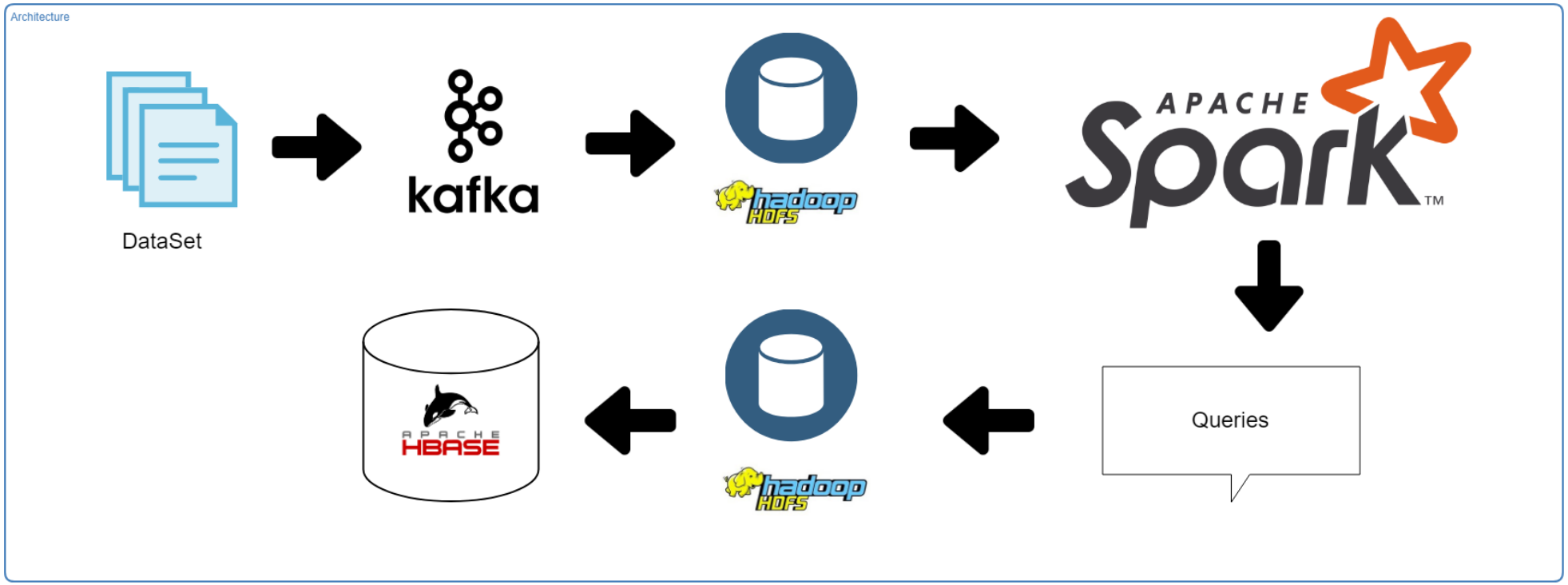
ANALISI DEL DATASET SULLE CONDIZIONI METEOROLOGICHE

