

# Massive Data Processing

## Taxi's data

Mariano Garralda Barrio  
Oscar Ujaque Perez

# 0. Introduction

- **Taxis from Porto (Portugal)**
- **Data from 13/07/13 to 30/06/14**
- **442 taxis running**



# 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 important attributes checked

Refine train CSV Permalink

Open... Export Help

Facet / Filter Undo / Redo 0

1710670 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: undefined

« 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	[[-8.618643,41.1411], [-8.618499,41.1411], [-8.620326,41.1425], [-8.622153,41.1438], [-8.623953,41.1444], [-8.62668,41.1447], [-8.627373,41.1444], [-8.630226,41.1452], [-8.632746,41.1468], [-8.631738,41.1485], [-8.629938,41.1503], [-8.62911,41.1512], [-8.629128,41.1511], [-8.628786,41.1522], [-8.628687,41.1522], [-8.628759,41.1522], [-8.630838,41.1524], [-8.632323,41.1536], [-8.631144,41.1544], [-8.630829,41.1545], [-8.630829,41.1545], [-8.630829,41.1544], [-8.630838,41.1544], [-8.630838,41.1544], [-8.639847,41.159], [-8.640351,41.1596], [-8.642196,41.1601], [-8.644455,41.1604], [-8.646921,41.1605], [-8.649999,41.1611], [-8.653167,41.1625], [-8.656434,41.1625], [-8.660178,41.1631]]]
2.	1372637303620000596	B		7	20000596	1372637303	A	False	[[-8.639847,41.159], [-8.640351,41.1596], [-8.642196,41.1601], [-8.644455,41.1604], [-8.646921,41.1605], [-8.649999,41.1611], [-8.653167,41.1625], [-8.656434,41.1625], [-8.660178,41.1631]]]

### 3. Map Reduce Tasks

- **Statistics computed:**
  - **Max Distance (km)**
  - **Distance average (km)**
  - **Max velocity (km/h)**
  - **Velocity average (km/h)**
  - **Max Trip time (hours)**
  - **Trip time Average (hours)**
  - **Number of trips (counter)**

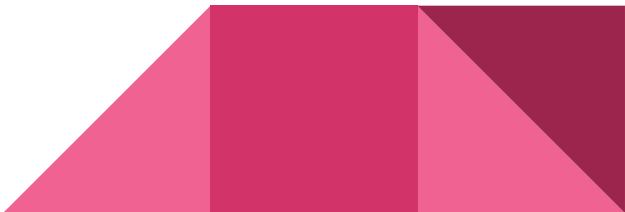


### 3. Map Reduce Tasks

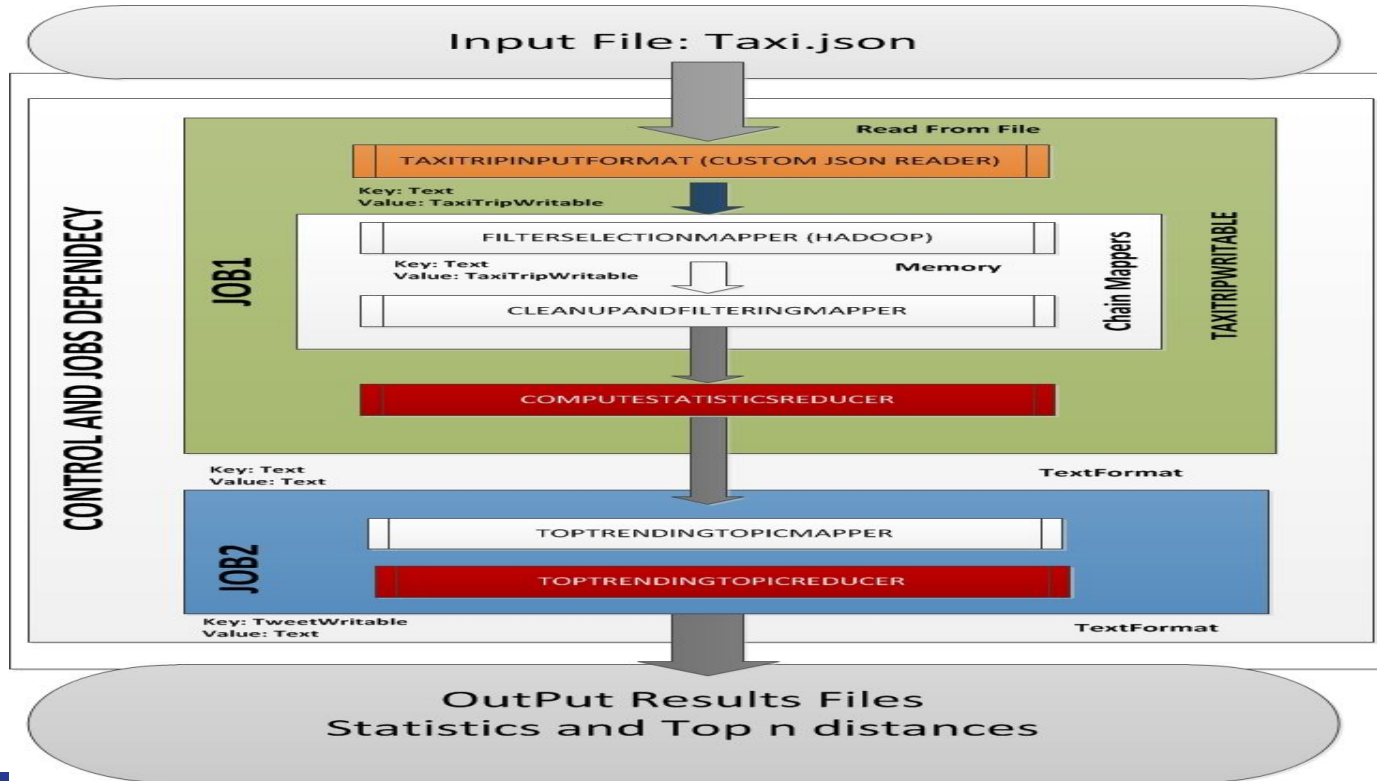
- Other computings

- Top N (Distance in Km)

Top:1	IDTAXI: 20000539	DistanceTOT: 673,33
Top:2	IDTAXI: 20000084	DistanceTOT: 659,35
Top:3	IDTAXI: 20000363	DistanceTOT: 370,22
Top:4	IDTAXI: 20000657	DistanceTOT: 359,22
Top:5	IDTAXI: 20000455	DistanceTOT: 298,05
Top:6	IDTAXI: 20000136	DistanceTOT: 273,92
Top:7	IDTAXI: 20000372	DistanceTOT: 273,17
Top:8	IDTAXI: 20000496	DistanceTOT: 265,24
Top:9	IDTAXI: 20000307	DistanceTOT: 260,62
Top:10	IDTAXI: 20000446	DistanceTOT: 256,01




## 4. Structure





## 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<GpsPositionWritable>:**
  - **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**

