**Spark & Cassandra Installation in Ubuntu 16.04**

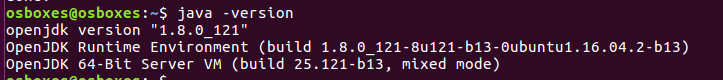
1. We need to make sure Java is installed:

$ java -version

If not, please follow below steps to install java on ubuntu 16.04 & verify.

$ sudo apt-get openjdk-8-jre-headless

$ java –version



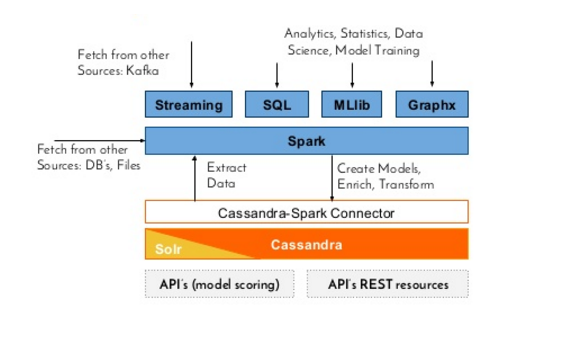
1. We will try to install below tools. Also, note that we're using Spark 2.0 and Scala 2.11 to avoid version mismatch:

Scala 2.11.6

Spark 2.0.2

Cassandra 3.0.2

Basic Pipeline



Apache Cassandra install

$ echo "deb http://www.apache.org/dist/cassandra/debian 36x main" | sudo tee -a /etc/apt/sources.list.d/cassandra.list

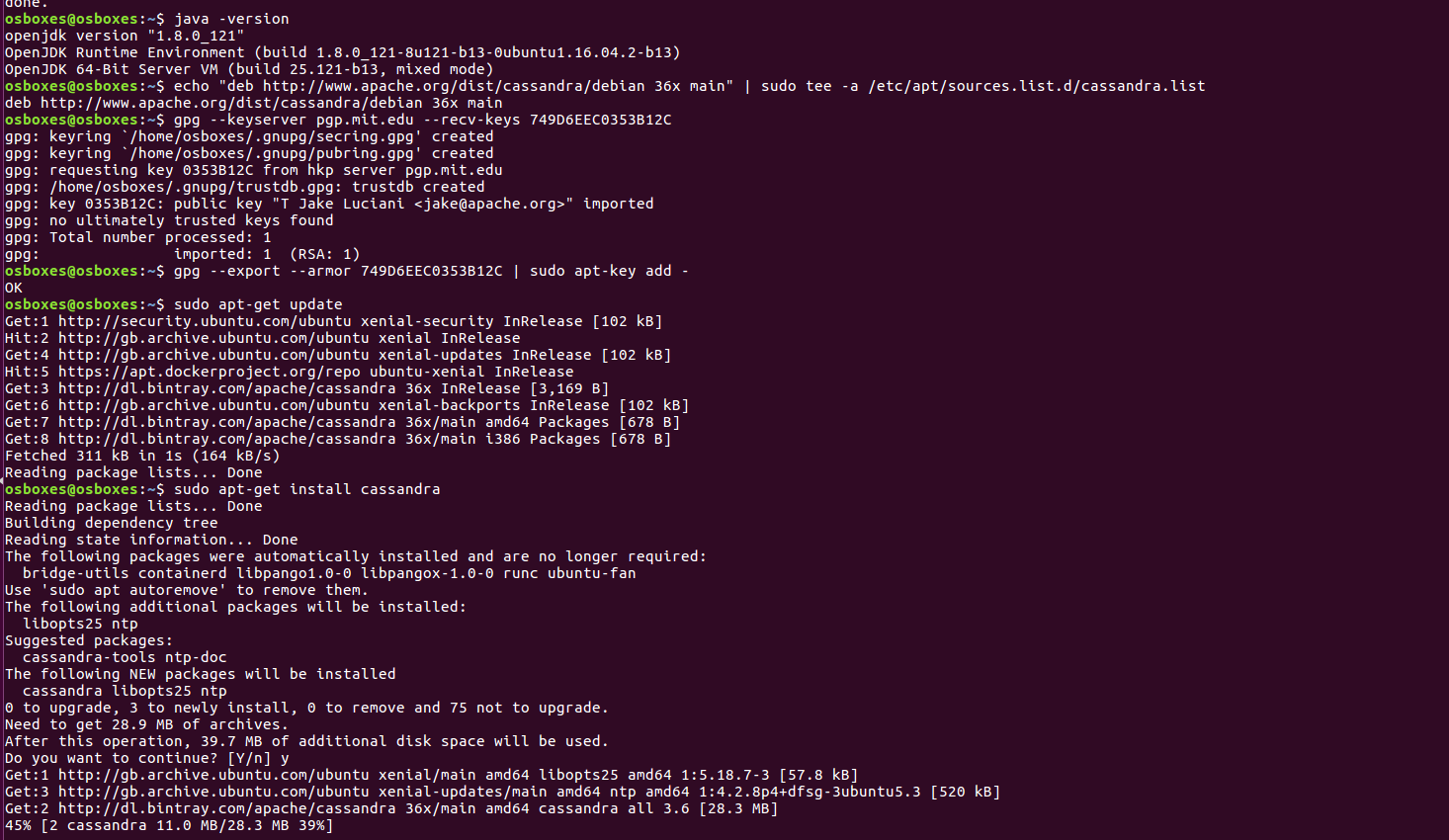
$ gpg --keyserver pgp.mit.edu --recv-keys 749D6EEC0353B12C

