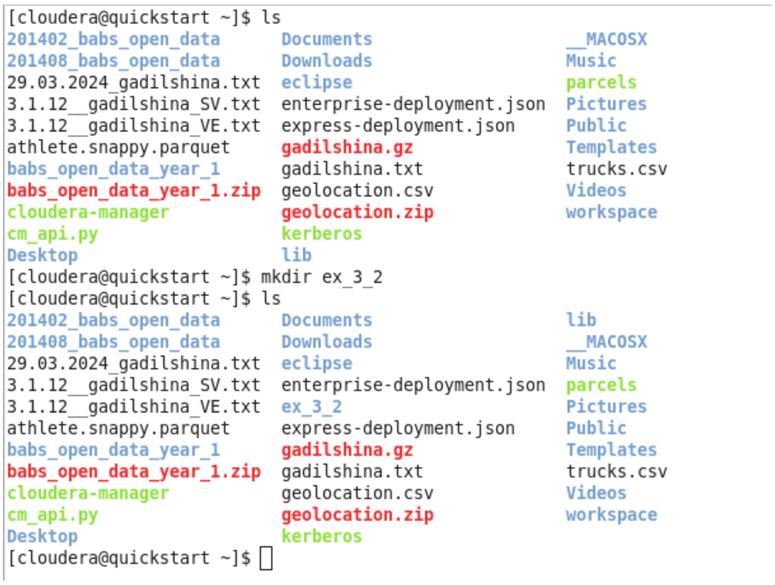
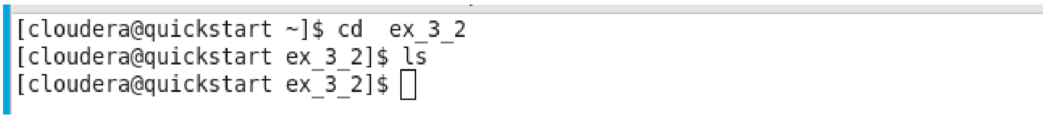
3-2.1 Скачать архив Geolocation github zip или Geolocation data из Cloudera.

Создать каталог ex_3_2:



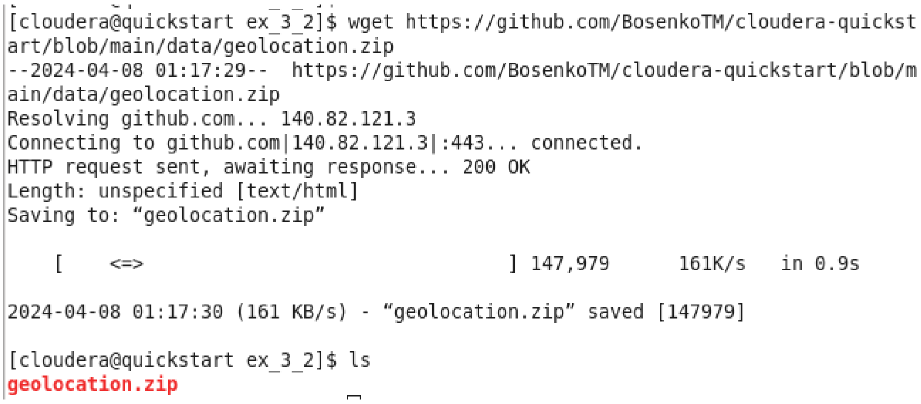
```
[cloudera@quickstart ~]$ ls
201402_babs_open_data      Documents                   __MACOSX
201408_babs_open_data      Downloads                   Music
29.03.2024_gadilshina.txt  eclipse                     parcels
3.1.12__gadilshina_SV.txt  enterprise-deployment.json  Pictures
3.1.12__gadilshina_VE.txt  express-deployment.json     Public
athlete.snappy.parquet     gadilshina.gz               Templates
babs_open_data_year_1      gadilshina.txt              trucks.csv
babs_open_data_year_1.zip  geolocation.csv             Videos
cloudera-manager           geolocation.zip             workspace
cm_api.py                  kerberos
Desktop                    lib
[cloudera@quickstart ~]$ mkdir ex_3_2
[cloudera@quickstart ~]$ ls
201402_babs_open_data      Documents                   lib
201408_babs_open_data      Downloads                   __MACOSX
29.03.2024_gadilshina.txt  eclipse                     Music
3.1.12__gadilshina_SV.txt  enterprise-deployment.json  parcels
3.1.12__gadilshina_VE.txt  ex_3_2                      Pictures
athlete.snappy.parquet     express-deployment.json     Public
babs_open_data_year_1      gadilshina.gz               Templates
babs_open_data_year_1.zip  gadilshina.txt              trucks.csv
cloudera-manager           geolocation.csv             Videos
cm_api.py                  geolocation.zip             workspace
Desktop                    kerberos
[cloudera@quickstart ~]$ 
```

