### **General Concepts**

1. **What is data collection?**Data collection is the process of gathering information from various sources to analyze and draw insights. It can involve retrieving data from files, databases, APIs, web scraping, or surveys.
2. **Why is data collection important in data analysis?**It ensures the availability of accurate, complete, and relevant data for analysis, enabling informed decision-making.
3. **What are the types of data used in analytics?**
   * Structured data (e.g., SQL tables)
   * Semi-structured data (e.g., JSON, XML)
   * Unstructured data (e.g., images, text, videos).
4. **What are the common sources of data collection?**
   * Databases
   * APIs
   * Web scraping
   * IoT devices
   * Files (CSV, Excel, JSON)
   * Surveys and experiments.
5. **What are the challenges in data collection?**
   * Missing or incomplete data
   * Data privacy concerns
   * Inconsistent formats
   * API rate limits or downtime
   * Data duplication.

### **Technical Questions**

**How do you read a CSV file in Python?**python  
Copy code  
import pandas as pd

df = pd.read\_csv("file.csv")

1. **How do you handle missing values in a dataset?**
   * Fill with default values (df.fillna())
   * Remove rows/columns (df.dropna())
   * Use interpolation.
2. **What Python libraries are commonly used for data collection?**
   * pandas for file handling
   * requests for APIs
   * BeautifulSoup/Selenium for web scraping
   * sqlalchemy for databases.

**How do you connect Python to a SQL database?**Using sqlalchemy:  
python  
Copy code  
from sqlalchemy import create\_engine

engine = create\_engine('mysql+pymysql://user:password@host/db')

1. **What is the difference between an API and web scraping?**

* **API**: A structured interface provided by a server to access its data.
* **Web scraping**: Extracting data directly from a website’s HTML.

### **APIs**

1. **What is an API?**An API (Application Programming Interface) allows applications to communicate and exchange data programmatically.
2. **How do you call an API in Python?**

python

Copy code

import requests

response = requests.get("https://api.example.com/data")

data = response.json()

1. **What is an API rate limit?**The maximum number of requests a client can make in a given time frame, enforced by the API provider.
2. **What is an API key?**A unique identifier used to authenticate API requests.
3. **How do you handle API pagination?**By iterating through pages using parameters like page or offset:

python

Copy code

while url:

response = requests.get(url)

data = response.json()

url = data.get('next') # Get the next page link

### **Web Scraping**

1. **What is web scraping?**Extracting data from websites using tools or code.
2. **What are the ethical considerations in web scraping?**

* Respect the robots.txt file.
* Avoid excessive requests to prevent server strain.
* Ensure compliance with legal regulations.

1. **What tools are used for web scraping?**

* BeautifulSoup for parsing HTML
* Selenium for dynamic content
* Scrapy for large-scale scraping.

1. **What is BeautifulSoup?**A Python library for parsing HTML and XML documents.
2. **How do you scrape a table from a webpage?**

python

Copy code

from bs4 import BeautifulSoup

import requests

url = "https://example.com"

response = requests.get(url)

soup = BeautifulSoup(response.content, "html.parser")

table = soup.find("table")

### **Data Validation**

1. **Why is data validation important?**To ensure the accuracy, completeness, and consistency of data before analysis.
2. **What are common data validation techniques?**

* Checking for missing values
* Ensuring correct data types
* Verifying ranges and constraints.

1. **How do you check for duplicates in a dataset?**

python

Copy code

duplicates = df.duplicated()

print(duplicates.sum())

1. **What is a data schema?**A structured representation of the data, defining tables, fields, and types.
2. **How do you handle outliers in a dataset?**

* Remove them
* Transform them (e.g., log-scaling)
* Use robust models like median-based statistics.

### **Automation and Scheduling**

1. **Why automate data collection?**To save time, reduce errors, and ensure timely updates.
2. **What tools are used for scheduling Python scripts?**

* cron (Linux)
* Windows Task Scheduler
* Python libraries like APScheduler.

1. **How do you automate daily data collection in Python?**Using APScheduler:

python

Copy code

from apscheduler.schedulers.blocking import BlockingScheduler

def job():

print("Running data collection...")

scheduler = BlockingScheduler()

scheduler.add\_job(job, 'interval', days=1)

scheduler.start()

### **Ethics and Security**

1. **What are the ethical considerations in data collection?**

* Obtain consent when required.
* Respect privacy laws (e.g., GDPR).
* Avoid collecting sensitive information without proper justification.

1. **How do you secure sensitive data (e.g., API keys)?**Use environment variables:

python

Copy code

import os

api\_key = os.getenv("API\_KEY")

### **Specific Use Cases**

1. **How do you collect data from a REST API?**

python

Copy code

response = requests.get("https://api.example.com/data")

print(response.json())

1. **What is JSON, and how do you handle it in Python?**JSON (JavaScript Object Notation) is a lightweight data format. Use json module:

python

Copy code

import json

with open("file.json") as f:

data = json.load(f)

1. **How do you scrape data from JavaScript-rendered websites?**Use Selenium or playwright to interact with dynamic content.
2. **How do you collect data from a database?**By querying using SQL and a library like pymysql or sqlalchemy.
3. **What is ETL in data collection?**ETL stands for Extract, Transform, Load, a process to gather, clean, and store data.

