**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 5 (Team)**

**Part 2 sandbox-hdp**

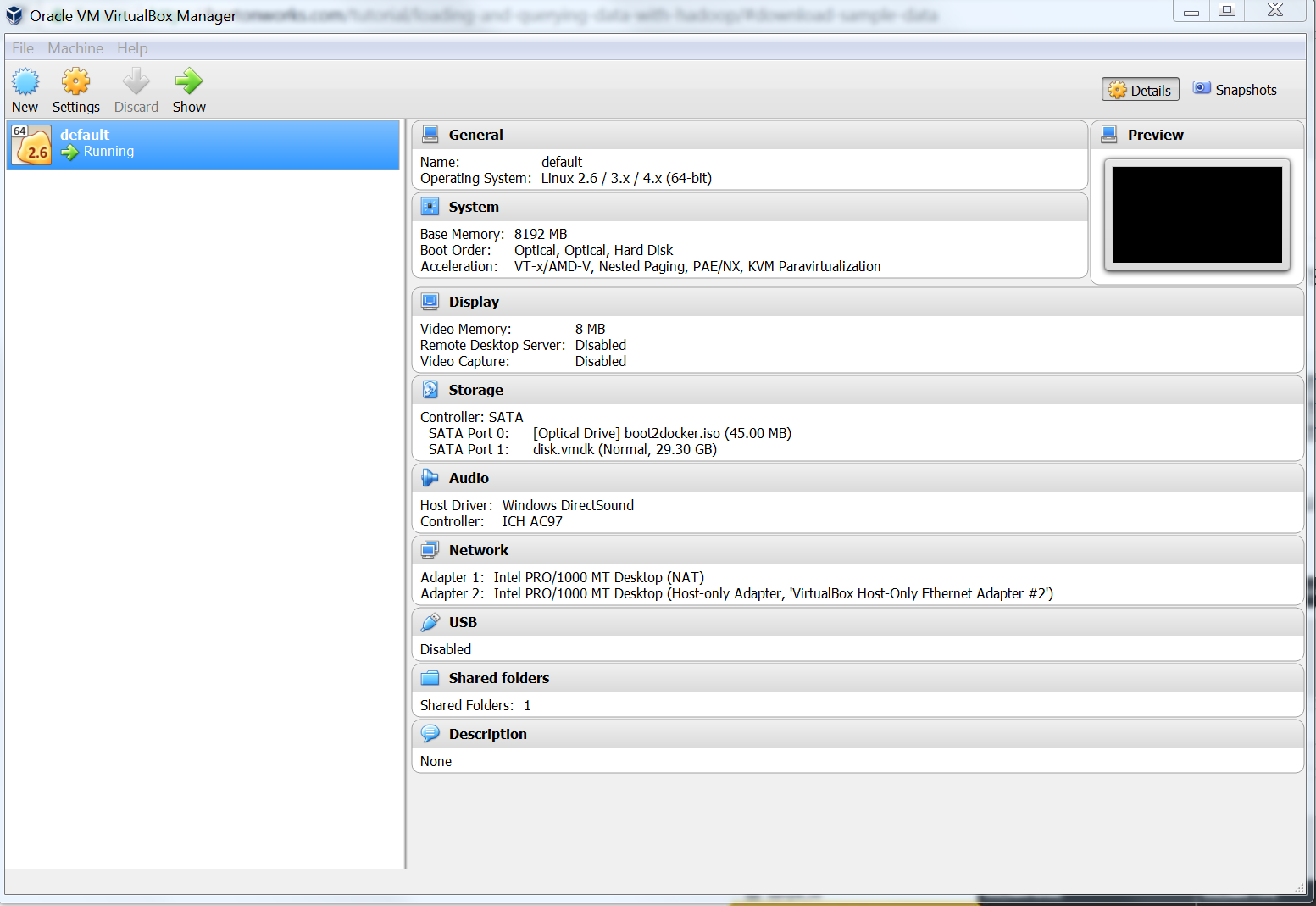
I downloaded and installed Docker-Toolbox that depends on VirtualBox since I have Windows7.

1. Remove old Docker machine “Default”.

*docker-machine rm default*

1. Create a new Docker machine using VirtualBox set memory to 8G and Disk about 30G.

*docker-machine create -d virtualbox --virtualbox-memory 8192 --virtualbox-disk-size 30000 default*



1. Generate new SSL Certs this allows Docker to run in cmd.exe. This is required to start Docker in cmd. See Reference [1].

*docker-machine regenerate-certs*

*docker-machine env --shell cmd default*

*@FOR /f "tokens=\*" %i IN ('docker-machine env --shell cmd default') DO @%i*

1. Verifies IP address of VitualBox is 192.168.99.100. This will be used to login to HortonWorks Sandbox.

*PS C:\Users\lj015625\Desktop\Big Data\Week5> docker-machine ls*

*NAME ACTIVE DRIVER STATE URL SWARM DO*

*CKER ERRORS*

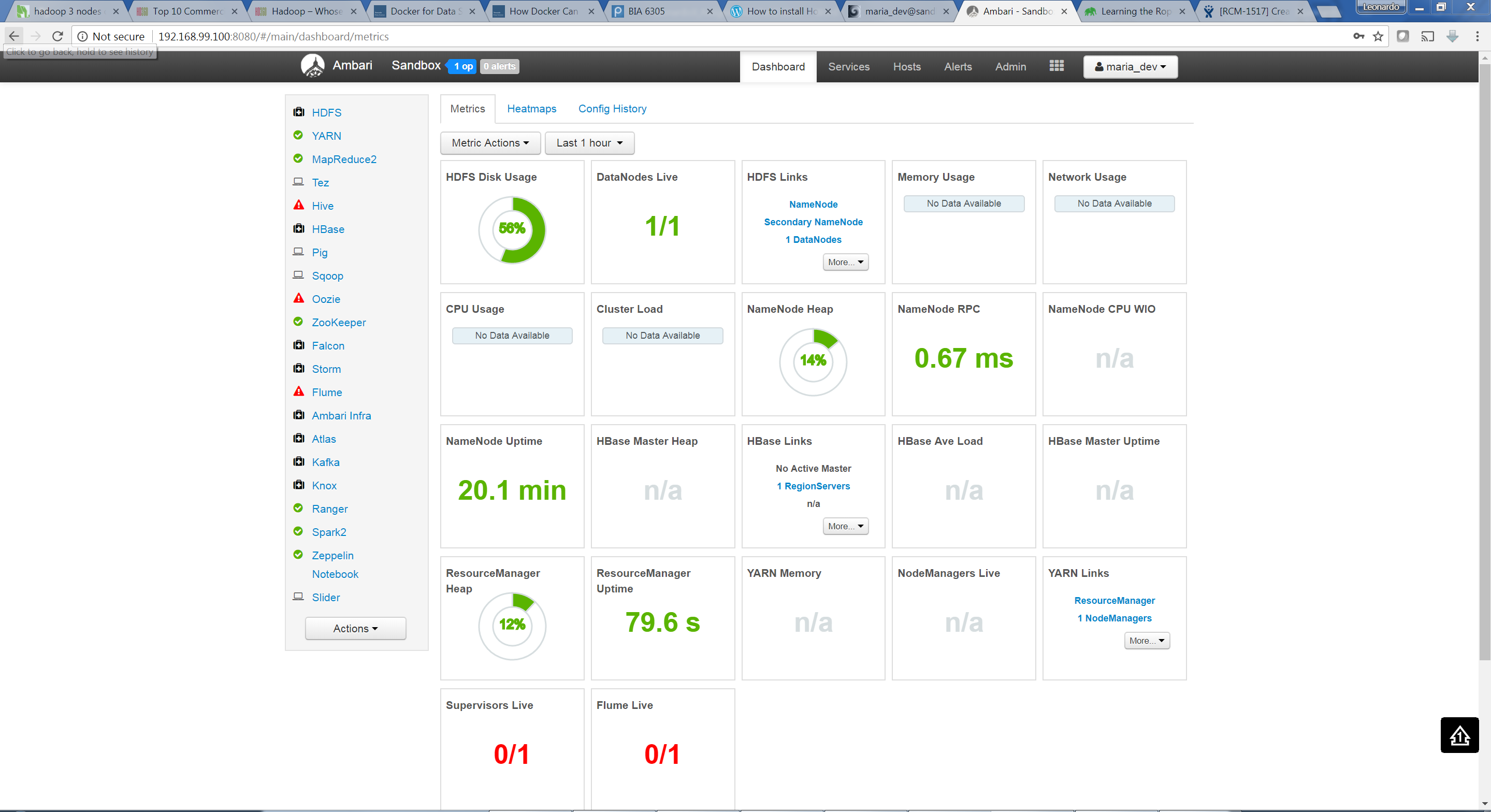
*default \* virtualbox Running tcp://192.168.99.100:2376 v1*

