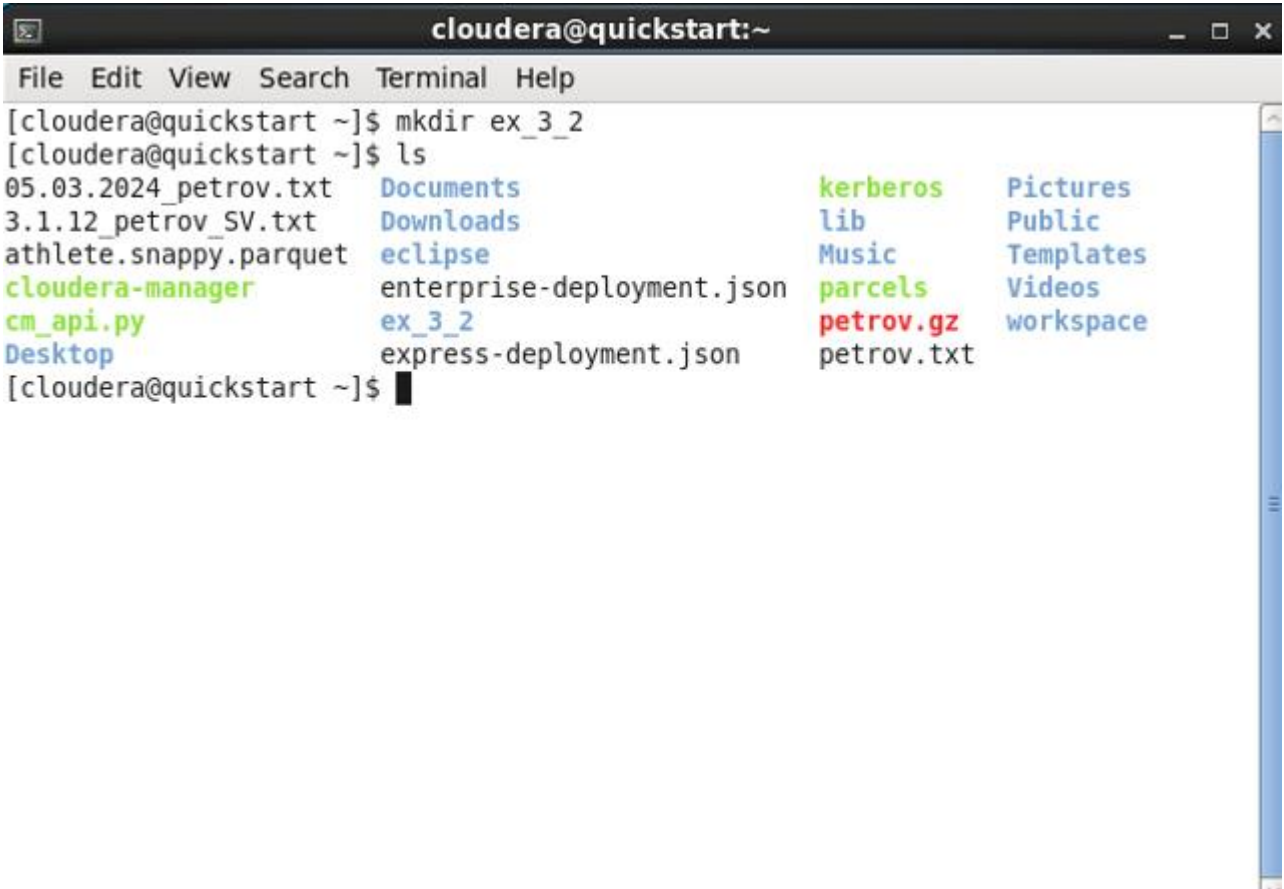


Кузьменко Сергей БД-231м

3-2.1 Скачать архив [Geolocation github zip](#) или [Geolocation data из Cloudera](#).

Создать каталог `ex_3_2`:



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ mkdir ex_3_2  
[cloudera@quickstart ~]$ ls  
05.03.2024_petrov.txt  Documents  kerberos  Pictures  
3.1.12_petrov_SV.txt  Downloads  lib       Public  
athlete.snappy.parquet eclipse     Music     Templates  
cloudera-manager      enterprise-deployment.json parcels    Videos  
cm_api.py             ex_3_2    petrov.gz workspace  
Desktop              express-deployment.json  petrov.txt  
[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-quickstart/blob/main/data/geolocation.csv
--2024-04-06 03:47:18-- https://github.com/BosenkoTM/cloudera-quickstart/blob/main/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-quickstart/blob/main/data/trucks.csv
--2024-04-06 03:47:38-- https://github.com/BosenkoTM/cloudera-quickstart/blob/main/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/.+**

