Кузьменко Сергей БД-231м

3-2.1 Скачать архив Geolocation github zip или Geolocation data из Cloudera.

Создать каталог ex_3_2:

```
[cloudera@quickstart ~]$ mkdir ex_3_2
[cloudera@quickstart ~]$ ls
201402_babs_open_data      cloudera-manager  eclipse                     kerberos       __MACOSX  Templates
201408_babs_open_data      cm_api.py         enterprise-deployment.json  kuzmenko       Music     trucks.csv
3.1.12__kuzmenko_SV.txt    Desktop           ex_3_2                      kuzmenko.gz    parcels   Videos
athlete.snappy.parquet     Documents         express-deployment.json     kuzmenko.txt   Pictures  workspace
babs_open_data_year_1.zip  Downloads         JrboaizPXSh0Mg              lib            Public
[cloudera@quickstart ~]$
```

Пререйти в каталог ex_3_2:+

Скачать данные Geolocation data:+

разархивировать данные+

(По разрешению преподавателя csv файл скачан github)

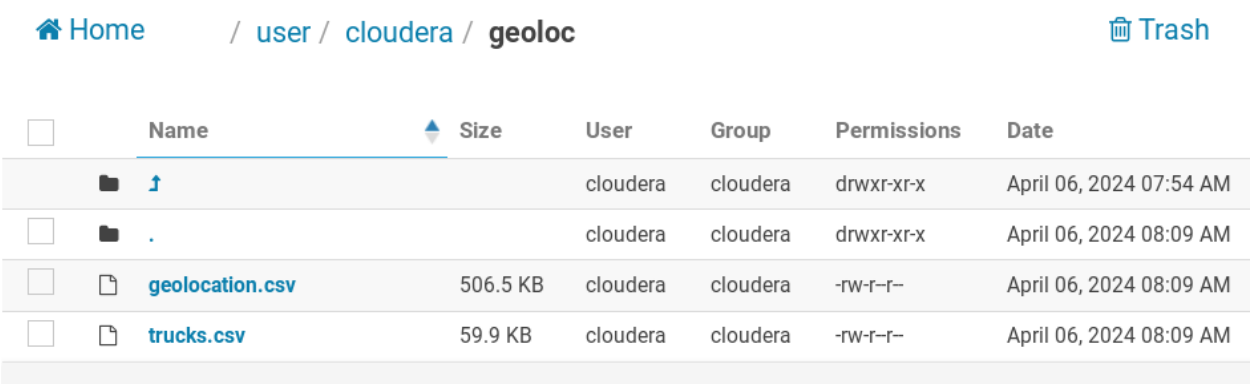
```
[cloudera@quickstart ex_3_2]$ wget https://github.com/BosenkoTM/cloudera-quickst
art/blob/main/data/geolocation.csv
--2024-04-06 03:47:18--  https://github.com/BosenkoTM/cloudera-quickstart/blob/m
ain/data/geolocation.csv
Resolving github.com... 140.82.121.4
Connecting to github.com|140.82.121.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: “geolocation.csv”

    [   <=>                                 ] 842,734     1.41M/s   in 0.6s

2024-04-06 03:47:20 (1.41 MB/s) - “geolocation.csv” saved [842734]

[cloudera@quickstart ex_3_2]$ wget https://github.com/BosenkoTM/cloudera-quickst
art/blob/main/data/trucks.csv
--2024-04-06 03:47:38--  https://github.com/BosenkoTM/cloudera-quickstart/blob/m
ain/data/trucks.csv
Resolving github.com... 140.82.121.4
Connecting to github.com|140.82.121.4|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: “trucks.csv”

    [  <=>                                  ] 231,905     1.00M/s   in 0.2s

2024-04-06 03:47:39 (1.00 MB/s) - “trucks.csv” saved [231905]
```

```
[cloudera@quickstart ex_3_2]$ ls
geolocation.csv  geolocation.zip  trucks.csv
```

3-2.2 В Hue, выбрать Browsers > Files.