*8.02.0-ce*

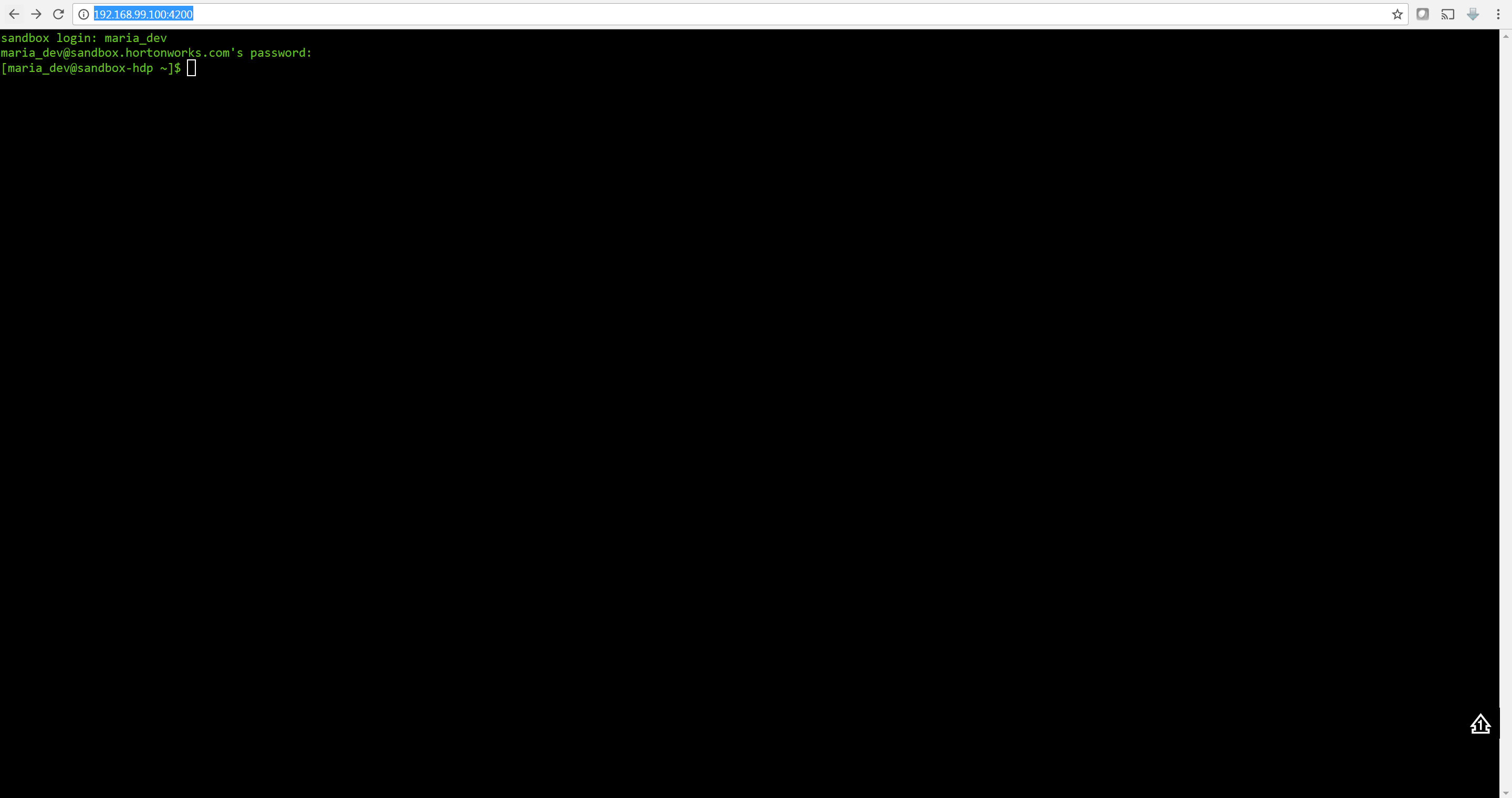
1. Loads the HDP Docker image.

*docker load -i HDP\_2.6.3\_docker\_10\_11\_2017.tar*

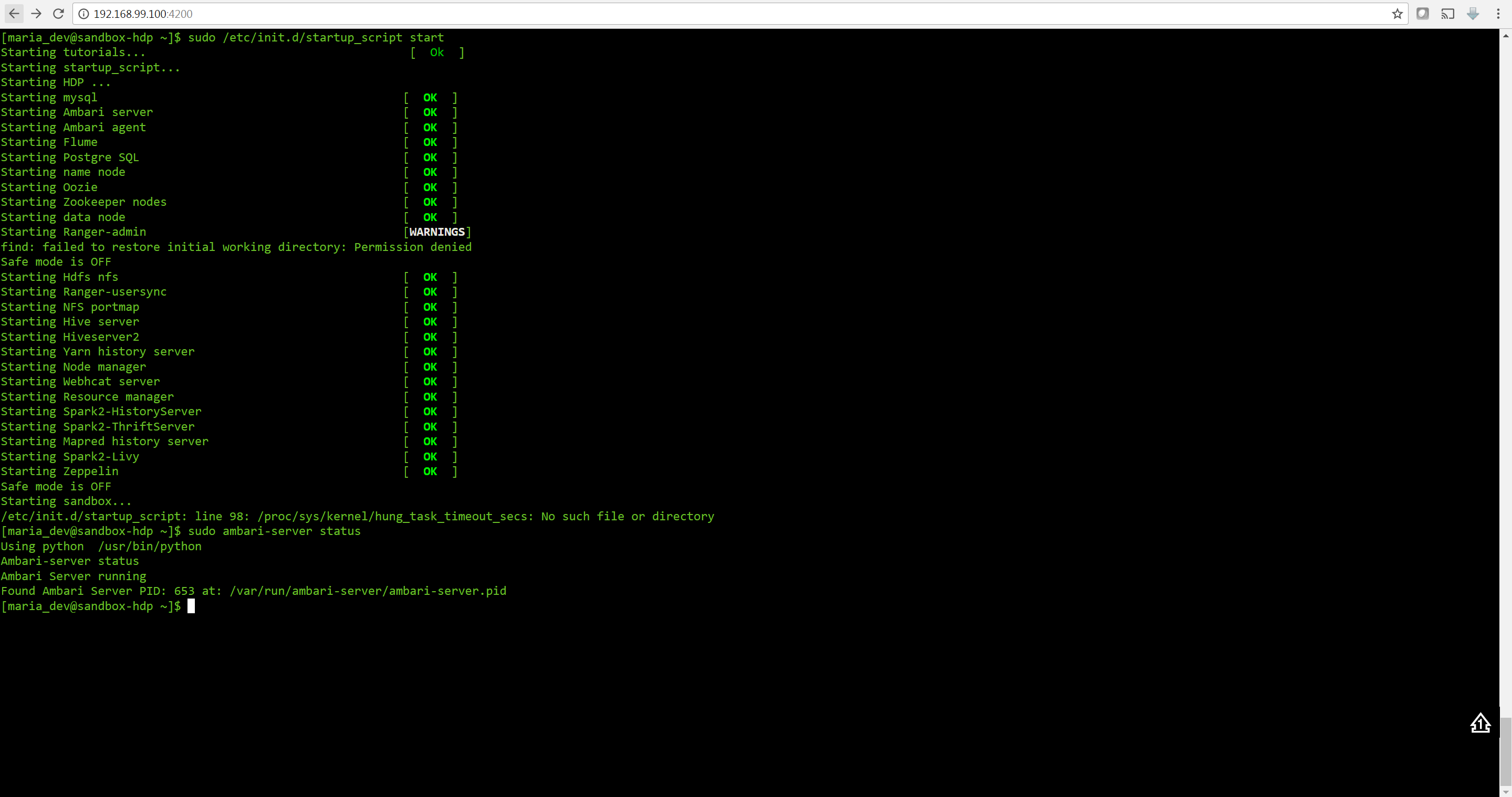
1. Execute the start\_sandbox-hdp script. docker run, docker start…etc. This script also starts mysqld, postgresql, ambari-server, ambari-agent, shellinaboxd, execute put requests to ambari, start tutorial.sh script. See reference [2].
2. Login to Web Dashboard (<http://192.168.99.100:8080/#/main/dashboard/metrics>) with maria\_dev/maria\_dev.