SISTEMI E ARCHITETTURE PER BIG DATA - A.A. 2018/19

Di Cosmo Giuseppe
Nedia Salvatore

TECNOLOGIE UTILIZZATE

ARCHITETTURA

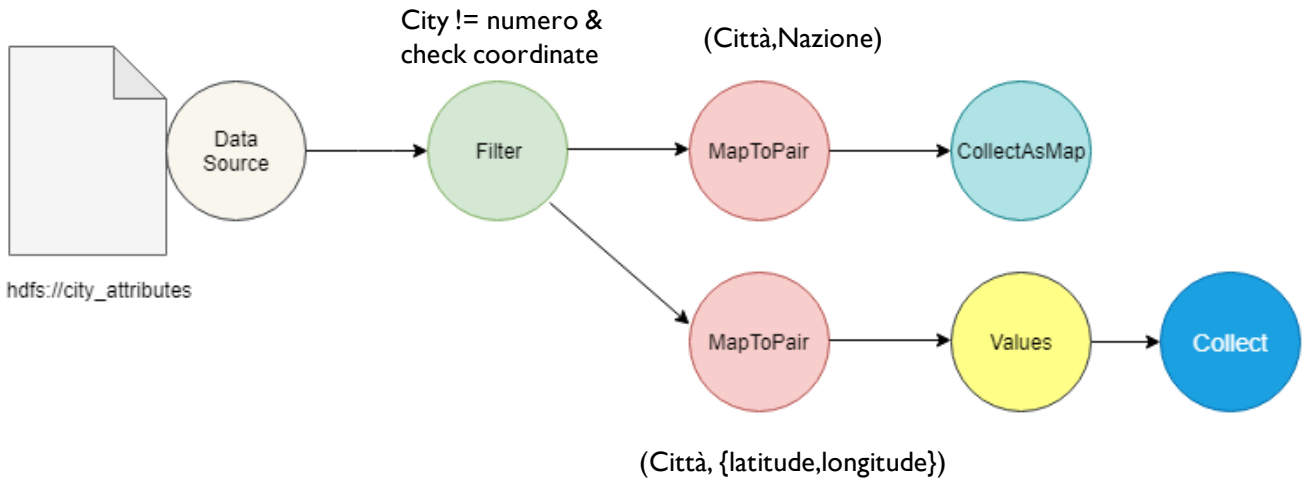


DATA INGESTION

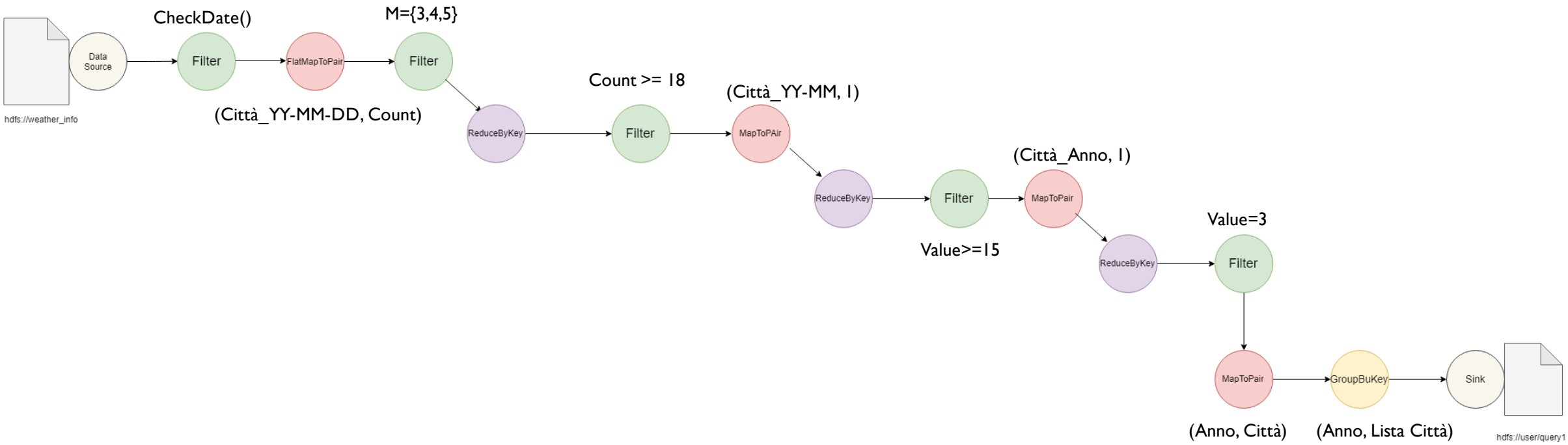
- Un topic per ogni file, con successiva produzione e consumazione
- Tre tipi di formato
  - Csv
  - Avro
  - Parquet
- Scrittura su HDFS

