Assignment 3

- 1)what is the process for loading a dataset from an external source ?
- Ans: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 ?
- Ans: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.

- 3) Describe the significance of DASK
- Ans: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.
- Dask emphasizes the following virtues: Familiar: Provides parallelized NumPy array and Pandas DataFrame objects. Flexible: Provides a task scheduling interface for more custom workloads and integration with other projects. Native: Enables distributed computing in pure Python with access to the PyData stack.

- 4) Describe the functions of DASK
- Ans:The Dask delayed function decorates your functions so that they operate lazily. Rather than executing your function immediately, it will defer execution, placing the function and its arguments into a task graph.
- We used the dask.delayed function to wrap the function calls that we want to turn into tasks. None of the inc, double, add, or sum calls have happened yet. Instead, the object total is a Delayed result that contains a task graph of the entire computation. Looking at the graph we see clear opportunities for parallel execution. The Dask schedulers will exploit this parallelism, generally improving performance (although not in this example, because these functions are already very small and fast.)

- 5) Describe the cassandra’s features .
- Ans: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
- 2)Supports replication & Multi data center replication