Интеллектуальные информационные системы

Apache Hive

Кафедра управления и интеллектуальных технологий НИУ «МЭИ»

2023 г.

Apache Hive

Hive — система управления базами данных на основе платформы Hadoop с SQL-подобным языком запросов HiveQL

Разработан в 2007 году в Facebook

Работает поверх Hadoop

https://hive.apache.org/

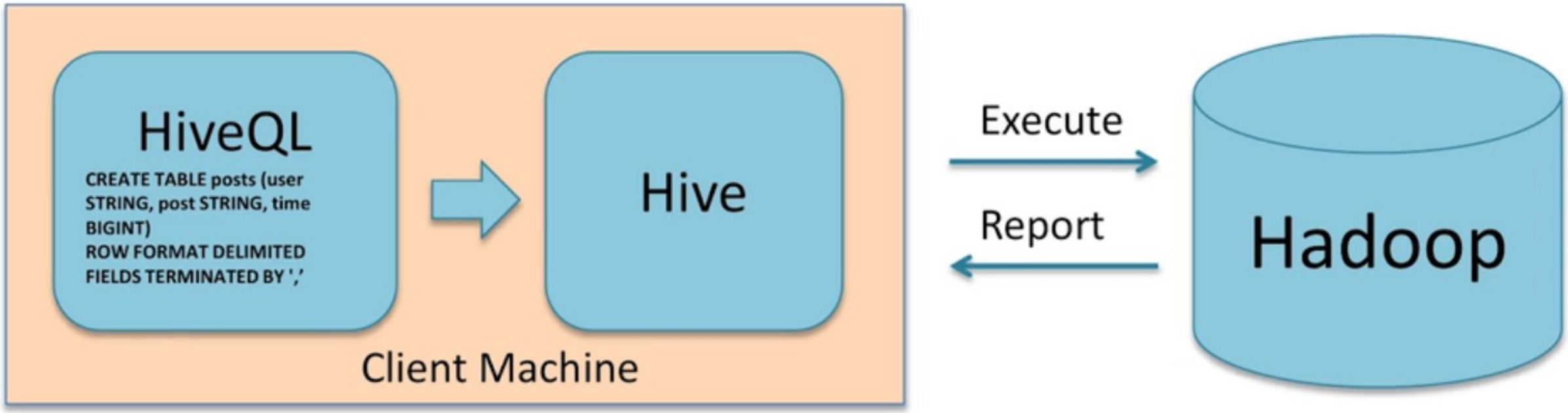
Hive предоставляет

- Возможность структурировать разные форматы данных
- Интерфейс для написания запросов, анализа и обобщения больших объемов данных
- Доступ к данным, хранящихся в HDFS и Hbase.
- Разработан с учетом масштабируемости

