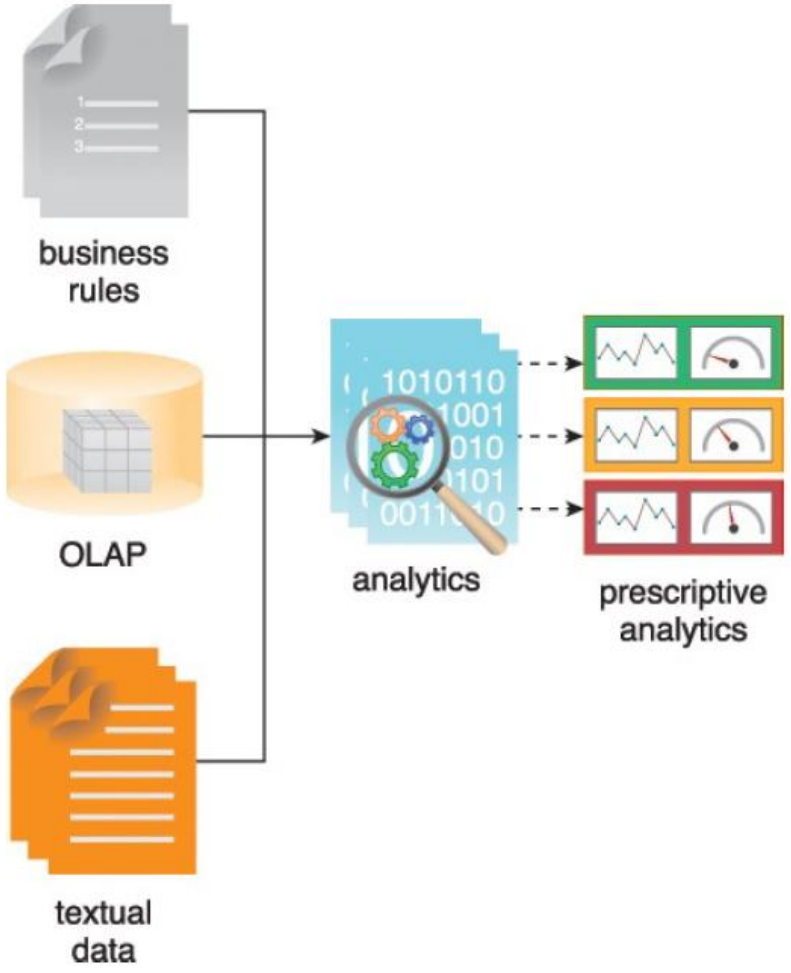


Predictive analytics tools can provide user-friendly front-end interfaces.

- Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future.
  - What are the chances that a customer will default on a loan if they have missed a monthly payment?
  - What will be the patient survival rate if Drug B is administered instead of Drug A?
  - If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?
- Predictive analytics try to predict the outcomes of events, and predictions are made based on patterns, trends and exceptions found in historical and current data.
- The typical approach adopted while developing prediction models is to divide the existing data into training and test data sets (for example 75% of the data is used for training and 25% data is used for testing the prediction model).

Concepts and Terminology - Prescriptive Analytics
What can we do to make it happen?

