Skills for Hire
Data Analytics
Week 1 – Data Analytics 101

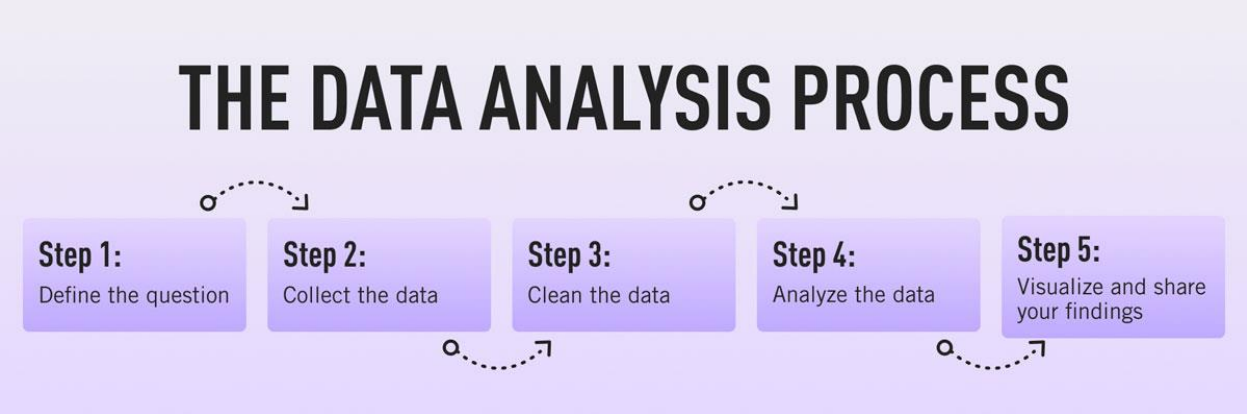
What is Data Analytics?

Data analytics is the process of analyzing raw data in order to draw out meaningful, actionable insights, which are then used to inform and drive smart business decisions.

Types of Data Analytics

- Descriptive analysis purely describes what has happened and presents it in a digestible snapshot.
- Diagnostic analysis seeks to establish why those things may have happened.
- Predictive analysis makes use of past patterns and trends in data to estimate the likelihood of a future outcome or event.
- Prescriptive analysis is the conclusion of the other forms of analysis: now that we've found out what happened, why it happened, and what may happen in the future, what should be done next?

Data Analytics Process



The steps are as follows:

- The data analyst will first need to define their objective, otherwise known as a 'problem statement.'
- Once the analyst has established their objective for the analysis, they'll need to design a strategy for collecting the appropriate data. Firstly, they'll need to determine what kind of data they'll need: quantitative (numeric) data such as sales figures, or qualitative (descriptive) data, which may include customer surveys.
- The data analyst will need to clean the data to make sure it's of high quality. This cleaning—or "scrubbing"—process involves:
    - Removing unwanted data points

    - Removing major errors, duplicates, and outliers
    - Filling in any missing data
    - Bringing structure to the data
- The data analyst will apply the methodologies associated with the analysis type that will best "solve" their problem statement.
- The data analyst must now present their findings in a way that's clear and easily understood by key stakeholders. In order to do this, an analyst may use visualization software—such as Tableau or Microsoft Power BI—that will generate reports, dashboards, or interactive visualizations.