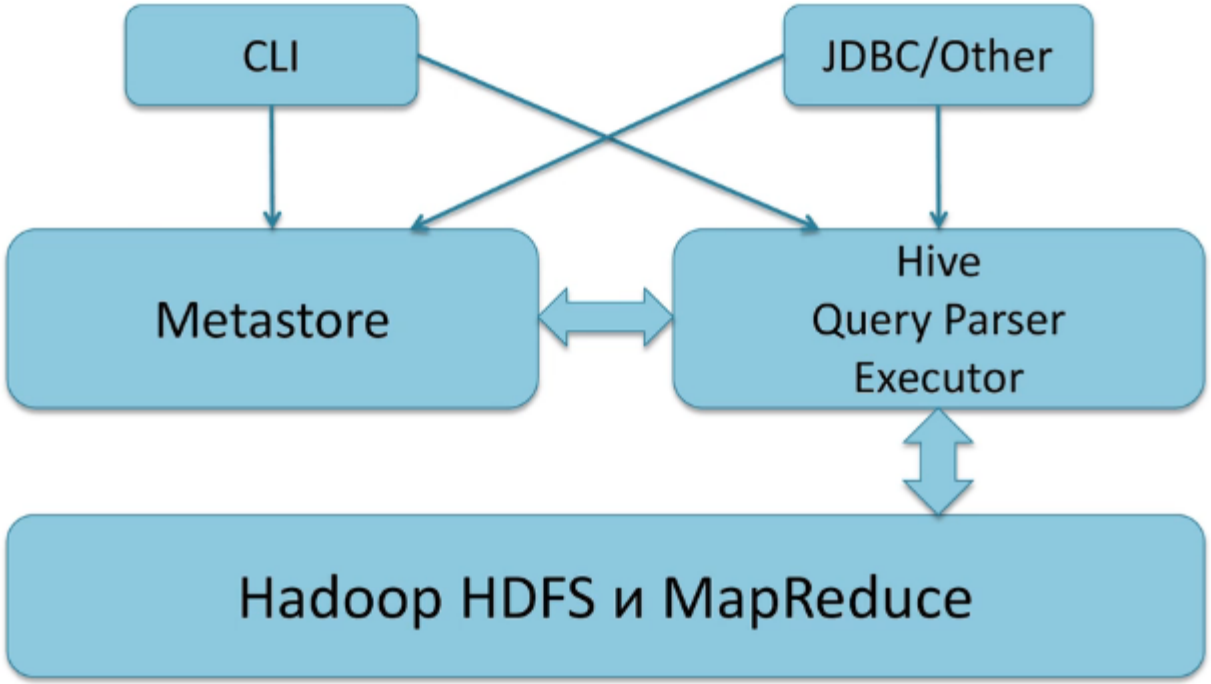
Не подходит для:

- Low-latency, real-time запросов – запрос даже небольшого объема данных может занимать длительное время