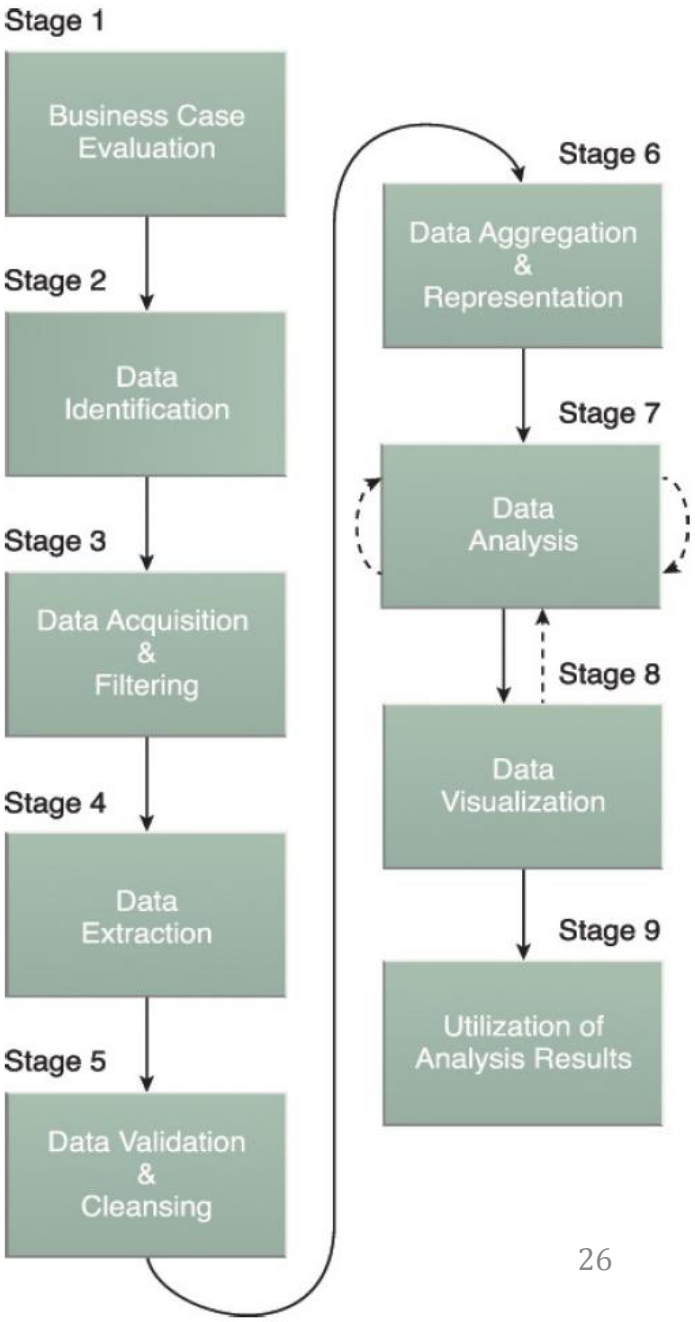
- While predictive analytics uses prediction models to predict the likely outcome of an event, prescriptive analytics uses multiple prediction models to predict various outcomes and the best course of action for each outcome.
  - For example, prescriptive analytics
  - can be used to prescribe the best medicine for treatment of a patient based on the outcomes of various medicines for similar patients.
  - Another example of prescriptive analytics would be to suggest the best mobile data plan for a customer based on the customer's browsing patterns.



Prescriptive analytics involves the use of business rules and internal and/or external data to perform an in-depth analysis.