PREPROCESSAMENTO CITY_ATTRIBUTES

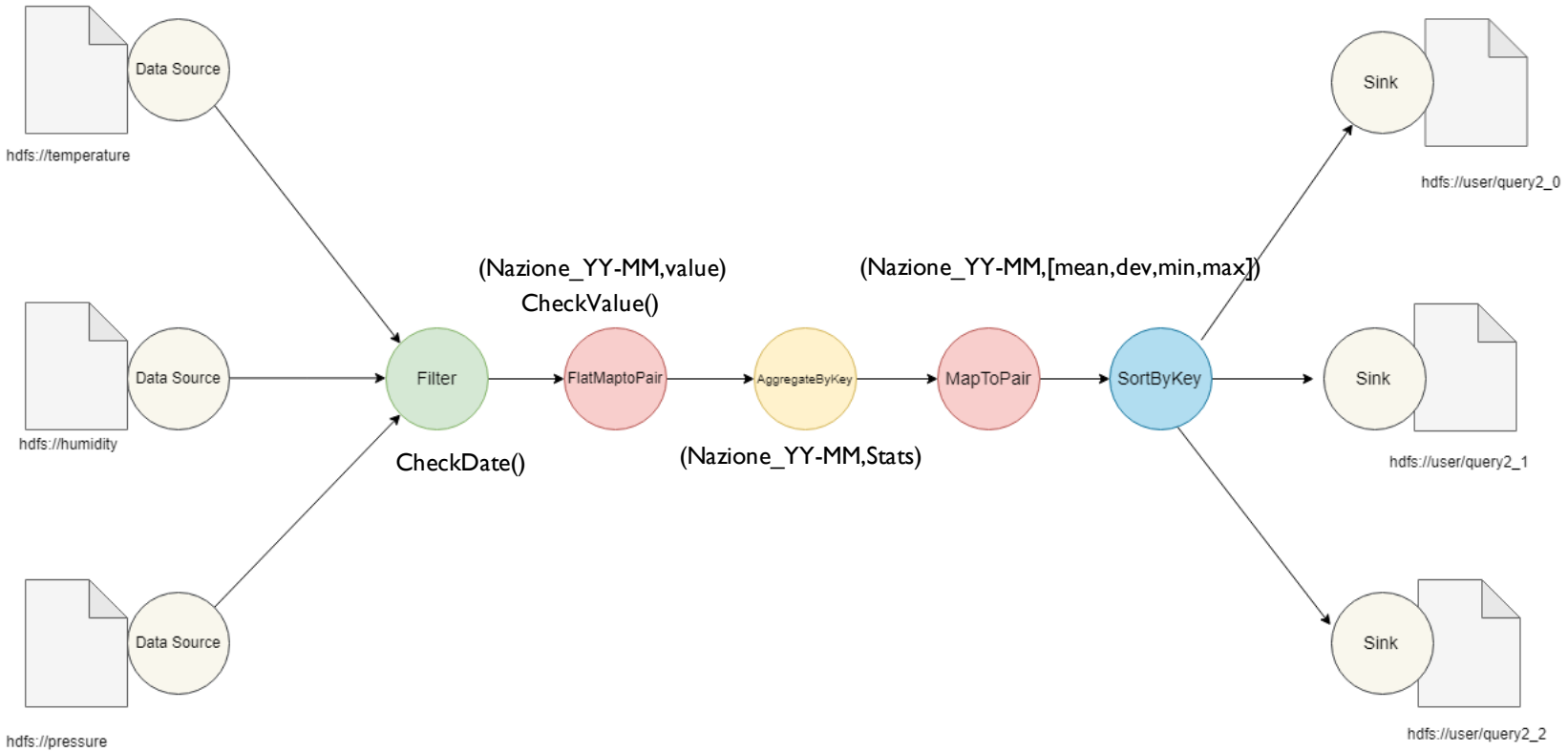
- Necessario avere mapping città/nazione, nazione/UTC, utile per tutte le query.
- Uso libreria *Nominatim* e *Timeshape*



