Introduction to Big Data Analytics

Notes by Mannan Ul Haq (BDS-4A)

Data vs. Information:

Data is like building blocks – numbers, words, raw facts and figures. When we process data and put it together in a smart way, we get information.

Big Data:

Big Data is like a huge amount of information. It's so big that regular tools (hardware devices) can't handle it. We need special tools and tricks to make sense of this massive amount of data.

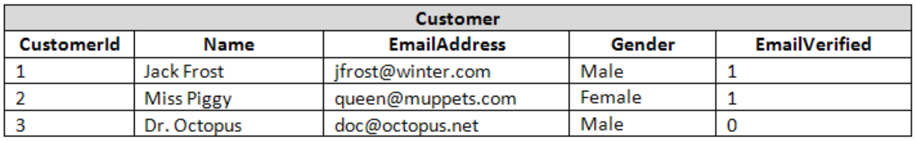
Five Characteristics of Big Data:

- **Volume:** How much data we have collected.
- **Velocity:** How fast data is coming at high speed.
- **Variety:** Different types of data (like numbers, words, pictures, voice-records).
- **Veracity:** How much we can trust the data (accuracy, precision, integrity, reliability).
- **Value:** How useful the data is to make useful decisions.

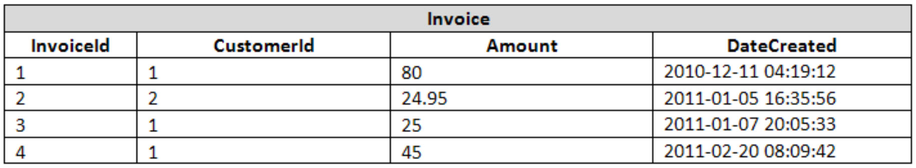
Types of Data (Based on their Structure):

Structured Data:

Structured data is like information clearly arranged in rows and columns, just like a spreadsheet. It's highly organized and follows a fixed format. Examples of structured data include databases, spreadsheets, and tables.



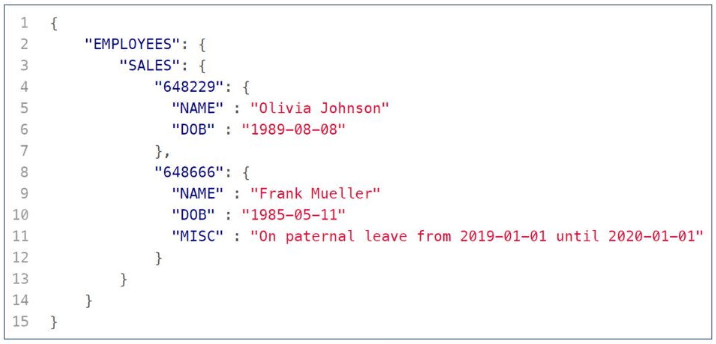
Customer				
CustomerId	**Name**	**EmailAddress**	**Gender**	**EmailVerified**
1	Jack Frost	jfrost@winter.com	Male	1
2	Miss Piggy	queen@muppets.com	Female	1
3	Dr. Octopus	doc@octopus.net	Male	0



Invoice			
InvoiceId	**CustomerId**	**Amount**	**DateCreated**
1	1	80	2010-12-11 04:19:12
2	2	24.95	2011-01-05 16:35:56
3	1	25	2011-01-07 20:05:33
4	1	45	2011-02-20 08:09:42

Semi-Structured Data:

Semi-structured data is a bit like a mix between structured and unstructured data. It doesn't fit neatly into rows and columns, but it has some level of structure. Think of it as a collection of documents where each document might have a title, author, and date, but the content itself might not follow a strict structure. Examples of semi-structured data include XML files and JSON files.



```
{
    "EMPLOYEES": {
        "SALES": {
            "648229": {
              "NAME" : "Olivia Johnson"
              "DOB" : "1989-08-08"
            },
            "648666": {
              "NAME" : "Frank Mueller"
              "DOB" : "1985-05-11"
              "MISC" : "On paternal leave from 2019-01-01 until 2020-01-01"
            }
        }
    }
}
```