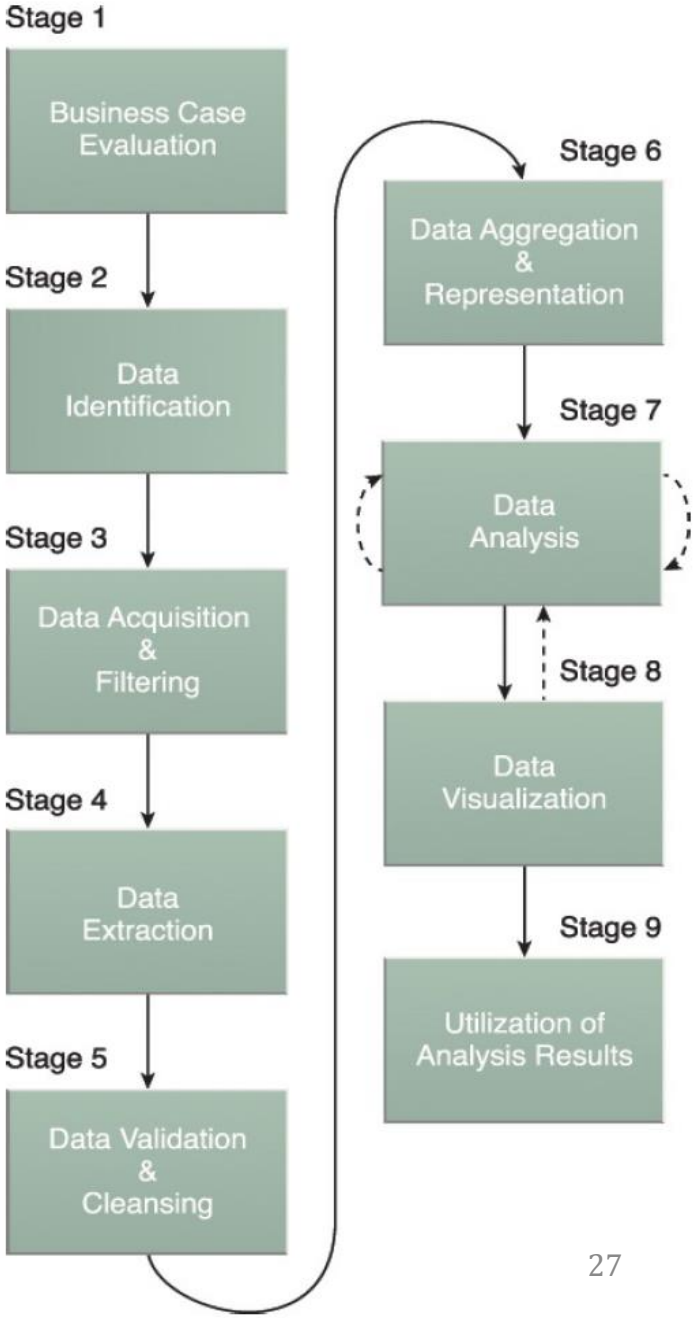
Big Data Analytics Lifecycle

- Each Big Data analytics lifecycle must begin with a well-defined **Business Case Evaluation** that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.
- The **Data Identification** stage is dedicated to identifying the datasets required for the analysis project and their sources. The sources can be internal (sales) and/or external (from blogs).
- 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 may subjected to automated filtering for the removal of corrupt data (missing or invalid or nonsensical values) or data that has been deemed to have no value to the analysis objectives.
  - It is advisable to store a verbatim (raw) copy of the original dataset before proceeding with the filtering.
- Metadata can be added via automation to data from both internal and external data sources to improve the classification and querying.



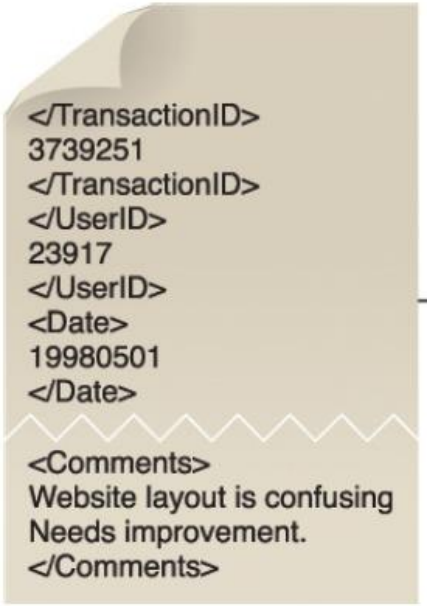
Big Data Analytics Lifecycle - Data Extraction

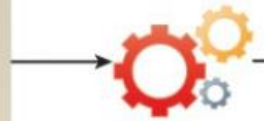
- Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution (especially from external source).
- The extent of **Data Extraction** and Transformation (Data Wrangling / Munging or ETL - Extract Transform Load or Pre-processing) required depends on the types of analytics and capabilities of the Big Data solution.
- Extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.



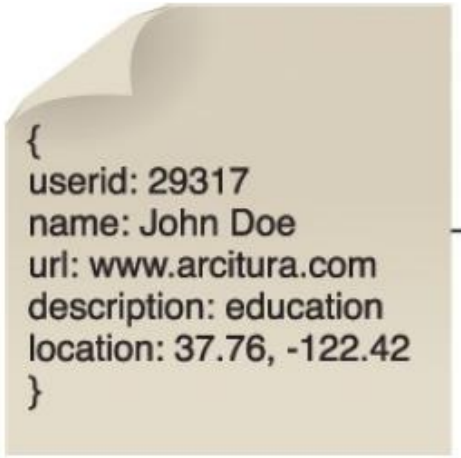
Big Data Analytics Lifecycle - Data Extraction (cont.)

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





User ID	Comments
23917	Website layout is confusing Needs improvement.