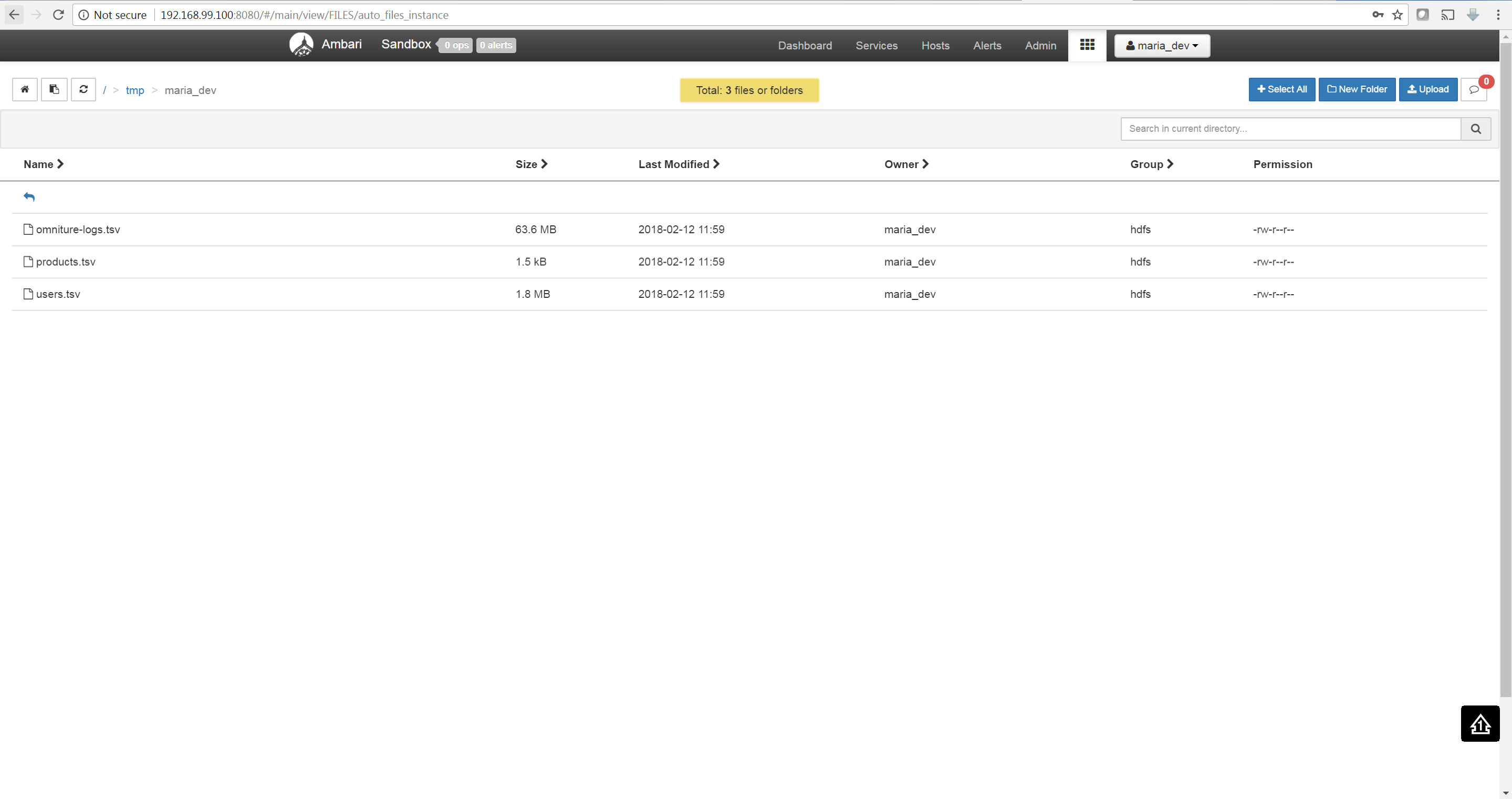
1. Login to Sandbox (<http://192.168.99.100:4200/>) with root/Hadoop. See reference [3].

