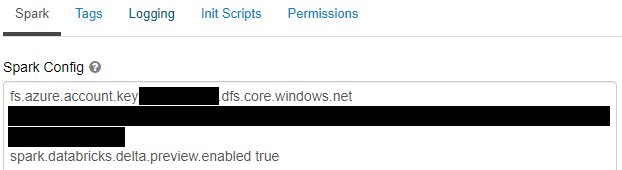
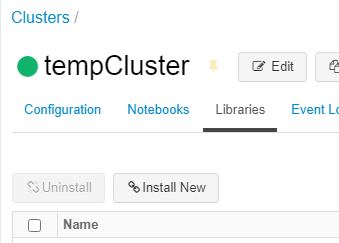
# Steps to add the precompiled jar to the cluster

1. Create new Databricks cluster.
2. In the advanced options, add the following:

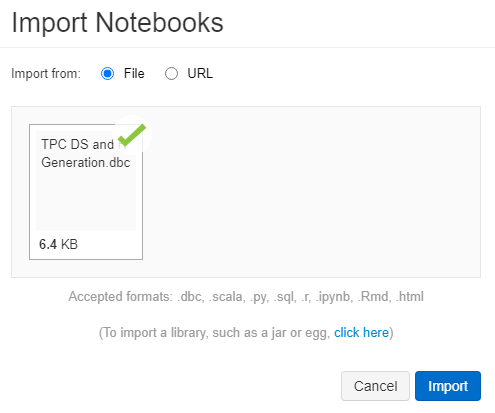
fs.azure.account.key.{StorageAccount}.dfs.core.windows.net {StorageAccountKey}



1. Once created, go the Libraries tab in the cluster and click Install New

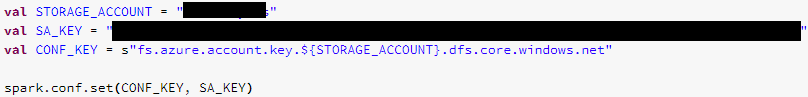


1. Upload the ‘spark-sql-perf\_2.12-0.5.1-SNAPSHOT.jar’ file that is in this repo and click Install. It will take a few seconds to install. Note, this jar is compatible with scala 2.12 so make sure pick a cluster type that has scala 2.12.
2. Once installed, import and open the TPC DS and H Generation notebook.

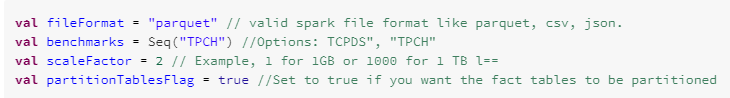


1. Set the configurations in cell 1 and 2.

Cell 1- Update to the storage account and key that the data will be generated.



Cell 2 – Define the file format, benchmark type, scale factor, and if the data should be partitioned or not.



1. Run the notebook. The data generation duration can vary (minutes to hours) based on the scale factor defined and the size of the cluster being used.

# 

# Steps to compile the Databricks Spark Sql Perf project

1. Java JDK must be installed

<https://www.oracle.com/java/technologies/javase-downloads.html>

1. Download the sbt tool to compile scala code to jars.

<https://docs.scala-lang.org/getting-started/sbt-track/getting-started-with-scala-and-sbt-on-the-command-line.html>

1. Download the databricks spark-sql-perf project and build the scala script

<https://github.com/databricks/spark-sql-perf>

cd to package location

sbt +package

1. Upload the compiled jar file to the Databricks library and install it on the cluster.