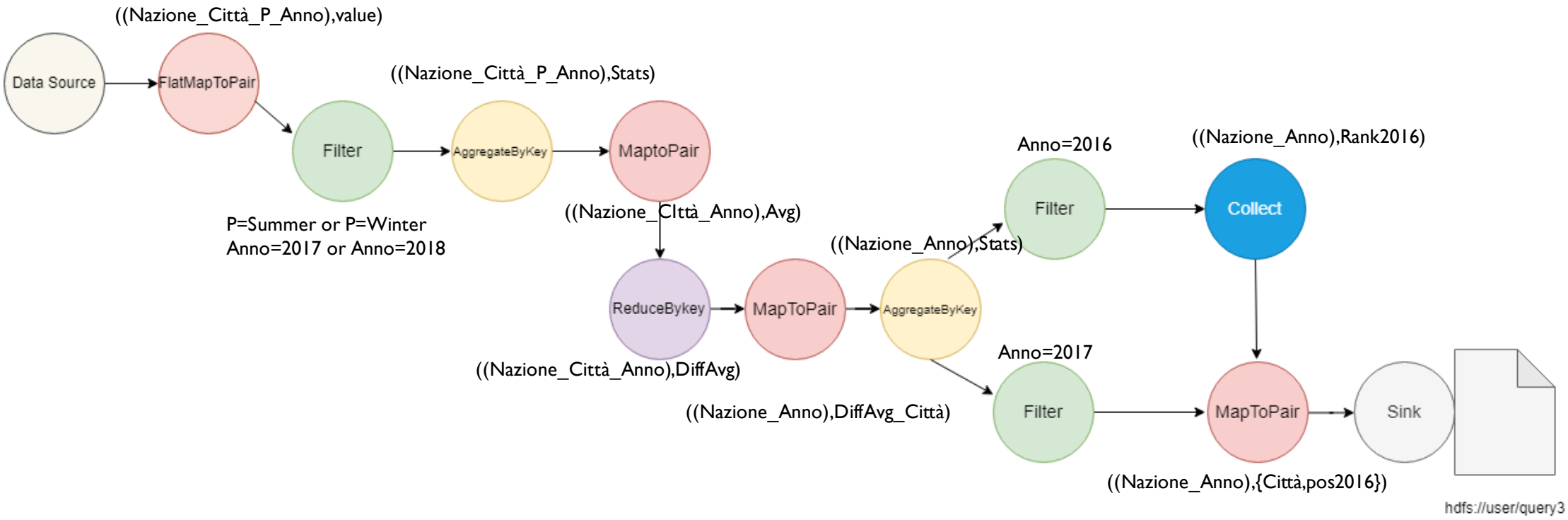
QUERY 1


Data Source
hdfs://weather_info
CheckDate()
Filter
FlatMapToPair
(Città_YY-MM-DD, Count)
M={3,4,5}
Filter
ReduceByKey
Count >= 18
Filter
(Città_YY-MM, 1)
MapToPair
ReduceByKey
Filter
Value>=15
(Città_Anno, 1)
MapToPair
ReduceByKey
Value=3
Filter
MapToPair
(Anno, Città)
GroupBuKey
(Anno, Lista Città)
Sink
hdfs://user/query1

