Цель модуля “SQL поверх больших данных (Hive)”

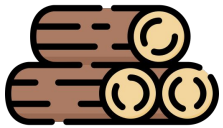
Драль Алексей, study@bigdatateam.org
CEO at BigData Team, https://bigdatateam.org
https://www.facebook.com/bigdatateam

BIGDATA TEAM

1. Наиболее популярные регионы

BIGDATA TEAM

1. Наиболее популярные регионы

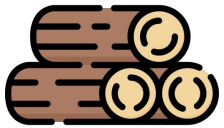


Web-service access logs

BIGDATA TEAM

1. Наиболее популярные регионы

Web-service access logs



Geobase

IPv4: 109.188.67.224

↓

area: Moscow City Center

↓

city: Moscow

↓

country: Russia

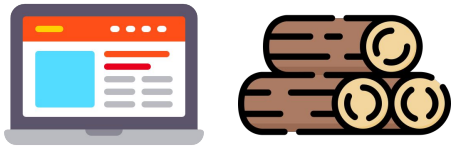
↓

Earth

109.188.67.224, Moscow City Centre, Moscow, Russia, Earth
109.188.67.221, Moscow City Centre, Moscow, Russia, Earth
...

BIGDATA TEAM

1. Наиболее популярные регионы



Web-service access logs

Geobase

`format: ip, request, status_code,...` `format: ip, region`$_{city}$`, region`$_{country}$`,...`

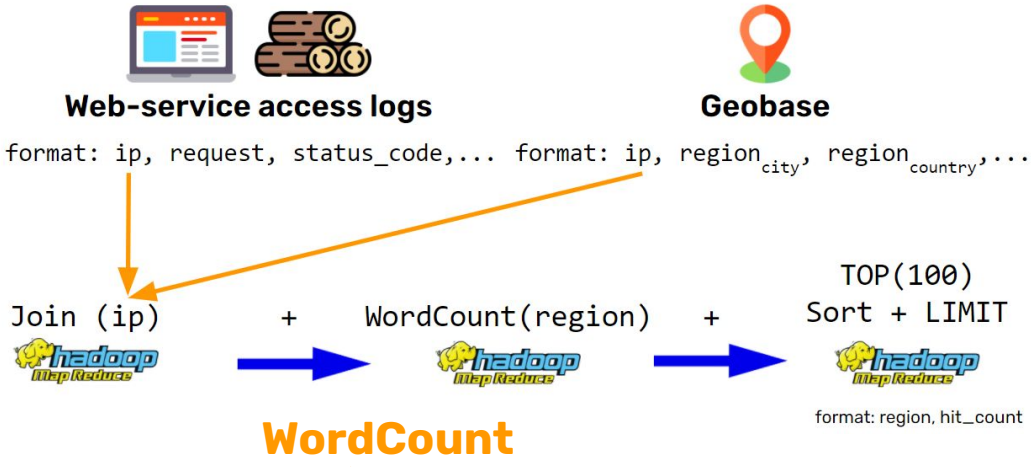
`Join (ip)` + `WordCount(region)` + `TOP(100)` `Sort + LIMIT`

hadoop Map Reduce → hadoop Map Reduce → hadoop Map Reduce

format: region, hit_count

BIGDATA TEAM

1. Наиболее популярные регионы



WordCount

Hive QL

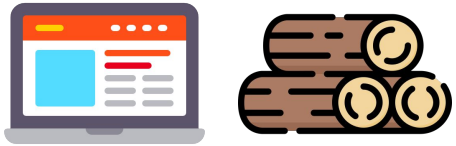
```
SELECT regioncity, COUNT(1) AS hit_count
FROM access_log JOIN geo_base
ON (access_log.host = geo_base.host)
GROUP BY regioncity ORDER BY hit_count LIMIT 100
```

join

TOP(100)

BIGDATA TEAM

2. Доля роботных запросов



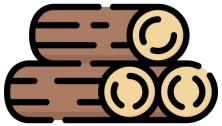
Web-service access logs

bot name + $user\_agent_1$ $user\_agent_2$ ... + ip_1 ip_2 ...