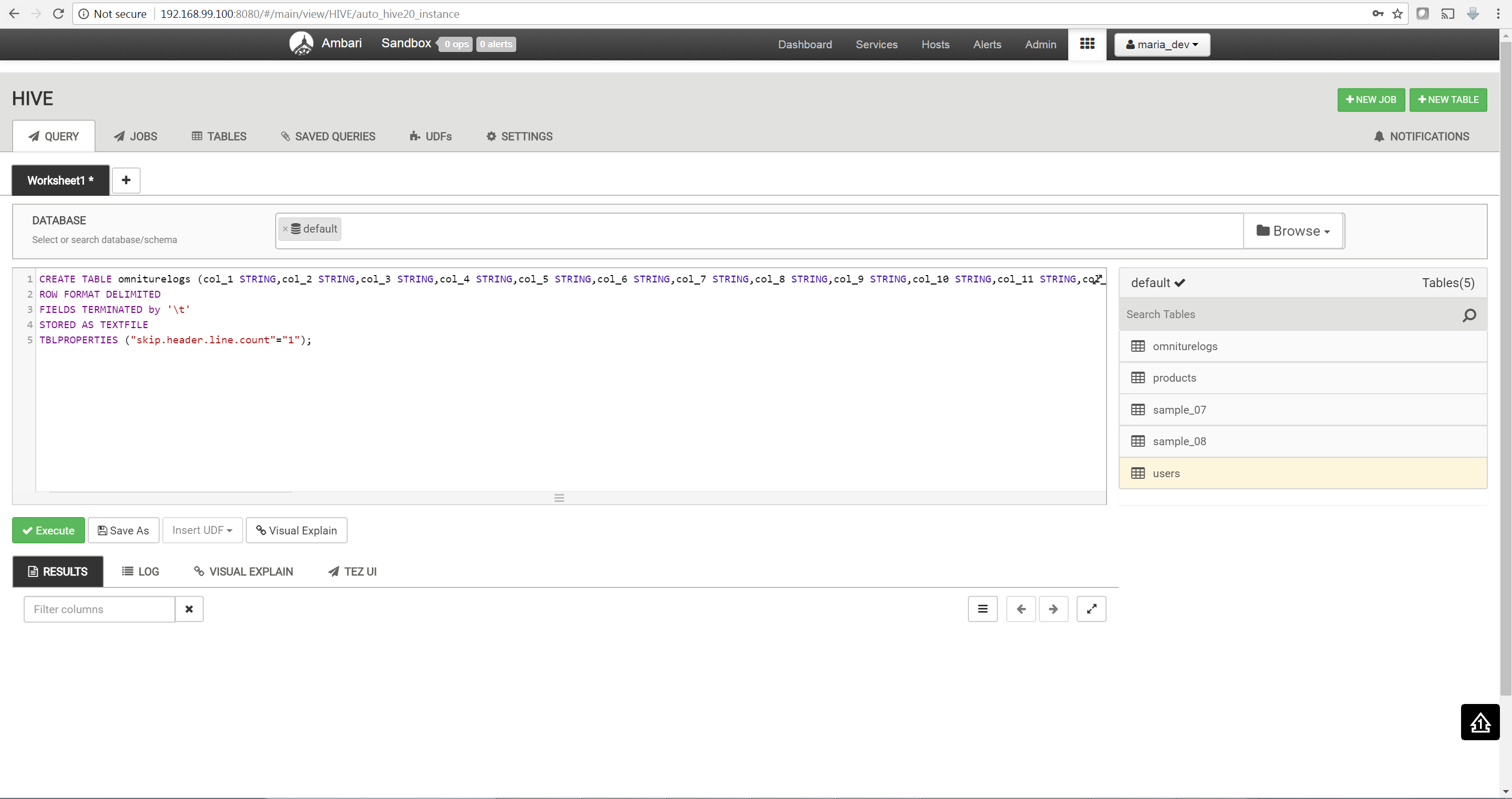


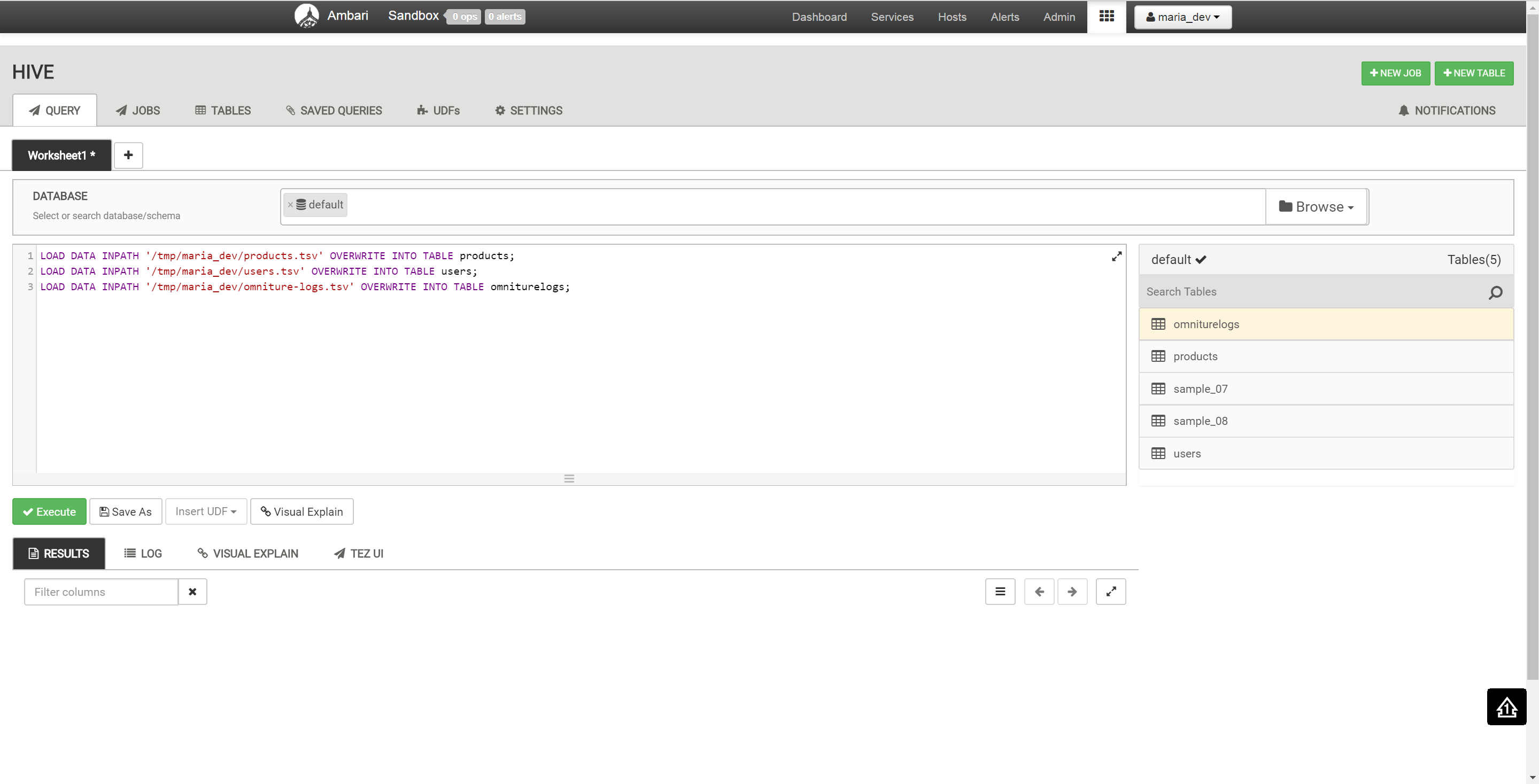
1. Verifies ambari is running. If not starts it.

