**Assignment 3**

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

**A)**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**.**

1. **How can we use pandas to read JSON files?**

**A)**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 1: Reading JSON file

Import pandas as pd

Df = pd.read\_json(“FILE\_JSON.json”)

Df.head()

Output:

One Two

1. 60 110
2. 1 60 117
3. 60 103
4. 3 45 109
5. 45 117
6. 5 60 102

Example 2: Creating JSON data and reading in dataframe

Import pandas as pd

Data = {

“One”: {

“0”: 60,

“1”: 60,

“2”: 60,

“3”: 45,

“4”: 45,

“5”: 60

},

“Two”: {

“0”: 110,

“1”: 117,

“2”: 103,

“3”: 109,

“4”: 117,

“5”: 102

}

}

Df = pd.DataFrame(data)

Print(df)

Output:

One Two

1. 60 110
2. 1 60 117
3. 60 103
4. 3 45 109
5. 45 117
6. 5 60 102
7. **Describe the significance of DASK.**

**A)**The DASK was the first computer in Denmark. It was commissioned in 1955, designed and constructed by Regnecentralen, and began operation in September 1957. DASK is an acronym for Dansk Aritmetisk Sekvens Kalkulator or Danish Arithmetic Sequence Calculator.Dask can enable efficient parallel computations on single machines by leveraging their multi-core CPUs and streaming data efficiently from disk. It can run on a distributed cluster, but it doesn’t have to.

1. **Describe the functions of DASK.**

**A)**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.

1. **Describe Cassandra’s features.**

**A)**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:

1.Distributed

2Supports replication & Multi data center replication.

3.MapReduce Support

4.Scalability

5.Fault-tolerance

6.Query Language