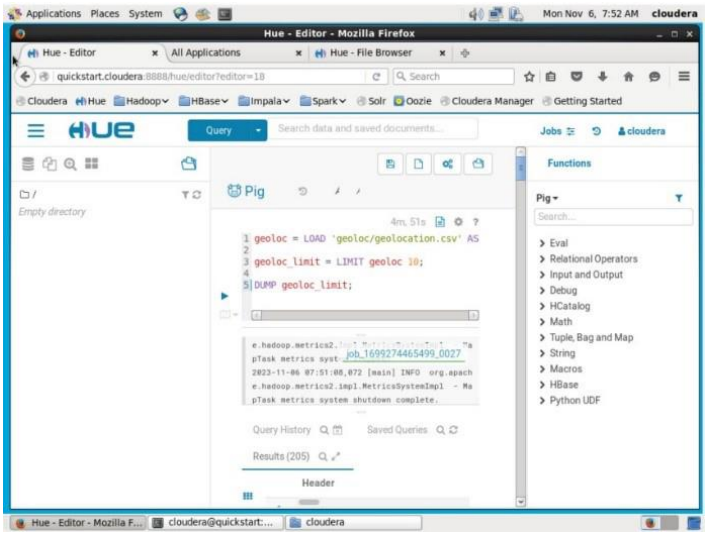
Самостоятельная работа 3.2

Pig

1. Запустите следующий скрипт/команды, чтобы загрузить и отобразить первые десять строк из файла геолокации:

```
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid,driverid,event,latitude,longitude,city,state,velocity,event_ind,idling_ind);
geoloc_limit = LIMIT geoloc 10;
DUMP geoloc_limit;
```



2. Посчитать статистику по этому файлу.

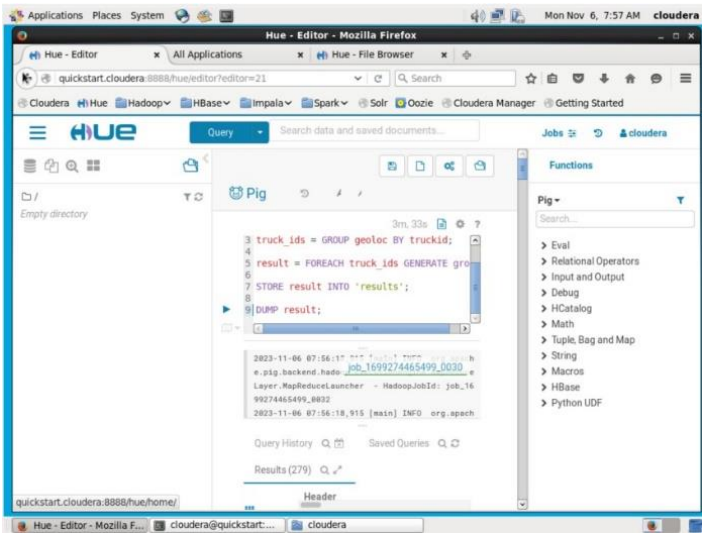
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:chararray, driverid:chararray, event:chararray, latitude:double, longitude:double, city:chararray, state:chararray, velocity:double, event_ind:long, idling_ind:long);

truck_ids = GROUP geoloc BY truckid;

result = FOREACH truck_ids GENERATE group AS truckid, COUNT(geoloc) as count;

STORE result INTO 'results';

DUMP result;



Проверьте папку «results», хранящуюся в HDFS, по строке STORE result. Что вы можете сказать по сравнению с количеством слов MapReduce?

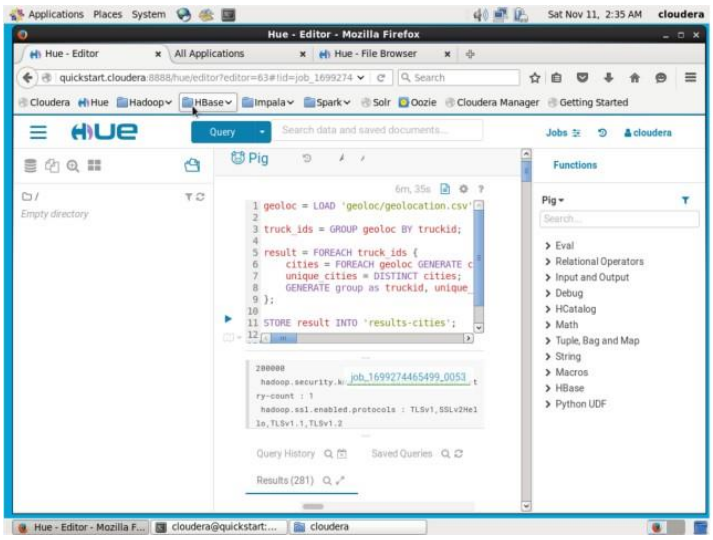
MapReduce принял пробел за разделить и посчитал количество строк, разделенных пробелами. А Pig посчитал количество записей, которые содержат определенный truckID (80).

