Task 5

Notebook:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1092176685531650/3530701261005494/6776489139542437/latest.html

Deal with different file formats

- **json**

```
x = spark.read.json("/home/s_kante/spark/data/emp.json")
x.printSchema()
x.show()

x.write.format("json").save("<Path>”)
```

- **csv**

```
x = spark.read.load("/home/s_kante/spark/data/emp.csv",
format='com.databricks.spark.csv',header='true',inferSchema='true')
x.show()

x.write.csv(“<Path>”)
```

- **pipe delimited**

```
dat = spark.read.load("/home/s_kante/spark/data/emp.dat", format="csv", sep='|',
inferSchema="true", header="true")
dat.show()
dat.printSchema()

dat.write.csv(“<Path>”,”|”)
```

- **Parquet**

```
x = spark.read.parquet("/home/s_kante/spark/data/emp.parquet")
x.printSchema()
x.show()

x.write.parquet(“<Path>”)
```

Save modes while saving the data

```
df.write.save("/home/s_kante/spark/data/emp.parquet", mode="append")
```

Querying data directly from file

```
df = spark.sql("SELECT * FROM parquet.`/home/s_kante/spark/data/emp.parquet/`")
df.show()
```

Register dataframe as temporary table

```
df.createOrReplaceTempView(“<Table Name>”)
```

Register dataframe as global temporary table

```
df.createGlobalTempView("accounts")
spark.sql("SELECT * FROM global_temp.accounts").show()
spark.newSession().sql("SELECT * FROM global_temp.accounts").show()
```

Query temporary metadata

```
spark.catalog.listTables().show()
spark.catalog.listColumns("employee").show()
```