Data Science
MODULE-2
(Data collection and management)

Presented by

Dr. N. Ramana
Associate Professor
Kakatiya University

Data Science
MODULE-2
(Data collection and management)

Agenda

- ✓ Introduction
- ✓ Sources of data
- ✓ Data collection and APIs
- ✓ Exploring and fixing data
- ✓ Data storage and management Using multiple data Sources

Introduction

➢Data collection and management encompasses identifying the data you need, exploring it, and processing it to be suitable for analysis. This stage is often the most time-consuming step in the process. We discuss key points such as:

- ✓What data is available to me?
- ✓Will it help me solve the problem?
- ✓Is it enough?
- ✓Is the data quality good enough?

Sources of Data

- A data source can be a database, a flat file, and real-time measurements from physical equipment, scraped online data, or any of the numerous static and streaming data providers available on the internet.
- Data sources can be internal or external to the company, and they can be primary or secondary, depending on whether we are getting data directly from the source, accessing it from external data sources, or buying it from data aggregators.
- Different Sources of Data Sources classified based on its collection methods are :

1. Internal Sources of Data
2. External Sources of Data

Sources of Data

Different Sources of Data Sources are :

- Internal Sources of Data: data is collected from records, archives, and various other sources within the organization itself. Such sources of data are termed internal sources of data.

Example: A school is performing an analysis to figure out the highest marks achieved in class 8 science subjects for the last 10 years.

- External Sources of Data: Data may also be collected from various sources outside the organization for analytical purposes. Such sources of data collection are known as external sources of data.

Example: As a patient, you are analyzing the price charts of your nearby hospitals for the treatment of ulcers.

Multiple Data Sources

- Databases, the web, social media, interactive platforms, sensor devices, data exchanges, surveys, and observation studies are some of the data sources we may be using.
- Data from diverse data sources is recognized and acquired, then merged using a range of tools and methodologies to create a unified interface for querying and manipulating data.
- The data we identify, the source of that data, and the methods we use to collect it all have quality, security, and privacy concerns that must be considered at this time.

Multiple Data Sources

Databases

- Relational databases, such as SQL Server, Oracle, MySQL, and IBM DB2, are used to store data in an organized manner in these systems. Data from databases and data warehouses can be utilized as an analysis source.
- Data from a retail transaction system, for example, can be used to analyze sales in different regions, while data from a customer relationship management system can be used to forecast sales.
- There are additional publicly and privately available datasets outside of the organization.

Multiple Data Sources

APIs

- APIs, or Application Program Interfaces, and Web Services are provided by many data providers and websites, allowing various users or programmes to communicate with and access data for processing or analysis.
- APIs and Web Services often listen for incoming requests from users or applications, which might be in the form of web requests or network requests, and return data in plain text, XML, HTML, JSON, or media files.
- APIs are used by apps that demand data and access an end-point that contains the data.
- Databases, online services, and data markets are examples of end-points.

Web Scraping

- Web scraping is a technique for obtaining meaningful data from unstructured sources.
- Online scraping, also known as screen scraping, web harvesting, and web data extraction, allows you to retrieve particular data from websites depending on predefined parameters.
- Web scrapers may harvest text, contact information, photos, videos, product items, and other information from a website.
- Web scraping is commonly used for a variety of purposes, including gathering product details from retailers, manufacturers, and eCommerce websites to provide price comparisons, generating sales leads from public data sources, extracting data from posts etc.

Data Streams

- Data streams are another popular method for collecting continuous streams of data from sources such as instruments, IoT devices and apps, GPS data from cars, computer programmes, websites, and social media posts.
- This information is often time stamped and geotagged for geographic identification.
- Stock and market tickers for financial trading, retail transaction streams for projecting demand and supply chain management,
- surveillance and video feeds for danger detection, social media feeds for emotion research, and so on

Data collection

- 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.
- Collected data used in later stages to analyze different types of information using a set of standard validated techniques.
- The main objective of data collection is to gather information-rich and reliable data, and analyze them to make critical business decisions.
- Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses.