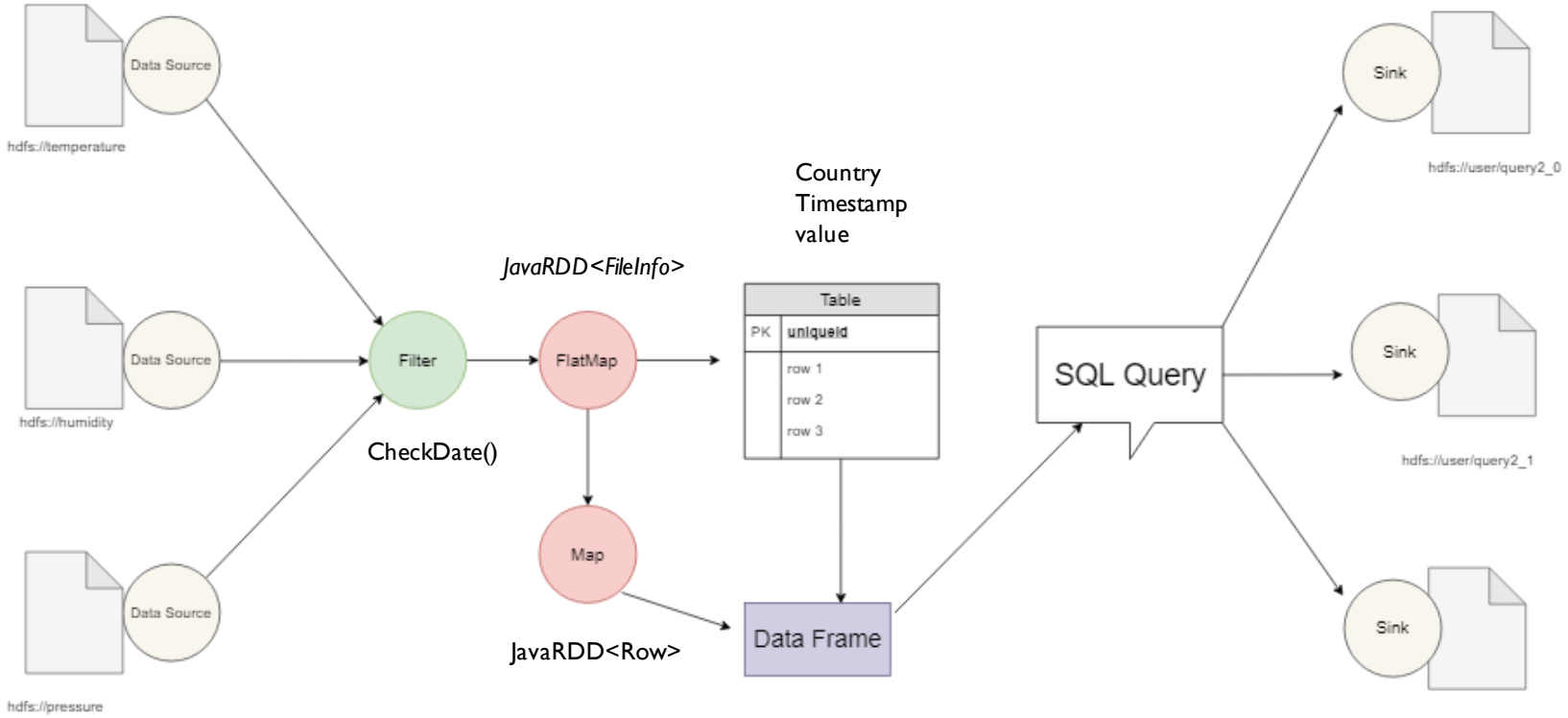
QUERY 2


Data Source
hdfs://temperature
Data Source
hdfs://humidity
Data Source
hdfs://pressure
Filter
CheckDate()
(Nazione_YY-MM,value)
CheckValue()
FlatMaptoPair
AggregateByKey
(Nazione_YY-MM,Stats)
MapToPair
(Nazione_YY-MM,[mean,dev,min,max])
SortByKey
Sink
hdfs://user/query2_0
Sink
hdfs://user/query2_1
Sink
hdfs://user/query2_2

