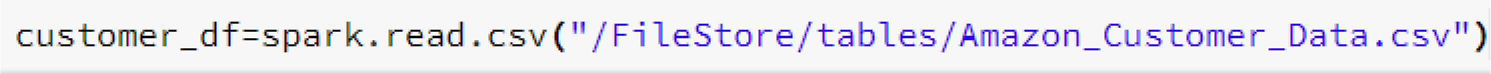
www.linkedin.com/in/naresh-neelam

Databricks-Spark DataFrame Reader

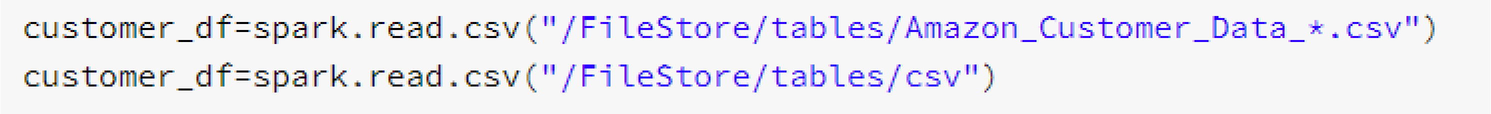
DataFrame Reader

Reading Single csv Files



```
customer_df=spark.read.csv("/FileStore/tables/Amazon_Customer_Data.csv")
```

Reading Multiple csv Files



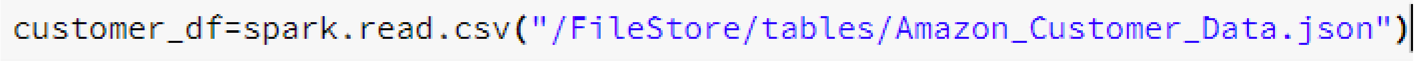
```
customer_df=spark.read.csv("/FileStore/tables/Amazon_Customer_Data_*.csv")
customer_df=spark.read.csv("/FileStore/tables/csv")
```

Important Options:

header=true/false
schema=<<schema Definition>>
inferSchema=true/false
sep=<<delimeter>>
dateFormat=yyyy-MM-dd
Mode=PERMISSIVE/DROPMALFORMED/FAILFAST

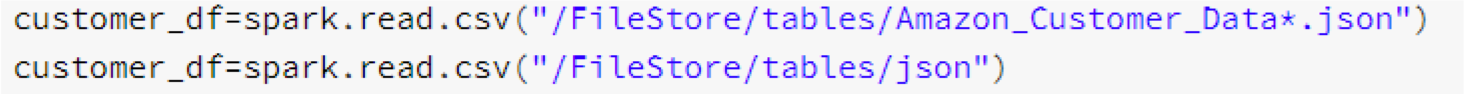
DataFrame Reader

Reading Single JSON Files



```
customer_df=spark.read.csv("/FileStore/tables/Amazon_Customer_Data.json")
```

Reading Multiple JSON Files



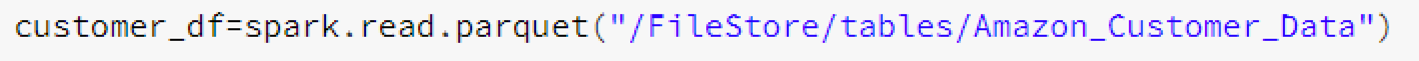
```
customer_df=spark.read.csv("/FileStore/tables/Amazon_Customer_Data*.json")
customer_df=spark.read.csv("/FileStore/tables/json")
```

Important Options:

schema=<<schema Definition>>
multiLine=true/false
dateFormat=yyyy-MM-dd
Mode=PERMISSIVE/DROPMALFORMED/FAILFAST

