ANALISI DEL DATASET DELLE VACCINAZIONI ANTI COVID-19

TIZIANA MANNUCCI (0285727)

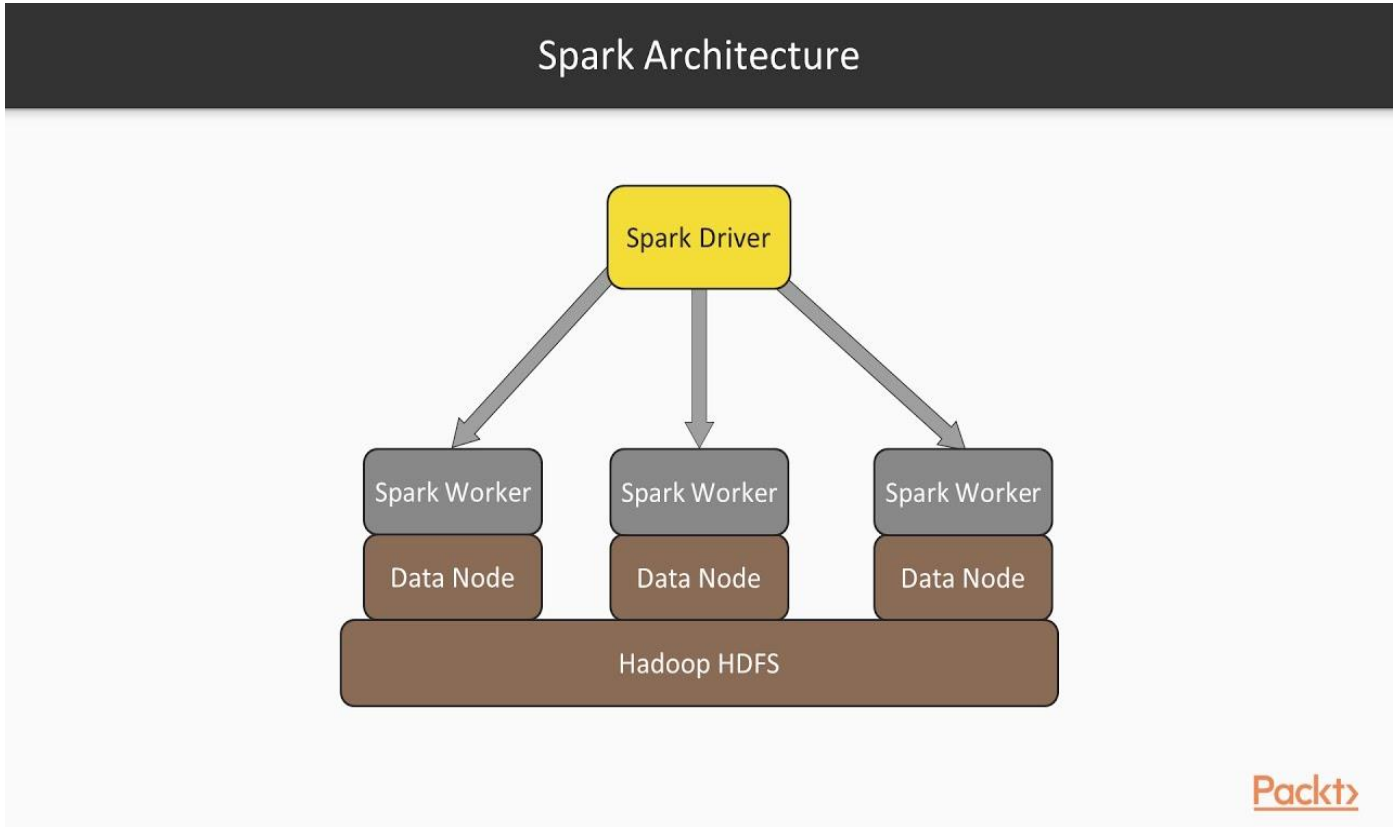
STRUTTURA

- Introduzione
- Architettura del sistema
- Analisi del Dataset
- Query 1
- Analisi Risultati Query 1
- Query 2
- Analisi Risultati Query 2
- Analisi Tempi di Esecuzione

INTRODUZIONE

- Processamento Batch
- Distribuito ed in-memory
- Programmazione basata su Trasformazioni e Azioni

