Федеральное агентство связи

Ордена Трудового Красного Знамени

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский технический университет связи и информатики»

Отчет по лабораторной работе №4

по дисциплине «Математические Методы в Больших данных»

«Hive and HBase»

Выполнил: студент группы БВТ1902

Адедиха Коффи Жермен

Руководитель: Мария Пугачева

Москва 2021

Цель работы:

Ознакомится с системой управления базами данных Hive и HBase.

Задачи:

Hive

1. Скачать любой датасет из списка ниже
   https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions
   https://www.kaggle.com/datasnaek/youtube-new
   https://www.kaggle.com/akhilv11/border-crossing-entry-data
   https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games
   https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil
2. Загрузить этот датасет в HDFS в свою домашнюю папку
3. Создать собственную базу данных в HIVE. (create database)
4. Создать таблицы внутри базы данных с использованием одного файла из загруженного датасета (предварительно создать таблицу с форматами аналогичными вашим данным в выбраной таблице, см.приложение).
5. Сделать любой простой отчет по загруженным данным используя груповые и агрегатные функции.

HBase https://hbase.apache.org/book.html

1. Создать таблицу
2. Посмотреть информацию о ней (list/describe оба варианта)
3. Положить в нее данные (3-5 строк)
4. Просканировать
5. Получить конкретную строку
6. Заблокировать/разблокировать таблицу
7. Удалить таблицу

