SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name –23 ADT202 FUNDAMENTAL OF DATA SCIENCE AND ANALYTICS
II Year / IV Semester

Unit 1 – Introduction to Data science
Retriving Data

Introduction to Data Retrieval

Data retrieval is a crucial step in the data science process.
The quality and relevance of data directly impact the analysis and outcomes.
This step involves gathering raw data from diverse sources.
These sources can range from structured databases to real-time data streams.

Types of Data Sources

Data can be sourced from various repositories, including internal and external systems.
Common sources include structured databases, APIs, and web scraping.
IoT devices and public datasets also provide valuable real-time and open data.
Each source has its strengths and is chosen based on project needs.

Internal Databases

Organizations store structured data in relational databases like MySQL or PostgreSQL.
NoSQL databases like MongoDB are also commonly used for unstructured data.
These databases typically contain transactional records and customer details.
Internal data is often prioritized due to its reliability and accuracy.

APIs (Application Programming Interfaces)

APIs provide access to data from third-party platforms such as Twitter or Google Maps.
They are especially useful for retrieving real-time or dynamic data.
API access often requires an API key for authentication and usage.
Common examples include financial services, weather, and social media APIs.

Web Scraping

Web scraping tools like BeautifulSoup and Selenium extract data from websites.
This technique is often used to collect data from e-commerce, news, or social media sites.
Web scraping is essential for gathering unstructured data not provided via APIs.
It allows for large-scale data extraction directly from web pages.

Public Datasets

Public datasets are often freely available through government and open-source platforms.

Examples include Kaggle, UCI Machine Learning Repository, and data.gov.

These datasets cover areas like demographics, climate, and economics.

They provide a valuable resource for building models and conducting research.

IoT Devices

IoT devices generate real-time sensor data in fields like healthcare and smart cities.
These devices collect data such as temperature, humidity, and motion, among others.
IoT data is valuable for monitoring systems in real-time and performing predictive analysis.
Data from IoT devices is typically streamed for immediate use or future analysis.

.

Steps for Retrieving Data

1. **Defining Data Requirements**: Identify necessary variables for analysis.
2. **Selecting Data Sources**: Choose reliable and relevant sources based on project needs.
3. **Accessing Data**: Use tools like SQL queries, APIs, or web scraping frameworks.
4. **Storing Data**: Securely store data in formats like CSV, JSON, or data warehouses.

Challenges and Example Code

Data retrieval can face challenges such as missing values, access restrictions, and scalability.

Access permissions and API limits may hinder data collection efforts.

Scalability challenges can require distributed systems like Hadoop.

Example: Using an API to retrieve weather data (see code snippet in the content).

THANK YOU