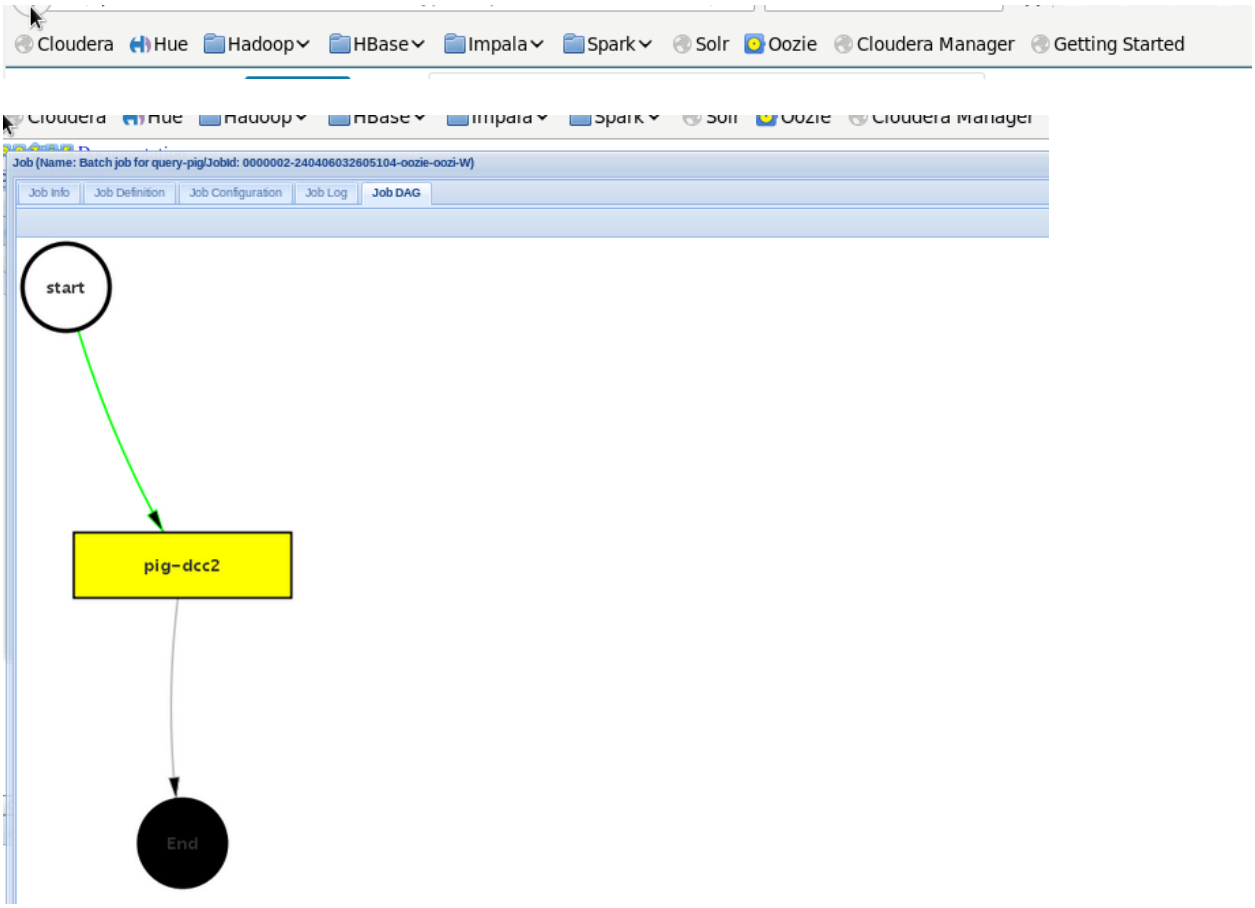
Создайте новый каталог в HDFS с именем geoloc внутри HDFS из Hue. По умолчанию это должно быть создано под hdfs:///user/cloudera/.+

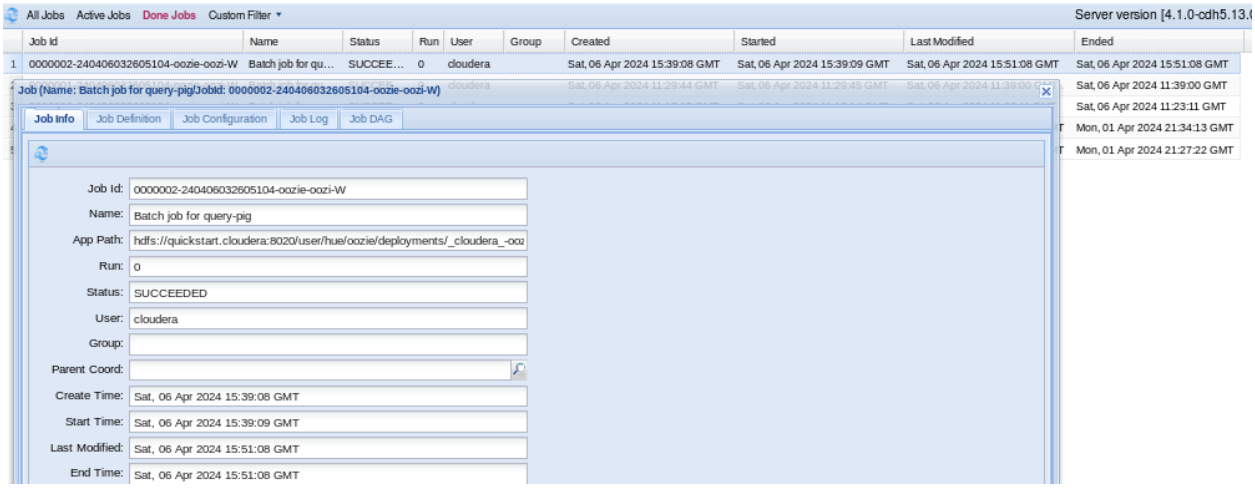
Загрузите Geolocation.csv и trucks.csv в только что созданную папку geoloc/.+