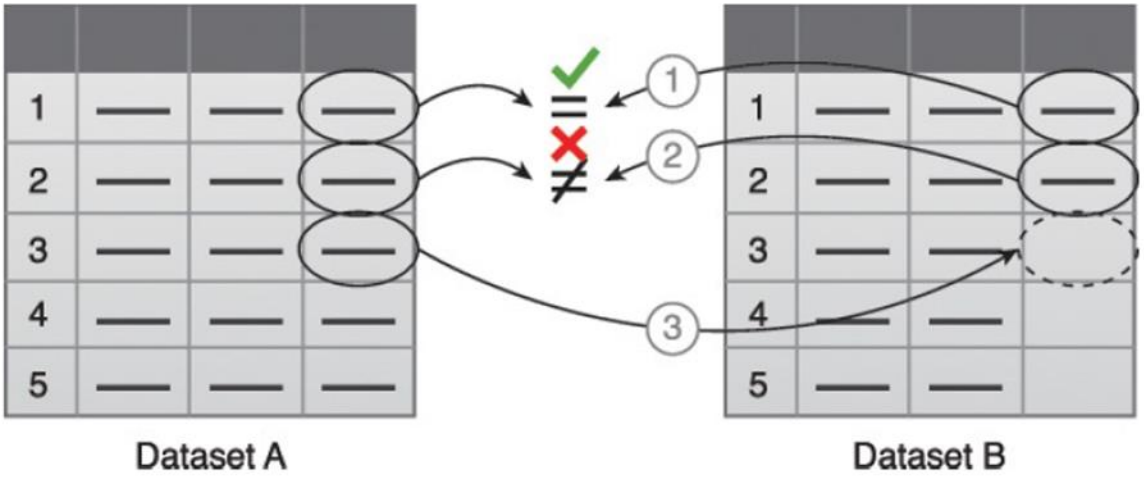


User ID	Latitude	Longitude
23917	37.75	-122.42

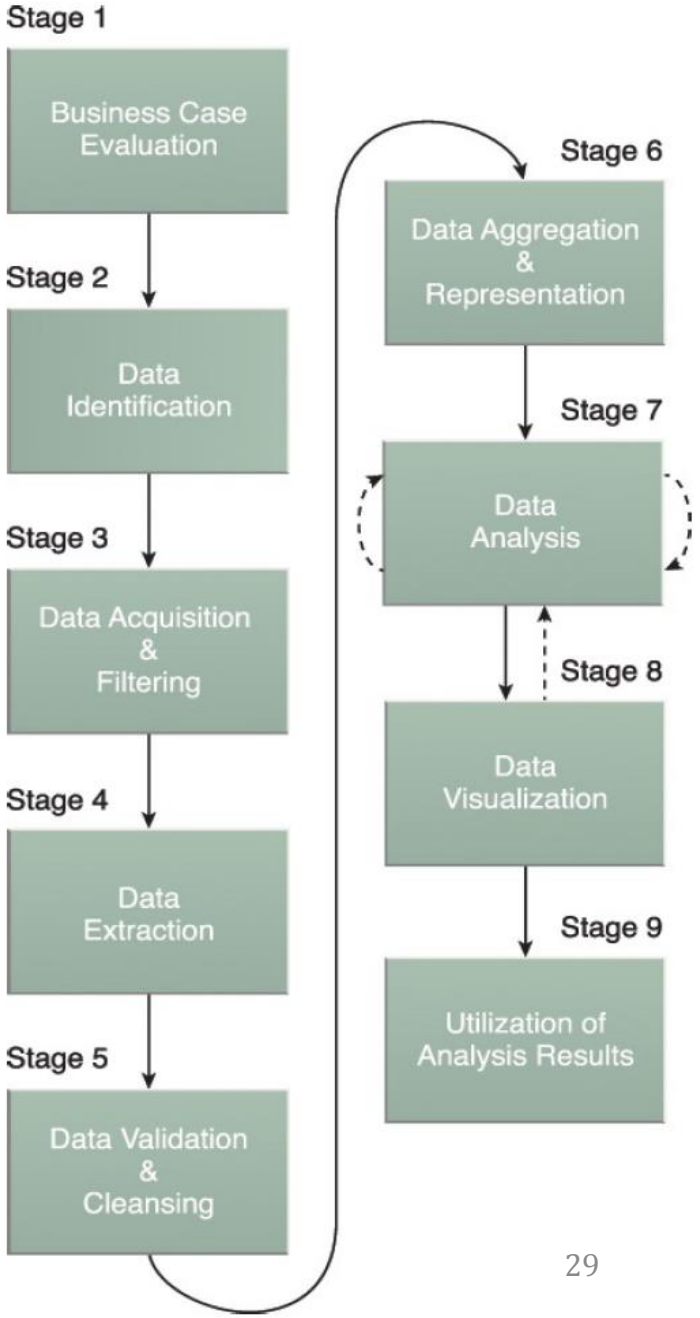
Demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

