Python Advance Assignment 3

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

For loading a dataset into a python kernel we need to:

- Import pandas.
- Read the files by specifying the file location and file type.

Ex:

pd.read_csv(“ ”) ### for csv file types.
pd.read_excel(“ ”) ### for excel file types.

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

For reading a JSON file.
pd.read_JSON(“ specify the file location and file type”)

3. Describe the significance of DASK.

- DASK is flexible library for parallel computations on single machines by leveraging their multi-core CPUs and streaming data efficiently from disk.
- It can run on a distributed cluster. It allows the user to replace clusters with single-machine scheduler which would bring down the overhead.
- It helps you scale you data science and machine learning workflow.

4. Describe the functions of DASK.

- **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.
- **Fast**: Operates with low overhead, low latency, and minimal serialization necessary for fast numerical algorithms
- **Scales up**: Runs resiliently on clusters with 1000s of cores
- **Scales down**: Trivial to set up and run on a laptop in a single process
- **Responsive**: Designed with interactive computing in mind, it provides rapid feedback and diagnostics to aid humans

5. Describe Cassandra's features.

Cassandra is a free and open-source, distributed wide-column store, NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.

1. **Distributed:**
   Each node in the cluster 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.
2. **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.
3. **Scalability:**
   It is designed to r/w throughput, Increase gradually as new machines are added without interrupting other applications.
4. **Fault-tolerance:**
   Data is automatically stored & replicated for fault-tolerance. If a node Fails, then it is replaced within no time.
5. **MapReduce Support:**
   It supports Hadoop integration with MapReduce support.Apache Hive & Apache Pig is also supported.
6. **Query Language:**
   Cassandra has introduced the CQL (Cassandra Query Language). Its a simple interface for accessing the Cassandra.