Massive Data Processing

Taxi’s data

Mariano Garralda Barrio
Oscar Ujaque Perez

1. Parsing Data

- **Initial data → CSV**
- **Final data → JSON**
- **Weight → 11MB and 1.9 GB**
- **How? → Using Open Refine**

Data output in json

Trip ID : String

Call Type : Char

- Type A: Trip dispatched from the central
- Type B: Trip dispatched from a stand
- Type C: Trip dispatched randomly in a street

Origin Call: Integer if Call Type = ‘A’. Null otherwise.

Origin Stand: Integer if Call Type = ‘B’. Null otherwise

Taxi Id : Integer

TimeStamp: Integer

Day Type: Char

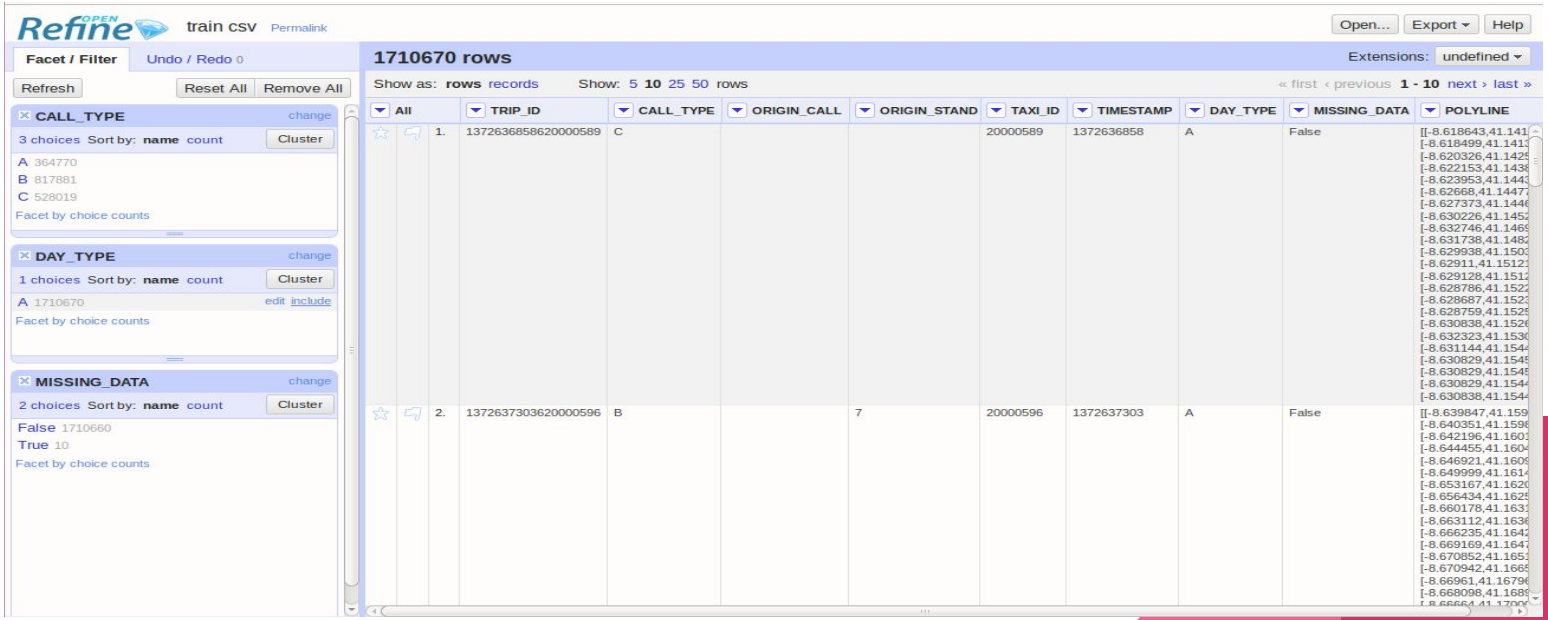
- Type A: normal day
- Type B: on holidays
- Type C: the day before holidays

Missing Data: Boolean. False if there is no missing data in polyline. True otherwise.

Polyline: String with GPS coordinates for each 15 seconds of the trip.

2. Cheking Attributes

- **All attributes checked**



CALL_TYPE change
3 choices Sort by: name count Cluster
A 364770
B 817881
C 528019
Facet by choice counts

DAY_TYPE change
1 choices Sort by: name count Cluster
A 1710670 edit include
Facet by choice counts

MISSING_DATA change
2 choices Sort by: name count Cluster
False 1710660
True 10
Facet by choice counts

