**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("<file dir>/emp.json")

x.printSchema()

x.show()

x.write.format("json").save("<storage path>”)

* **csv**

x = spark.read.load("<file dir>/emp.csv", format='com.databricks.spark.csv',header='true',inferSchema='true')

x.show()

x.write.csv(“<storage path>”)

* **pipe delimited**

dat = spark.read.load("<file dir>/emp.dat", format="csv", sep='|', inferSchema="true", header="true")

dat.show()

dat.printSchema()

dat.write.csv(“<storage path>”,sep=”|”) # sep is not the second param

* **Parquet**

x = spark.read.parquet("<file dir>/emp.parquet")

x.printSchema()

x.show()

x.write.parquet(“<Path>”)

**Save modes while saving the data**

df.write.save("<storage path>/emp.parquet", mode="append")

**Querying data directly from file**

df = spark.sql("SELECT \* FROM parquet.`<file dir>/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() #the database for global temporary view is fixed

**Query temporary metadata**

spark.catalog.listTables().show()

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