Big Data Analytics Lifecycle - **Data Validation and Cleansing**

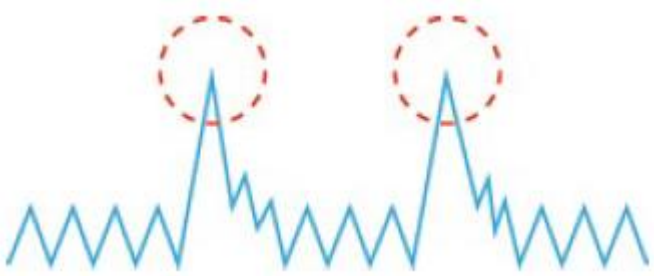
- Unlike RDBMS, where the data structure is pre-defined and data is pre-validated, Big Data solutions often receive redundant data across different datasets.
- The first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against 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.



Big Data Analytics Lifecycle - Data Validation and Cleansing (cont.)

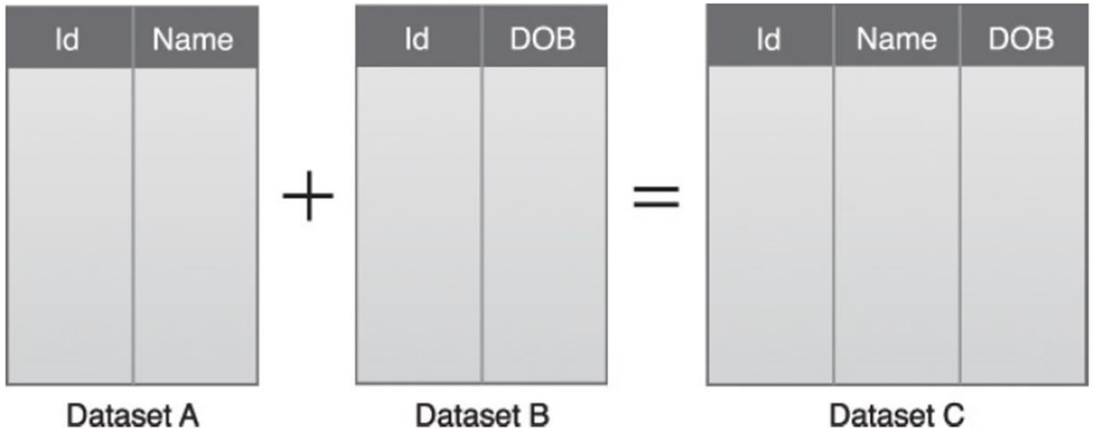


The presence of invalid data is resulting in spikes.