ARCHITETTURA DEL SISTEMA



- **$ $SPARK_HOME/sbin/start-master.sh**
- **$ $SPARK_HOME/sbin/start-slave.sh <master-spark-URL>**
- **Usare il docker-compose.yml fornito per creare i Container per l'HDFS**

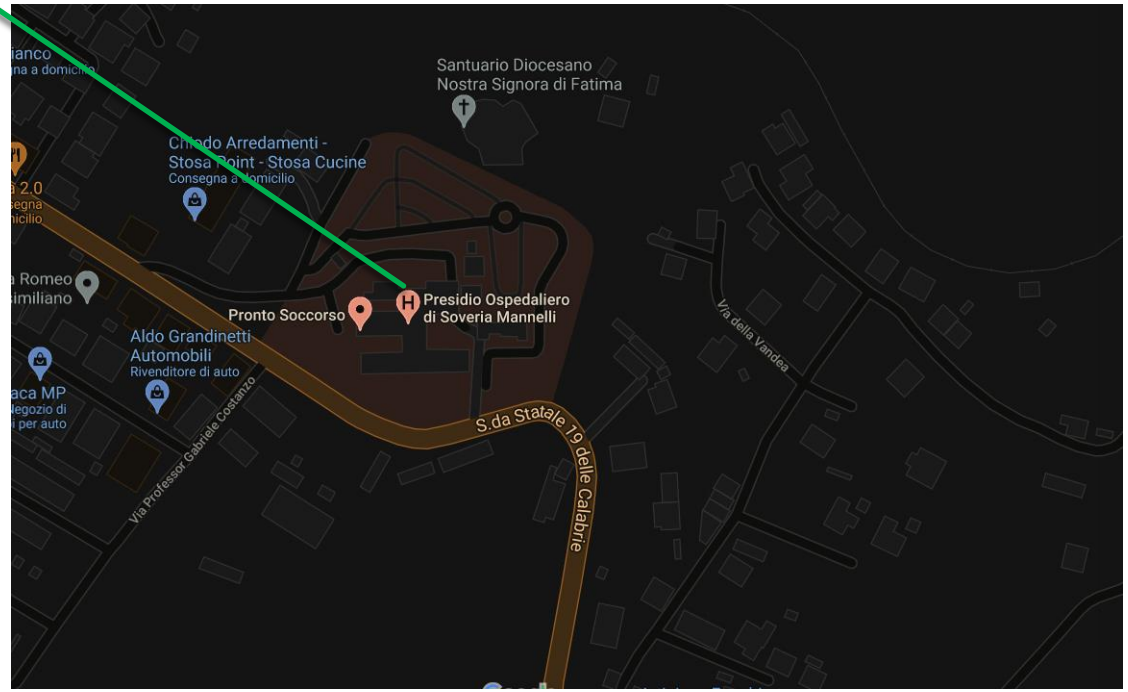
ANALISI DEL DATASET

TUTTI I FILE UTILIZZATI SONO AGGIORNATI AL 1/06/2021

- punti-somministrazione-tipologia.csv

CAL	OSPEDALE DI SOVERIA MANNELLI	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE DI SOVERIA MANNELLI	Territoriale	ITF	ITF6	18 Calabria
CAL	OSPEDALE GUDO CHIDICHIMO TREBISACCE	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE PAOLA	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE PUGLIESE	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE T.EVOLI MELITO P.S.	Territoriale	ITF	ITF6	18 Calabria
CAL	P.O. 'GIOVANNI XXIII'	Ospedaliero	ITF	ITF6	18 Calabria
CAL	P.O. 'GIOVANNI XXIII'	Territoriale	ITF	ITF6	18 Calabria

Tali **centri duplicati** vengono considerati come un'istanza singola.