Перейти в каталог ex_3_2:



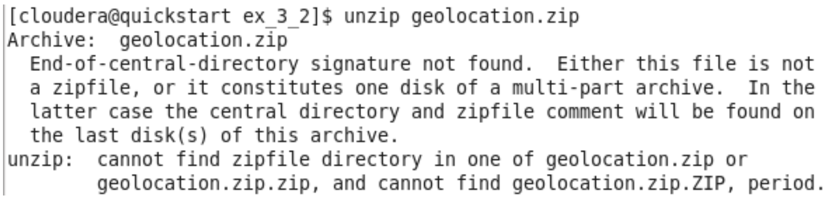
```
[cloudera@quickstart ~]$ cd  ex_3_2
[cloudera@quickstart ex_3_2]$ ls
[cloudera@quickstart ex_3_2]$ 
```

Скачать данные Geolocation data:

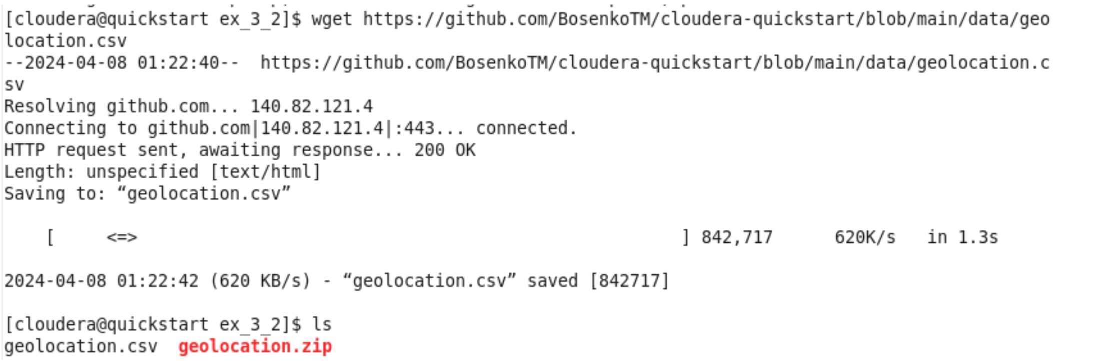


```
[cloudera@quickstart ex_3_2]$ wget https://github.com/BosenkoTM/cloudera-quickst
art/blob/main/data/geolocation.zip
--2024-04-08 01:17:29--  https://github.com/BosenkoTM/cloudera-quickstart/blob/m
ain/data/geolocation.zip
Resolving github.com... 140.82.121.3
Connecting to github.com|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: “geolocation.zip”

    [    <=>                                ] 147,979      161K/s   in 0.9s

2024-04-08 01:17:30 (161 KB/s) - “geolocation.zip” saved [147979]

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

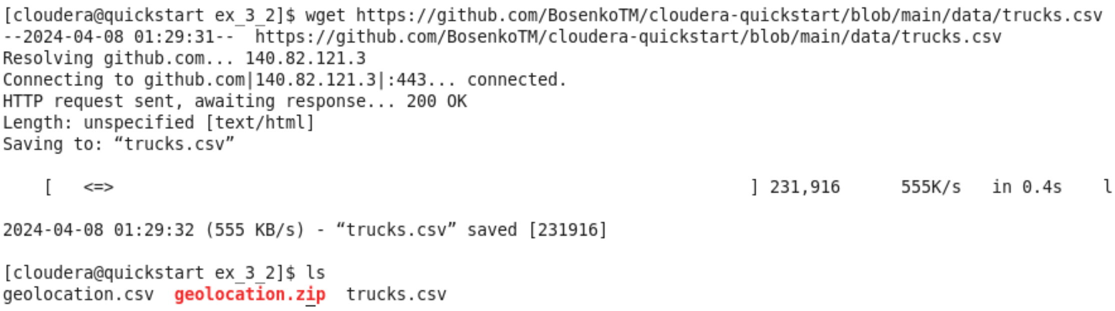
Разархивировать данные:



```
[cloudera@quickstart ex_3_2]$ unzip geolocation.zip
Archive:  geolocation.zip
  End-of-central-directory signature not found.  Either this file is not
  a zipfile, or it constitutes one disk of a multi-part archive.  In the
  latter case the central directory and zipfile comment will be found on
  the last disk(s) of this archive.
unzip:  cannot find zipfile directory in one of geolocation.zip or
        geolocation.zip.zip, and cannot find geolocation.zip.ZIP, period.
```



```
[cloudera@quickstart ex_3_2]$ wget https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/geo
location.csv
--2024-04-08 01:22:40--  https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/geolocation.c
sv
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”

    [     <=>                                            ] 842,717      620K/s   in 1.3s

2024-04-08 01:22:42 (620 KB/s) - “geolocation.csv” saved [842717]