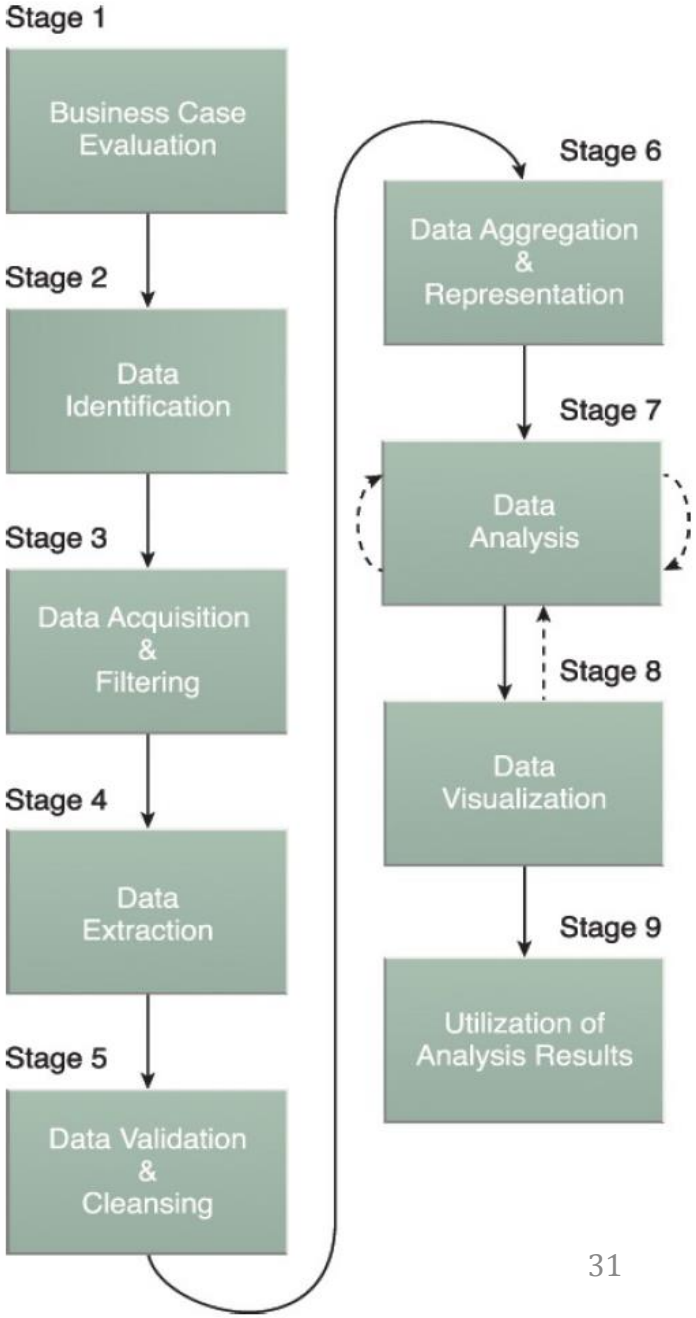
Although the data appears abnormal, it may be indicative of a new pattern.

Big Data Analytics Lifecycle - **Data Aggregation and Representation**

- Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID.
- The large volumes processed by Big Data solutions can make data aggregation a time and effort-intensive operation.