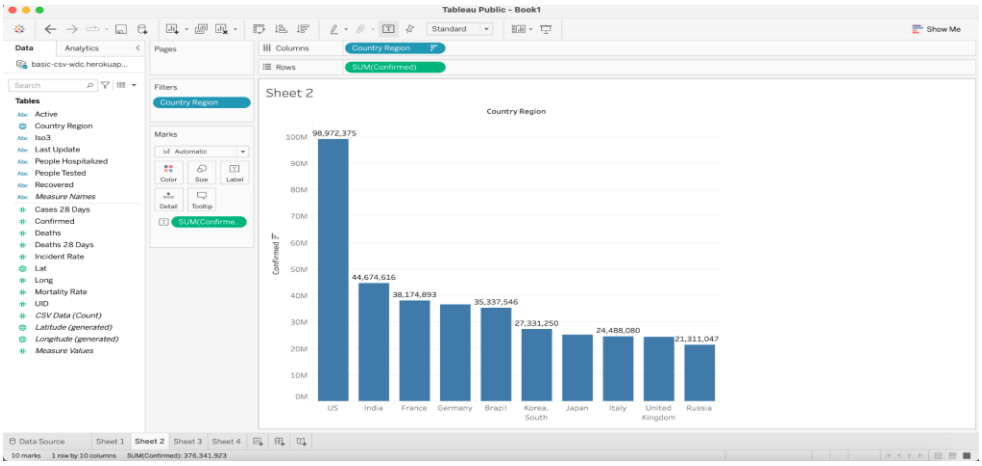
Basic Data Analytics Tools

Excel

- Excel is a spreadsheet and a simple yet powerful tool for data collection and analysis.
- Excel is not free; it is a part of the Microsoft Office "suite" of programs.
- Excel does not need a UI to enter data; you can start right away.
- It is readily available, widely used and easy to learn and start on data analysis
- The Data Analysis Toolpak in Excel offers a variety of options to perform statistical analysis of your data. The charts and graphs in Excel give a clear interpretation and visualization of your
- Data, which helps in decision-making as they are easy to understand.
- Demo – Basic Excel Functions. Live Demo)
    - Arithmetic Operations (+, -, *,/)
    - Sorting and Filtering
    - VLOOKUP and IF function-
    - Visualization - Charts.
    - Pivot Table - A Pivot Table is a summary of a large dataset that usually includes the total figures, average, minimum, maximum, etc. let's say you have sales data for different regions, with a pivot table, you can summarize the data by region and find the average sales per region, the maximum and minimum sale per region, etc. Pivot tables allow us to analyze, summarize and show only relevant data in our reports.

Tableau

- Tableau is a BI (Business Intelligence) tool developed for data analysts where one can visualize, analyze, and understand data.
- Tableau is not free software, and the pricing varies as per different data needs
- Tableau provides fast analytics; it can explore any type of data – spreadsheets, databases, data on Hadoop, and cloud services
- It is easy to use as it has powerful drag-and-drop features that anyone with an intuitive mind can handle.
- The data visualization with smart dashboards can be shared within seconds.



**Optional Training material available on Pluralsight*.

Python

- Python was initially designed as an Object-Oriented Programming language for software and web development and later enhanced for data science. Python is the fastest-growing programming language today.
- It is a powerful Data Analysis tool and has a great set of friendly libraries for any aspect of scientific computing.
- Python is free, open-source software, and it is easy to learn.
- Python's data analysis library Pandas was built over NumPy, which is one of the earliest libraries in Python for data science.

**Python will be covered in detail in the coming weeks*

Databases

- Databases provide a tremendous amount of capacity and flexibility in working with our data far beyond that of a spreadsheet. So when we need to store large quantities of data of we have complex relationships in the data or we just simply need to interact with it in more sophisticated or advanced ways, databases provide all the capabilities we need, whereas a spreadsheet is very limited to what you can do kind of by hand working with it in Excel.

Two major types of databases

  - **Relational:** In relational database, every piece of information has a relationship with every other piece of information. This is on account of every data value in the database having a unique identity in the form of a record.
    - Note that all data is tabulated in this model. Therefore, every row of data in the database is linked with another row using a primary key. Similarly, every table is linked with another table using a foreign key.
    - Refer to the diagram below and notice how the concept of 'Keys' is used to link two tables:

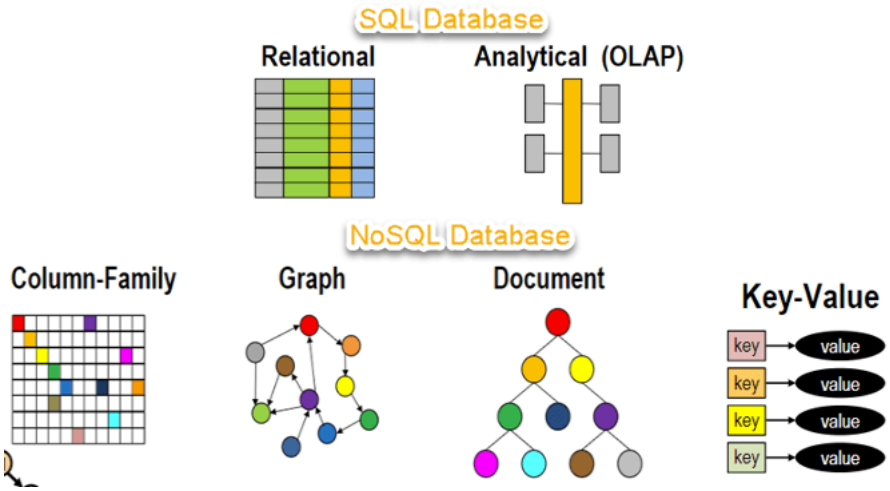
Roll no.	Student Name	Marks Awarded
1	Raman Triphati	86
2	Rajan Govindan	94
3	Mahesh Nandalal	94