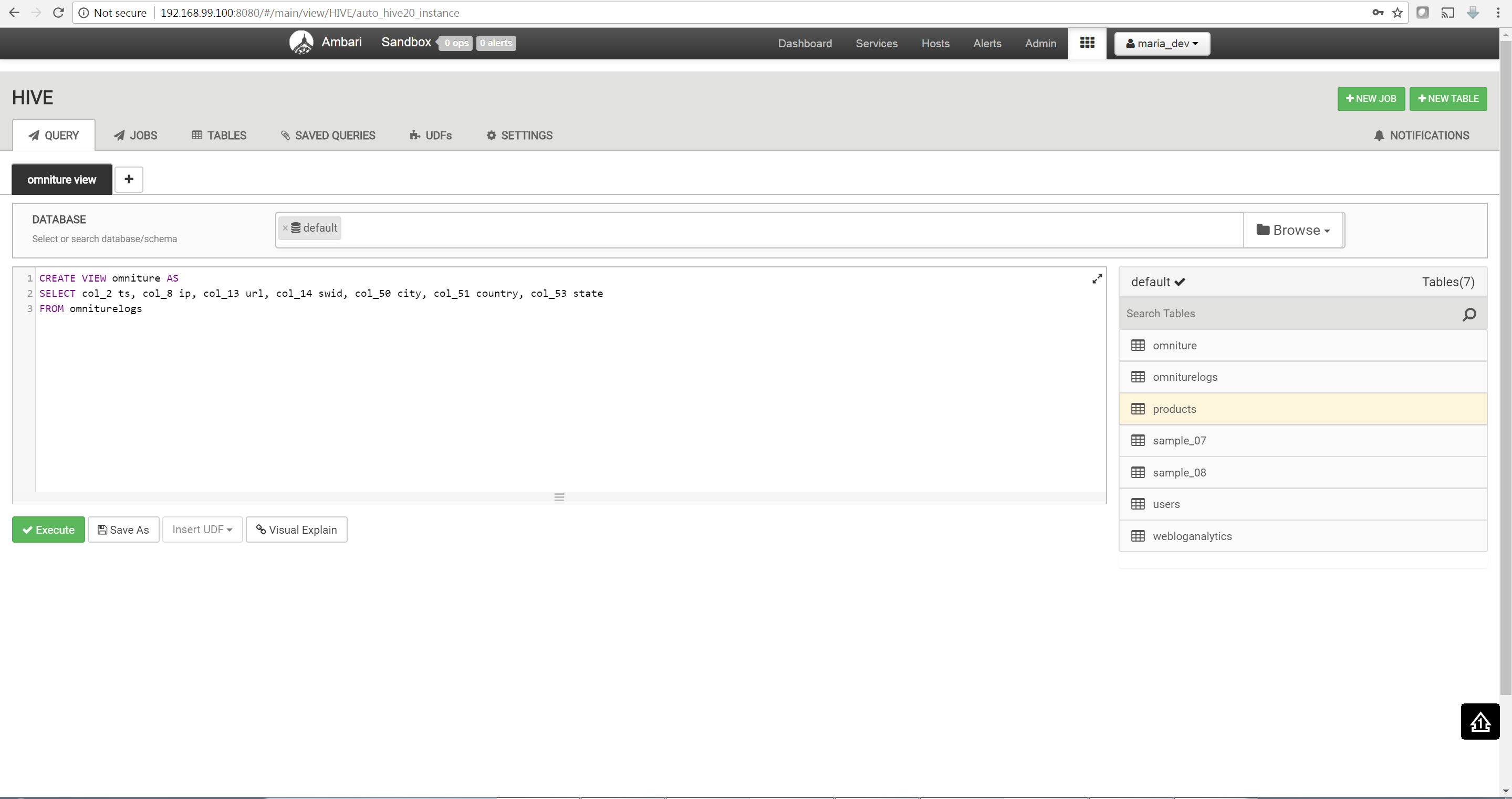


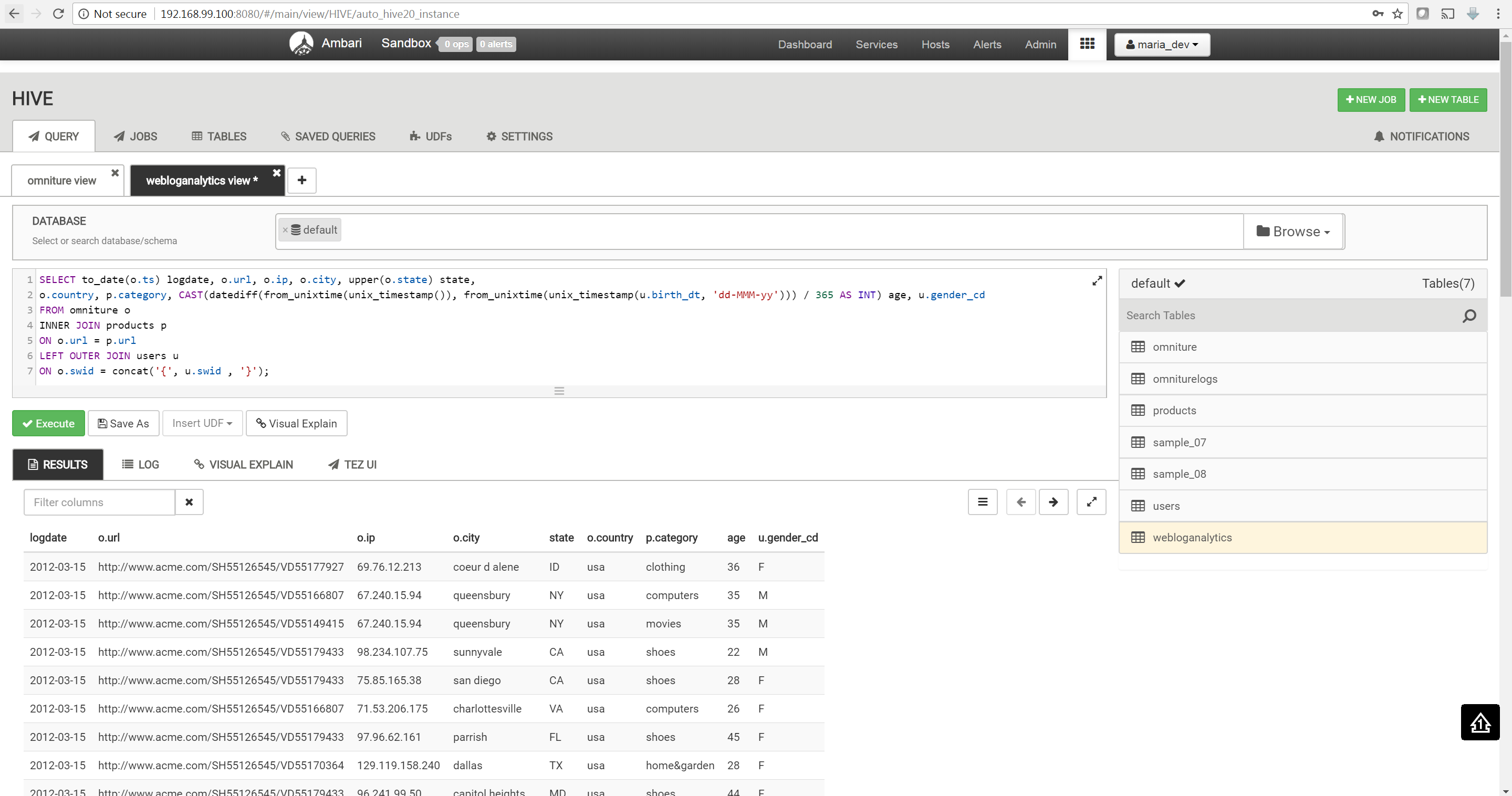
1. Tutorial retail-store-logs-sample-data.zip. [4].