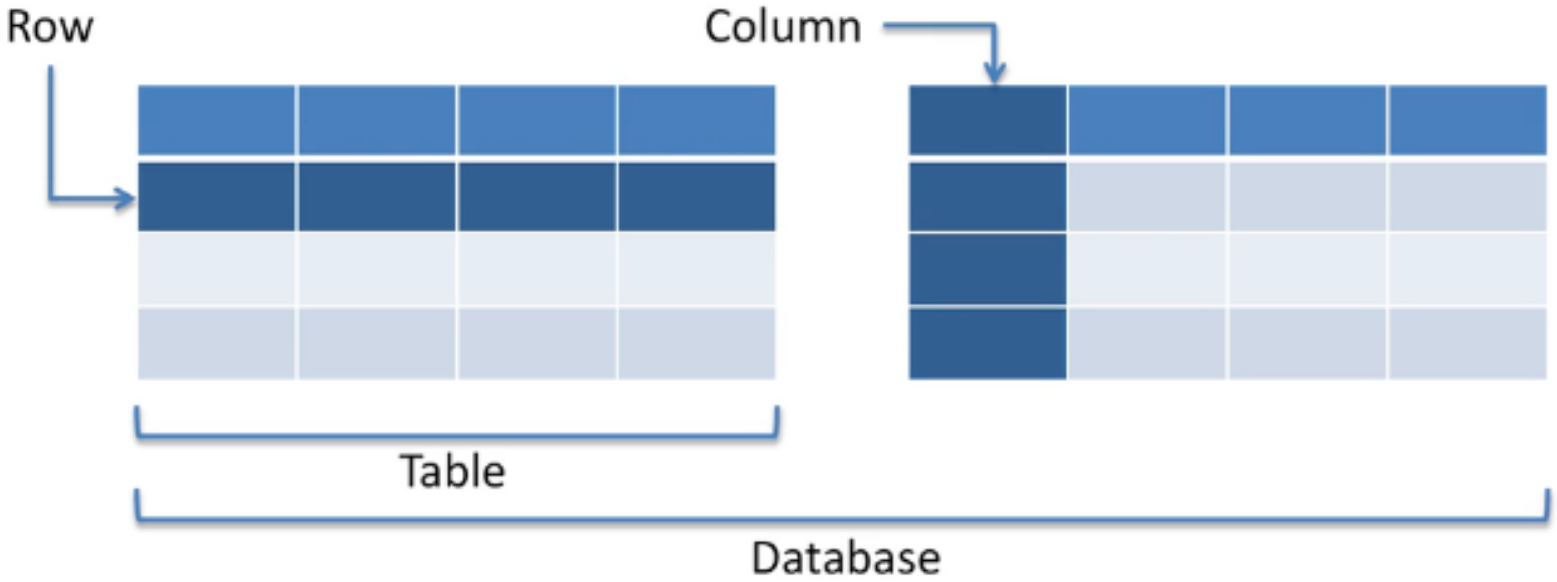
Схема работы



Архитектура



Концепция



Похожа на реляционную модель данных.
Колонка определяет тип данных.
Разные строки имеют одинаковый набор колонок – отличие от Pig, где кортежи могут иметь разный набор полей)

Создание таблицы

Hive работает в интерактивной оболочке, по аналогии с *grunt* для *Pig*.

```
$ hive
hive> !cat posts.txt;
1       My new post     134318456372
2       Funny story     134318456849
4       New story in blog   1343184571192
1       One more post   134318458874
```

Создание таблицы

```
hive> CREATE TABLE posts (user string, post STRING, time BIGINT)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.04 seconds

hive> show tables;
OK
employee
posts

hive> describe posts;
OK
user                    string
post                    string
time                    bigint
Time taken: 0.108 seconds, Fetched: 3 row(s)
```

Загрузка данных в таблицу

```
hive> LOAD DATA LOCAL INPATH 'posts.txt'
    > OVERWRITE INTO TABLE posts;
Loading data to table default.posts
Table default.posts stats: [numFiles=1, numRows=0, totalSize=118, rawDataSize=0]
OK
Time taken: 0.29 seconds
```

Запросы к таблицам

```
hive> select * from posts;
OK
1       My new post         134318456372
2       Funny story         134318456849
4       New story in blog   1343184571192
1       One more post       134318458874
Time taken: 0.028 seconds, Fetched: 4 row(s)
```

В HDFS таблица хранится в папке:

```
[cloudera@quickstart lab02]$ hdfs dfs -cat /user/hive/warehouse/posts/posts.txt
1       My new post         134318456372
2       Funny story         134318456849
4       New story in blog   1343184571192
1       One more post       134318458874
```

Запросы к таблицам

hive> **select count(*) from posts;**
Query ID = cloudera_20231113102828_ed04e2b3-0a2c-4595-ab91-f44a12f985ac
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
Starting Job = **job_1695730907710_0024**, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_1695730907710_0024/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1695730907710_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2023-11-13 10:28:23,287 Stage-1 map = 0%, reduce = 0%
2023-11-13 10:28:32,194 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.06 sec
2023-11-13 10:28:42,150 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.75 sec
MapReduce Total cumulative CPU time: 2 seconds 750 msec
Ended Job = job_1695730907710_0024
OK
4
Time taken: **30.918 seconds**, Fetched: 1 row(s)

Запросы к таблицам

```
hive> select * from posts where user = '1' ;
OK
1       My new post   134318456372
1       One more post         134318458874
Time taken: 0.934 seconds, Fetched: 2 row(s)
```

```
hive> select count(*) from posts where time > 134318456849;
OK
2
Time taken: 36.678 seconds, Fetched: 1 row(s)
```

GROUP BY

Синтаксис схож со стандартным SQL-запросом:

```
SELECT user, AVG(time)
FROM posts
GROUP BY user;
```

Функции аггрегации:
SUM()
COUNT()
AVG()
MIN()
MAX()

JOIN

Синтаксис схож со стандартным SQL-запросом:

```
SELECT column1,column2
FROM table1
INNER JOIN table2
ON table1.column1 = table2.column1;
```

```
SELECT column1,column2
FROM table1
LEFT JOIN table2
ON table1.column1 = table2.column1;
```

Удаление таблицы

```
hive> DROP TABLE posts;
OK
Time taken: 0.176 seconds

hive> show tables;
OK
employee
Time taken: 0.049 seconds, Fetched: 1 row(s)


[cloudera@quickstart lab02]$ hdfs dfs -ls /user/hive/warehouse/
Found 1 items
drwxrwxrwx   - cloudera supergroup          0 2023-10-22 20:56 /user/hive/warehouse/employee
```