Different Sources of Data for Data Analysis

Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis. In the process of big data analysis, "Data collection" is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as "knowledge". Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.

Data collection starts with asking some questions such as what type of data is to be collected and what the source of collection is. Most of the data collected are of two types known as "qualitative data" which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is "quantitative data" which is in numerical forms and can be calculated using different scientific tools and sampling data.

The actual data is then further divided mainly into two types known as:

1. **Primary data**
2. **Secondary data**

1. Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:
The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. Observation method:

The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.

4. Experimental method:

The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

- **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
- **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
- **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
- **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

2. Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source:

These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

External source:

The data which can’t be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Other sources:

- **Sensors data:** With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
- **Satellites data:** Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
- **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.