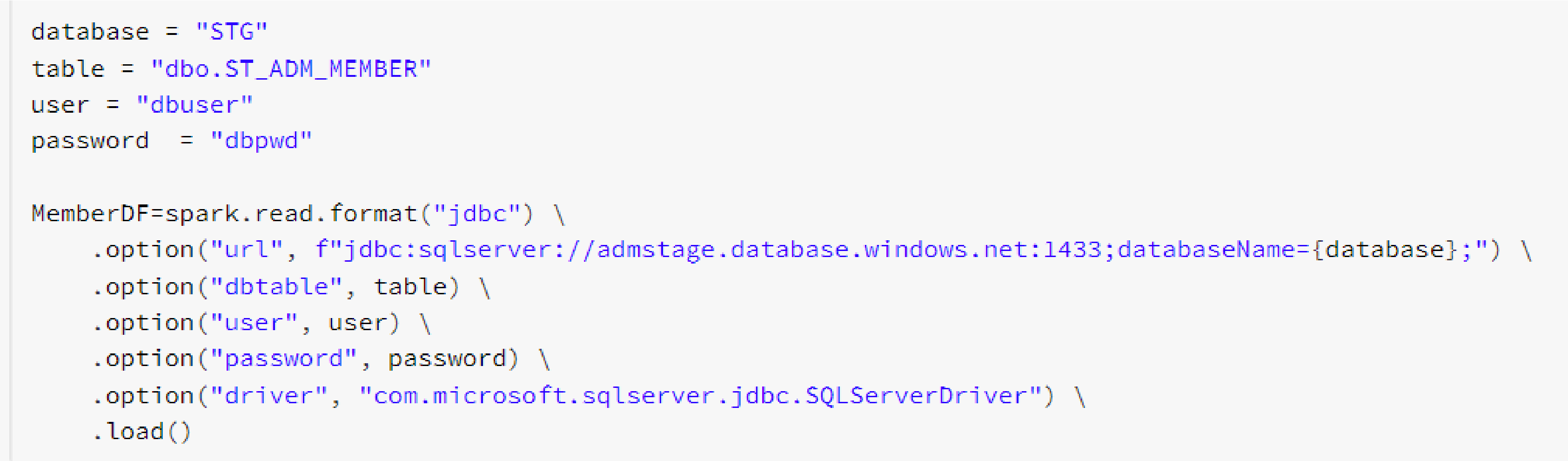
DataFrame Reader

Reading Parquet Files



```
customer_df=spark.read.parquet("/FileStore/tables/Amazon_Customer_Data")
```

Reading RDBMS Tables

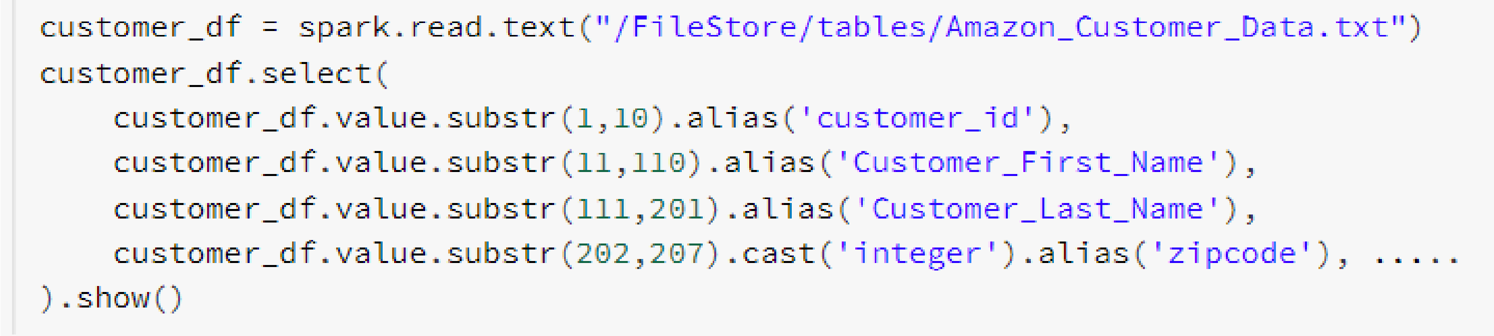


```
database = "STG"
table = "dbo.ST_ADM_MEMBER"
user = "dbuser"
password  = "dbpwd"

MemberDF=spark.read.format("jdbc") \
    .option("url", f"jdbc:sqlserver://admstage.database.windows.net:1433;databaseName={database};") \
    .option("dbtable", table) \
    .option("user", user) \
    .option("password", password) \
    .option("driver", "com.microsoft.sqlserver.jdbc.SQLServerDriver") \
    .load()
```

DataFrame Reader

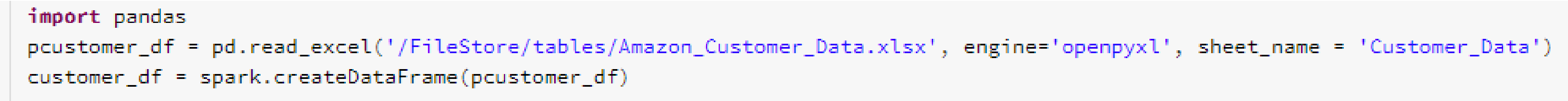
Reading Fixed Width Files



```
customer_df = spark.read.text("/FileStore/tables/Amazon_Customer_Data.txt")
customer_df.select(
    customer_df.value.substr(1,10).alias('customer_id'),
    customer_df.value.substr(11,110).alias('Customer_First_Name'),
    customer_df.value.substr(111,201).alias('Customer_Last_Name'),
    customer_df.value.substr(202,207).cast('integer').alias('zipcode'), .....
).show()
```

Reading Excel Files

Need to user python pandas, install xlrd and openpyxl package in your Databricks cluster



```
import pandas
pcustomer_df = pd.read_excel('/FileStore/tables/Amazon_Customer_Data.xlsx', engine='openpyxl', sheet_name = 'Customer_Data')
customer_df = spark.createDataFrame(pcustomer_df)
```