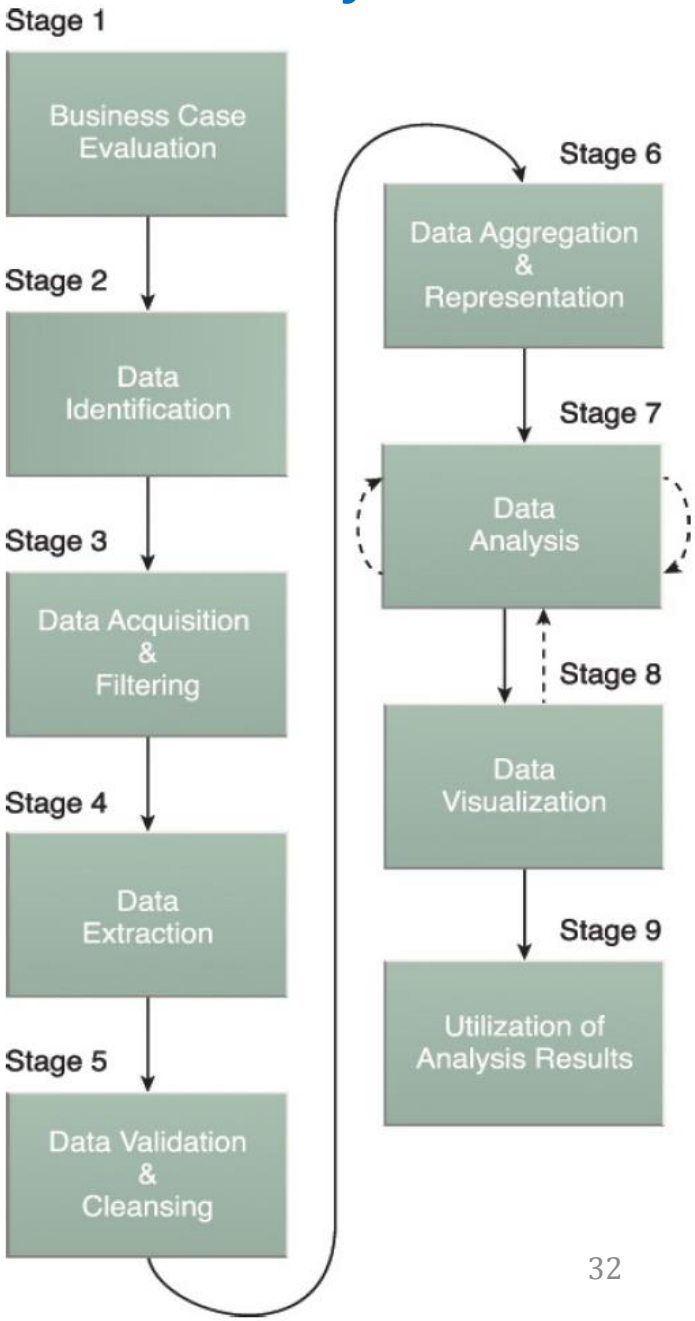


Simple example of data aggregation where two datasets are aggregated together using the Id field.



Big Data Analytics Lifecycle - Data Analysis

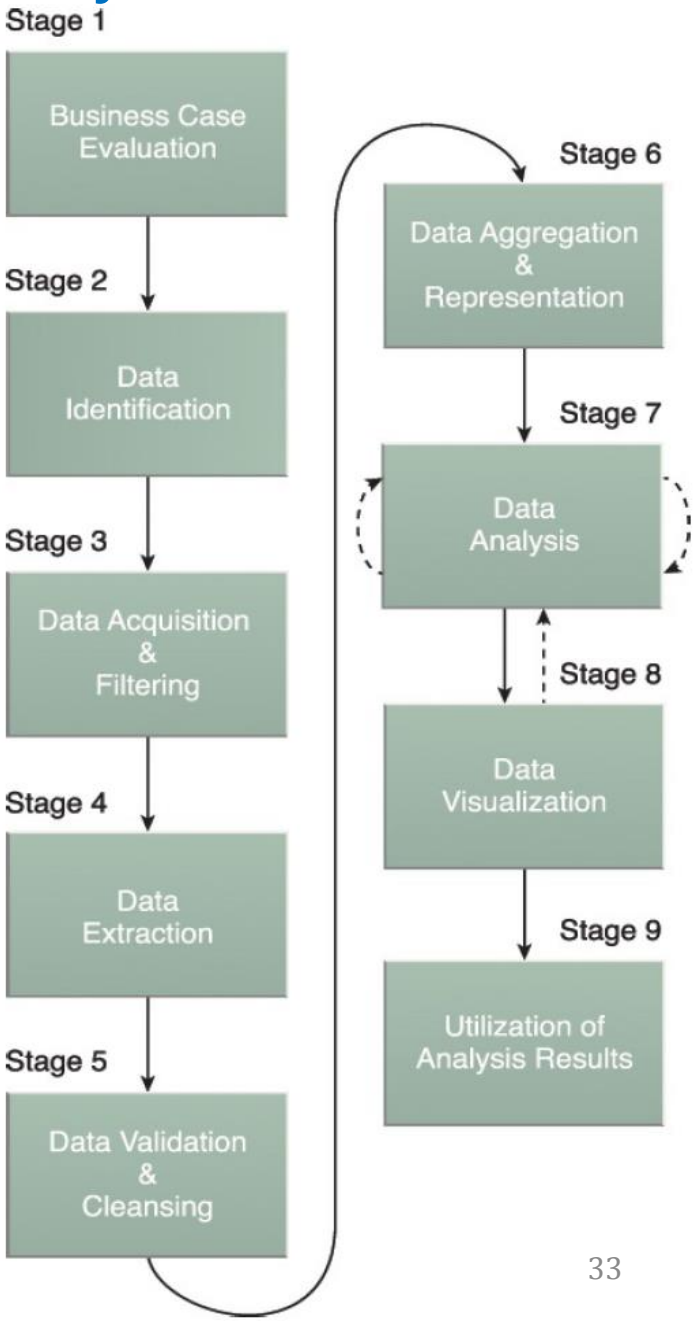
- The **Data Analysis** stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics.
- This stage can be iterative in nature, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
- This stage can be as challenging 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.

