Gathering Data

Identifying Data for Analysis:

- Process for identifying data includes following steps, -
  (i) Determine the information you want to collect
  (ii) Define a plan for collecting data.
  (iii) Determine data collection methods.
- The data we identify, the source of that data and the practices we employ for gathering the data have implications for- quality, security, privacy. All of these are relevant throughout the life cycle of data analysis process.

Data Sources:

- **Primary Data-** it refers to information obtained directly from the source, for e.g., data from organization's HR, data we gather directly from surveys.
- **Secondary Data-** it refers to information retrieved from existing sources, i.e., external databases, articles, publications, externally conducted surveys etc.
- **Third Party Data-** it refers to data purchased from aggregators who collect data from various sources and combine it into comprehensive datasets for the purpose of selling the data.

How to Gather and Import Data:

- **SQL-** offers simple command to specify what is retrieved from database, table from which it needs to be extracted, grouping records with matching values, dictating sequence in which query results are displayed, limiting the number of results that can be returned by the query.
- **API's-** used for extracting data from a variety of data sources. API's are invoked from applications that require the data and access an endpoint containing the data. It is also used for data validation, e.g., a data analyst may utilize API to validate postal address and zip codes.
- **Extracting Data from Web-** downloading specific data from web pages.
- **Data Exchange-** exchanging data between data providers and consumers. It provides data licensing workflows, de-identification and protection of personal information, legal frameworks and a quarantined analytics environment, e.g., AWS.
- Data that are gathered needs to be loaded or imported to a data repository before it can be wrangled, mined and analyzed.