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

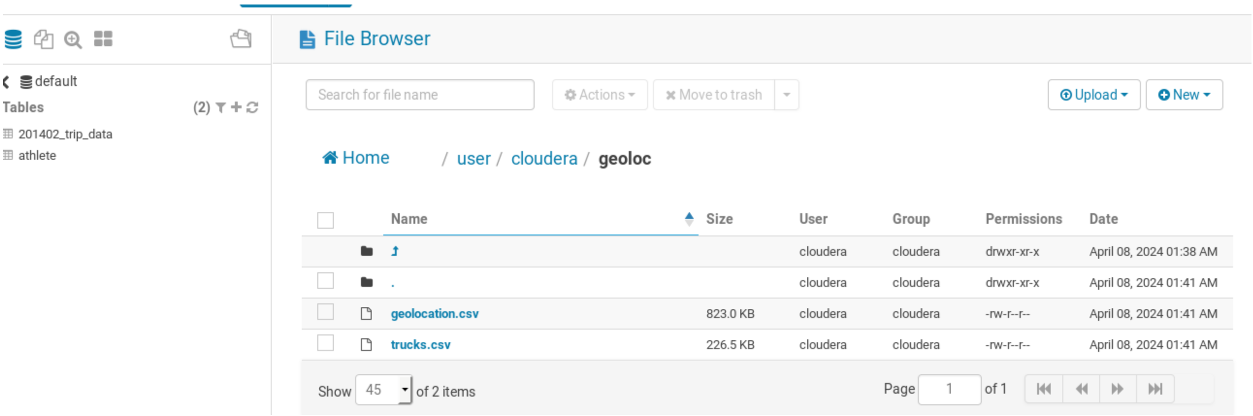


```
[cloudera@quickstart ex_3_2]$ wget https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/trucks.csv
--2024-04-08 01:29:31--  https://github.com/BosenkoTM/cloudera-quickstart/blob/main/data/trucks.csv
Resolving github.com... 140.82.121.3
Connecting to github.com|140.82.121.3|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: “trucks.csv”

    [   <=>                                                    ] 231,916      555K/s   in 0.4s    l

2024-04-08 01:29:32 (555 KB/s) - “trucks.csv” saved [231916]

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

3-2.2 В Hue, выбрать Browsers > Files.

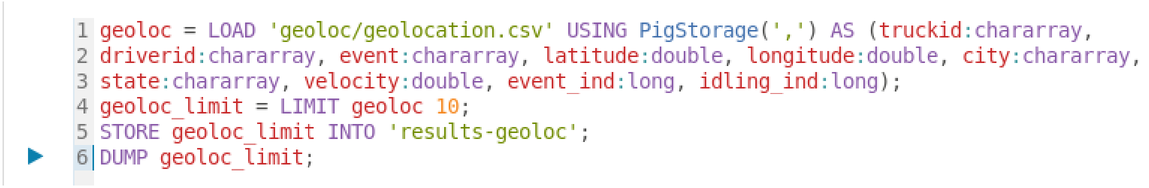
Создайте новый каталог в HDFS с именем data внутри HDFS из Hue. По умолчанию это должно быть создано под hdfs:///user/cloudera/.

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

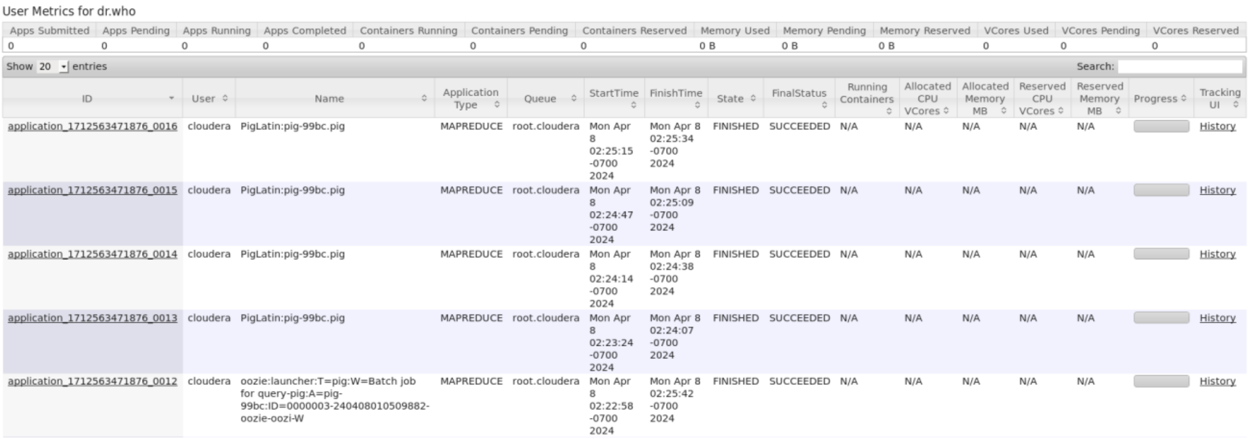


Name	Size	User	Group	Permissions	Date
↥		cloudera	cloudera	drwxr-xr-x	April 08, 2024 01:38 AM
.		cloudera	cloudera	drwxr-xr-x	April 08, 2024 01:41 AM
geolocation.csv	823.0 KB	cloudera	cloudera	-rw-r--r--	April 08, 2024 01:41 AM
trucks.csv	226.5 KB	cloudera	cloudera	-rw-r--r--	April 08, 2024 01:41 AM

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:



```
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);
geoloc_limit = LIMIT geoloc 10;
STORE geoloc_limit INTO 'results-geoloc';
DUMP geoloc_limit;
```



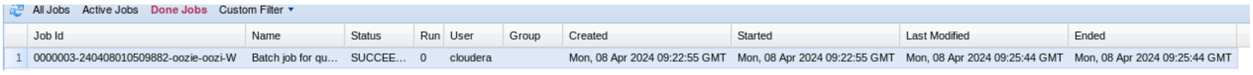
User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0	0 B	0 B	0 B	0	0	0

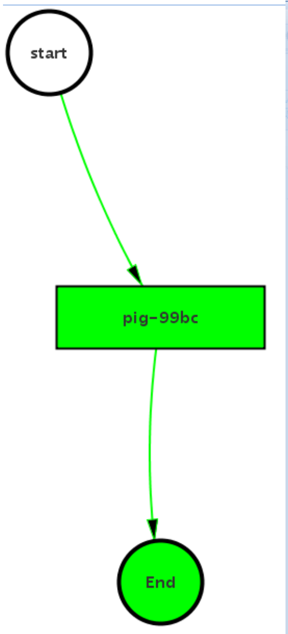
Show 20 entries

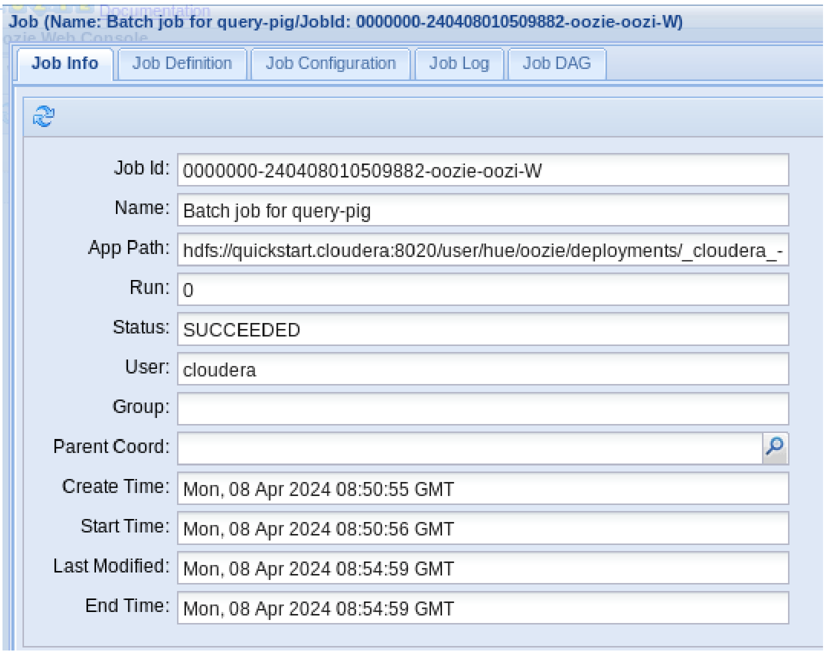
Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress	Tracking UI
application_1712563471876_0016	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:25:15 -0700 2024	Mon Apr 8 02:25:34 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0015	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:24:47 -0700 2024	Mon Apr 8 02:25:09 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0014	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:24:14 -0700 2024	Mon Apr 8 02:24:38 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0013	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:23:24 -0700 2024	Mon Apr 8 02:24:07 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0012	cloudera	oozie:launcher:T=pig:W=Batch job for query-pig:A=pig-99bc:ID=0000003-240408010509882-oozie-oozi-W	MAPREDUCE	root.cloudera	Mon Apr 8 02:22:58 -0700 2024	Mon Apr 8 02:25:42 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History



	Job Id	Name	Status	Run	User	Group	Created	Started	Last Modified	Ended
1	0000003-240408010509882-oozie-oozi-W	Batch job for qu…	SUCCEE…	0	cloudera		Mon, 08 Apr 2024 09:22:55 GMT	Mon, 08 Apr 2024 09:22:55 GMT	Mon, 08 Apr 2024 09:25:44 GMT	Mon, 08 Apr 2024 09:25:44 GMT

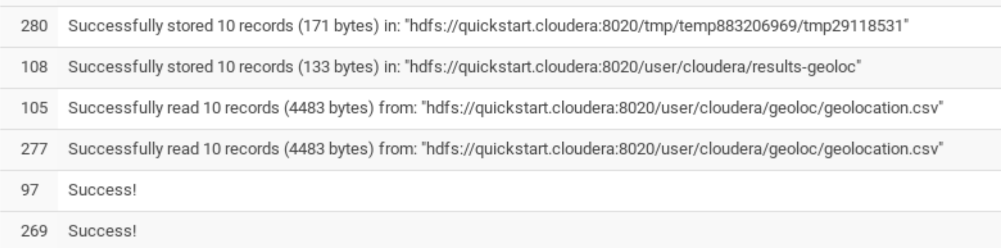




Job (Name: Batch job for query-pig/JobId: 0000000-240408010509882-oozie-oozi-W)

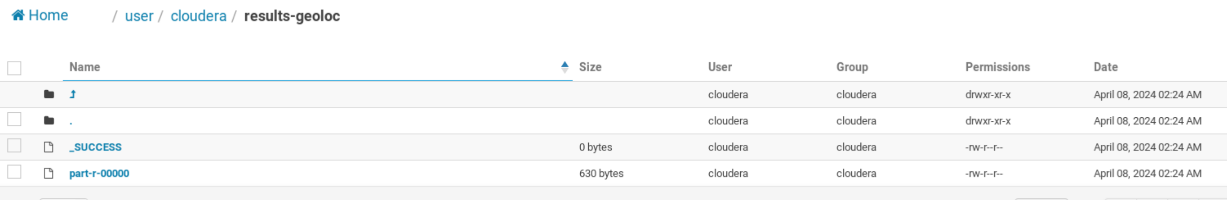
Job Id:	0000000-240408010509882-oozie-oozi-W
Name:	Batch job for query-pig
