**1)what is the process for loading a dataset from an external source?**

**A) loading data from external data source. dataframe reader is a fluent API to describe the input data source that will be used to load data from an external source Dataframe reader is created exclusively using spark session. Dataframe reader API methods are CSV,format,jdbc,json,load, option etc Dataframe reader supports many file formats natively and offers the interface to define custom formals**

**2)how can we use pandas to read json files?**

**A) prepare your JSON data for pandas .In most cases people often fail to create a proper JSON data format and thus they aren’t able to read JSON files in pandas.make sure the Datastructure is actually JSON .used read\_json( )function in pandas .once you have yours JSON file ready you can easily read it in as a data frame in pandas using the read JSON () function in pandas .**

**3)Describe the significance of DASK?**

**A)DASK is convenient on laptop it installs trivially with conda or pip and extends the size of convenient data from “fits in memory” to “fits on disk”.Dask can scale to a cluster of 100s of machines it is resilient elastic datalocal and low latency .DASK can enable efficient parallel computations on single machines by leve raging their multi-core CPUs and streaming data efficiently from disk it can run on distributed cluster,but it doesn’t have to**

**4)Describe the functions of DASK?**

**A)DASK is a free and open-source library for parallel computing in python DASK help you scale your datascience and machine learning workflow 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 been used with dozens of other systems like xgboost,pytorch, perfect etc**

**5)Describe Cassandra’s features?**

**A)Apache Cassandra is an open source ,user-available distributed,nosal DBMS which is designed to handle large amount of data across many servers it provides zero point of failure Cassandra spanning multiple datacentres**