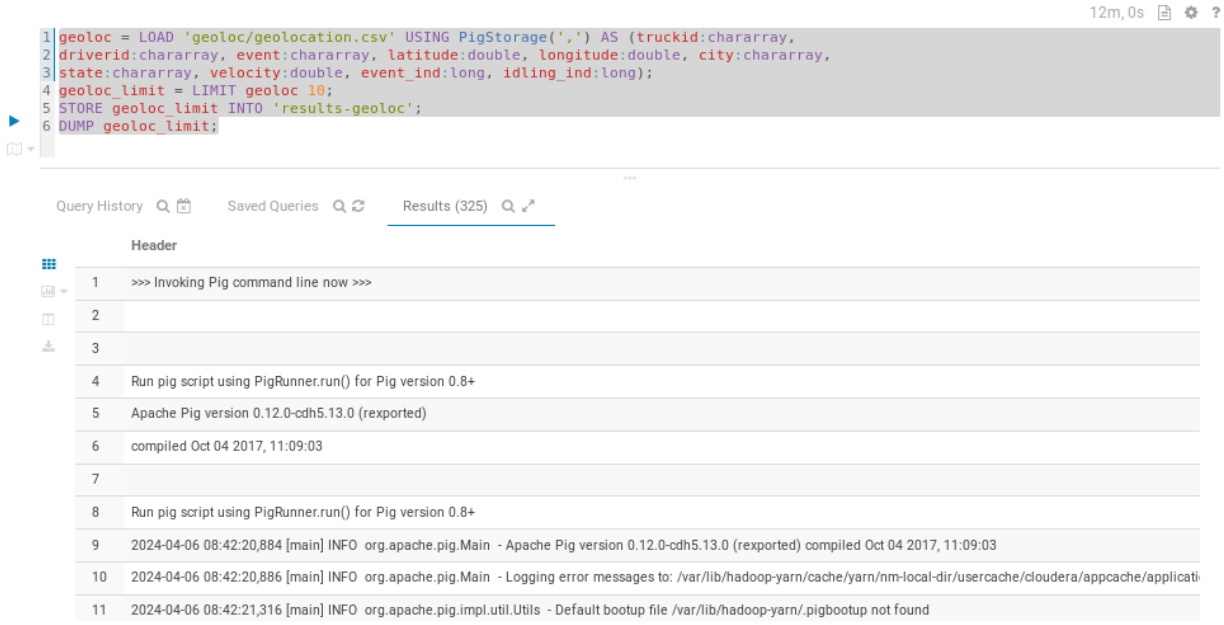


3-2.3 Запустить скрипт/команды, чтобы загрузить и отобразить первые десять строк из файла 'geoloc/geolocation.csv' в каталог 'results-geoloc'(он будет создан автоматически, после выполнения скрипта) в редакторе `Pig` через Hue: `Query > Editor > Pig`:

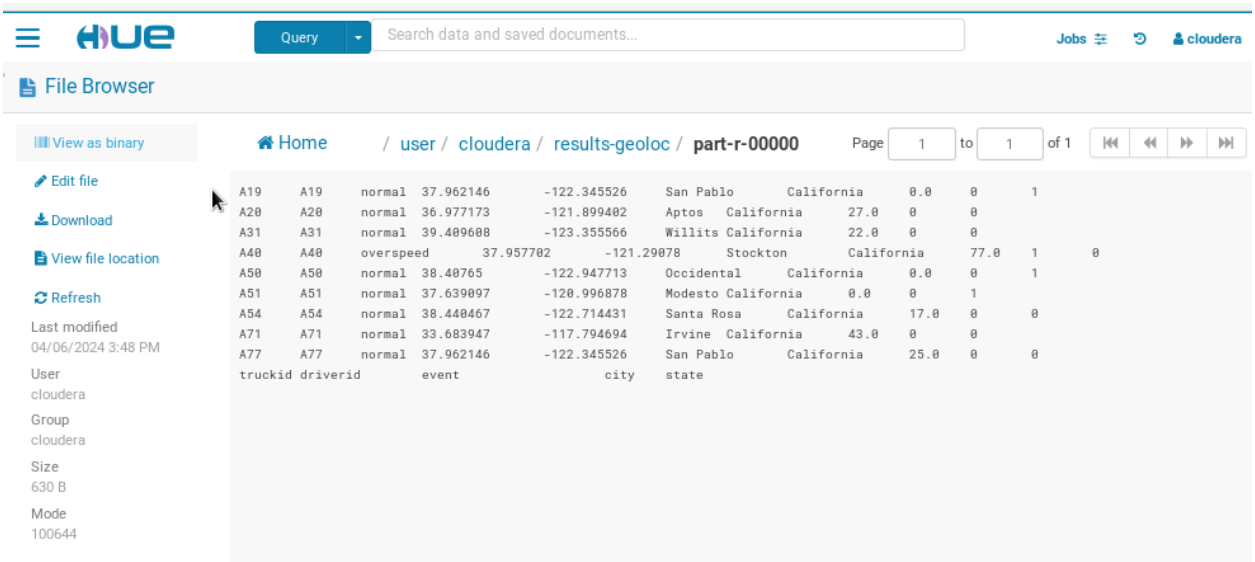
Сначала убедиться в запуске `Dataflow` инструмента `Oozie`. На панели инструментов линк `Oozie` должен быть активным.+




