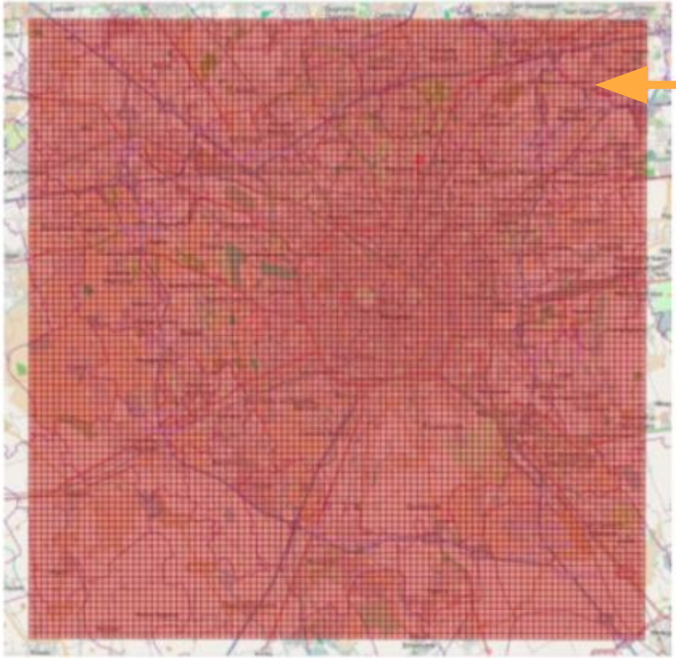
Map-Side Join

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

Telecommunications Dataset



Milano Grid

- Square ID
- Time Interval
- Country Code
- SMS-in Activity
- SMS-out Activity
- Call-in Activity
- Call-out Activity
- Internet Traffic Activity

Schema

https://dandelion.eu/datagems/SpazioDati/telecom-sms-call-internet-mi

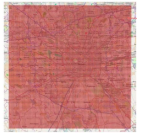
Telecommunications Dataset

BIG

- Square ID
- Time Interval
- Country Code
- SMS-in Activity
- SMS-out Activity
- Call-in Activity
- Call-out Activity
- Internet Traffic Activity

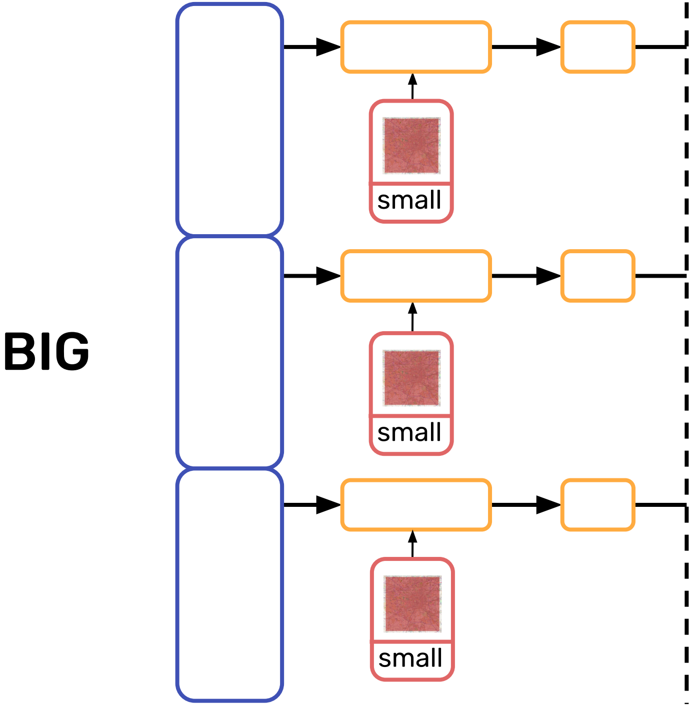
```
1  1383260400000 0 0.08136262351125882
1  1383260400000 39 0.14186425470242922
0.1567870050390246 0.16093793691701822
0.052274848528573205 11.028366381681026
1  1383261000000 0 0.13658782275823106
0.02730046487718618
1  13832610000000 33
0.026137424264286602
```

small



{‘type’: ‘Polygon’, ‘coordinates’:
[[[9.0114910478323, 45.35880131440966],
[9.014491488013135, 45.35880097314403],
[9.0144909480813, 45.35668565341486],
[9.011490619692509,
45.356685994655464], [9.0114910478323,
45.35880131440966]]]}
...

Map- Side Join



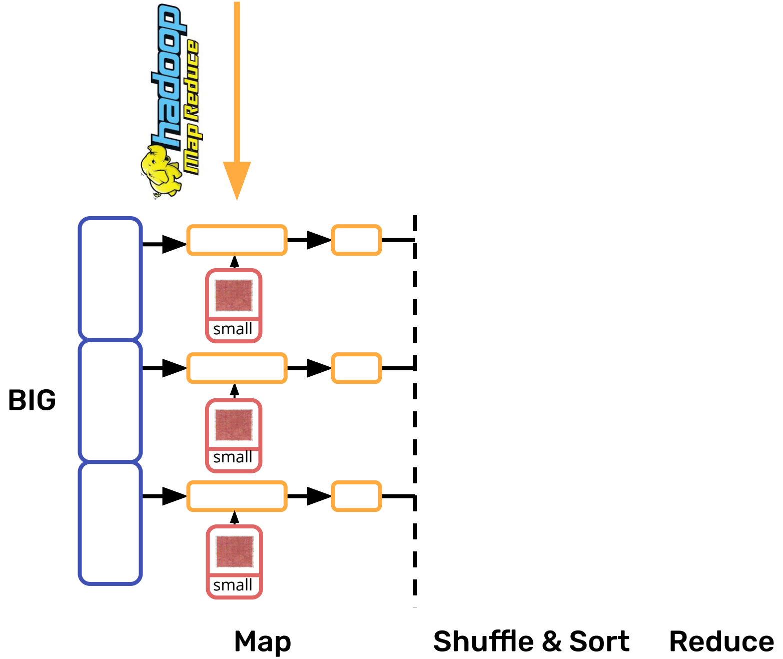
Map- Side Join

```
yarn jar $HADOOP_STREAMING_JAR \
        -files map_side_mapper.py,hdfs:///user/adral/milano-grid.geojson \
        -mapper "python3 map_side_mapper.py" \
        -numRediceTasks 0 \
        -input /data/telecommunication \
        -output telecom-joins
```

HDFS data
Distributed Cache

Hive Map- Side Join

SELECT region$_{\text{city}}$, **COUNT(1)** AS hit_count
FROM access_log **JOIN** geo_base
ON (access_log.host = geo_base.host)
GROUP BY region$_{\text{city}}$ **ORDER BY** hit_count **LIMIT 100**



Hive Map- Side Join

```
yarn jar $HADOOP_STREAMING_JAR \
        -files map_side_mapper.py,hdfs:///user/adral/milano-grid.geojson \
        -mapper "python3 map_side_mapper.py" \
        -numRediceTasks 0 \
        -input /data/telecommunication \
        -output telecom-joins
```

HDFS data
Distributed Cache

1. Клиент: загружает датасет из HDFS
2. Строит hashtable
3. Загружает hashtable в Distributed Cache