**DOCKER IMAGE TO RUN APACHE HIVE**

Docker is an easy-to deploy environment for building, assembling and shipping applications. Docker containers wrap a piece of software in a complete filesystem that contains everything needed to run: code, runtime, system tools, system libraries – anything that can be installed on a server. This guarantees that the software will always run the same, regardless of its environment. We can simply use Docker Images for building and running Apache Hadoop and Apache Hive on HDFS. Below is the sample example for the same.

**Steps:**

1. Pull the docker image which has the setup for Apache Hadoop and Apache Hive. For instance:

docker pull cdecl/hadoop-hive:2.6.0

The above image has the following installed:

* Hadoop 2.6.0
* Hive 1.0
* Ubuntu 14.10

1. Start the Docker container using the below command

docker run -i -t cdecl/hadoop-hive:2.6.0 /etc/bootstrap.sh -bash

1. Update the packages and install Git and Maven using the below commands

apt-get update

apt-get install git

apt-get install maven

1. Clone Beetest from Github. Beetest is a simple utility for testing of Apache Hive scripts locally

Find more details regarding Beetest from the following link:

<https://github.com/kawaa/Beetest>

git clone https://github.com/kawaa/Beetest.git

cd Beetest

mvn -P full package

1. To execute the hive commands we would need to navigate to /usr/local/hive/bin folder and then execute the command ./hive. This tells your shell to look into the local directory instead of using the $PATH. To permanently add Hive to the $PATH, run the following command

Note: The following must be the HADOOP\_HOME and HIVE\_HOME variable values. You can check them by using the following commands

echo $HADOOP\_HOME

/usr/local/hadoop

echo $HIVE\_HOME

/usr/local/hive

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

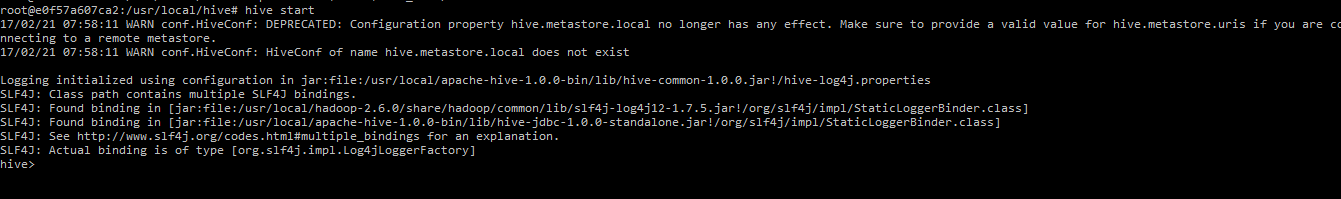
1. Start MySQL (Hive’s metastore backend) by using the following command

service mysql start

1. Start Hive using the following command

hive start

On successfully starting Hive, the following should be displayed in the terminal



1. Create a Hive table and load the data into it from a CSV file

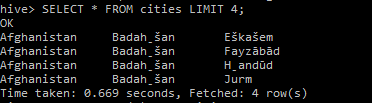
CREATE TABLE cities (country string, r1 string, r2 string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA LOCAL INPATH '/root/data/cities.csv' OVERWRITE INTO TABLE cities;

1. Select data from cities table using the following command

SELECT \* FROM cities LIMIT 4;

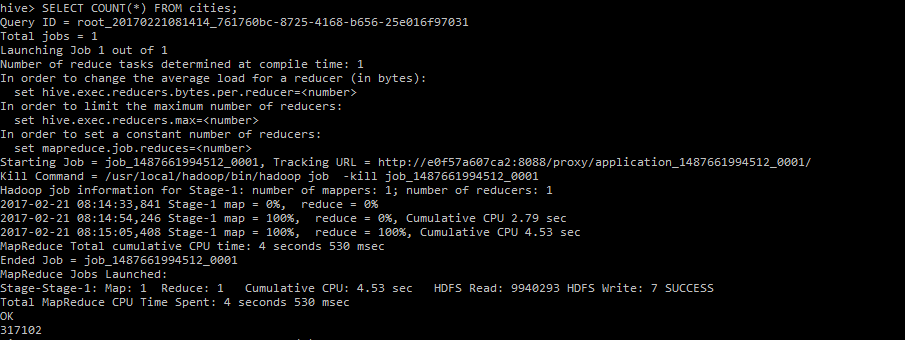
The following must be the output:



1. Count how many rows we have in this table/file

SELECT COUNT(\*) FROM cities;

The following must be the output. You can see the Map Reduce jobs running beneath which computes the result for us.



1. Get the count of instances for every Country. We can perform this by using the group by function with the count aggregate function

SELECT country, COUNT(\*) cnt FROM cities GROUP BY country;

The following must be the output. You can see the Map Reduce jobs running beneath which computes the result for us.