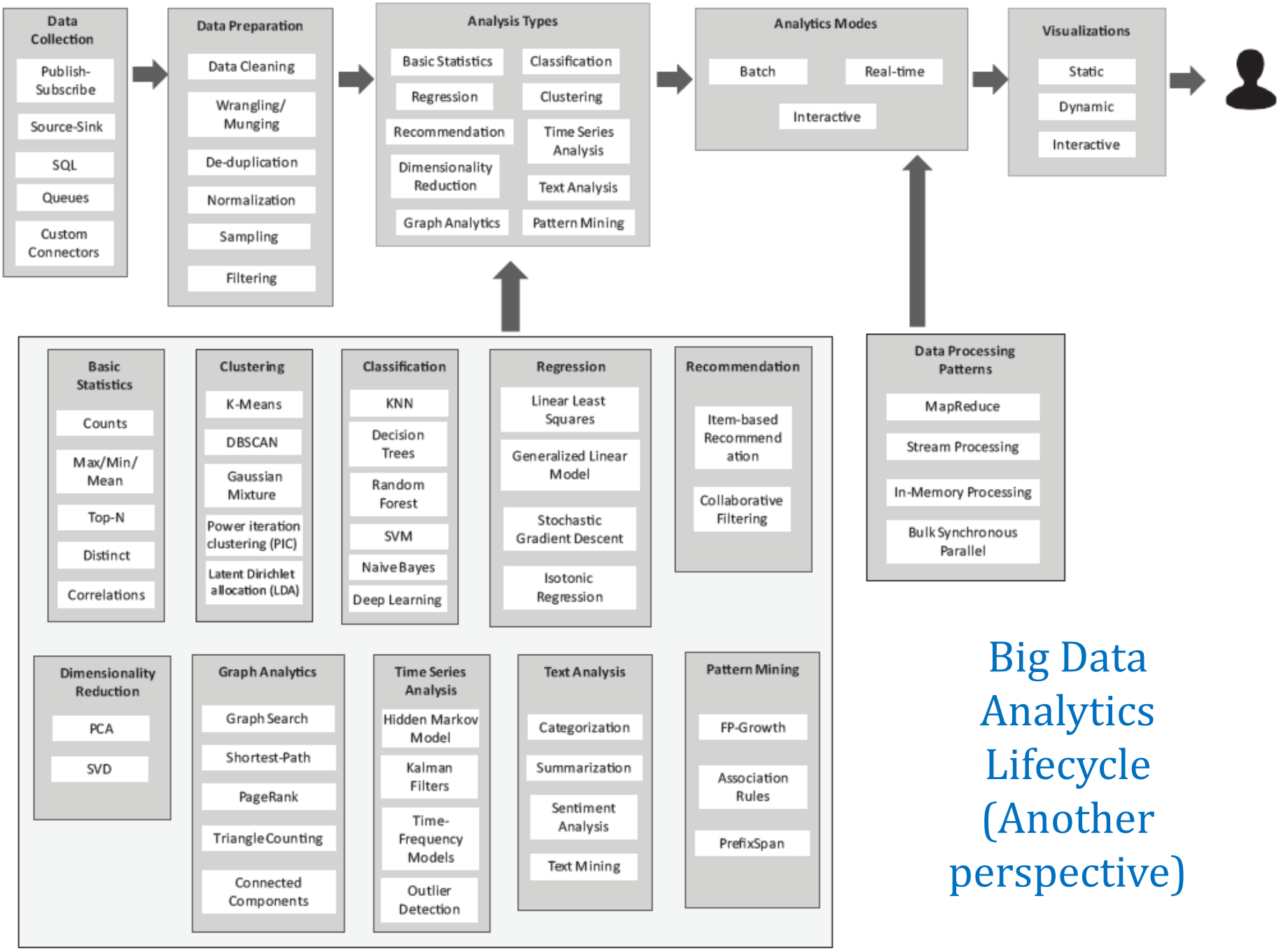


Big Data Analytics Lifecycle

- 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.
- 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, leading from stage 8 back to stage 7.

- The **Utilization of Analysis Results** stage is dedicated to determining how and where processed analysis data can be further leveraged.





Big Data Analytics Lifecycle (Another perspective)