академия больших данных

mail.ru group

SQL поверх больших данных

Дмитрий Музалевский

03.10.2020

План занятия

1. Hive
2. HQL
3. Athena
4. Практика

MapReduce и SQL

Масштабируемость MapReduce:

- Гибкость SQL
- Легкость диалекта

Решение: Комбинация MapReduce и SQL

Hive

База данных/Хранилище данных поверх Hadoop:

- Использование Rich Data Types (листы и мапы)
- Эффективность использования where, joins и groupby в MapReduce
- Использование SerDe (интерфейса сериализации/десериализации)
- Возможность гибко писать свои версии SerDe
- Эволюция схем в Hive

MetaStore

Опции таблиц и партишнов:

- Схема таблицы и SerDe
- Локация таблицы на HDFS
- Логические партишны ключей и типов
- Другая информация

Thrift API:

- Python interface
- Java interface
- PHP interface
- Perl interface

Hive CLI

DDL:

- Create Table/Drop Table/Rename Table
- Alter Table Add Column

Browsing:

- Show Tables
- Describe Table
- Cat Table

Загрузка Данных

Запросы

Web UI Hive

MetaStore UI:

- Обзор и навигация по всем таблицам системы
- Комментирование по каждой таблице и колонке
- Работа с зависимостями данных

HiPal:

- Интерактивное построение SQL запросов с помощью мышки
- Поддерживает фильтеринг, группинг и джоины

Hive Query Language (HQL)

Философия:

- SQL
- MapReduce с кастомными скриптами

Команды запросов:

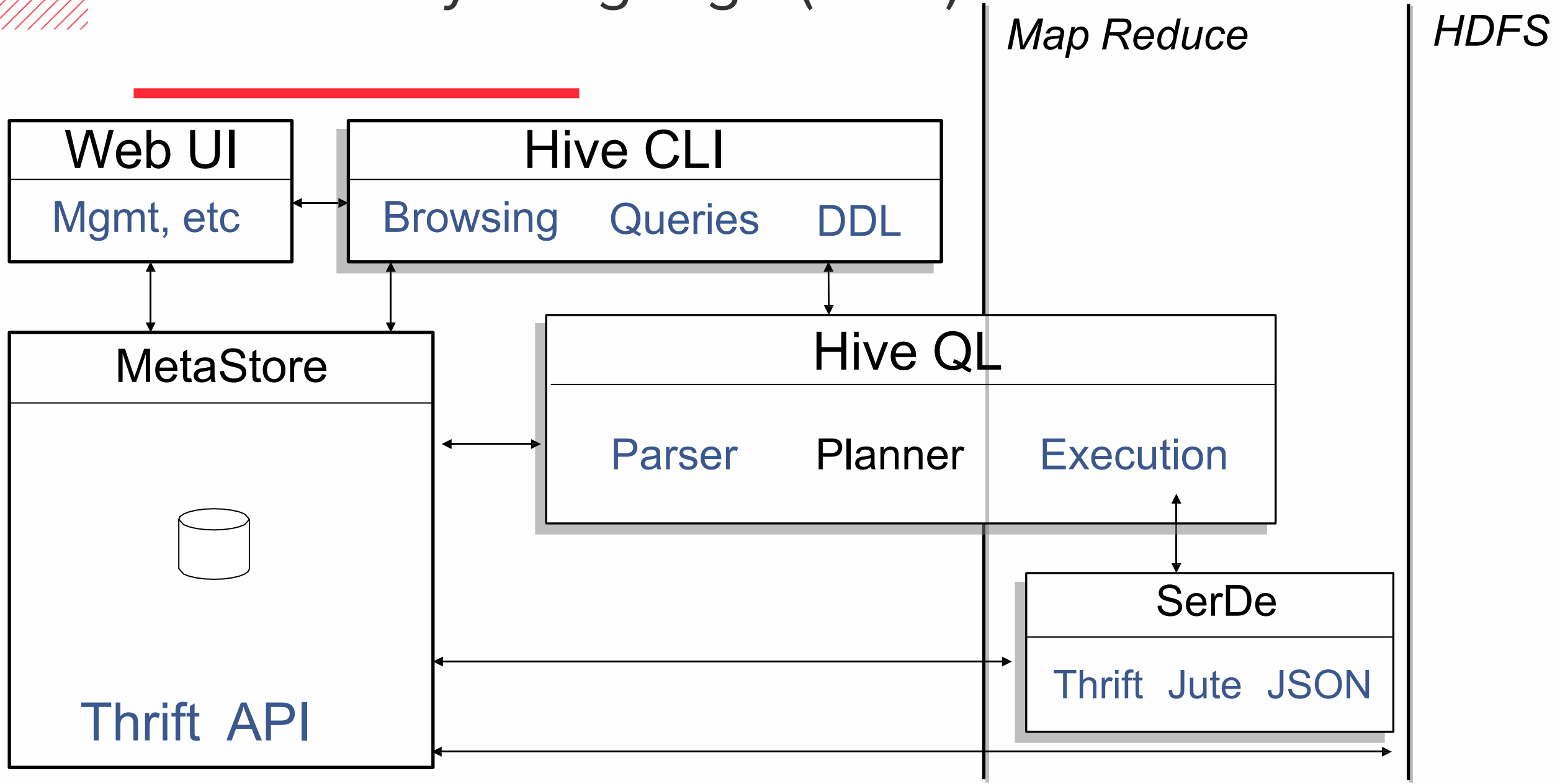
- Фильтрация
- Group by
- Sampling
- Order by
- Joins