[ro]bot database

BIGDATA TEAM

2. Доля роботных запросов



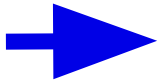
Web-service access logs

format: ip, request, user_agent,...

[ro]bot database

format: bot_name, user_agents, ips

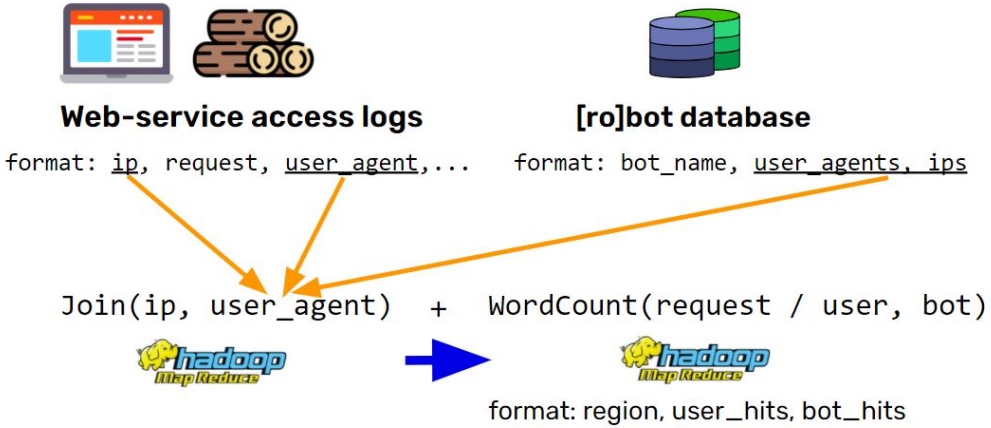
Join(ip, user_agent) + WordCount(request / user, bot)



format: region, user_hits, bot_hits

BIGDATA TEAM

2. Доля роботных запросов

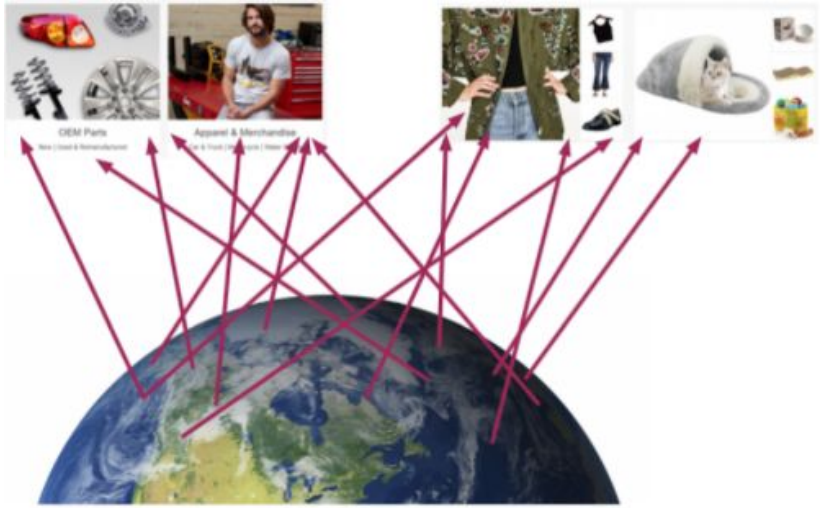


Hive QL


```
SELECT request,
  SUM(IF(robot.bot_name IS NULL, 1, 0)) as user_hit_count,
  SUM(IF(robot.bot_name IS NOT NULL, 1, 0)) as bot_hit_count
FROM access_log LEFT OUTER JOIN robot ON (
  access_log.host = robot.host
  AND access_log.user_agent = robot.user_agent
)
GROUP BY request
```

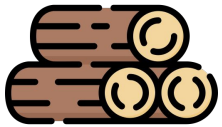
BIGDATA TEAM

3. Гендерное распределение



user_id	age	gender	occupation	zipcode
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
6	42	M	executive	98101
7	57	M	administrator	91344
8	36	M	administrator	05201
9	29	M	student	01002
10	53	M	lawyer	90703

Web-service access logs

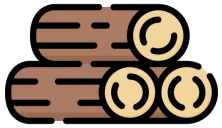


Geobase

User personal data

BIGDATA TEAM

3. Гендерное распределение в регионах

Web-service access logs

User personal data

Geobase

format: user_id, gender, age,...

format: ip, request, user_agent,...

format: ip, region$_{city}$, region$_{country}$,...