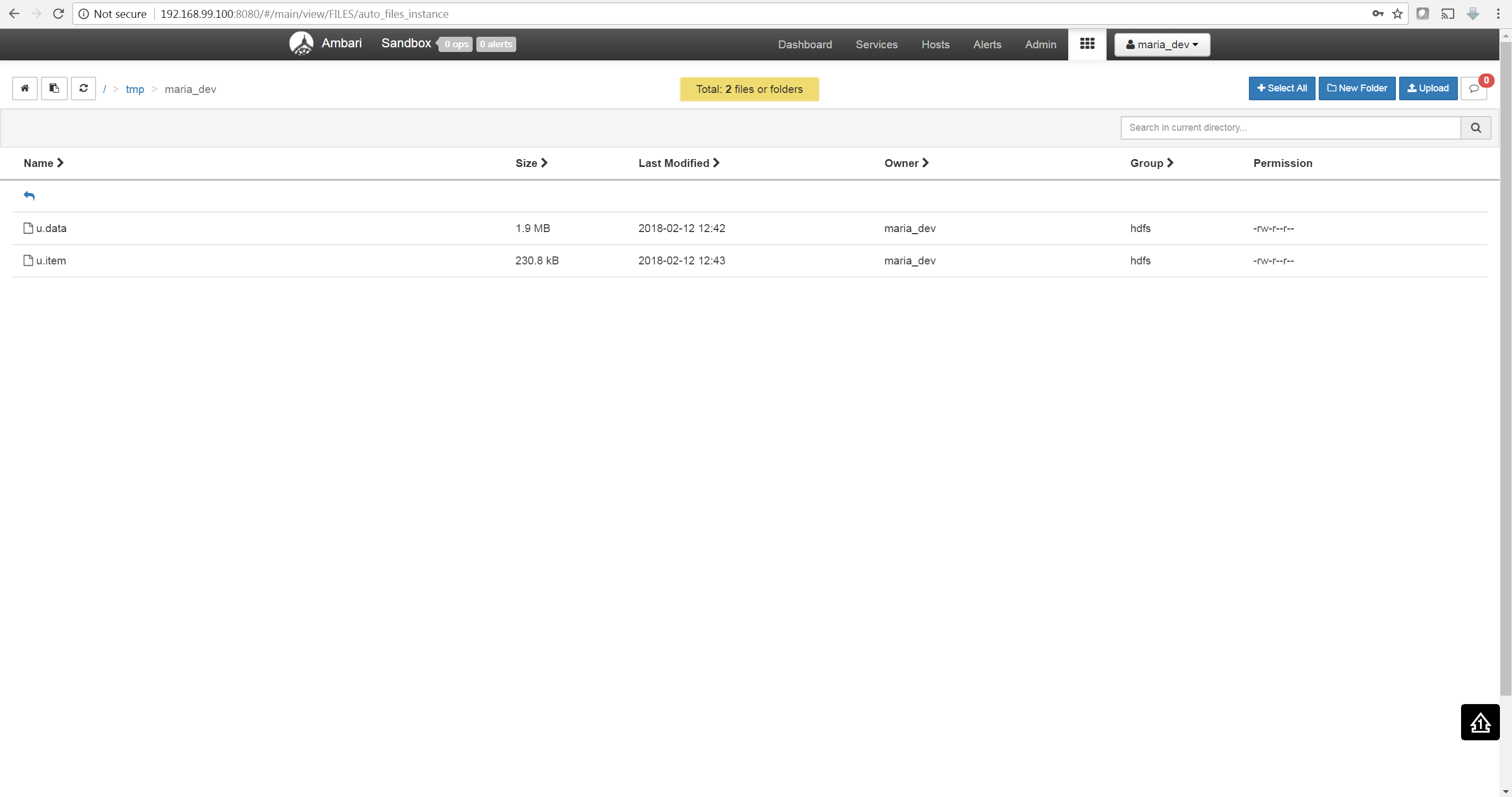








1. Movie review dataset Hive [5].



CREATE TABLE data (userId INT, movieId INT, rating INT, ratingtime INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

TBLPROPERTIES ("skip.header.line.count"="1");

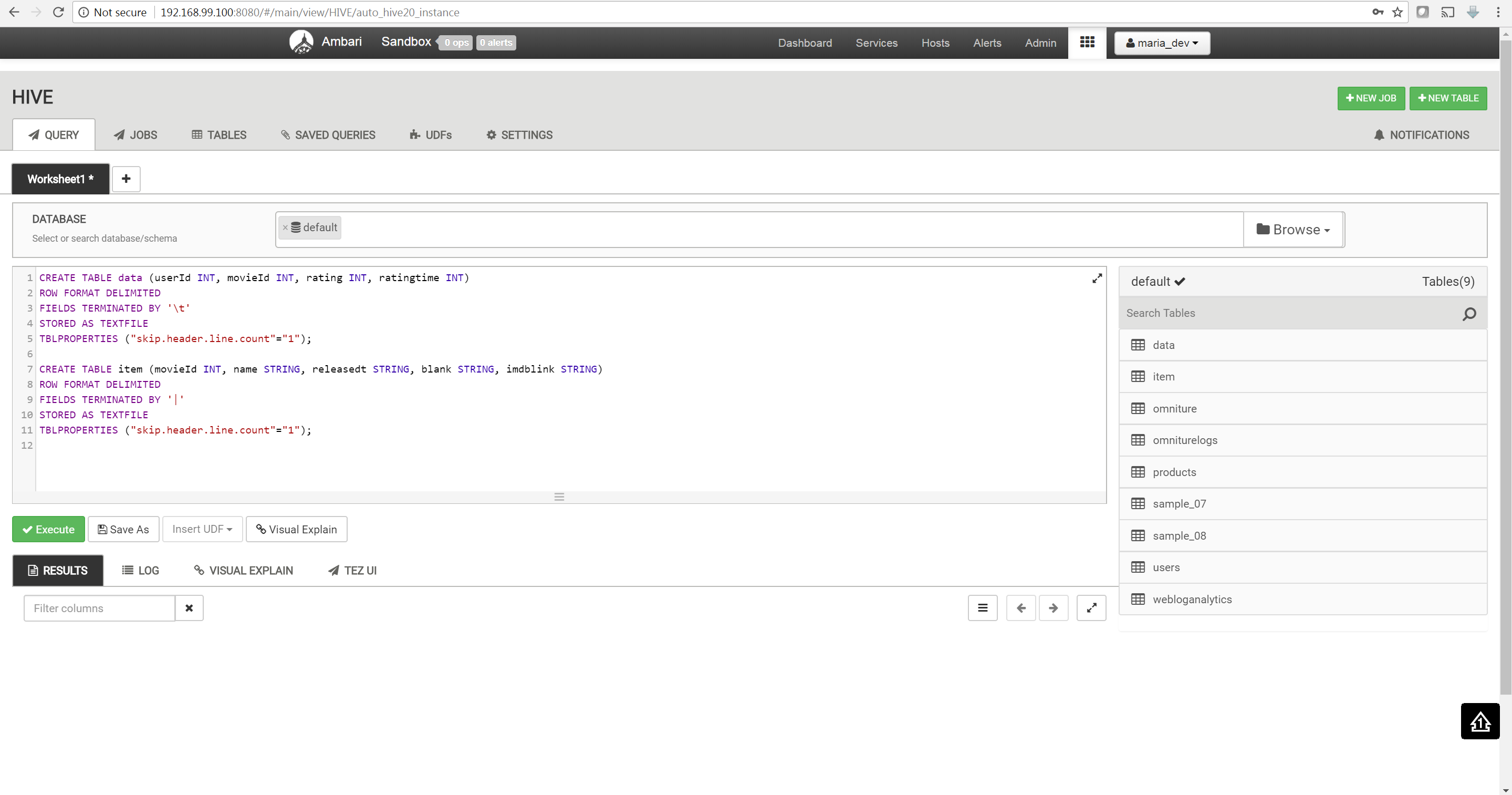
CREATE TABLE item (movieId INT, name STRING, releasedt STRING, blank STRING, imdblink STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '|'