Выполнение

Скачал датасет https://www.kaggle.com/akhilv11/border-crossing-entry-data

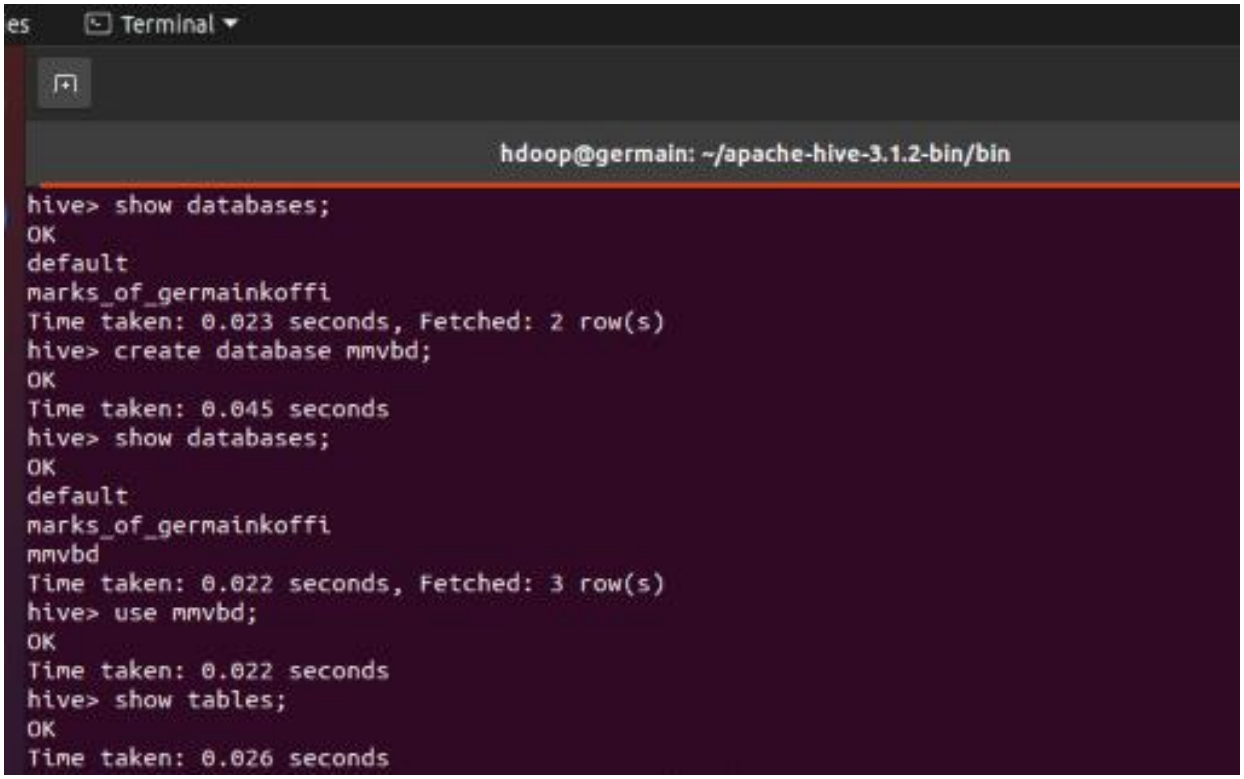

Рис-1 Создание database в Hive

```
hdoop@germain: ~/apache-hive-3.1.2-bin/bin
hive> show databases;
OK
default
marks_of_germainkoffi
mmvbd
Time taken: 0.019 seconds, Fetched: 3 row(s)
hive> use mmvbd;
OK
Time taken: 0.021 seconds
hive> show tables;
OK
Time taken: 0.025 seconds
hive> create table bd_instruments(id int, instrument_name string)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.09 seconds
hive> show tables;
OK
bd_instruments
Time taken: 0.025 seconds, Fetched: 1 row(s)
```

Рис-2 Создание таблицы в database в Hive

```
hive> create table border_crossing
    > (
    > Port string,
    > State string,
    > Code int,
    > Border string,
    > Measure string,
    > Value int,
    > Location string
    > )
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat'
    > OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
    > TBLPROPERTIES(
    > 'serialization.null.format'='',
    > 'skip.header.line.count'='1')
    > ;
OK
Time taken: 0.577 seconds
hive> show databases;
OK
default
marks_of_germainkoffi
mmvbd
Time taken: 0.035 seconds, Fetched: 3 row(s)
hive> use mmvbd
    > ;
OK
Time taken: 0.024 seconds
hive>
```

Рис -3 создание таблицы для загрузки датасет в HDFS в свою домашнюю папку

```
hive>
hive> load data local inpath '/home/hdoop/Downloads/archive/Border_Crossing_Entry_Data.csv' into table border_crossing;
Loading data to table mmvbd.border_crossing
OK
Time taken: 0.368 seconds
```

Рис -4 Загрузка датасет в HDFS в свою домашнюю папку

Отчет по загруженным данным используя груповые и агрегатные функции.

```
hive>
hive> load data local inpath '/home/hdoop/Downloads/archive/Border_Crossing_Entry_Data.csv' into table border_crossing;
Loading data to table mmvbd.border_crossing
OK
Time taken: 0.368 seconds
hive> select * from border_crossing
    > where Port='Douglas';
OK
Douglas Arizona 2601    US-Mexico Border        Trucks  2175    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Full   1639    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicles       144173  POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Bus Passengers  575     POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicle Passengers     255070  POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Empty  536     POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Buses   57      POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Pedestrians     75746   POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicles       125735  POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Empty  517     POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Full   1622    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Pedestrians     65680   POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Trucks  2140    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicle Passengers     222245  POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Buses   44      POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Bus Passengers  1961    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Bus Passengers  1961    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Buses   34      POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Full   1786    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Pedestrians     70359   POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Trucks  2346    POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Truck Containers Empty  560     POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicles       132376  POINT (-109.54472 31.344439999999995)
Douglas Arizona 2601    US-Mexico Border        Personal Vehicle Passengers     236767  POINT (-109.54472 31.344439999999995)
```

Рис-5 Operation Select над таблицей border_crossing

```
Time taken: 0.201 seconds, Fetched: 38204 row(s)
hive> select Port , Code  from border_crossing
    > where value =1;
OK
Hansboro        3415
Raymond 3301
Del Rio 2302
Fort Kent       110
Laurier 3016
Metaline Falls  3025
Porthill        3308
Skagway 3103
Carbury 3421
Boundary        3015
Hansboro        3415
Noonan  3420
Wildhorse       3323
Vanceboro       105
Van Buren       108
Del Bonita      3322
Skagway 3103
Hansboro        3415
Sherwood        3414
Hansboro        3415
Eastport        103
Turner  3306
Vanceboro       105
Roseau  3426
Frontier        3020
Maida   3416
Metaline Falls  3025
Del Bonita      3322
Madawaska       109
Danville        3012
Port Angeles    3007
Eastport        103
Northgate       3406
Noonan  3420
Del Bonita      3322
Wildhorse       3323
Madawaska       109
Raymond 3301
Roosville       3318
Eastport        103
Fortuna 3417
Madawaska       109
Porthill        3308
```

