## LESSON 10

**Big data** - extremely large and complex data sets that are generated from a variety of sources, including social media, internet searches, online transactions, sensors, and more.

**Characteristics of Big Data**

1. **Volume** - refers to the vast amount of data that is generated and collected. Big Data sets are typically so large that traditional data storage and processing technologies are inadequate to handle them.
2. **Velocity** - refers to the speed at which data is generated, collected, and processed. In Big Data environments, data can arrive at fast speeds, and enormous datasets can accumulate within a short time.
3. **Variety** - refers to the diversity of data types. An organization might obtain data from a number of different data sources
4. **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.
5. **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.

**Human-generated data** - the result of human interaction with systems, such as online services and digital devices(e.g. social media, blog posts, emails, photo sharing and messaging)

**Machine-generated data** is generated by software programs and hardware devices in response to real-world events (e.g. weblogs, sensor data, telemetry data, smart meter data, and appliance usage data)

**Structured data** refers to data that has a well-defined format or schema, making it easy to search, analyze, and process using automated tools.

**Unstructured data** refers to data that does not have a pre-defined format or structure, making it difficult to organize and process using automated tools, because it

can be in different forms such as text, image, audio, videos, etc.

**Semi-structured data** - refers to data that has some structure, but not enough to be classified as fully structured data.

## LESSON 11

**Provenance** - refers to information about the source of the data and how it has been processed.

- helps determine the authenticity and quality of data, and it can be used for auditing purposes.

**Methodology** - required to control how data flows into and out of Big Data solutions.

**Challenges of big data storage**

1. **Storage** - Unstructured data cannot be stored in traditional databases.
2. **Processing** - refers to the reading, transforming, extraction, and formatting of useful information from raw information.
3. **Security** - Non-encrypted information is at risk of theft or damage by cyber-criminals.

**Big Data Analytics Life Cycle**

1. **Business Case/Problem Definition** -problem is identified, and assumptions are made that how much potential gain a company will make after carrying out the analysis.
2. **Data Identification** - identifying the datasets required for the analysis project and their sources.
3. **Data Acquisition and Filtration** - data is gathered from all of the data sources that were identified during the previous stage, then it is subjected to automated filtering for the removal of corrupt data or data w/ no value
4. **Data Extraction** - extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.
5. **Data Validation and Cleaning** - dedicated to establishing often complex validation rules and removing any known invalid data.
6. **Data Aggregation & Representation** - a method of data reconciliation is required or the dataset representing the correct value needs to be determined.

7. **Data Analysis** - dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics.
8. **Data Visualization** - dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation
9. **Utilization of Analysis Results** - The results can be used for optimization, to refine the business process. It can also be used as an input for the systems to enhance performance.

## LESSON 12

### Big Data Components

Data must first be **ingested from sources**, **translated**, and **stored**, then **analyzed** before the final **presentation**

1. **Source/Ingestion**
    - the very first step in pulling in raw data. It comes from internal sources, relational databases, non-relational databases, and others.
    - **Batch** - large groups of data are gathered and delivered together. Conditions, schedule launches, or ad hoc can trigger data collection.
    - **Streaming** - continuous flow of data; necessary for real-time analytics. Constantly monitors and pulls data as generated, requiring more resources

**ETL (Extract, transform, load)**

- data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.
- simplifies the ingestion and storage of large amounts of data for analysis
- used primarily for the structured type of data
- **Extract** - getting data from sources
- **Transform** - cleaning and formatting data for analysis preparation
- **Load** - moving the transformed data into the target data warehouse.

2. **Storage** - loading activity of ETL Process
    - **Data lake**- homogenous pool of uniformly organized data.

3. **Analysis** - data gets passed through several tools, shaping it into actionable Insights. It uses the 4 types of analytics: Diagnostic, Descriptive, Predictive, Prescriptive
4. **Consumption**-The final big data component involves presenting the information in a format digestible to the end user.

**Big Data Analytics Tools**

1. **APACHE Hadoop**
    - Open-source, Highly Secured, Distributed data processing software
    - Used in batch processing, though very complicated and requires expertise

2. **Spark** - Open-source cluster computing framework
    - Process real-time data, from real time events like Twitter and Facebook
    - Easier to use, but less secure than Hadoop
3. **Tableau** - Data visualization and BI software
    - Excels in self-service visual analysis with a friendly user interface
    - Commercial, subscription-based, and has a tentative security

4. **MySQL** - open-source analytic engine
    - Primarily used in Data storage, retrieval, and management. It's security features data encryption, SSH, and SSL support
    - Knowledge of query processing is required.

## Lesson 13

**Data warehouse** - a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.

- solely intended to perform queries and analysis and often contain large amounts of historical data.
- Contains a relational database, an extraction, loading, and transformation (ELT) solution, Statistical analysis, Client analysis tools, and other more sophisticated analytical applications.

When to Use Big Data Storage?

- Large volume of data – big data storage can deliver the highly required processing performance and scalability when dealing with extremely huge amounts of data.
- Multiple data sources – Big data storage helps companies in collecting classifying, formatting, and processing various forms of data coming from multiple sources
- Real-time processing – Executives and decision-makers in the business will be able to give timely conclusions on the information brought by analyzing the big data as a big data storage process and analyzing data in real time.
- Predictive analytics – big data storage supports models of predictive analytics for better extraction of insights
- Economical storage–great for business is looking for cost-effective storage solutions
- High-availability – The data stored from this storage are always highly available against downtimes.
- Machine-learning – big data storage supports machine-learning algorithms for accurate information generation of large data sets.

**Cluster** - a tightly coupled collection of servers or nodes.

- Each node in the cluster has its own dedicated resources, such as memory, a processor, and a hard drive.
- can execute a task by splitting it into small pieces and distributing their execution onto different computers that belong to the cluster

A **file system** is a method of storing and organizing data on a storage device, such as flash drives, DVDs, and hard drives.

A **distributed file system** is a file system that can store large files spread across the nodes of a cluster.

- Examples include the Google File System (GFS) and Hadoop Distributed File System (HDFS).

Not-only SQL (**NoSQL**) database

- a non-relational database that is highly scalable, fault-tolerant, and specifically designed to house semi-structured and unstructured data.
- Supports query languages other than Structured Query Language (SQL) because SQL was designed to query structured data stored within a relational database.

**Sharding** is the process of horizontally partitioning a large dataset into a collection of smaller, more manageable datasets called shards.

- allows the distribution of processing loads across multiple nodes to achieve horizontal scalability.

**Shards** - distributed across multiple nodes, where a node is a server or a machine.

- Each shard is stored on a separate node and each node is responsible for only the data stored on it.

**Replication** - stores multiple copies of a dataset, known as replicas, on multiple nodes. It ensures that data isn't lost when an individual node fails.

## LESSON 15

**Big Data Processing** - is the collection of methodologies or frameworks enabling access to enormous amounts of information and extracting meaningful insights.

**Parallel data processing** - involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

- reduce the execution time by dividing a single larger task into multiple smaller tasks that run concurrently.
- One advantage of parallel processing is its ability to **scale horizontally.** This means that it can add more computing resources to increase processing power.

**Distributed Data Processing** - achieved through physically separate machines that are networked together as a cluster.

- closely related to parallel data processing in that the same principle of "divide-and-conquer" is applied.
- has several advantages, including the ability to process large datasets in parallel, which can improve performance and reduce processing time.

**Processing workload** - defined as the amount and nature of data that is processed within a certain amount of time.

**Batch processing / offline processing** - involves processing data in batches and usually imposes delays, which in turn results in high-latency responses.

- involve large quantities of data with sequential read/writes and comprise groups of read or write queries.

**Transactional processing/online processing** - follows an approach whereby data is processed interactively without delay, resulting in low-latency responses.

## LESSON 16

### Big Data Analysis Techniques

1. **A/B testing** - also known as split or bucket testing, compares two versions of an element to determine which version is superior based on a pre-defined metric.
    - The current version of the element is called the **control version**, whereas the modified version is called the **treatment**. Both versions are subjected to an experiment simultaneously.

2. **Data mining** extracts patterns from large data sets by combining methods from statistics and machine learning, within database management.
    - can be used to predict future illnesses or outbreaks based on public health data.

3. **Machine learning** - similar methods to data mining , but involves the application of algorithms or a set of

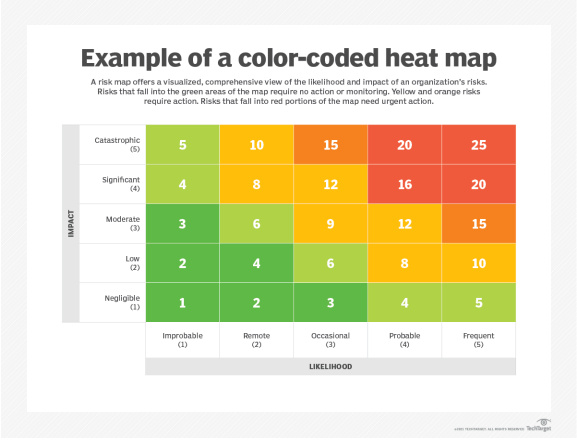
instructions to automate the process.

- it works with computer algorithms to produce assumptions based on data.

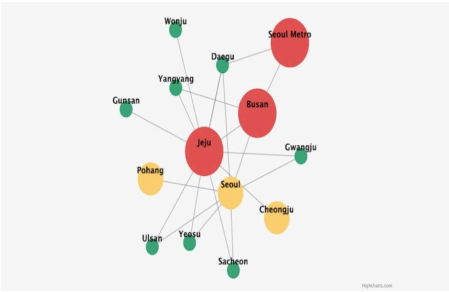
4. **Visual analysis** - form of data analysis that involves the graphic representation of data to enable or enhance its visual perception.

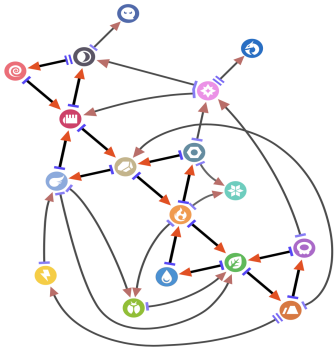
## Types of Visual Analysis

1. **Heat maps** - a visual, color-coded representation of data values. Each value is given a color according to its type or the range that it falls under.




2. **Network graph**- depicts an interconnected collection of entities.
   - It involves plotting entities as nodes and connections as edges between nodes.
   - **Entity** - can be a person, a group, or some other business domain object such as a product. Entities may be connected with one another directly or indirectly.



Graphic by /u/dazmax
Icons by Andreuvv - CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?oldid=41847398

3. **Spatial or geospatial data** - commonly used to identify the geographic location of individual entities that can then be mapped.
   - manipulated through a Geographic Information

System (GIS) that plots spatial data on a map generally using its longitude and latitude coordinates.