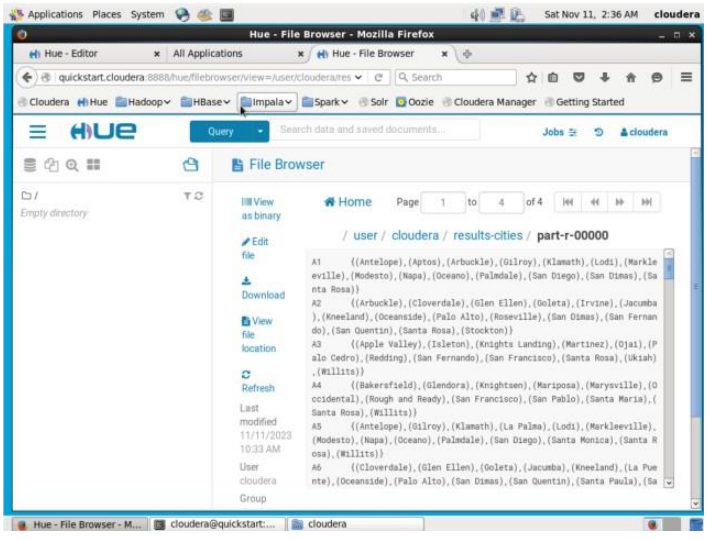
Также просмотрите журналы на новой вкладке Hadoop > YARN Resource Manager в Firefox. Поясните список логов в журнале.

Появилось 3 лога: лог задачи, ставящей Pig скрипты в очередь выполнения, и 2 лога задач, выполняющих скрипты.

Можете ли вы подсчитать список различных городов, посещенных каждым идентификатором грузовика, и среднюю скорость для каждого идентификатора грузовика?

```
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:chararray, driverid:chararray, event:chararray, latitude:double, longitude:double, city:chararray, state:chararray, velocity:double, event_ind:long, idling_ind:long);
truck_ids = GROUP geoloc BY truckid;
result = FOREACH truck_ids {
   cities = FOREACH geoloc GENERATE city;
   unique_cities = DISTINCT cities;
   GENERATE group as truckid, unique_cities;
};
STORE result INTO 'results-cities';
DUMP result;
```





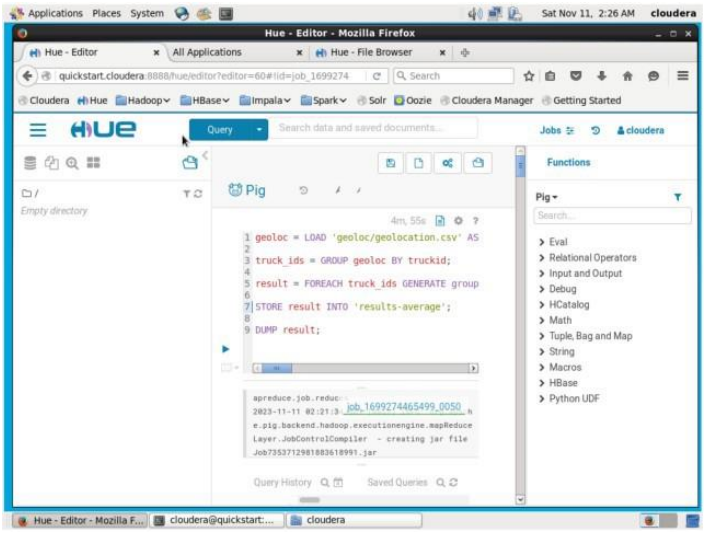
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:chararray, driverid:chararray, event:chararray, latitude:double, longitude:double, city:chararray, state:chararray, velocity:double, event_ind:long, idling_ind:long);