Рис-6 Operation Select над таблицей border_crossing

```
hive> select State
    > from border_crossing
    > where Measure='Trains'
    > group by state;
Query ID = hdoop_20211110023901_2cc81cb4-0c58-459a-94ab-b25ea2d4621e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1636481052635_0004, Tracking URL = http://germain:8088/proxy/application_1636481052635_0004/
Kill Command = /home/hdoop/hadoop-3.3.1/bin/mapred job  -kill job_1636481052635_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-11-10 02:39:09,682 Stage-1 map = 0%,  reduce = 0%
2021-11-10 02:39:20,050 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.98 sec
2021-11-10 02:39:26,257 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.47 sec
MapReduce Total cumulative CPU time: 13 seconds 470 msec
Ended Job = job_1636481052635_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 13.47 sec   HDFS Read: 58172533 HDFS Write: 378 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 470 msec
OK
Alaska
Arizona
California
Idaho
Maine
Michigan
Minnesota
Montana
New Mexico
New York
North Dakota
Texas
Vermont
Washington
Time taken: 26.476 seconds, Fetched: 14 row(s)
hive>
```

Рис-7 Operation «Select n Group By» над таблицей border_crossing

```
hive> SELECT SUM(Value)
    > FROM border_crossing
    > WHERE Port = 'Del Rio';
Query ID = hdoop_20211111003457_f071b280-0001-4bc0-8278-040c6333cd31
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1636578229251_0002, Tracking URL = http://localhost:8088/proxy/application_1636578229251_0002/
Kill Command = /home/hdoop/hadoop-3.3.1/bin/mapred job  -kill job_1636578229251_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-11-11 00:35:05,913 Stage-1 map = 0%,  reduce = 0%
2021-11-11 00:35:16,302 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.25 sec
2021-11-11 00:35:22,526 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.54 sec
MapReduce Total cumulative CPU time: 12 seconds 540 msec
Ended Job = job_1636578229251_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.54 sec   HDFS Read: 58168758 HDFS Write: 112 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 540 msec
OK
2.68269034E8
Time taken: 26.537 seconds, Fetched: 1 row(s)
hive>
```

Рис-8 Запрос с функцей SUM

```
Time taken: 28.192 seconds, Fetched: 1 row(s)
hive> SELECT MIN(Value),MAX(value)
    > FROM border_crossing;
Query ID = hdoop_20211111003934_a490cf6b-f2fe-4367-8f27-83a10d3ce881
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1636578229251_0004, Tracking URL = http://localhost:8088/proxy/application_1636578229251_0004/
Kill Command = /home/hdoop/hadoop-3.3.1/bin/mapred job  -kill job_1636578229251_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-11-11 00:39:44,310 Stage-1 map = 0%,  reduce = 0%
2021-11-11 00:39:54,698 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.71 sec
2021-11-11 00:40:01,953 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.11 sec
MapReduce Total cumulative CPU time: 13 seconds 110 msec
Ended Job = job_1636578229251_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 13.11 sec   HDFS Read: 58173017 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 110 msec
OK
0       99999
Time taken: 28.805 seconds, Fetched: 1 row(s)
hive>
```

Рис-9 Запрос с функцями MAX и MIN

```
hive> SELECT AVG(Value)
    > FROM border_crossing
    > Where State = 'California';
Query ID = hdoop_20211111004231_dad35f42-7c2f-4942-9cf5-6f2fcd426ce4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1636578229251_0005, Tracking URL = http://localhost:8088/proxy/application_1636578229251_0005/
Kill Command = /home/hdoop/hadoop-3.3.1/bin/mapred job  -kill job_1636578229251_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-11-11 00:42:42,706 Stage-1 map = 0%,  reduce = 0%
2021-11-11 00:42:52,076 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.71 sec
2021-11-11 00:42:59,332 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.82 sec
MapReduce Total cumulative CPU time: 13 seconds 820 msec
Ended Job = job_1636578229251_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 13.82 sec   HDFS Read: 58169899 HDFS Write: 117 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 820 msec
OK
130851.3096010889
Time taken: 30.193 seconds, Fetched: 1 row(s)
hive>
```