App Path:	hdfs://quickstart.cloudera:8020/user/hue/oozie/deployments/_cloudera_-
Run:	0
Status:	SUCCEEDED
User:	cloudera
Group:	
Parent Coord:	
Create Time:	Mon, 08 Apr 2024 08:50:55 GMT
Start Time:	Mon, 08 Apr 2024 08:50:56 GMT
Last Modified:	Mon, 08 Apr 2024 08:54:59 GMT
End Time:	Mon, 08 Apr 2024 08:54:59 GMT

После получения метаданных выполнения запроса:



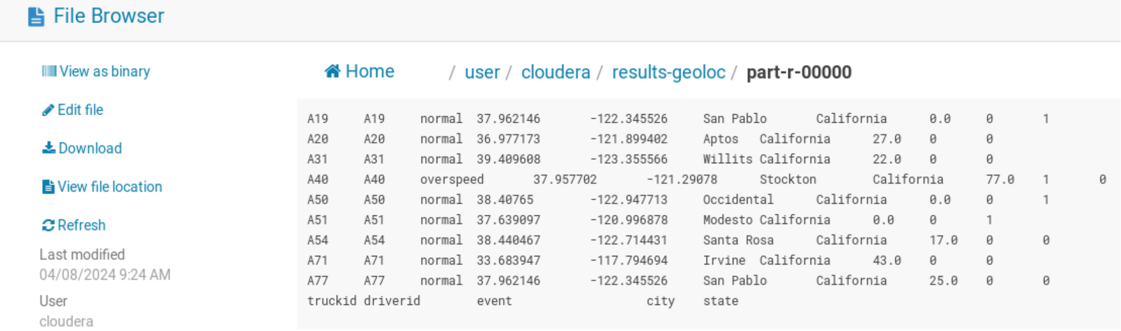
```
280  Successfully stored 10 records (171 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp883206969/tmp29118531"
108  Successfully stored 10 records (133 bytes) in: "hdfs://quickstart.cloudera:8020/user/cloudera/results-geoloc"
105  Successfully read 10 records (4483 bytes) from: "hdfs://quickstart.cloudera:8020/user/cloudera/geoloc/geolocation.csv"
277  Successfully read 10 records (4483 bytes) from: "hdfs://quickstart.cloudera:8020/user/cloudera/geoloc/geolocation.csv"
97   Success!
269  Success!
```

Проверить в каталоге 'results-geoloc' файл 'part-r-00000'



Home / user / cloudera / results-geoloc

□		Name	Size	User	Group	Permissions	Date
		↥		cloudera	cloudera	drwxr-xr-x	April 08, 2024 02:24 AM
□		.		cloudera	cloudera	drwxr-xr-x	April 08, 2024 02:24 AM
□		_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	April 08, 2024 02:24 AM
□		part-r-00000	630 bytes	cloudera	cloudera	-rw-r--r--	April 08, 2024 02:24 AM



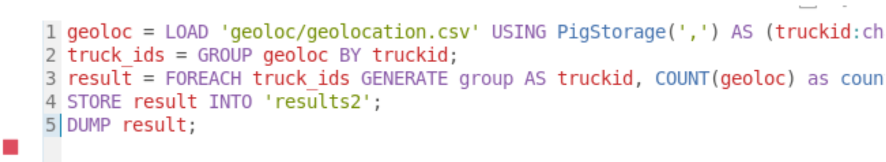
Last modified
04/08/2024 9:24 AM

User
cloudera

Home / user / cloudera / results-geoloc / part-r-00000

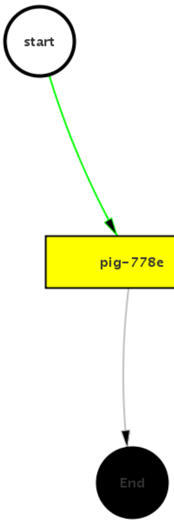
```
A19     A19     normal  37.962146       -122.345526     San Pablo       California      0.0     0       1
A20     A20     normal  36.977173       -121.899402     Aptos   California      27.0    0       0
A31     A31     normal  39.409608       -123.355566     Willits California      22.0    0       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
