Python Advance Assignment- 3

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

Answer:-

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.

```
    # underscore to omit the label arrays
(train_images, train_labels), (_, _) = tf.keras.datasets.mnist.load_data()



train_images = train_images.reshape(train_images.shape[0], 28, 28, 1).astype('float32')
train_images = (train_images - 127.5) / 127.5 # Normalize the images to [-1, 1]


BUFFER_SIZE = 60000
BATCH_SIZE = 256


# Batch and shuffle the data
train_dataset = tf.data…
```

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

Answer:-

Pandas / Python pandas read_json () function can be used to read JSON file or string into DataFrame.

It supports JSON in several formats by using orient param.

JSON is shorthand for JavaScript Object Notation which is the most used file format that is used to

exchange data between two systems or web applications.

To read a JSON file via Pandas, we'll utilize the read_json () method and pass it the path to the file we'd

like to read. The method returns a Pandas DataFrame that stores data in the form of columns and rows.

```
# Import pandas

import pandas as pd


# Read json from String

json_str = '{"Courses":{"r1":"Spark"},"Fee":{"r1":"25000"},"Duration":{"r1":"50 Days"}}'

df = pd.read_json(json_str)

print(df)


# Outputs

#   Courses    Fee Duration

#r1   Spark  25000  50 Days
```

3. Describe the significance of DASK.

Answer:-

Dask is a library that supports parallel computing in python. It provides features like-

Dynamic task scheduling which is optimized for interactive computational workloads

Big data collections of dask extends the common interfaces like NumPy, Pandas etc.

Pandas Read Json Example: In the next example we are going to use Pandas read_json method to

read the JSON file we wrote earlier (i.e., data.json). It's fairly simple we start by importing pandas

as pd: import pandas as pd # Read JSON as a dataframe with Pandas: df = pd.read_json ('data.json') dfPandas

is one of the most commonly used Python libraries for data handling and visualization. The Pandas library provides

classes and functionalities that can be used …

4.Describe the functions of DASK.

Answer:-

Dask is a library that supports parallel computing in python. It provides features like- Dynamic task

scheduling which is optimized for interactive computational workloads Big data collections of dask extends

the common interfaces like NumPy, Pandas etc

Dask is a library that supports parallel computing in python. It provides features like- Big data collections

of dask extends the common interfaces like NumPy, Pandas etc. Why Dask? Most of the BigData analytics will be

using Pandas, NumPy for analyzing big data.

How to use Dask Bag for various operations? *Dask provides efficient parallelization for data analytics in python.

Dask Dataframes allows you to work with large datasets for both data manipulation and building ML models with only

minimal code changes. It is open source and works well with python libraries like NumPy, scikit-learn, et

5. Describe Cassandra's features.

Answer:-

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: