

For Spark DataFrames to work correctly and access tables stored in Hive, it is necessary to copy the Hive configuration file `hive-site.xml` to the spark configuration folder:

```
sudo cp /etc/hive/conf.dist/hive-site.xml /usr/lib/spark/conf/
```

Check this is working by opening the PySpark shell and executing:

```
sqlCtx.createDataFrame([("somekey", 1)])
```

This should print logging messages and then:

```
Out[]: DataFrame[_1: string, _2: bigint]
```