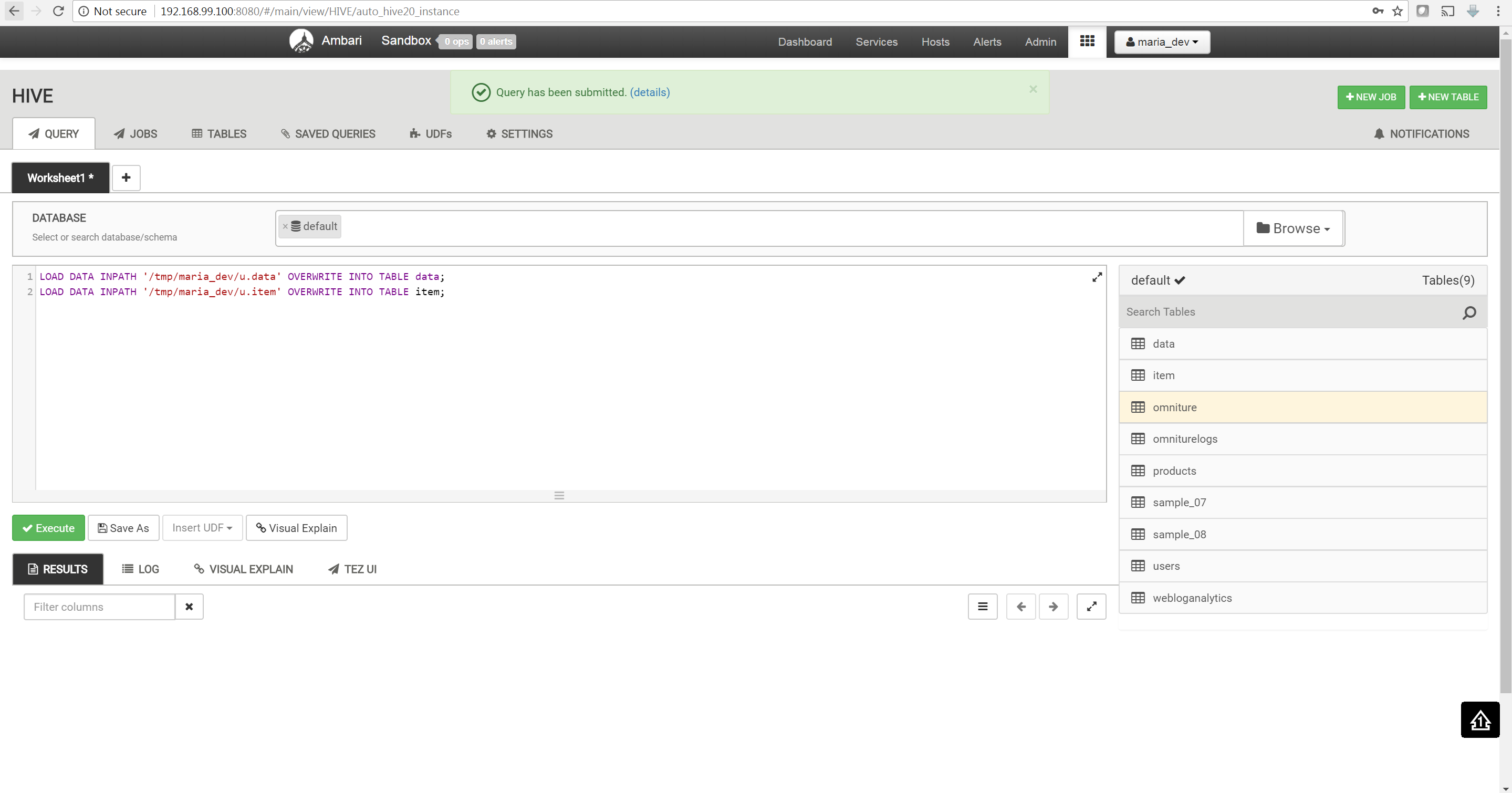
STORED AS TEXTFILE

TBLPROPERTIES ("skip.header.line.count"="1");



LOAD DATA INPATH '/tmp/maria\_dev/u.data' OVERWRITE INTO TABLE data;

LOAD DATA INPATH '/tmp/maria\_dev/u.item' OVERWRITE INTO TABLE item;

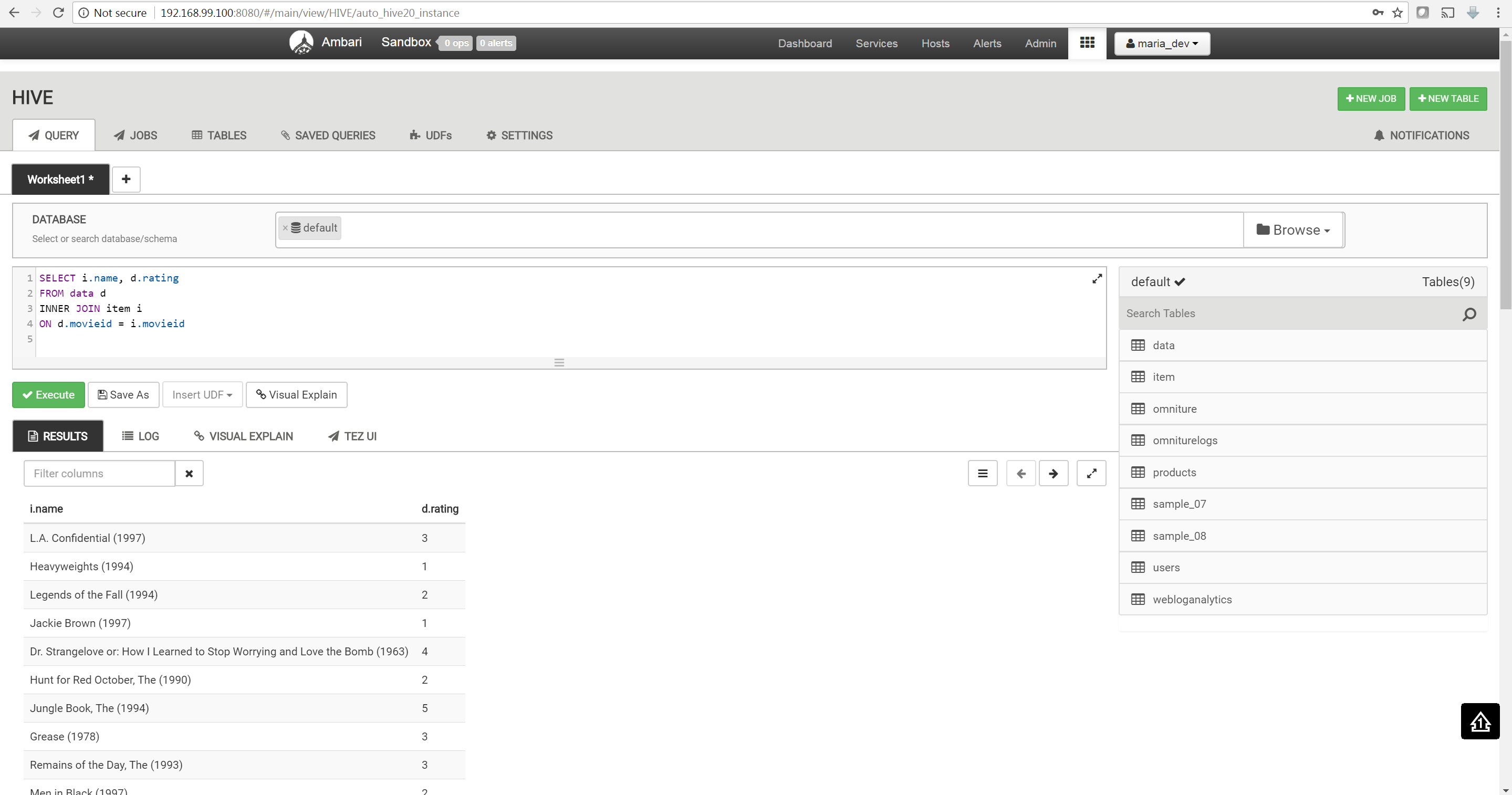


SELECT i.name, d.rating

FROM data d

INNER JOIN item i

ON d.movieid = i.movieid



DROP VIEW movie\_rating;