```

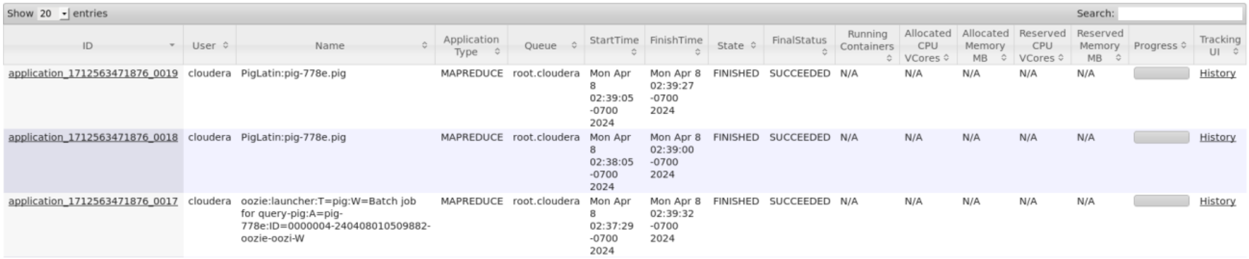
3-2.4 Посчитать статистику по файлу.



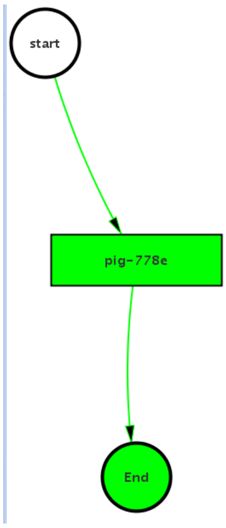
```
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:ch
truck_ids = GROUP geoloc BY truckid;
result = FOREACH truck_ids GENERATE group AS truckid, COUNT(geoloc) as coun
STORE result INTO 'results2';
DUMP result;
```

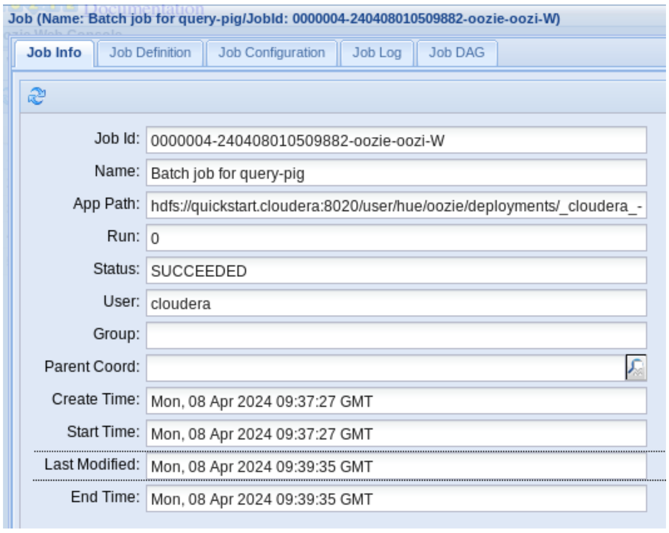
Просмотреть Job DAG Oozie во время выполнения запроса к Pig





ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress	Tracking UI
application_1712563471876_0019	cloudera	PigLatin:pig-778e.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:39:05 -0700 2024	Mon Apr 8 02:39:27 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0018	cloudera	PigLatin:pig-778e.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:38:05 -0700 2024	Mon Apr 8 02:39:00 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0017	cloudera	oozie:launcher:T=pig:W=Batch job for query-pig:A=pig-778e:ID=0000004-240408010509882-oozie-oozi-W	MAPREDUCE	root.cloudera	Mon Apr 8 02:37:29 -0700 2024	Mon Apr 8 02:39:32 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History

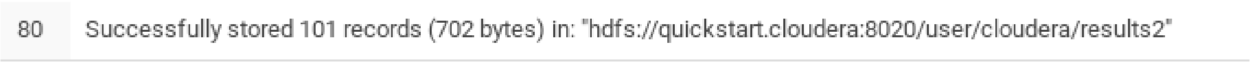




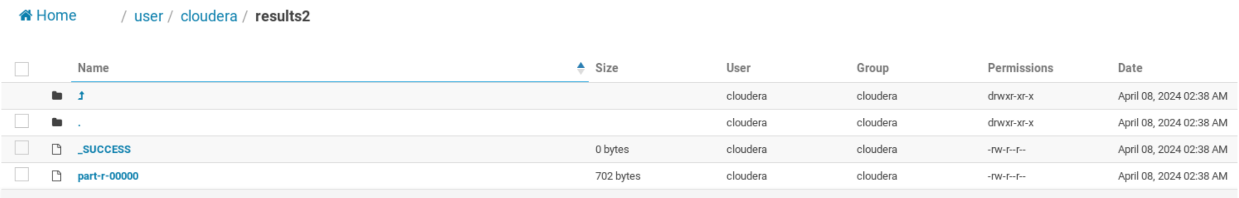
Job (Name: Batch job for query-pig/JobId: 0000004-240408010509882-oozie-oozi-W)

Job Id:	0000004-240408010509882-oozie-oozi-W
Name:	Batch job for query-pig
App Path:	hdfs://quickstart.cloudera:8020/user/hue/oozie/deployments/_cloudera_-
Run:	0
Status:	SUCCEEDED
User:	cloudera
Group:	
Parent Coord:	
Create Time:	Mon, 08 Apr 2024 09:37:27 GMT
Start Time:	Mon, 08 Apr 2024 09:37:27 GMT
Last Modified:	Mon, 08 Apr 2024 09:39:35 GMT
End Time:	Mon, 08 Apr 2024 09:39:35 GMT

Провирить в каталоге 'results2' файл 'part-r-00000'

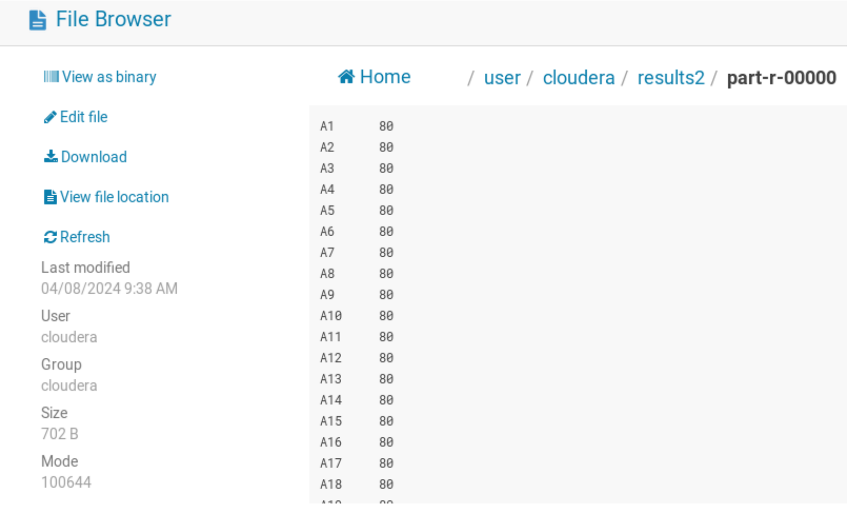