Hive Query Language (HQL)

Расширенный SQL:

- FROM (
  - FROM pv_users
  - **MAP** pv_users.userid, pv_users.date
  - **USING** 'map_script' AS (dt, uid)
  - **CLUSTER BY** dt) map
- INSERT INTO TABLE pv_users_reduced
  - **REDUCE** map.dt, map.uid
  - **USING** 'reduce_script' AS (date, count);

Hive Query Language (HQL)



Hive QL - Join

page_view

page id	**user id**	time
1	**111**	9:08:01
2	**111**	9:08:13
1	**222**	9:08:14

X

user

user id	age	gender
111	25	female
222	32	male

=

pv_users

page id	age
1	25
2	25
1	32

```
INSERT INTO TABLE
pv_users  SELECT pv.pageid,
u.age
FROM page_view pv JOIN user u ON (pv.userid = u.userid);
```

Hive QL – Join MapReduce

page id	**user id**	time
1	**111**	9:08:01
2	**111**	9:08:13
1	**222**	9:08:14

user id	age	gender
111	25	female
222	32	male

Map

key	value
111	<**1**,1>
111	<**1**,2>
222	<**1**,1>

key	value
111	<**2**,25>
222	<**2**,32>

Shuffle
Sort

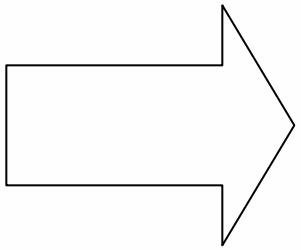
key	value
111	<**1**,1>
111	<**1**,2>
111	<**2**,25>

Reduce

key	value
222	<**1**,1>
222	<**2**,32>

Hive QL – GroupBY

page id	age
1	25
2	25
1	32
2	25



pageid_age_sum

pageid	age	Count
1	25	1
2	25	2
1	32	1

- INSERT INTO TABLE pageid_age_sum
- SELECT pageid, age, count(1)
- FROM pv_users
- GROUP BY pageid, age;