Home

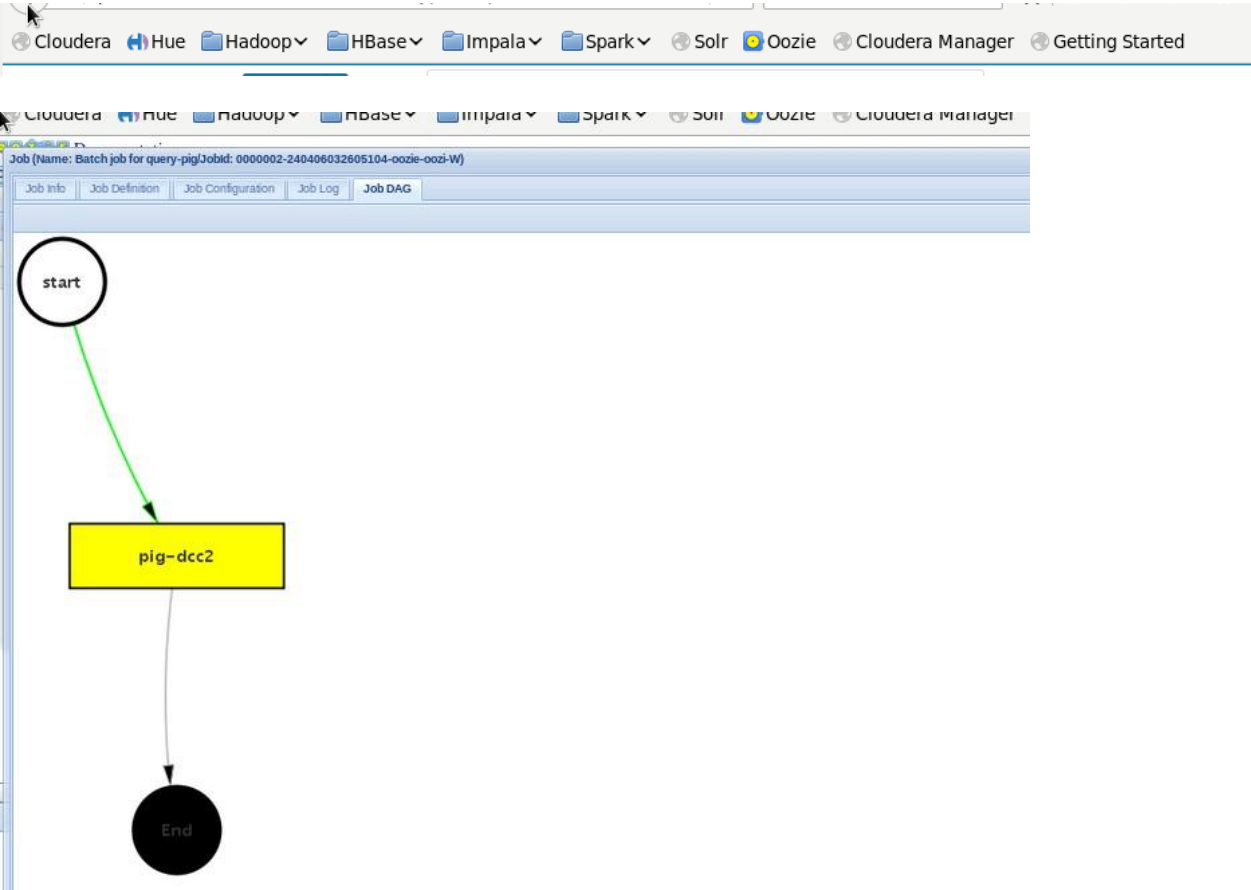
/ user / cloudera / geoloc

Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		cloudera	cloudera	drwxr-xr-x	April 06, 2024 07:54 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	April 06, 2024 08:09 AM
<input type="checkbox"/>	geolocation.csv	506.5 KB	cloudera	cloudera	-rw-r--r--	April 06, 2024 08:09 AM
<input type="checkbox"/>	trucks.csv	59.9 KB	cloudera	cloudera	-rw-r--r--	April 06, 2024 08:09 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**:

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



All JobsActive JobsDone JobsCustom Filter

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

Job (Name: Batch job for query-pig;JobId: 0000002-240406032605104-oozie-oozi-W)

Job InfoJob DefinitionJob ConfigurationJob LogJob 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 HistorySaved QueriesResults (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 (rexpoted)
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 (rexpoted) 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 bootstrap file /var/lib/hadoop-yarn/pigbootstrap not found

Результат

File Browser

View as binary

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

Page 1 to 1 of 1

Edit file

Download

View file location

Refresh

Last modified 04/06/2024 3:48 PM

User cloudera

Group cloudera

Size 630 B

Mode 100644

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	39.489608	-123.355566	Willits	California	22.0	0	0
A40	A40	overspeed	37.957702	-121.29878	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	1	0
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			