```
80   Successfully stored 101 records (702 bytes) in: "hdfs://quickstart.cloudera:8020/user/cloudera/results2"
```



Home / user / cloudera / results2

□		Name	Size	User	Group	Permissions	Date
		↥		cloudera	cloudera	drwxr-xr-x	April 08, 2024 02:38 AM
□		.		cloudera	cloudera	drwxr-xr-x	April 08, 2024 02:38 AM
□		_SUCCESS	0 bytes	cloudera	cloudera	-rw-r--r--	April 08, 2024 02:38 AM
□		part-r-00000	702 bytes	cloudera	cloudera	-rw-r--r--	April 08, 2024 02:38 AM



Home / user / cloudera / results2 / part-r-00000

Last modified
04/08/2024 9:38 AM

User
cloudera

Group
cloudera

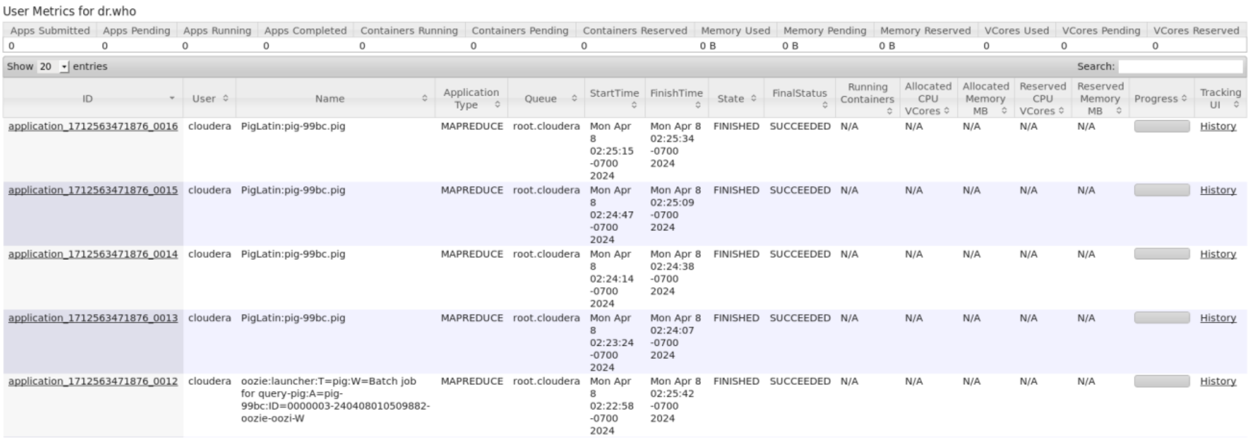
Size
702 B

Mode
100644