All Jobs
Active Jobs
Done Jobs
Custom Filter
Server version [4.1.0-cdh5.13.
Job Id
Name
Status
Run
User
Group
Created
Started
Last Modified
Ended
0000002-240406032605104-oozie-oozi-W
Batch job for qu...
SUCCEE...
0
cloudera
Sat, 06 Apr 2024 15:39:08 GMT
Sat, 06 Apr 2024 15:39:09 GMT
Sat, 06 Apr 2024 15:51:08 GMT
Sat, 06 Apr 2024 15:51:08 GMT
Sat, 06 Apr 2024 11:39:00 GMT
Sat, 06 Apr 2024 11:23:11 GMT
Mon, 01 Apr 2024 21:34:13 GMT
Mon, 01 Apr 2024 21:27:22 GMT
Job (Name: Batch job for query-pig/JobId: 0000002-240406032605104-oozie-oozi-W)
Job Info
Job Definition
Job Configuration
Job Log
Job DAG
Job Id: 0000002-240406032605104-oozie-oozi-W
Name: Batch job for query-pig
App Path: hdfs://quickstart.cloudera:8020/user/hue/oozie/deployments/_cloudera_-ooz
Run: 0
Status: SUCCEEDED
User: cloudera
Group:
Parent Coord:
Create Time: Sat, 06 Apr 2024 15:39:08 GMT
Start Time: Sat, 06 Apr 2024 15:39:09 GMT
Last Modified: Sat, 06 Apr 2024 15:51:08 GMT
End Time: Sat, 06 Apr 2024 15:51:08 GMT


12m, 0s
1 geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:chararray,
2 driverid:chararray, event:chararray, latitude:double, longitude:double, city:chararray,
3 state:chararray, velocity:double, event_ind:long, idling_ind:long);
4 geoloc_limit = LIMIT geoloc 10;
5 STORE geoloc_limit INTO 'results-geoloc';
6 DUMP geoloc_limit;
Query History
Saved Queries
Results (325)
Header
1 >>> Invoking Pig command line now >>>
2
3
4 Run pig script using PigRunner.run() for Pig version 0.8+
5 Apache Pig version 0.12.0-cdh5.13.0 (rexported)
6 compiled Oct 04 2017, 11:09:03
7
8 Run pig script using PigRunner.run() for Pig version 0.8+
9 2024-04-06 08:42:20,884 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (rexported) compiled Oct 04 2017, 11:09:03
10 2024-04-06 08:42:20,886 [main] INFO org.apache.pig.Main - Logging error messages to: /var/lib/hadoop-yarn/cache/yarn/nm-local-dir/usercache/cloudera/appcache/applicati
11 2024-04-06 08:42:21,316 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /var/lib/hadoop-yarn/.pigbootup not found

Результат


HUE
Query
Search data and saved documents...
Jobs
cloudera
File Browser
View as binary
Edit file
Download
View file location
Refresh
Last modified
04/06/2024 3:48 PM
User
cloudera
Group
cloudera
Size
630 B
Mode
100644
Home / user / cloudera / results-geoloc / part-r-00000
Page 1 to 1 of 1
A19 A19 normal 37.962146 -122.345526 San Pablo California 0.0 0 1
A20 A20 normal 36.977173 -121.899402 Aptos California 27.0 0 0
A31 A31 normal 39.409608 -123.355566 Willits California 22.0 0 0
A40 A40 overspeed 37.957702 -121.29078 Stockton California 77.0 1 0
A50 A50 normal 38.40765 -122.947713 Occidental California 0.0 0 1
A51 A51 normal 37.639097 -120.996878 Modesto California 0.0 0 1
A54 A54 normal 38.440467 -122.714431 Santa Rosa California 17.0 0 0
A71 A71 normal 33.683947 -117.794694 Irvine California 43.0 0 0
A77 A77 normal 37.962146 -122.345526 San Pablo California 25.0 0 0
truckid driverid event city state