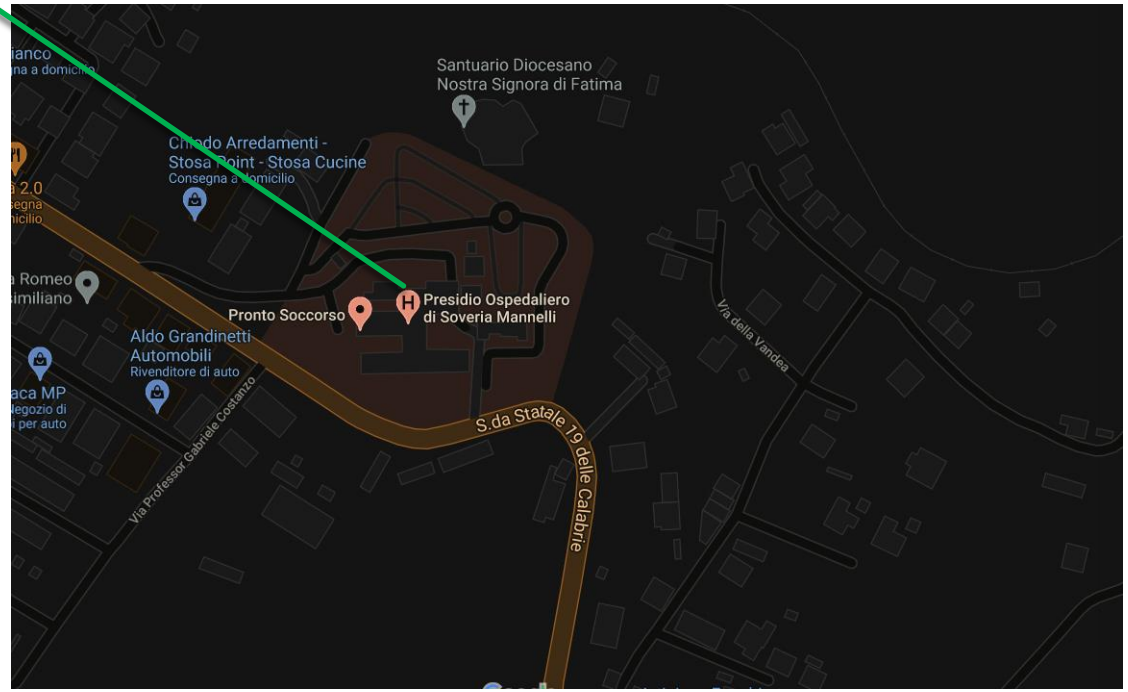
ANALISI DEL DATASET

TUTTI I FILE UTILIZZATI SONO AGGIORNATI AL 1/06/2021

- punti-somministrazione-tipologia.csv

CAL	OSPEDALE DI SOVERIA MANNELLI	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE DI SOVERIA MANNELLI	Territoriale	ITF	ITF6	18 Calabria
CAL	OSPEDALE GUDO CHIDICHIMO TREBISACCE	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE PAOLA	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE PUGLIESE	Ospedaliero	ITF	ITF6	18 Calabria
CAL	OSPEDALE T.EVOLI MELITO P.S.	Territoriale	ITF	ITF6	18 Calabria
CAL	P.O. 'GIOVANNI XXIII'	Ospedaliero	ITF	ITF6	18 Calabria
CAL	P.O. 'GIOVANNI XXIII'	Territoriale	ITF	ITF6	18 Calabria

Tali **centri duplicati** vengono considerati come un'istanza singola.



- somministrazioni-vaccini-latest.csv
- somministrazioni-vaccini-summary-latest.csv

QUERY 1

$$\text{avg}_{wz} = \frac{1}{y_{wz}} \sum_{i} \frac{d_i}{x_w} \qquad \forall \;\; w \equiv regione, z \equiv mese, i \equiv giorni\ vaccinazione$$

x_w *centri vaccinali nella regione* w
d_i*sono le vaccinazioni effettuate in un giorno* i *fissate la regione ed il mese*

QUERY 1

$$\text{avg}_{wz} = \frac{1}{y_{wz}} \sum_i \frac{d_i}{x_w} \quad \forall \; w \equiv regione, z \equiv mese, i \equiv giorni\ vaccinazione$$

$$\text{avg}_{wz} = \frac{1}{y_{wz} * x_w} \sum_i d_i \quad \forall \; w \equiv regione, z \equiv mese, i \equiv giorni\ vaccinazione$$

x_w *centri vaccinali nella regione* w
d_i*sono le vaccinazioni effettuate in un giorno* i *fissate la regione ed il mese*

QUERY 1

$$\text{avg}_{\text{wz}} = \frac{1}{\text{y}_{\text{wz}}} \sum_i \frac{d_i}{x_w} \quad \forall \; w \equiv regione, z \equiv mese, i \equiv giorni\ vaccinazione$$