```
A1    80
A2    80
A3    80
A4    80
A5    80
A6    80
A7    80
A8    80
A9    80
A10   80
A11   80
A12   80
A13   80
A14   80
A15   80
A16   80
A17   80
A18   80
```

3-2.5 Анализ.

- **Провести анализ журналов Hadoop > YARN Resource Manager в Firefox. Продоставить в виде скринов все задачи, выполенные в п 3-2.1 - 3-2.4.**

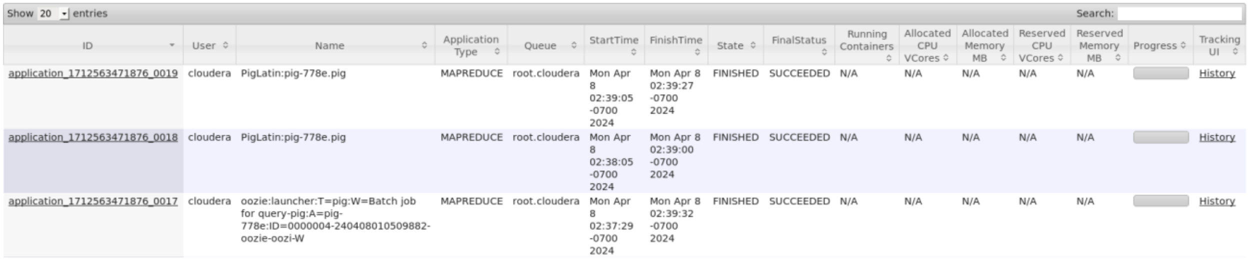


User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0	0 B	0 B	0 B	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress	Tracking UI
application_1712563471876_0016	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:25:15 -0700 2024	Mon Apr 8 02:25:34 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0015	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:24:47 -0700 2024	Mon Apr 8 02:25:09 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0014	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:24:14 -0700 2024	Mon Apr 8 02:24:38 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0013	cloudera	PigLatin:pig-99bc.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:23:24 -0700 2024	Mon Apr 8 02:24:07 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0012	cloudera	oozie:launcher:T=pig:W=Batch job for query-pig:A=pig-99bc:ID=0000003-240408010509882-oozie-oozi-W	MAPREDUCE	root.cloudera	Mon Apr 8 02:22:58 -0700 2024	Mon Apr 8 02:25:42 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History



Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	Progress	Tracking UI
application_1712563471876_0019	cloudera	PigLatin:pig-778e.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:39:05 -0700 2024	Mon Apr 8 02:39:27 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0018	cloudera	PigLatin:pig-778e.pig	MAPREDUCE	root.cloudera	Mon Apr 8 02:38:05 -0700 2024	Mon Apr 8 02:39:00 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History
application_1712563471876_0017	cloudera	oozie:launcher:T=pig:W=Batch job for query-pig:A=pig-778e:ID=0000004-240408010509882-oozie-oozi-W	MAPREDUCE	root.cloudera	Mon Apr 8 02:37:29 -0700 2024	Mon Apr 8 02:39:32 -0700 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A		History

- **Подсчитать количество уникальных городов, в которых был уникальный грузовик по его truckid, среднюю скорость грузовика из файла trucks.csv?**