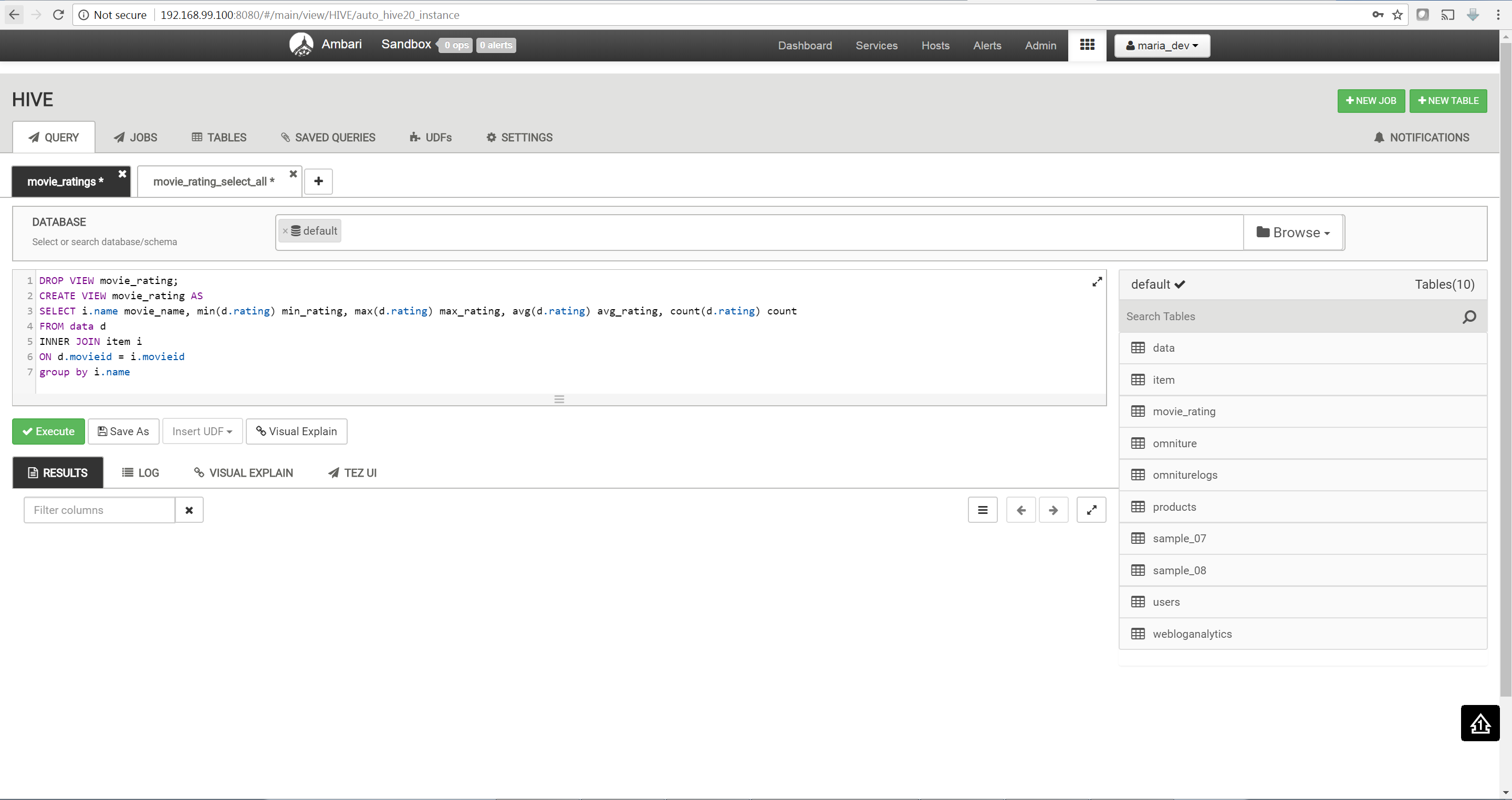
CREATE VIEW movie\_rating AS

SELECT i.name movie\_name, min(d.rating) min\_rating, max(d.rating) max\_rating, avg(d.rating) avg\_rating, count(d.rating) count

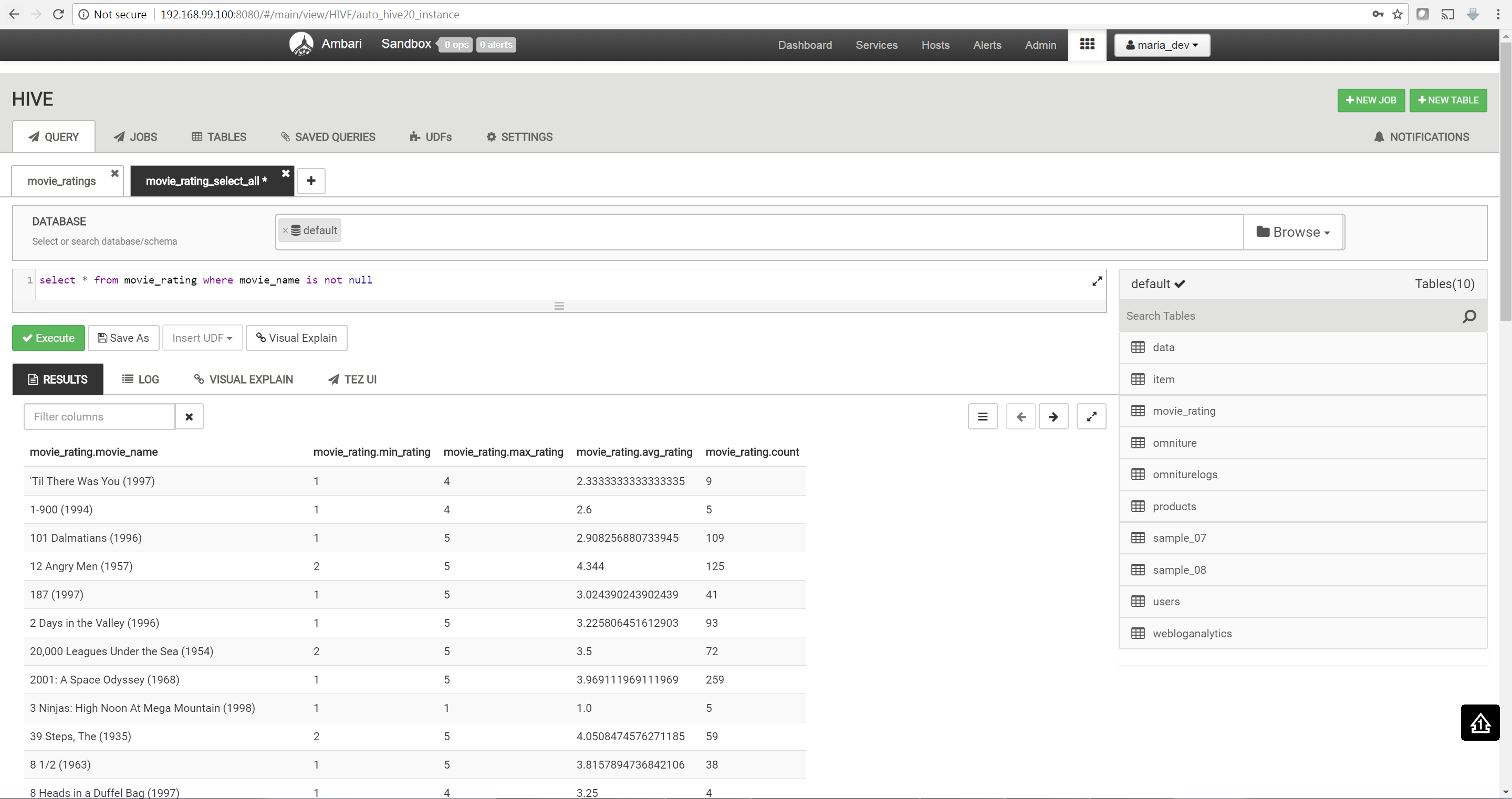
FROM data d

INNER JOIN item i

ON d.movieid = i.movieid

group by i.name 

select \* from movie\_rating where movie\_name is not null



**Reference:**

[1] <https://github.com/docker/toolbox/issues/636>

[2] <https://mohdnaeem.wordpress.com/2018/01/27/how-to-install-hortonworks-sandbox-using-docker/>

[3] <https://hortonworks.com/tutorial/learning-the-ropes-of-the-hortonworks-sandbox/#terminal-access>

[4] <https://hortonworks.com/tutorial/loading-and-querying-data-with-hadoop/#download-sample-data>

[5] <https://grouplens.org/datasets/movielens/>