Hive QL – OrderBY

pageid	userid	time
2	111	9:08:13
1	111	9:08:01

pageid	userid	time
2	111	9:08:20
1	222	9:08:14

Shuffle

key	v
<1,111>	9:08:01
<2,111>	9:08:13

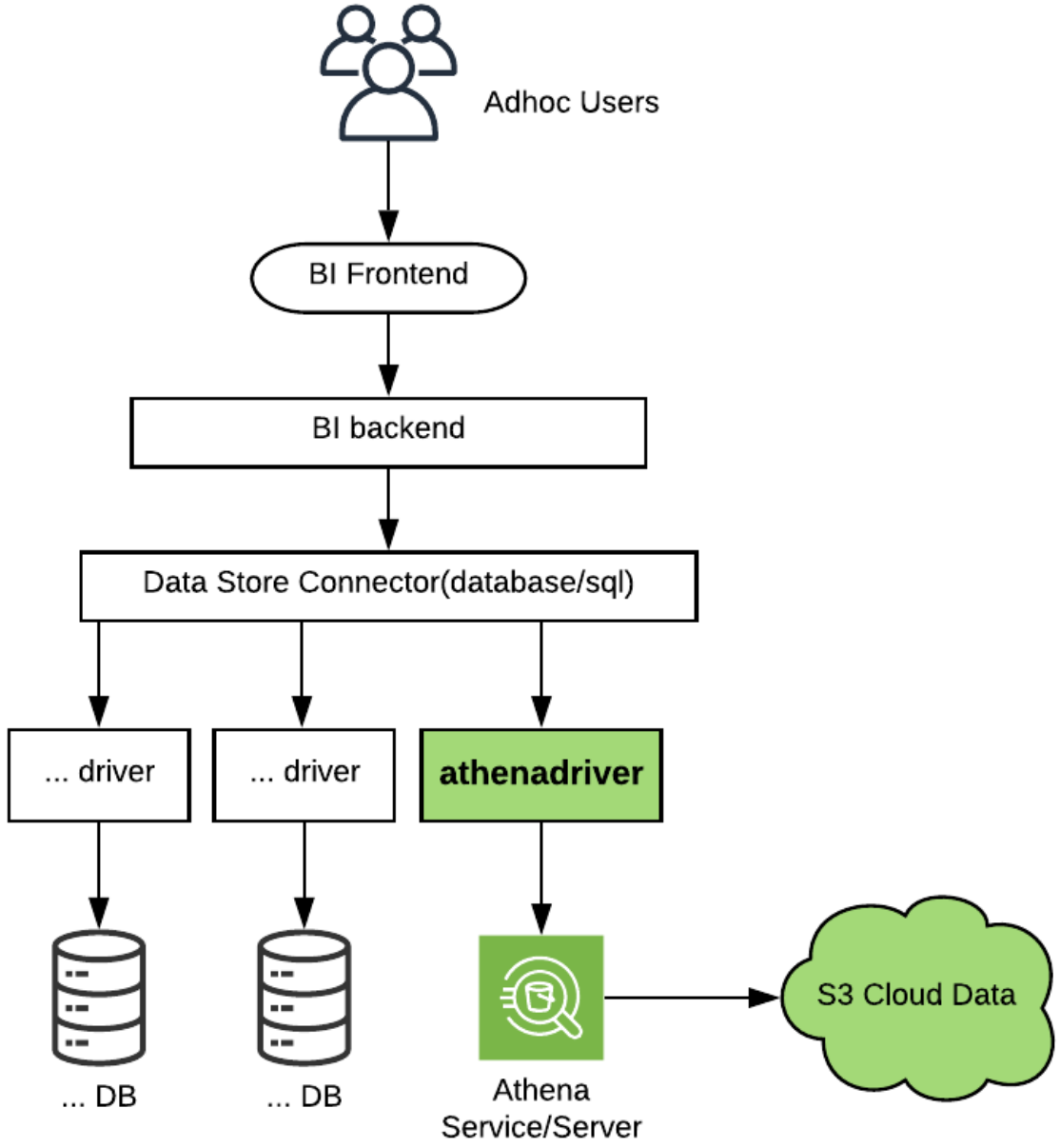
key	v
<1,222>	9:08:14
<2,111>	9:08:20

Reduce

pageid	userid	
1	111	
2	111	

pageid	userid	
1	222	
2	111	

Athena


Adhoc Users
BI Frontend
BI backend
Data Store Connector(database/sql)
... driver
... driver
athenadriver
... DB
... DB
Athena
Service/Server
S3 Cloud Data

Athena


Source 1
Source 2
Source 3
Source N
Data load into S3
S3 Staging
Athena Transformation
Event
S3 Reporting
Athena Reporting
Event
Amazon QuickSight Dashboard


Hello?
is it me
you're looking for?
and I want to tell you so much
I love you
are you somewhere feeling lonely
or is someone loving you
tell me how to win your heart
but we can start by saying