Key = 94

Marks Awarded	Student Name	Rank	Scholarship
94	Rajan Govindan	17	Yes
94	Mahesh Nandlal	16	Yes

Section	Student Name	Marks Awarded	Rank
A	Raman Tripathi	86	43
B	Rajan Govindan	94	17
C	Mahesh Nandlal	94	16

- **NoSQL**: NoSQL Database is a non-relational Data Management System, that does not require a fixed schema. It avoids joins, and is easy to scale. Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.



**We will cover SQL in detail in the upcoming weeks*

Data Modelling

Data modeling (data modeling) is the process of creating a data model for the data to be stored in a database. This data model is a conceptual representation of Data objects, the associations between different data objects, and the rules. Data modeling helps in the visual representation of data and enforces business rules, regulatory compliances, and government policies on the data.

Types

- Conceptual Data Model: This Data Model defines WHAT the system contains. This model is typically created by Business stakeholders and Data Architects. The purpose is to organize, scope, and define business concepts and rules.
- Logical Data Model: Defines HOW the system should be implemented regardless of the DBMS. This model is typically created by Data Architects and Business Analysts. The purpose is to develop a technical map of rules and data structures.
- Physical Data Model: This Data Model describes HOW the system will be implemented using a specific DBMS system. This model is typically created by DBA and developers. The purpose is actual implementation of the database.

Key Terms

- Relationship: How tables will relate to each other so that you can do your transactional processing or do your analytics or whatever you're using your database for. The different types of relationships: There's the one-to-one, the one-to-many, and then conversely the many-to-one, and the many-to-many.

- Normalization is used to minimize the redundancy from a relation or set of relations. It is also used to eliminate undesirable characteristics like Insertion, Update, and Deletion Anomalies.
    - Insertion Anomaly: Insertion Anomaly refers to when one cannot insert a new tuple into a relationship due to a lack of data.
    - Deletion Anomaly: The delete anomaly refers to the situation where the deletion of data results in the unintended loss of some other important data.
    - Updatation Anomaly: The update anomaly is when an update of a single data value requires multiple rows of data to be updated.

Data Warehouse

Data Warehousing (DW) is a process for collecting and managing data from varied sources to provide meaningful business insights. It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is a process of transforming data into information and making it available to users in a timely manner to make a difference.

Data may be: Structured Semi-structured and Unstructured data

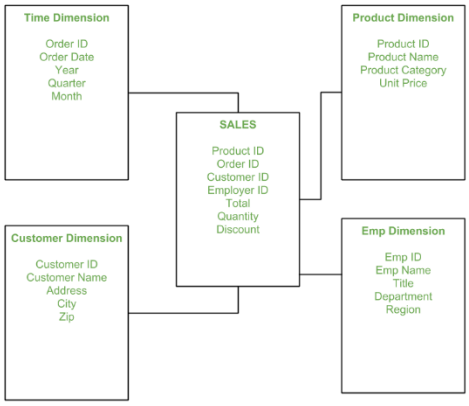
The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.

Star Schema – the foundation of data warehouse Star schema is the fundamental schema among the data mart schema, and it is simplest. This schema is widely used to develop or build a data warehouse

and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables.

It is said to be the star as its physical model resembles the star shape having a fact table at its center and the dimension tables at its periphery representing the star's points. Below is an example to demonstrate the Star Schema:



Big Data

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is data with so large a size and complexity that none of the traditional data management tools can store it or process it efficiently. Big data is also data but with a huge size. For example – Social media 500+terabytes of new data get ingested into the databases of the social media site Facebook, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments, etc.

Big data can be described by the following characteristics:

- Volume refers to the size of the data.
- Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- Velocity refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines the real potential of the data.
- Variability refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.