QUERY 3


((Nazione_Città_P_Anno),value)
Data Source
FlatMapToPair
Filter
P=Summer or P=Winter
Anno=2017 or Anno=2018
((Nazione_Città_P_Anno),Stats)
AggregateByKey
MaptoPair
((Nazione_CIttà_Anno),Avg)
ReduceBykey
MapToPair
((Nazione_Città_Anno),DiffAvg)
AggregateByKey
((Nazione_Anno),Stats)
Anno=2016
Filter
((Nazione_Anno),Rank2016)
Collect
Anno=2017
Filter
((Nazione_Anno),DiffAvg_Città)
MapToPair
Sink
((Nazione_Anno),{Città,pos2016})
hdfs://user/query3

QUERY 2 SQL




```
SELECT DISTINCT
country,timestamp,AVG(value),STD(value),MIN(value),MAX(value)
FROM query2
GROUP BY country,timestamp
```

SCRITTURA SU HBASE E RISULTATI

Query 1

RowKey	CityList
2013	Fam1:[Eilat,Las Vegas]
2014	Fam1:[Las Vegas]
2016	Fam1:[Phoenix,Eilat,LasVegas]
2017	Fam1:[Phoenix,Eilat,LasVegas]

Query 2

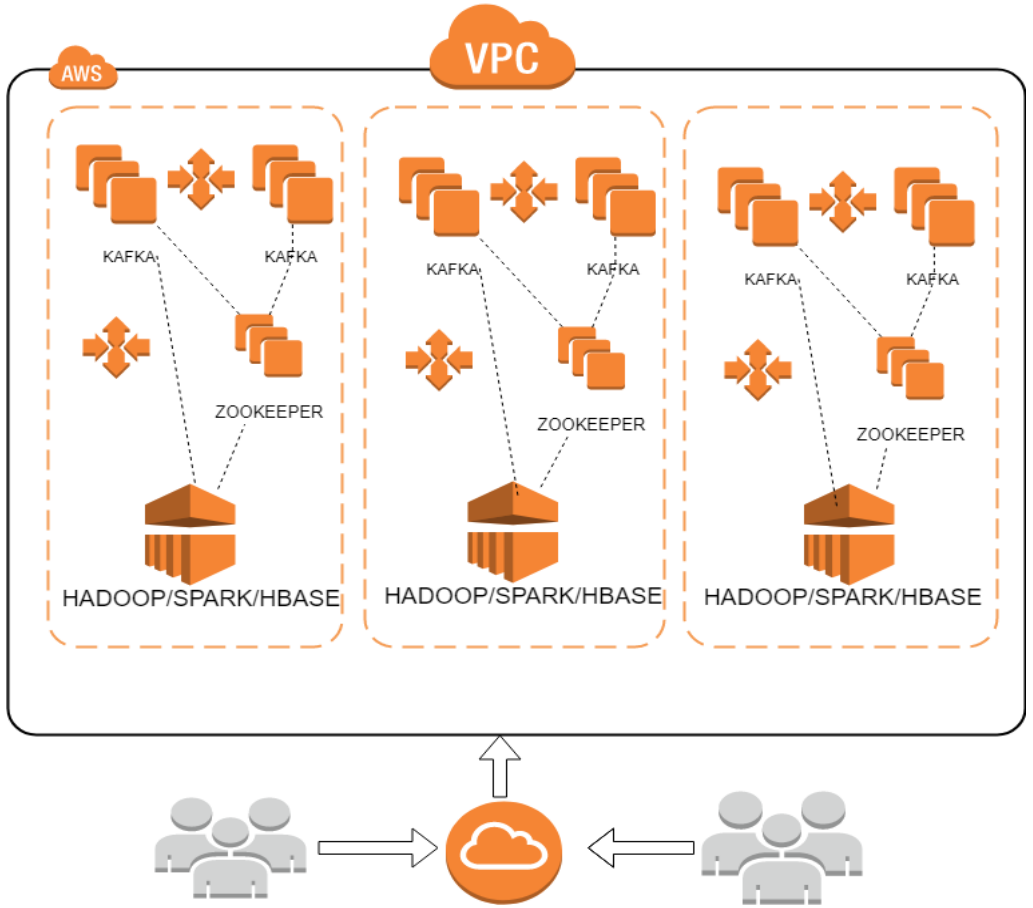
Row Key	mean	dev	min	max
Israel_2012-10	Temp:298.88 Hum:57.63 Press:1002.97	Temp:3.956 Hum:16.48 Press:13.34	Temp:298.15 Hum:12 Press:959	Temp:314.82 Hum:100 Press:1020
USA_2017-07	Temp:298.88 Hum:66.14 Press:1016.66	Temp:3.956 Hum:21.89 Press:3.87	Temp:298.15 Hum:5 Press:995	Temp:314.82 Hum:100 Press:1030