Join(user_id) + Join(ip) + WordCount(region/gender)



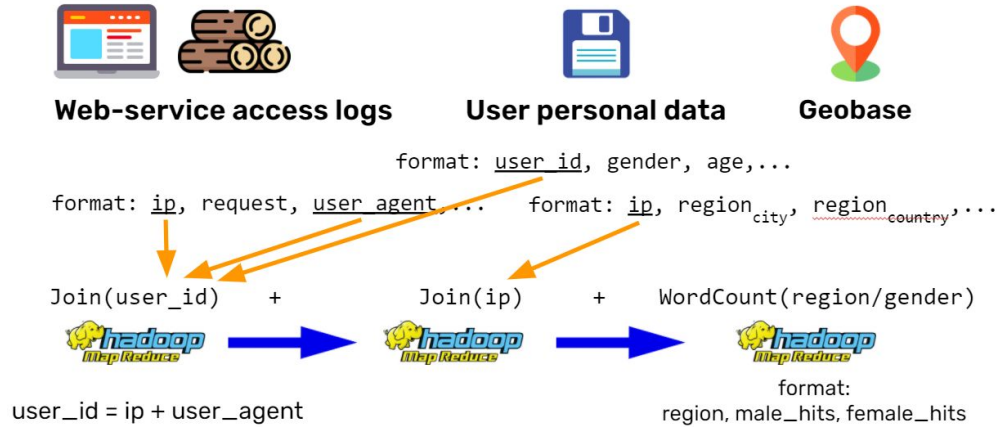


user_id = ip + user_agent

format:
region, male_hits, female_hits

BIGDATA TEAM

3. Гендерное распределение в регионах



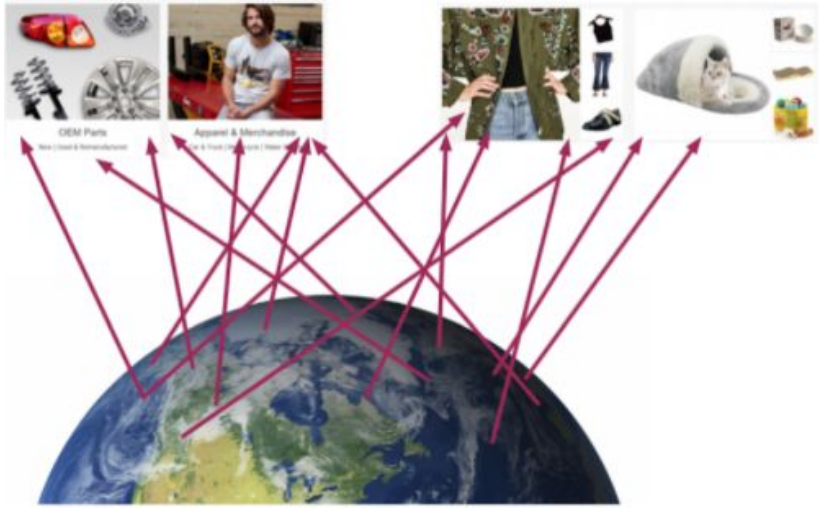
Hive QL


```
SELECT region_city,
  SUM(IF(user.gender = "M",1,0)) as male_hit_count,
  SUM(IF(user.gender = "F",1,0)) as female_hit_count
FROM access_log
  JOIN geo_base ON (access_log.host = geo_base.host)
  JOIN user ON (access_log.host = user.host
    AND access_log.user_agent = user.user_agent
  )
GROUP BY region_city
```

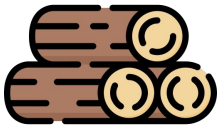
two joins

BIGDATA TEAM

4. Средний возраст клиента



user_id	age	gender	occupation	zipcode
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
6	42	M	executive	98101
7	57	M	administrator	91344
8	36	M	administrator	05201
9	29	M	student	01002
10	53	M	lawyer	90703