1710670 rows

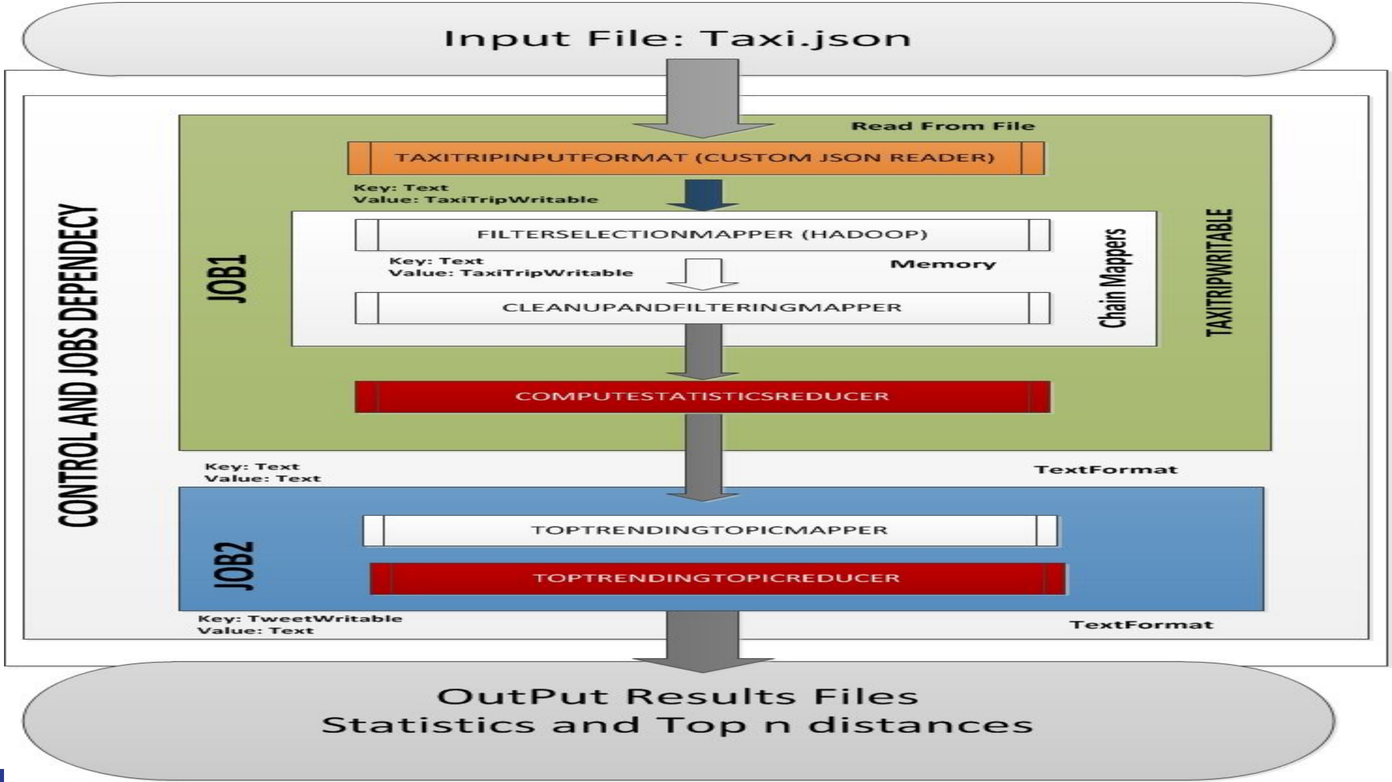
Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next › last »

All	TRIP_ID	CALL_TYPE	ORIGIN_CALL	ORIGIN_STAND	TAXI_ID	TIMESTAMP	DAY_TYPE	MISSING_DATA	POLYLINE
1.	1372636858620000589	C			20000589	1372636858	A	False	[[-8.618643,41.141 [-8.618499,41.1413 [-8.620326,41.1425 [-8.622153,41.1438 [-8.623953,41.1443 [-8.62668,41.1447 [-8.627373,41.1446 [-8.630226,41.1452 [-8.632746,41.1469 [-8.631738,41.1482 [-8.629938,41.1503 [-8.62911,41.1512 [-8.629128,41.1512 [-8.628786,41.1522 [-8.628687,41.1523 [-8.628759,41.1525 [-8.630838,41.1526 [-8.632323,41.1530 [-8.631144,41.1544 [-8.630829,41.1545 [-8.630829,41.1545 [-8.630829,41.1544 [-8.630838,41.1544
2.	1372637303620000596	B		7	20000596	1372637303	A	False	[[-8.639847,41.159 [-8.640351,41.1598 [-8.642196,41.1601 [-8.644455,41.1604 [-8.646921,41.1609 [-8.649999,41.1614 [-8.653167,41.1620 [-8.656434,41.1625 [-8.660178,41.1631 [-8.663112,41.1636 [-8.666235,41.1642 [-8.669169,41.1647 [-8.670852,41.1651 [-8.670942,41.1665 [-8.66961,41.16796 [-8.668098,41.1689 [-8.66664,41.1700

3. Map Reduce Tasks

- **Statistics computed:**
  - **Max Distance**
  - **Distance average**
  - **Max velocity**
  - **Velocity average**
  - **Max Trip time**
  - **Trip time Average**
  - **Number of trips**

4. Structure



Input File: Taxi.json
CONTROL AND JOBS DEPENDECY
JOB1
Read From File
TAXITRIPINPUTFORMAT (CUSTOM JSON READER)
Key: Text
Value: TaxiTripWritable
FILTERSELECTIONMAPPER (HADOOP)
Key: Text
Value: TaxiTripWritable
Memory
CLEANUPANDFILTERINGMAPPER
Chain Mappers
TAXITRIPWRITABLE
COMPUTESTATISTICSREDUCER
Key: Text
Value: Text
TextFormat
JOB2
TOPTRENDINGTOPICMAPPER
TOPTRENDINGTOPICREDUCER
Key: TweetWritable
Value: Text
TextFormat
OutPut Results Files
Statistics and Top n distances

5. Advanced structures used

- **Custom Input File Format: For reading the json file.**
- **Custom Writables:**
  - **TaxiTripWritable: For storing the json data after reading**
  - **GpsPositionWritable: For storing the gps coordinates after reading.**
  - **ArrayWritable<GpsPositioWritable>:**
- **Chain Mappers: For chaining two mappers in first job.**
- **Mappers from hadoop: FieldSelectionMappers: For treating the input data.**
- **Job Control and dependencies: For running two jobs.**

Thank You