Table Data: 1

(1,"Smith",-1,"2018","10","M",3000),

(2,"Rose",1,"2010","20","M",4000),

(3,"Williams",1,"2010","10","M",1000),

(4,"Jones",2,"2005","10","F",2000),

(5,"Brown",2,"2010","40","",-1),

(6,"Brown",2,"2010","50","",-1)

"emp\_id","name","superior\_emp\_id","year\_joined",

"emp\_dept\_id","gender","salary"

Table Data: 2

("Finance",10),

("Marketing",20),

("Sales",30),

("IT",40)

"dept\_name","dept\_id"

1. Create Dataframe1 using Table Data-1 and Dataframe2 using Table Data - 2 using PySpark
2. Left-anti join two DF using similarity check (i.e. from the similar columns)
3. Store the final table in csv file locally, upon using Outer join
4. Store the final table in parquet file locally, upon using Outer join

from pyspark.sql import SparkSession

from pyspark.sql import Row

from pyspark.sql.types import StructField,StringType,IntegerType,StructType

from pyspark.sql import SQLContext

from pyspark.sql.types import \*

# Create Spark session

spark = SparkSession.builder.master("local").appName("demo").getOrCreate()

print(spark)

# print("APP Name :"+spark.sparkContext.appName)

# print("Deploy Mode :"+spark.sparkContext.deployMode)

# print("Master :"+spark.sparkContext.master)

# Create a schema for the dataframe

schema1 = StructType([

StructField("emp\_id",IntegerType(),True),

StructField("name",StringType(),True),

StructField("superior\_emp\_id",IntegerType(),True),

StructField("year\_joined", StringType(), True),

StructField("emp\_dept\_id", StringType(), True),

StructField("gender",StringType(), True),

StructField("salary",IntegerType(), True)

])

# List

data1 = [

(1,"Smith",-1,"2018","10","M",3000),

(2,"Rose",1,"2010","20","M",4000),

(3,"Williams",1,"2010","10","M",1000),

(4,"Jones",2,"2005","10","F",2000),

(5,"Brown",2,"2010","40","",-1),

(6,"Brown",2,"2010","50","",-1)

]

# Convert list to RDD

rdd1 = spark.sparkContext.parallelize(data1)

print(rdd1)

# Create data frame

emp = spark.createDataFrame(data1,schema1)

# df1.printSchema()

display(emp)

#list

data2=[

("Finance",10),

("Marketing",20),

("Sales",30),

("IT",40),

]

#Schema

schema2 = StructType([

StructField("dept\_name",StringType(),True),

StructField("dept\_id",IntegerType(),True),

])

#List to rdd

rdd2=spark.sparkContext.parallelize(data2)

dept = spark.createDataFrame(rdd2,schema2)

display(dept)

#Let anti join

df=emp.join(dept,[emp.emp\_dept\_id == dept.dept\_id],how='left\_anti')

display(df)

#Outer join on emp and dept

result = emp.join(dept,[emp.emp\_dept\_id == dept.dept\_id],how='outer')

display(result)

#CSV file

result.write.format('com.databricks.spark.csv').save("dbfs:/FileStore/shared\_uploads/kashif.ahmed@diggibyte.com/result.csv",header="true", inferSchema="true")

data\_result1 = sqlContext.read.format("csv").options(header='true',inferSchema = 'true',delimeter = '|').load("dbfs:/FileStore/shared\_uploads/kashif.ahmed@diggibyte.com/result.csv")

data\_result1.show()

#Parpuet File

result.write.parquet("dbfs:/FileStore/shared\_uploads/kashif.ahmed@diggibyte.com/result2.parquet")

result2=spark.read.parquet('dbfs:/FileStore/shared\_uploads/kashif.ahmed@diggibyte.com/result2.parquet')

result2.show()