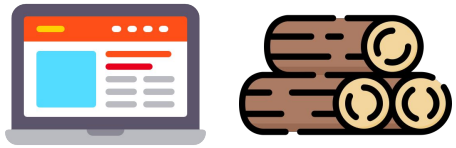
Web-service access logs

Geobase

User personal data

BIGDATA TEAM

4. Средний возраст клиента



Web-service access logs

User personal data

Geobase

format: user_id, gender, age,...

format: ip, request, user_agent,...

format: ip, region$_{city}$, region$_{country}$,...

Join(user_id) + Join(ip) + Average(age)

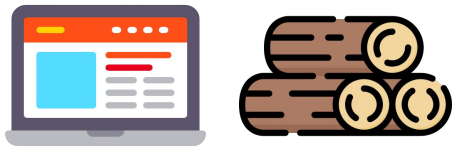
hadoop Map Reduce → hadoop Map Reduce → hadoop Map Reduce

format: region, average_age

user_id = ip + user_agent

BIGDATA TEAM

4. Средний возраст клиента



Web-service access logs

User personal data

Geobase

format: user_id, gender, age,...

format: ip, request, user_agent,...

format: ip, region$_{city}$, region$_{country}$,...

Join(user_id) + Join(ip) +

Average(age)

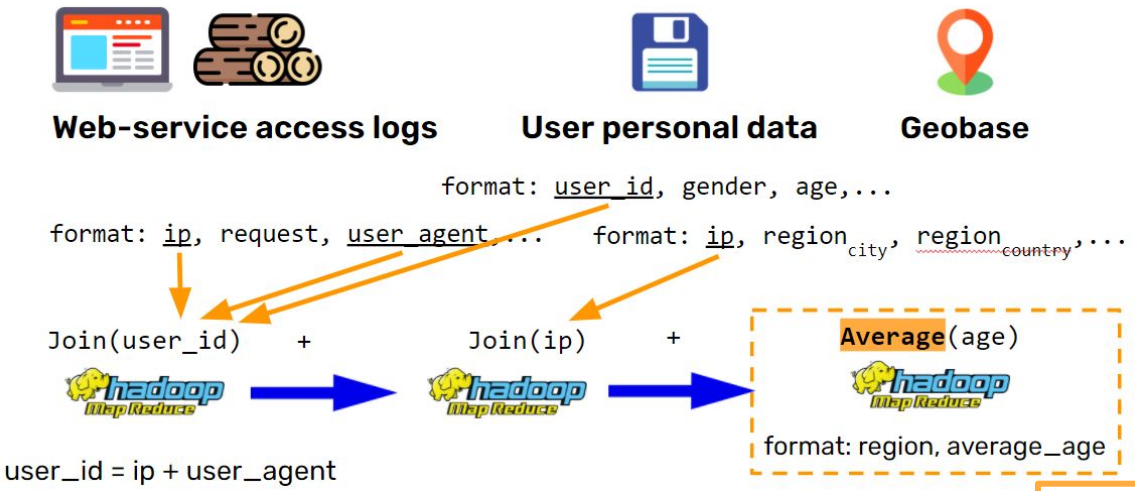
format: region, average_age

user_id = ip + user_agent

use Combiner for optimisation

BIGDATA TEAM

4. Средний возраст клиента



GenericUDAFAverage
Hive Java source code:
https://rebrand.ly/hql-avg

Hive QL


```
SELECT region_city, AVG(user.age)
FROM access_log
      JOIN geo_base ON (access_log.host = geo_base.host)
      JOIN user ON (access_log.host = user.host
            AND access_log.user_agent = user.user_agent
      )
GROUP BY region_2
```

План

План

- Map-Side Join
- Reduce-Side Join
- Bucket Map-Side и SMB Join

План

- Map-Side Join
- Reduce-Side Join
- Bucket Map-Side и SMB Join

А также:

- Как теряют production данные
- Как правильно пользоваться RegExpSerDe