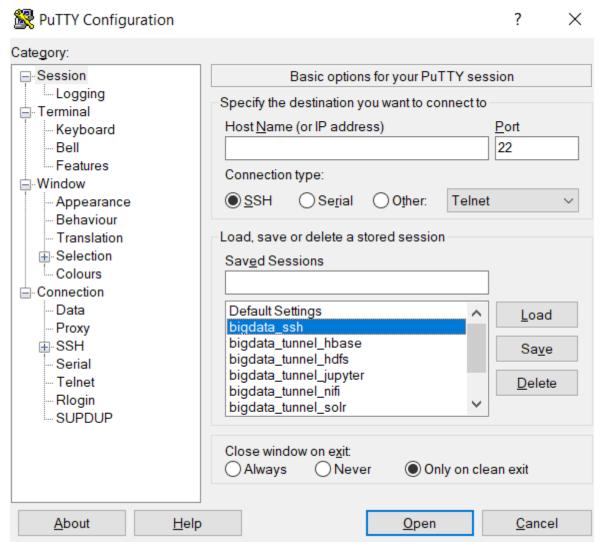
SparkSQL and Scala with Google Cloud Instance and Putty

• Putty Configuration



• Navigating to Directory and Starting Docker Containers:

```
louis@bigdata:~/dsc650-infra/bellevue-bigdata$ cd hadoop-hive-spark-hbase louis@bigdata:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose up -d Creating network "hadoop-hive-spark-hbase_spark_net" with the default driver Creating hadoop-hive-spark-hbase_zoo2_1 ... done Creating hadoop-hive-spark-hbase_hivemetastore_1 ... done Creating hadoop-hive-spark-hbase_zoo1_1 ... done Creating hadoop-hive-spark-hbase_zoo3_1 ... done Creating hadoop-hive-spark-hbase_worker2_1 ... done Creating hadoop-hive-spark-hbase_master_1 ... done Creating hadoop-hive-spark-hbase_master_1 ... done Creating hadoop-hive-spark-hbase_worker1_1 ... done louis@bigdata:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$
```

Accessing the Master Container:

```
louis@bigdata:~/dsc650-infra/bellevue-bigdata/hadoop-hive-spark-hbase$ docker-compose exec ma ster bash
bash-5.0# hdfs dfs -mkdir /data
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/program/hadoop/share/hadoop/common/lib/slf4j-log4j12-1
.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/tez/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/program/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4
j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2024-08-18 23:50:38,738 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
bash-5.0#
```

Load CSV into HDFS:

```
hdfs dfs -mkdir /data
hdfs dfs -put /data/grades.csv /data/grades.csv
```

• Enter Spark Shell & Load Data:

• Running Queries on Data:

```
> spark.sql("SHOW TABLES").show()
110176 [main] WARN org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.strict.mana
ged.tables does not exist
110176 [main] WARN org.apache.hadoop.hive.conf.HiveConf - HiveConf of name hive.create.as.i
nsert.only does not exist
110210 [main] WARN org.apache.spark.sql.hive.client.HiveClientImpl - Detected HiveConf hive
execution.engine is 'tez' and will be reset to 'mr' to disable useless hive logic.
|database|tableName|isTemporary|
                 dfl
                           true|
 cala> spark.sql("SELECT * FROM df WHERE Final > 50").show()
 Last name|First name|
                              SSN|Test1|Test2|Test3|Test4|Final|Grade|
               Andrew|223-45-6789|
                                                              831
                                                                     Αl
  Airpump|
   Backus|
                  Jim|143-12-1234|
                                                                     A+I
                                                        881
                                                              77|
                                                                     B-|
                Benny|234-56-2890|
                                                        801
                                                                     B-|
 cala> spark.sql("SELECT * FROM df").show()
 Last name|First name|
                          SSN|Test1|Test2|Test3|Test4|Final|Grade|
                                                                     D-I
                                                        971
                                                                     D+|
               Gramma|567-89-0123|
                                                        40|
             Electric|087-65-4321|
  Android|
                                                                     B-|
                 Fred | 456-78-9012|
  Bumpkinl
                                                                     A-|
   Rubble
                                                                     F
    Noshow|
                  Bif|632-79-9939|
                                                        40|
                                                              50|
                                                                     B+|
               Andrew | 223-45-6789 |
  Airpump|
                                      49|
                  Jim|143-12-1234|
                                                  97
                                                              97|
   Backusl
                                                                     A+1
 Carnivore
    Dandy|
                                      47|
                  Ima|456-71-9012|
                                                        88|
                                                                     B-|
                                                        801
                                                                     B-|
                  Boy|345-67-3901|
   George
               Harvey|632-79-9439|
|Heffalump|
```

```
cala> spark.sql("select * from df where Test3>80").show()
|Last name|First name|
                             SSN|Test1|Test2|Test3|Test4|Final|Grade|
  Alfalfa| Aloysius|123-45-6789|
                                                                   D-|
                                                                   D+|
               Fred|456-78-9012|
  Bumpkin|
                                                                   A-|
               Andrew|223-45-6789|
  Airpump|
  Backus|
                                                                   A+|
               Benny|234-56-2890|
                                                                   B-|
cala> spark.sql("select Test4 from df where Test4<90").show()
|Test4|
   77|
scala> spark.sql("select Test1,Test2 from df where(Test1<50 and Test2>90 )").show()
|Test1|Test2|
```