SCRITTURA SU HBASE E RISULTATI

Query 3

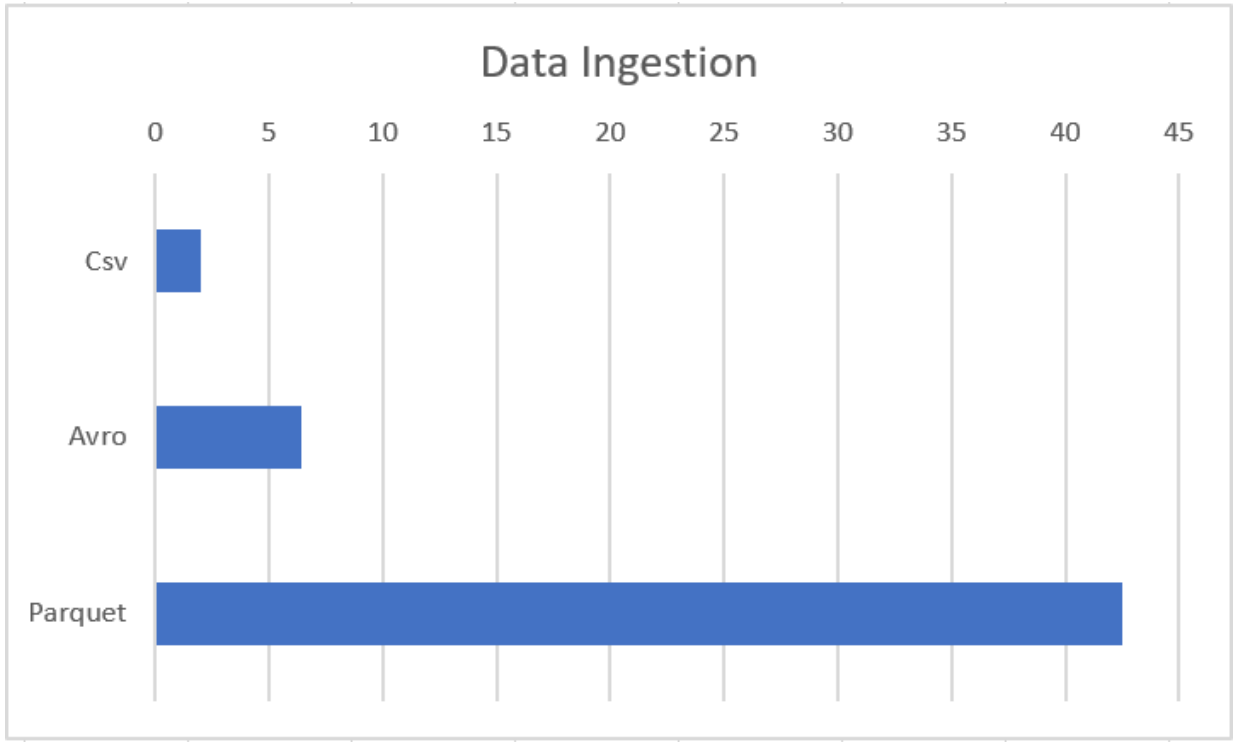
Row Key	city	pos2016
USA_2017	Pos1:Minneapolis Pos2:Chicago Pos3:Detroit	2 3 1
Israel_2017	Pos1:Beersheba Pos2:Eilat Pos3:Haifa	1 4 2