- During data collection, the researchers must identify the data types, the sources of data, and what methods are being used.
- Before an analyst begins collecting data, they must answer three questions first:
  - ✓What's the goal or purpose of this research?
  - ✓What kinds of data are they planning on gathering?
  - ✓What methods and procedures will be used to collect, store, and process the information?
- There are two main methods of data collection in research based on the information:
  1. Primary Data Collection
  2. Secondary Data Collection

Primary Data Collection Methods

- Primary data refers to data collected directly from the main source. It refers to data that has never been used in the past. The data gathered by primary data collection methods are generally regarded as the best kind of data in research.
- Here are some of the most common primary data collection methods:
  - ✓ Interviews
  - ✓ Observations
  - ✓ Surveys and Questionnaires
  - ✓ Focus Groups
  - ✓ Oral Histories

Interviews

- Interviews are a direct method of data collection. It is simply a process in which the interviewer asks questions and the interviewee responds to them.
- It provides a high degree of flexibility because questions can be adjusted and changed anytime according to the situation.

Observations

- In this method, researchers observe a situation around them and record the findings.
- It can be used to evaluate the behaviour of different people in controlled (everyone knows they are being observed) and uncontrolled (no one knows they are being observed) situations.
- This method is highly effective because it is straightforward and not directly dependent on other participants.

- For example, a person looks at random people that walk their pets on a busy street, and then uses this data to decide whether or not to open a pet food store in that area.

Surveys and Questionnaires

- Surveys and questionnaires provide a broad perspective from large groups of people.
- They can be conducted face-to-face, mailed, or even posted on the Internet to get respondents from anywhere in the world.
- The answers can be yes or no, true or false, multiple choice, and even open-ended questions.
- However, a drawback of surveys and questionnaires is delayed response and the possibility of ambiguous answers.

Focus Groups

- A focus group is similar to an interview, but it is conducted with a group of people who all have something in common.
- The data collected is similar to in-person interviews, but they offer a better understanding of why a certain group of people thinks in a particular way.
- However, some drawbacks of this method are lack of privacy and domination of the interview by one or two participants.
- Focus groups can also be time-consuming and challenging, but they help reveal some of the best information for complex situations.

Oral Histories

- Oral histories also involve asking questions like interviews and focus groups.
- However, it is defined more precisely and the data collected is linked to a single phenomenon.
- It involves collecting the opinions and personal experiences of people in a particular event that they were involved in.
- For example, it can help in studying the effect of a new product in a particular community.

Secondary Data Collection Methods

- Secondary data refers to data that has already been collected by someone else. It is much more inexpensive and easier to collect than primary data.
- While primary data collection provides more authentic and original data, there are numerous instances where secondary data collection provides great value to organizations.
- Here are some of the most common secondary data collection methods:

Internet

- The use of the Internet has become one of the most popular secondary data collection methods in recent times.
- There is a large pool of free and paid research resources that can be easily accessed on the Internet.

- While this method is a fast and easy way of data collection, you should only source from authentic sites while collecting information.

Government Archives

- There is lots of data available from government archives that you can make use of. The most important advantage is that the data in government archives are authentic and verifiable.
- The challenge, however, is that data is not always readily available due to a number of factors.
- For example, criminal records can come under classified information and are difficult for anyone to have access to them

Libraries

- Most researchers donate several copies of their academic research to libraries. You can collect important and authentic information based on different research contexts.
- Libraries also serve as a storehouse for financial statements, Retailer / Distributor/Deal Feedback, Sales annual reports, business directories, and other similar documents that help businesses in their research.

Well-designed data collection processes include the following steps:

1. Identify a business or research issue that needs to be addressed and set goals for the project.
2. Gather data requirements to answer the business question or deliver the research information.
3. Identify the data sets that can provide the desired information.
4. Set a plan for collecting the data, including the collection methods that will be used.
5. Collect the available data and begin working to prepare it for analysis

1. What are the key skills required for a Data Scientist?
2. List two common sources of data used in Data Science.
3. How can web scraping be useful for data collection in Data Science?
4. Name one advantage of using APIs for data collection.
5. What is data cleaning, and why is it important in Data Science?
6. What is the difference between using an API and web scraping for data collection?
7. Describe one technique to handle outliers in a dataset.
8. What is meant by "missing data," and how can it affect your analysis?