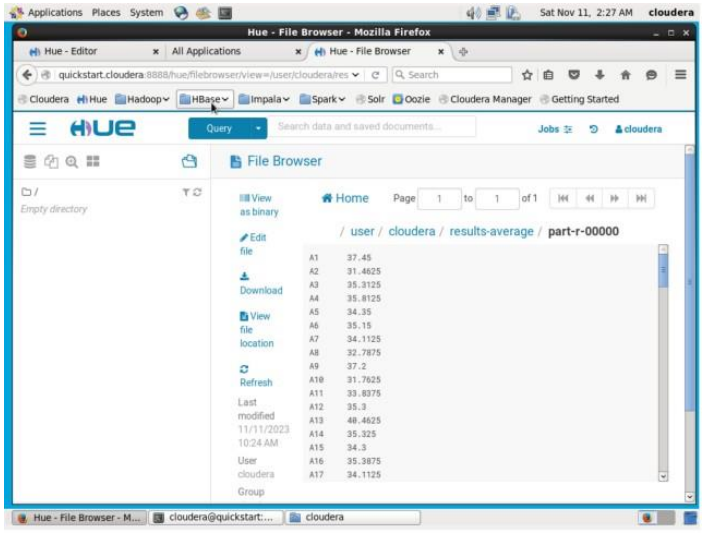
truck_ids = GROUP geoloc BY truckid;

result = FOREACH truck_ids GENERATE group as truckid, AVG(geoloc.velocity) as avg_velocity;

STORE result INTO 'results-average';

DUMP result;

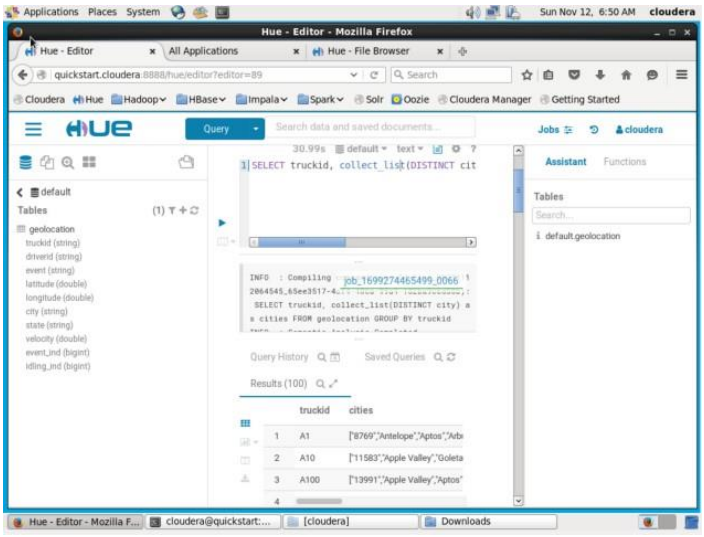




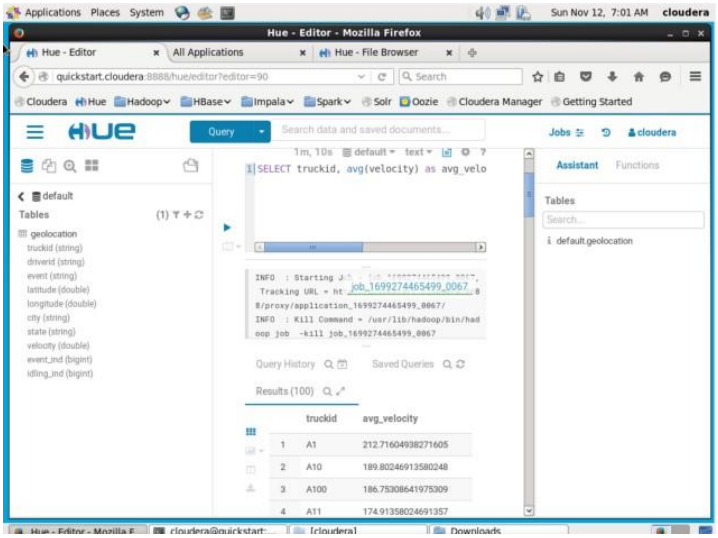
Hive

Можете ли вы снова подсчитать список различных городов, посещенных каждым идентификатором грузовика, и среднюю скорость для каждого идентификатора грузовика?