$ gpg --export --armor 749D6EEC0353B12C | sudo apt-key add -

$ sudo apt-get update

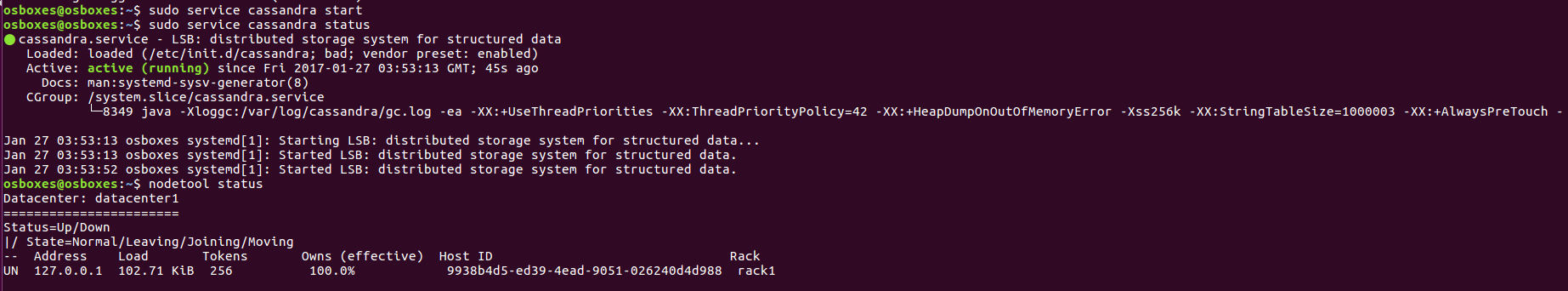
$ sudo apt-get install cassandra

$ sudo service cassandra start



Need to verify whether the Cassandra cluster installed or not.

$ nodetool status



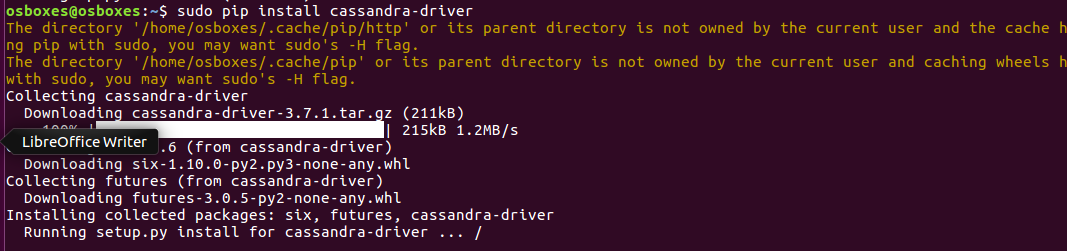
We see that Cassandra is up now and now we will try to execute basic Cassandra queries using cqlsh interface.