Рис-10 Запрос с функцей AVG

HBase https://hbase.apache.org/book.html

1. Создать таблицу
2. Посмотреть информацию о ней (list/describe оба варианта)
3. Положить в нее данные (3-5 строк)
4. Просканировать
5. Получить конкретную строку
6. Заблокировать/разблокировать таблицу
7. Удалить таблицу

```
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
Version 1.4.14, re7cbc2debc11a01dd4f3e6f6d6992b7bd307bbcb, Thu Oct 21 00:05:07 CST 2021

hbase(main):001:0> list
TABLE
0 row(s) in 0.2280 seconds

=> []
hbase(main):002:0> create 'bd_tools','id_tool','name_tool'
0 row(s) in 1.4470 seconds

=> Hbase::Table - bd_tools
hbase(main):003:0>
```

Рис -11 Создание таблицы

```
hbase(main):005:0> describe 'bd_tools'
Table bd_tools is ENABLED

bd_tools

COLUMN FAMILIES DESCRIPTION

{NAME => 'id_tool', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BL
OCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0'
, BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

{NAME => 'name_tool', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_
BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '
0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}

2 row(s) in 0.0300 seconds

hbase(main):006:0>
```

Рисунок 12 – Просмотр информации о таблице

```
hdoop@germain: ~

hbase(main):009:0> put 'bd_tools','row1','id_tool:1','Storm'
0 row(s) in 0.0190 seconds

hbase(main):010:0> put 'bd_tools','row2','id_tool:2','MongoDB'
0 row(s) in 0.0060 seconds

hbase(main):011:0> put 'bd_tools','row3','id_tool:3','Cassandra'
0 row(s) in 0.0060 seconds

hbase(main):012:0> put 'bd_tools','row4','id_tool:4','Cloudera'
0 row(s) in 0.0050 seconds

hbase(main):013:0> put 'bd_tools','row5','id_tool:5','OpenRefine'
0 row(s) in 0.0070 seconds

hbase(main):014:0> put 'bd_tools','row6','id_tool:6','Oracle'
0 row(s) in 0.0070 seconds

hbase(main):015:0> scan 'bd_tools'
ROW                                                    COLUMN+CELL

 row1                                                  column=id_tool:1, timestamp=1636574372908, value=Storm

 row2                                                  column=id_tool:2, timestamp=1636574428169, value=MongoDB

 row3                                                  column=id_tool:3, timestamp=1636574477889, value=Cassandra

 row4                                                  column=id_tool:4, timestamp=1636574500229, value=Cloudera

 row5                                                  column=id_tool:5, timestamp=1636574527750, value=OpenRefine

 row6                                                  column=id_tool:6, timestamp=1636574620803, value=Oracle

6 row(s) in 0.0260 seconds

hbase(main):016:0>
```

Рис- 13 Положили в таблице данные и просканировали

```
hbase(main):017:0> get 'bd_tools','row4'
COLUMN                          CELL
 id_tool:4                      timestamp=1636574500229, value=Cloudera
1 row(s) in 0.0210 seconds

hbase(main):018:0> get 'bd_tools','row3'
COLUMN                          CELL
 id_tool:3                      timestamp=1636574477889, value=Cassandra
1 row(s) in 0.0050 seconds
```

Рис-14 получили конкретную строку

```
hbase(main):019:0> disable 'bd_tools'
0 row(s) in 2.2760 seconds

hbase(main):020:0> get 'bd_tools','row3'
COLUMN                          CELL

ERROR: bd_tools is disabled.
```

```
hbase(main):021:0> enable 'bd_tools'
0 row(s) in 1.2780 seconds

hbase(main):022:0> get 'bd_tools','row3'
COLUMN                          CELL
 id_tool:3                      timestamp=1636574477889, value=Cassandra
1 row(s) in 0.0290 seconds
```

Рис-15 заблокировали/разблокировали таблицу

```
hbase(main):025:0> drop 'bd_tools'
0 row(s) in 1.2760 seconds

hbase(main):026:0> list
TABLE
0 row(s) in 0.0090 seconds
```

Рис -16 - удалили таблицу

Вывод

Данная лабораторная работа позволила нам ознакомится с Hive - система управления базами данных на основе платформы Hadoop.Система позволяет выполнять запросы, агрегировать и анализировать данные, хранящиеся в Hadoop. Получили базовые знания о Hbase