SELECT truckid, collect_list(DISTINCT city) as cities FROM geolocation GROUP BY truckid;



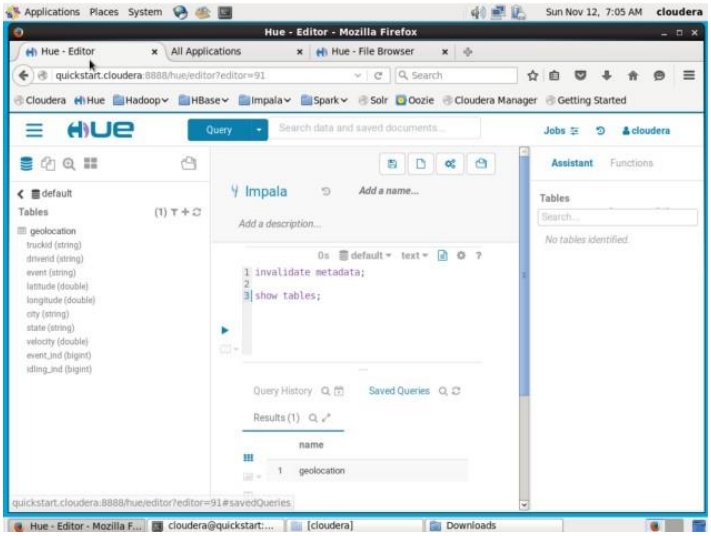
SELECT truckid, avg(velocity) as avg_velocity FROM geolocation GROUP BY truckid;



Impala

1. Перейдите в редактор Impala вместо редактора Hive и запустите

invalidate metadata;
show tables;

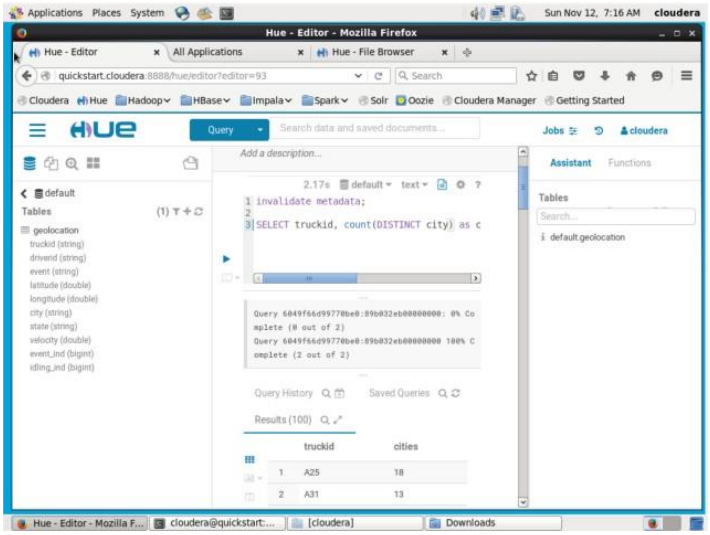


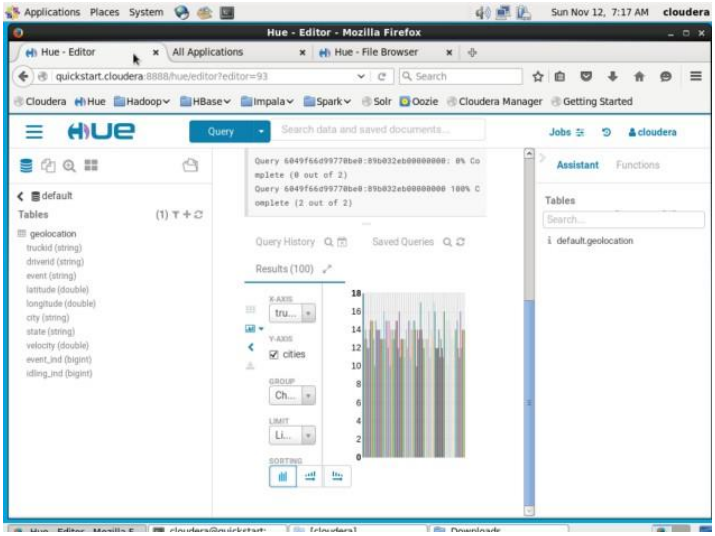
Перезапустите один из сценариев SQL, которые запускали ранее. Заметили разницу в скорости?

Быстрее.

2. Подсчитайте количество геолокаций для каждого грузовика и отобразите его на гистограмме.

SELECT truckid, count(DISTINCT city) as cities FROM geolocation GROUP BY truckid;





3. Выберите грузовик A80 и нанесите его координаты геолокации в редакторе Impala.

SELECT latitude, longitude from geolocation WHERE truckid = "A80";