Hive

План семинара

1. Кратко о Hive и интерфейсе Hive
2. Импорт данных. Regex
3. Custom mapper (Hive streaming)

Конспект

https://gitlab.com/VeLKerr/pardistrib/-/blob/master/distribute/practice/04-Hive.md

Интерфейс взаимодействия

Доступ через:

- **Hive CLI**. Проприетарный Thrift-based протокол (RPC для Java)
  `$ hive`

- **JDBC**
  На учебном кластере доступен **HUE**:
  `ssh <MIPT_HADOOP> -N -L 8888:mipt-node03.atp-fivt.org:8888`
  (затем — `http://localhost:8888/`)

Возможности Hive

Суть™:
Предоставляется SQL-***подобный*** язык (HiveQL).

Cheatsheet **SQL-to-HiveQL**:
http://hortonworks.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf

Ранее в Hive присутствовали **ограничения Hadoop** (напр., отсутствие транзакций). Сейчас они так или иначе **решены**

- OLTP = Online Transaction Processing ← Hive не годится
- OLAP = Online Analytical Processing ← Hive подходит

Запросы...

См. также: Создание базы данных.

```
-- Необходимая однократно строка
ADD JAR /opt/cloudera/parcels/CDH/lib/hive/lib/hive-contrib.jar;

-- Таблица для импорта данных («внешняя»). Различия
CREATE EXTERNAL TABLE Subnets (
    ip STRING,
    mask STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY  '\t'
STORED AS TEXTFILE
LOCATION '/data/subnets/variant1';
```

Запросы...

```
-- Секционированная (partitioned) таблица
CREATE EXTERNAL TABLE SubnetsPart (
    ip STRING
)
PARTITIONED BY (mask STRING)
STORED AS TEXTFILE;

-- OVERWRITE: Предварительно удалить все существующие данные
INSERT OVERWRITE TABLE SubnetsPart PARTITION (mask)
SELECT * FROM Subnets;
```

Запросы...

```
-- Hive SerDe: специальные форматы ввода
ADD JAR /opt/cloudera/parcels/CDH/lib/hive/lib/hive-serde.jar;

CREATE EXTERNAL TABLE SerDeExample (
    ip STRING,
    date STRING,
    request STRING,
    responseCode STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
    "input.regex" = '^(\\S*)\\t.*$'  -- \ нужно экранировать!
)
STORED AS TEXTFILE
LOCATION '/data/user_logs/user_logs_S';
```

Для проверки regex'ов: https://regex101.com/

Запросы…

```
-- Подсчитать число различных масок подсети: элементарно
SELECT COUNT(DISTINCT mask)
FROM Subnets;
```

Здесь секционирование создаёт **лишние** накладные расходы

Почему?

Запросы...

```
-- Подсчитать число различных масок подсети: элементарно
SELECT COUNT(DISTINCT mask)
FROM Subnets;
```

Здесь секционирование создаёт **лишние** накладные расходы

Потому что необходимо просмотреть **всю таблицу** (sequential scan).

Но зато **map** можно сделать «**более параллельным**»!

Поэтому **секционирование** на большом dataset
полезно в большинстве случаев.

Запросы...

```
-- Hive streaming (custom mapper; бывает и reducer)
SELECT TRANSFORM(ip)
USING 'cut -d . -f 1' AS ip
FROM Subnets
LIMIT 10;

-- То же, но код вынесен в отдельный файл
ADD FILE ./script.sh;

SELECT TRANSFORM(ip)
USING './script.sh' AS ip2
FROM Subnets
LIMIT 10;
```

Полезные материалы

Hive Language Manual
https://cwiki.apache.org/confluence/display/Hive/LanguageManual