DEPLOY SU AWS



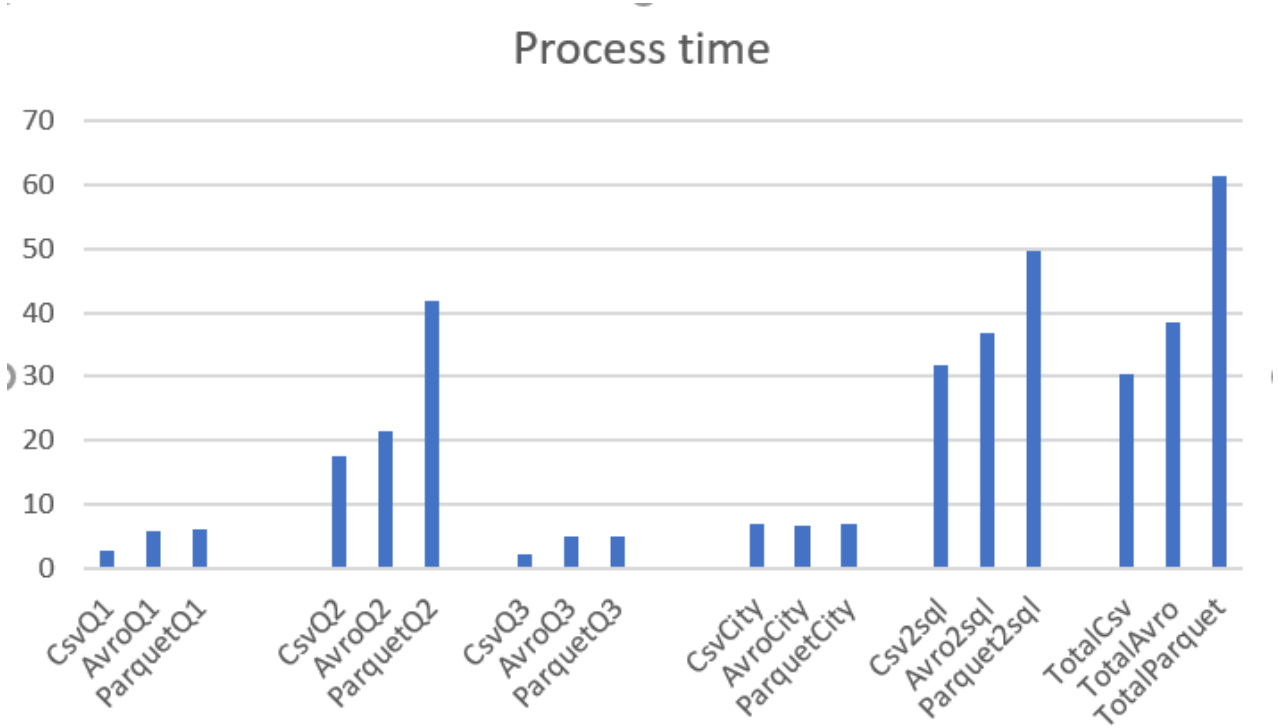
Istanze emr m4.2xlarge
4 slave

PERFORMANCE



	CSV	AVRO	PARQUET
Media(s)	2	6.4	42.5
Deviazione standard(s)	1.095	0.8	1.47

PERFORMANCE



	CSV (Media/dev)	AVRO (Media/dev)	PARQUET (media/dev)
City processing(s)	7/6	6.6/4.91	6.8/5.6
Query 1 (s)	2.6/1.2	5.8/0.4	6.2/0.4
Query 2 (s)	17.4/0.8	21.4/0.8	41.8/0.4
Query 3 (s)	2.2/0.4	5/0.2	5/0.2
Query 2 SQL(s)	31.8/2.1	36.9/2.1	49.6/1.8
Tempo totale(s)	30.4	38.4	61.4

Scrittura Hbase	Valore
Media(s)	5.2
Deviazione standard(s)	0.98