Experiment-9

Q Working with dataframe

1. load userdetails.json file from data (local) folder to Databrick file System.

2. Create the dataframe from userdetails.json file.

```
→ df = spark.read.json ("dbfs:/Filestore/tables/
                          userdetails.json")
  df.show()
```

```
→ 2. df1 = spark.read.format("json").load("dbfs:/
          FileStore/tables/userdetails.json")
  df1.show()
```

3. Print the schema in a tree format

```
→ df.printSchema()
```

4. Select only the first_name column.

```
→ df.select("first_name").show()
```

5. Select the people whose id is greater than 10.

```
→ a = df.filter (df['id'] >10)
  a.show()
```

6. Count the number of people by gender.
→ df.groupBy("gender").count().show()

Q Working with parquet format dataframe.

1. import SQLContext package.
→ from pyspark.sql import SQLContext

2. create sqlcontext
→ sqlcontext = SQLContext(sc)

3. Load 'episodes.parquet' file into Databricks tables folder.

4. Load the episodes.parquet file and create dataframe.
→ episodeDF = spark.read.parquet("dbfs:/FileStore/tables/episodes.parquet")

5. Print schema of the dataset.
→ episodeDF.printschema()

6. Show the contents of the dataframe
→ episodeDF.show()

7. Count total number of records.
→ episodeDF.count()

8. Filter the records where numbers of doctor. are greater than 5.

→ filter_episode = episodeDF.filter("doctor > 5")

9. Display filtered data using show()

→ filter_episode.show()

10. Save the above data in Databricks tables folder.

→ filter_episode.write.parquet("dbfs:/FileStore/tables/episode_file")

11. Verify the output in Databricks.

→ myfile = spark.read.parquet("dbfs:/FileStore/tables/episode-file")

myfile.show()

Q. Working with avro format dataframe.

1. Import SQLContext package.

→ from pyspark.sql import SQLContext

2. create SqlContext.

→ sqlContext = SQLContext(sc)

3. Load 'episodes.avro' file from local folder into Databricks tables folder.

4. load the episodes.avro file and create dataframe.

→ episode_avro = spark.read.format("com.databricks.spark.avro").load("dfs:/Filestore/tables/episodes.avro")

5. Print schema of dataset.

→ episode_avro.printSchema()

6. Show the contents of dataset.

→ episode_avro.show()

7. Count total number of records.

→ episode_avro.count()

8. Filter the records where no. of doctor are greater than 3.

→ episode_filter = episode_avro.filter("doctor>3")

9. Display filtered data using show function()

→ episode_filter.show()

Outputs: 9(a)

Out[1]:

city	email	first name	gender	id	ip_address	last_name	race	timestamp
Miami	wbell@tumblr.com	Wayne	Male	1	77.152.229.17	Bell	Costa Rican	1463468947
Xiejia	aallen1@state.gov	Anthony	Male	2	62.49.195.120	Allen	Comanche	1442538227

Out[3] root.

```
 |-- city : string (nullable = true)
 |-- email : string (nullable = true)
 |-- first_name : string (nullable = true)
 |-- gender : string (nullable = true)
 |-- id : long (nullable = true)
 |-- ip_address : string (nullable = true)
 |-- last_name : string (nullable = true)
 |-- race : string (nullable = true)
 |- timestamp : string (nullable = true)
```

Out[4]

first name
Wayne
Anthony
Eric
Jimmy
Diana

Teacher's Signature ....

Out[5]:

city	email	first_name	gender	id	ip_address	last_name	race	timestamp
Madison	lmccoy@jiathis.com	Larry	Male	12	31.136.32.15	Mccoy	Venezuelan	1474672688
idery	ewood@youtube.com	Emily	Female	13	7.219.[illegible]	wood	Peruvian	147819984,

Out[6]:

gender	count
Female	507
Male	493

9(b).

```
Out[5] root
 |-- title: String (nullable = true)
 |-- air_date: String (nullable = true)
 |-- doctor: Integer (nullable = true)
```

Out[6]

title	air_date	doctor
The Eleventh hour	3 april 2010	11
The Doctor's wife	14 May 2011	11
Horror of fang Rock	3 Sept. 1977	4
The Mysterious	6 sept 1986	6
Rose	26 Mar. 2005	9
Castrolava	4 Jan. 1982	5

Out[7]: 6

Teacher's Signature ....

out[9).

title	air_date	doctor
The Eleventh Hour	3 april 2010	11
The Doctor's Wife	14 May 2011	11
The Mysterious Rose	6 sept. 1986	6
	26 Mar. 2005	9

9(c)

out[5)
```
root
 |-- title : string (nullable = true)
 |-- air_date : string (nullable = true)
 |-- doctor : integer (nullable = true)
```

out[6)

Title	air_date	doctor
The Eleventh Hour.	3 April 2010	11
The Doctor's wife	14 May 2011	11

out(7) 8

out[9]

Title	air_date	doctor
Rosa	26 March 2005	9
Castrolava	4 Jan 1982	5