JSON

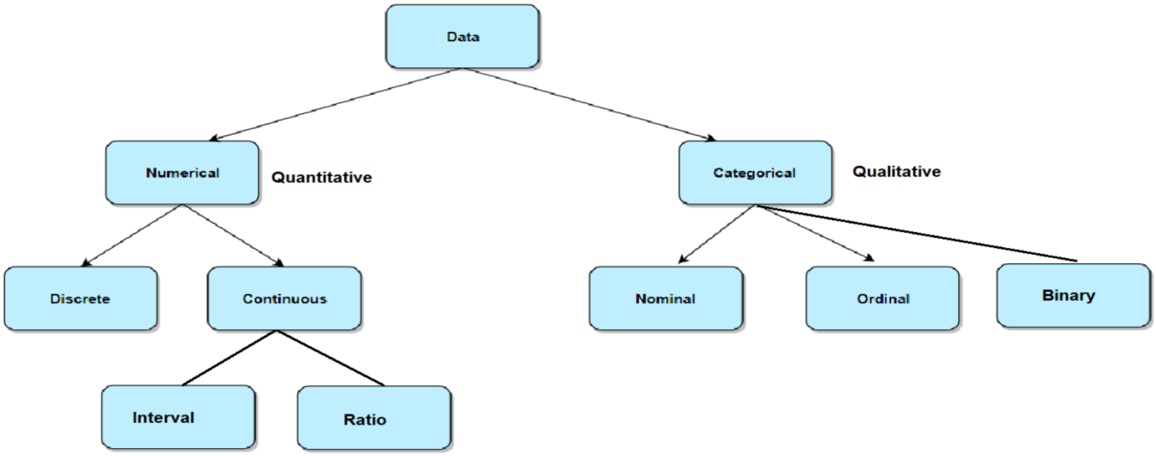
Unstructured Data:

Unstructured data is like a bunch of information without a specific order. It's more like the freeform text you find in a book or a social media post.

Examples:

- Images
- Videos
- Speeches

Data Types:



Categorical Data:

Categorical data represents categories or characteristics like gender, language, or movie genre. It's also called qualitative data. You can use numbers for them, but those numbers have no real math value (like 0/1 for male/female).

Types:

1. **Nominal Data**:
    - No order.
    - Examples: gender, language, eye color.
    - Analyze with frequencies, pie charts, etc.
2. **Ordinal Data**:
    - Has order.
    - Examples: happiness level, education level, movie ratings.
    - Summarize with median, mean, visualize with bar charts.
3. **Binary Data**:
    - Just two values: yes or no.
    - Represented as "True" and "False" or 1 and 0.

Numerical Data

Numerical data is expressed as numbers, allowing quantification. It represents values like integers or real numbers. It is also called quantitative data. Examples include a person's height, product prices, IQ scores, the number of lessons in a course, etc.

Types:

1. Discrete Data

- Has a finite or countably infinite set of values.
- Values are distinct and separate.
- Examples: zip codes, words in a document collection, number of coin toss heads, students in a classroom, cars in a showroom.
- Often represented as integer variables.
- Analyzed using mean, median, quartiles, box plots, and histograms.

2. Continuous Data

- Cannot be counted but can be measured.
- Represents measurements.
- Examples: market share price, height/weight of a person, amount of rainfall, car speed, Wi-Fi frequency.
- Can be divided into meaningful parts.
- Has real numbers as attribute values.

Types of Continuous Data:

a. Interval Data

- Categorized, ranked, and evenly spaced.
- Values have order and can be positive, zero, or negative.
- Allows comparison and quantification of differences.
- Examples: temperatures in Celsius or Fahrenheit, calendar dates.

b. Ratio Data

- Numeric attribute with an inherent zero-point.
- Values can be multiples or ratios of one another.
- Ordered values with computed differences.
- Examples: Kelvin temperature scale, years of experience, number of words.

Sources of Data:

1. **Machine-Generated Data:**
   - **Sensor Data:** Generated by various sensors, such as IoT devices, industrial sensors, and smart devices.
   - **Log Data:** Generated by software and systems, including server logs, application logs, and network logs.
2. **Human-Generated Data:**
   - **Social Media Data:** Data from platforms like Facebook, Twitter, and Instagram, including posts, comments, and user interactions.
   - **User-Generated Content:** Content created by users, such as reviews, comments, and forum posts.
3. **Business Generated Data:**
   - Business-Generated Data refers to data that is created, collected, or generated as a result of day-to-day business operations and activities. This may include various types of data that are essential for the functioning and decision-making within a business.

Big Data Analytics:

Analysis vs. Analytics:

Analysis: Making use of past data and deriving results from the data

Analytics: Using data to obtain future insights and details regarding trends etc..

Definition of Data Analytics:

The process of examining data in order to draw and communicate useful conclusions about the information it contains.

Tools that used in Big Data Analytics:

- Hadoop
- spark
- mongoDB
- talend
- kalfa
- Storm
- cassandr