To fix the issue, we need to define environment variable CQLSH\_NO\_BUNDLED and export it:

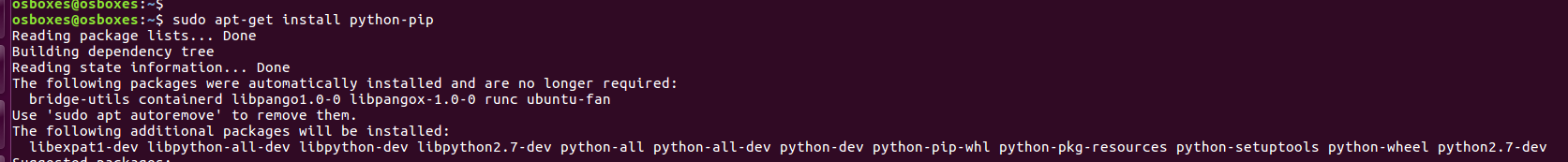
$ sudo pip install cassandra-driver

$ export CQLSH\_NO\_BUNDLED=true

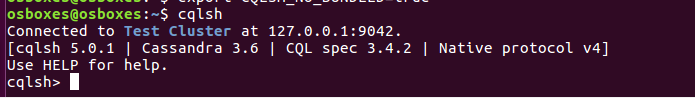




\*If pip is not installed, follow the below steps to install pip and execute above two steps.



We install the latest Python Cassandra driver and tell cqlsh (which is Python program) to use the external Cassandra Python driver, not the one bundled with the distribution.



We created a keyspace "demo" and a table ("user") in that keyspace. Then we stored (name, favorite\_food) pairs and displayed them.

create KEYSPACE demo WITH replication = {'class': 'SimpleStrategy', 'replication\_factor': 1};

use demo ;

Creating another table user:

create TABLE user ( name text PRIMARY KEY, favorite\_food text );

insert into user (name, favorite\_food) values ('jon', 'bacon');

insert into user (name, favorite\_food) values ('luke', 'pie');

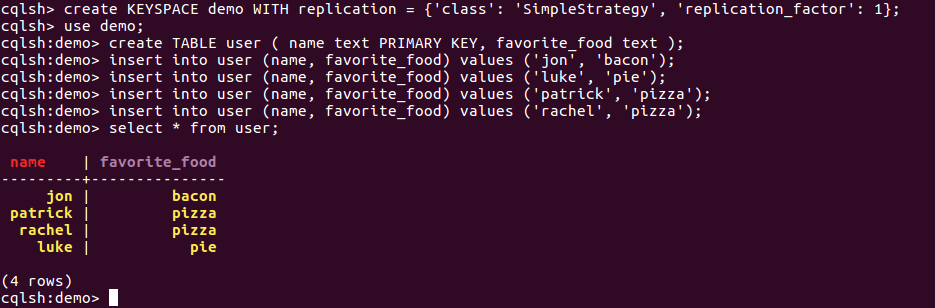
insert into user (name, favorite\_food) values ('patrick', 'pizza');

insert into user (name, favorite\_food) values ('rachel', 'pizza');

Creating another table favorite\_foods:

create TABLE favorite\_foods ( food text, name text,

primary key (food, name));

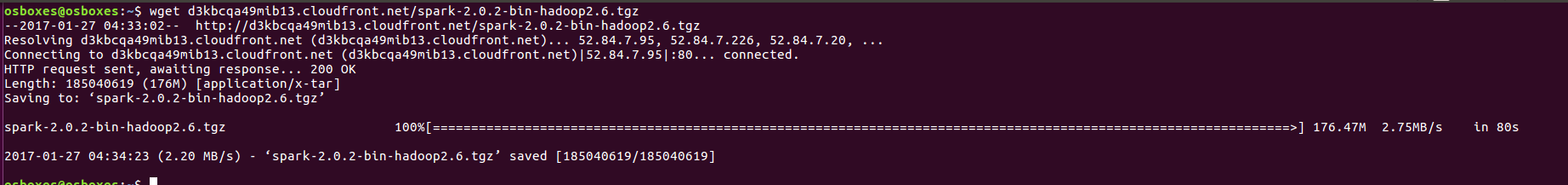


Apache Spark Install

Download the latest pre-built Apache spark version for Hadoop2.6:

$ wget d3kbcqa49mib13.cloudfront.net/spark-2.0.2-bin-hadoop2.6.tgz

$ sudo tar xvzf spark-2.0.2-bin-hadoop2.6.tgz -C /usr/local

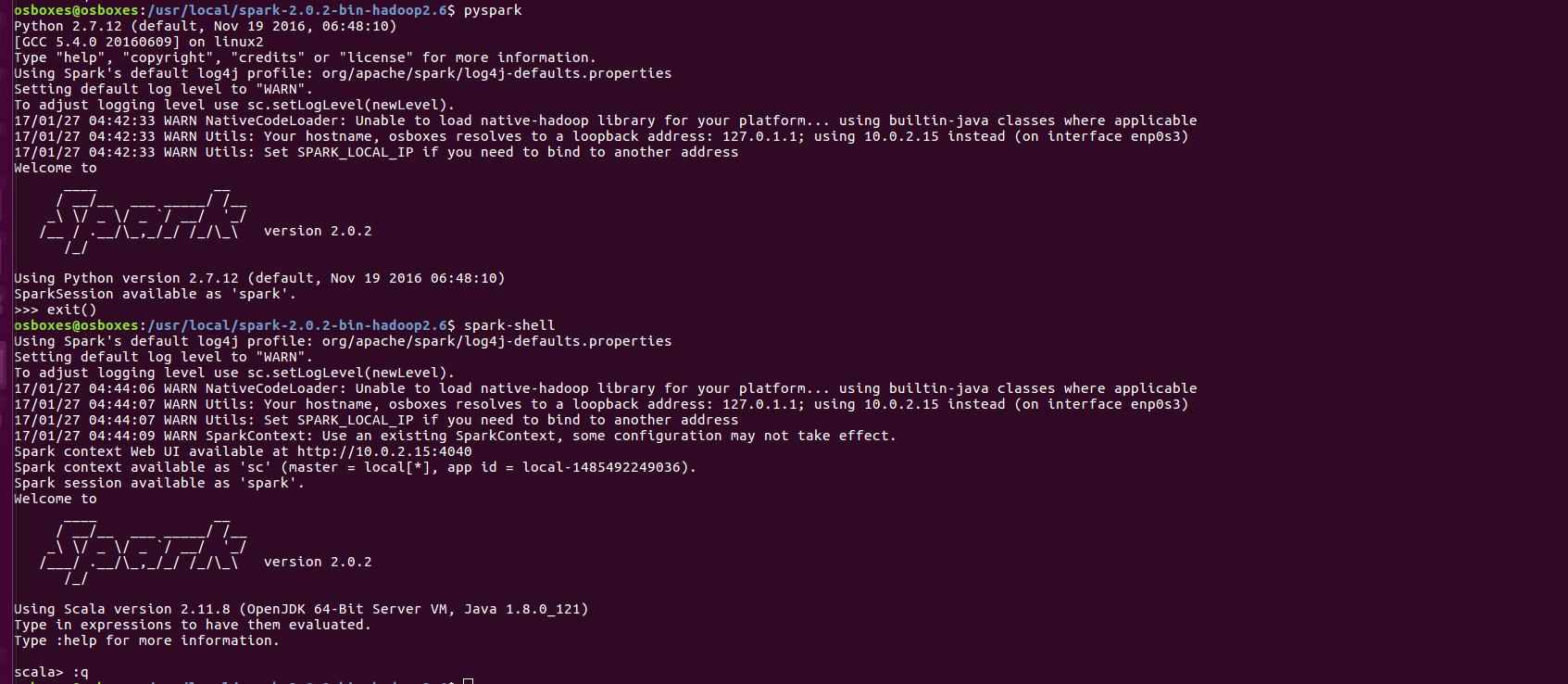




Need to update environmental variables in ~/.bashrc file.

export SPARK\_HOME=/usr/local/spark-2.0.2-bin-hadoop2.6

export PATH=$SPARK\_HOME/bin:$PATH



Let’s write a python program and fetch details from Cassandra table we have created.

**from** pyspark **import** SparkContext, SparkConf  
**from** pyspark.sql **import** SQLContext  
  
conf = SparkConf()\  
 .setAppName(**"Python Script for getting the table ‘favorite foods’"**)\  
 .setMaster(**"local[\*]"**) \  
 .set(**"spark.cassandra.connection.host"**, **"127.0.0.1"**)  
sc = SparkContext(conf=conf)  
sqlContext = SQLContext(sc)  
sqlContext.read.format(**"org.apache.spark.sql.cassandra"**).options(table=**"user"**, keyspace=**"demo"**).load().show()

Save the above code in sparkcassandra.py and execute below command

$ spark-submit --packages datastax:spark-cassandra-connector:2.0.0-M2-s\_2.11 sparkcassandra.py

