Python Advance Assignment-3

1. What is the process for loading a dataset from an external source?

Solution:- When we load data from an external source, we have to **load it into a suspense table**. We can then review the data in the suspense table and modify it. To load data into the suspense table, position the source file or tape, specify the location of the source, and run the appropriate load external data process.

2. How can we use pandas to read JSON files?

Solution:- To read the files, we use **read_json()** function and through it, we pass the path to the JSON file we want to read. Once we do that, it returns a "DataFrame" that stores data. If we want to read a file that is located on remote servers then we pass the link to its location instead of a local path.

```
import pandas as pd


df = pd.read_json("FILE_NAME.json")
```

3. Describe the significance of DASK.

solution-Analysts often use tools like Pandas, Scikit-Learn, Numpy, and the rest of the Python ecosystem to analyze data on their personal computer. They like these tools because they are efficient, intuitive, and widely trusted. However, when they choose to apply their analyses to larger datasets, they find that these tools were not designed to scale beyond a single machine. And so, the analyst rewrites their computation using a more scalable tool, often in another language altogether.

Dask provides ways to scale Pandas, Scikit-Learn, and Numpy workflows more natively, with minimal rewriting. It integrates well with these tools so that it copies most of their API and uses their data structures internally. Moreover, Dask is co-developed with these libraries to ensure that they evolve consistently, minimizing friction when transitioning from a local laptop, to a multi-core workstation, and then to a distributed cluster.

4. Describe the functions of DASK.

Solution:- Dask is a free and open-source library for parallel computing in Python. Dask helps you scale your data science and machine learning workflows. Dask makes it easy to work with Numpy, pandas, and Scikit-Learn, but that's just the beginning. Dask is a framework to build distributed applications that has since been used with dozens of other systems like XGBoost, PyTorch, Prefect, Airflow, RAPIDS, and more. It's a full distributed computing toolbox that fits comfortably in your hand.
Dask collections provide the API that you use to write Dask code. Collections create *task graphs* that define how to perform the computation in parallel. Each node in the task graph is a normal Python function and edges between nodes are normal Python objects. You can view the graph by calling visualize() on any collection object.

5. Describe Cassandra's features.

Solution:- It is an open source, user-available, distributed, NoSQL DBMS which is designed to handle large amounts of data across many servers. It provides zero point of failure. Cassandra offers massive support for clusters spanning multiple datacenters .

Some features of Cassandra are:-

- **Distributed:**
  Each node in the cluster has same role. There's no question of failure & the data set distributed across the cluster but one issue is there that is the master isn't present in each node to support request for service.
- **Supports replication & Multi data center replication:**
  Replication factor comes with best configurations in cassandra. Cassandra is designed to have a distributed system, for the deployment of large number of nodes for across multiple data centers and other key features too.
- **Scalability:**
  It is designed to r/w throughput, Increase gradually as new machines are added without interrupting other applications.