### **Practical Examples**

1. **How do you extract metadata from an image?**Use Pillow or ExifTool.
2. **How do you read data from a ZIP file?**

python

Copy code

import zipfile

with zipfile.ZipFile("file.zip", "r") as zip\_ref:

zip\_ref.extractall("path")

1. **How do you handle large datasets in Python?**Use libraries like Dask or chunked reading in pandas:

python

Copy code

chunks = pd.read\_csv("large\_file.csv", chunksize=1000)

1. **How do you collect real-time data?**By streaming APIs like Twitter API or using message queues like Kafka.
2. **How do you save data to a database in Python?**

python

Copy code

data.to\_sql("table\_name", engine, if\_exists="replace")

### **Advanced**

1. **How do you scrape data behind a login?**Use requests.Session or Selenium.
2. **What is data augmentation?**Enhancing data by adding transformations or synthetic data.
3. **What is the difference between GET and POST in APIs?**

* **GET**: Retrieve data.
* **POST**: Send data to the server.

1. **What is a webhook?**A mechanism to receive real-time data updates from an API.

### **Advanced Concepts**

1. **What is data streaming, and how is it different from batch processing?**

* **Data streaming** processes data in real-time as it becomes available.
* **Batch processing** handles data in chunks after it's been accumulated.

1. **What are examples of data streaming tools?**Apache Kafka, Apache Flink, and Spark Streaming.
2. **How do you handle data from IoT devices?**Collect data via MQTT protocols, APIs, or cloud platforms like AWS IoT Core.
3. **What is web scraping rate limiting, and how can it be managed?**It’s a restriction on the number of requests sent to a website. Manage it by implementing delays (time.sleep()) or exponential backoff.
4. **What is data ingestion?**The process of importing and transferring data into a system for analysis.

### **Data Wrangling and Cleaning**

1. **What is data wrangling?**The process of cleaning and transforming raw data into a structured format for analysis.
2. **What steps are involved in data cleaning?**

* Removing duplicates
* Filling or handling missing values
* Converting data types
* Handling outliers.

1. **How do you ensure data quality?**By performing data validation, monitoring integrity, and implementing constraints during ingestion.
2. **What is data imputation?**A technique to handle missing values by filling them with mean, median, mode, or using predictive models.
3. **What are common tools for data cleaning in Python?**

* pandas for structured data cleaning
* OpenRefine for semi-structured data
* NumPy for numerical operations.

### **Ethics and Privacy**

1. **What is GDPR, and how does it relate to data collection?**The General Data Protection Regulation governs data privacy in the EU, ensuring data collection respects user consent and rights.
2. **What is personally identifiable information (PII)?**Data that can identify an individual, like names, addresses, or Social Security numbers.
3. **How do you anonymize data?**By removing or masking identifiable information using techniques like hashing or tokenization.
4. **What is data sovereignty?**The concept that data is subject to the laws of the country in which it is collected or stored.
5. **What is the difference between encryption and hashing?**

* **Encryption**: Converts data into a secure format that can be reversed with a key.
* **Hashing**: Converts data into a fixed output, which is irreversible.

### **APIs and Databases**

1. **How do you handle time zone differences in data collection?**Normalize all timestamps to a common time zone, such as UTC, and store the original time zone if needed.
2. **What is a relational database?**A database structured in tables with rows and columns, using SQL for querying.
3. **What are NoSQL databases?**Non-relational databases like MongoDB, which store unstructured or semi-structured data.
4. **How do you query hierarchical data in a NoSQL database?**Use dot notation in MongoDB:

python

Copy code

db.collection.find({"field.subfield": "value"})

1. **What are ACID properties in databases?**

* **Atomicity**: Transactions are all-or-nothing.
* **Consistency**: Data must remain consistent.
* **Isolation**: Transactions occur independently.
* **Durability**: Changes persist after a commit.

### **Practical Use Cases**

1. **How do you extract data from an Excel file?**

python

Copy code

import pandas as pd

df = pd.read\_excel("file.xlsx")

1. **What is the difference between a flat file and a relational database?**

* **Flat file**: Stores data in plain text or a single table format.
* **Relational database**: Stores data in multiple related tables.

1. **What is the difference between structured and unstructured data?**

* **Structured data**: Organized in a predefined schema (e.g., SQL tables).
* **Unstructured data**: Lacks a predefined structure (e.g., videos, images).

1. **What is metadata, and why is it important?**Metadata is data about data, like file size or creation date, which helps in organizing and retrieving information.
2. **How do you scrape data from a website requiring a CAPTCHA?**Use CAPTCHA-solving services like 2Captcha or manual intervention in Selenium.

### **Data Storage**

1. **What are the common file formats for storing data?**

* **Text-based**: CSV, JSON, XML
* **Binary**: Parquet, Avro.

1. **What is the difference between JSON and XML?**

* **JSON**: Lightweight, easy-to-read data format.
* **XML**: More verbose, supports attributes and hierarchical structures.

1. **How do you decide between local and cloud storage for data?**

* Local for small-scale, sensitive, or offline projects.
* Cloud for scalability, accessibility, and collaborative projects.

1. **What is data partitioning in storage?**Splitting data into smaller chunks for better performance and easier management.
2. **What is the difference between OLAP and OLTP systems?**

* **OLAP (Online Analytical Processing)**: Used for analytics and querying large datasets.
* **OLTP (Online Transaction Processing)**: Used for transactional tasks like order processing.

### **Specialized Techniques**

1. **What is data scraping throttling?**Limiting the speed of data requests to avoid being blocked by the server.
2. **What are log files, and how can they be used in data analysis?**Log files record events or transactions, providing insights into system behavior or user activity.
3. **How do you collect geospatial data?**Use APIs like Google Maps API or geospatial datasets in formats like GeoJSON.
4. **How do you collect social media data?**Use platforms’ APIs (e.g., Twitter API, Facebook Graph API) or tools like Tweepy for Python.
5. **How do you process data from PDFs?**Use libraries like PyPDF2, pdfplumber, or OCR tools like Tesseract.
6. **What is data versioning, and why is it important?**Data versioning tracks changes to datasets, ensuring reproducibility and rollback capabilities during analysis.