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

When you load data from an external source, you load it into a suspense table. You 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. PeopleSoft Process Scheduler runs the process and stores the data in the suspense tables. When it is finished, the PeopleSoft Process Scheduler displays a process instance number in the lower left corner of the screen. Use this number to review the data on the appropriate Suspense Process Options page.

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

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”(A table of rows and columns) 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.

Example:

df = pd.read_json("FILE_JSON.json")

df.head()

Output:

	One	Two
0	60	110
1	60	117
2	60	103
3	45	109
4	45	117
5	60	102

3.Describe the significance of DASK

Dask is convenient on a laptop. It installs trivially with conda or pip and extends the size of convenient datasets from “fits in memory” to “fits on disk”.

Dask can scale to a cluster of 100s of machines. It is resilient, elastic, data local, and low latency.Dask represents parallel computations with task graphs. Dask.distributed is a lightweight library for distributed computing in Python. It extends both the concurrent.futures and dask APIs to moderate sized clusters.Dask.distributed is a centrally managed, distributed, dynamic task scheduler. The central dask-scheduler process coordinates the actions of several dask-worker processes spread across multiple machines and the concurrent requests of several clients.

4.Dedcribe the functions of DASK

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.If you have larger-than-memory data, you can use Dask to scale up your workflow to leverage all the cores of your local workstation, or even scale out to the cloud.Dask collections provide the API that you use to write Dask code. For example, Dask DataFrame used in the previous section is a Collection.

Dask has several collections:

High-Level collections:

Arrays: Parallel NumPy

Bags: Parallel lists

DataFrames: Parallel pandas

Machine Learning: Parallel Scikit-Learn

Others from external projects, like Xarray

Low-Level collections:

Delayed: Parallel function evaluation

Futures: Real-time parallel function evaluation

5.Describe Cassandra's features

Apache Cassandra 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 datacentres.There are some

massive features of Cassandra. Here are some of the features described below:

Distributed:

Each node in the cluster has has same role. There's no question of failure & the data set is 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.

Fault-tolerance:

Data is automatically stored & replicated for fault-tolerance. If a node Fails, then it is replaced within no time.

MapReduce Support:

It supports Hadoop integration with MapReduce support.Apache Hive & Apache Pig is also supported.

Query Language:

Cassandra has introduced the CQL (Cassandra Query Language). Its a simple interface for accessing the Cassandra.