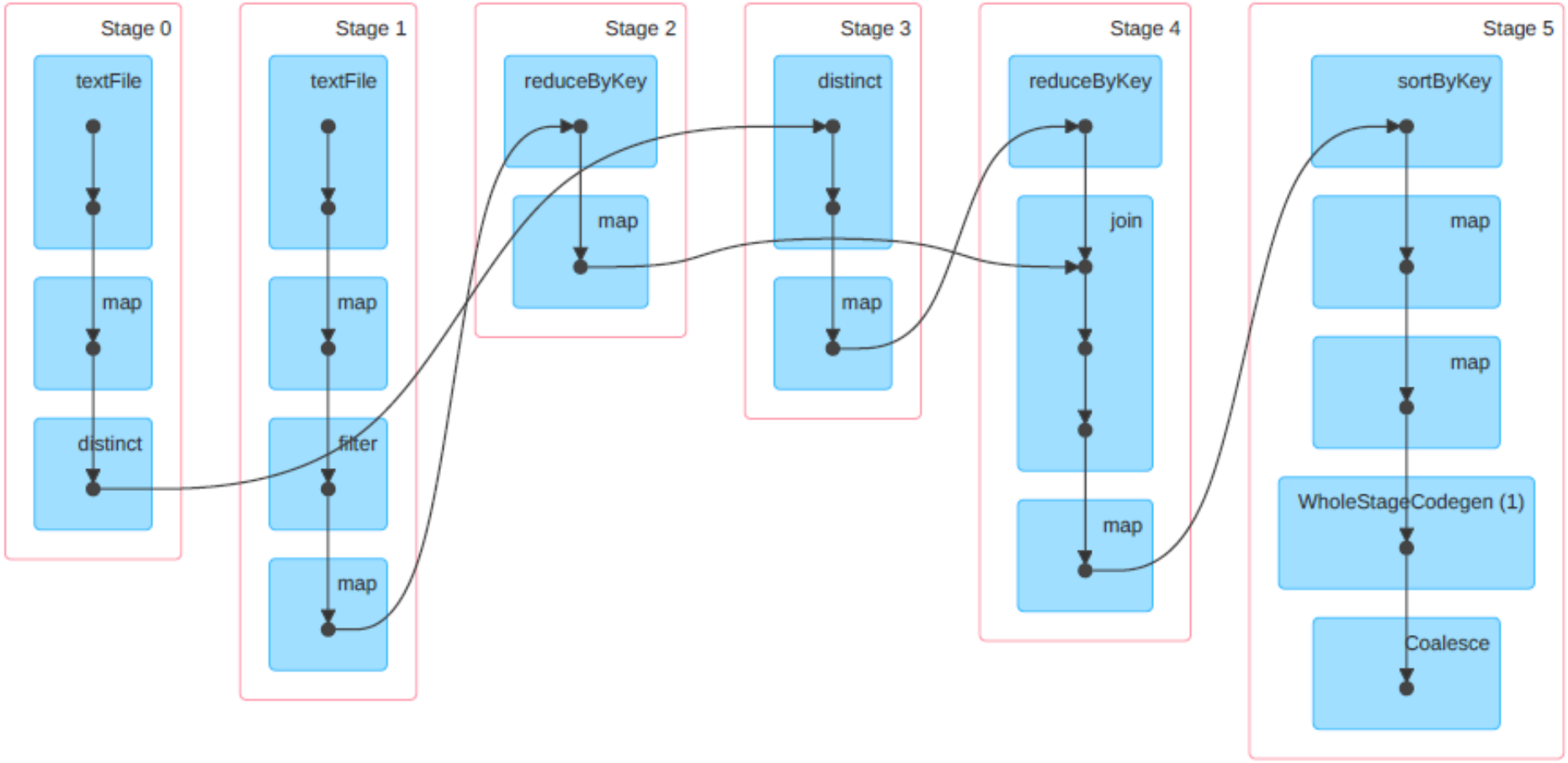
$$\text{avg}_{\text{wz}} = \frac{1}{\text{y}_{\text{wz}} * \text{x}_{\text{w}}} \sum_i d_i \quad \forall \; w \equiv regione, z \equiv mese, i \equiv giorni\ vaccinazione$$

$$tot_{wz} = \sum_i d_i$$

$$\text{avg}_{\text{wz}} = \frac{\text{tot}_{\text{wz}}}{\text{y}_{\text{wz}} * \text{x}_{\text{w}}}$$

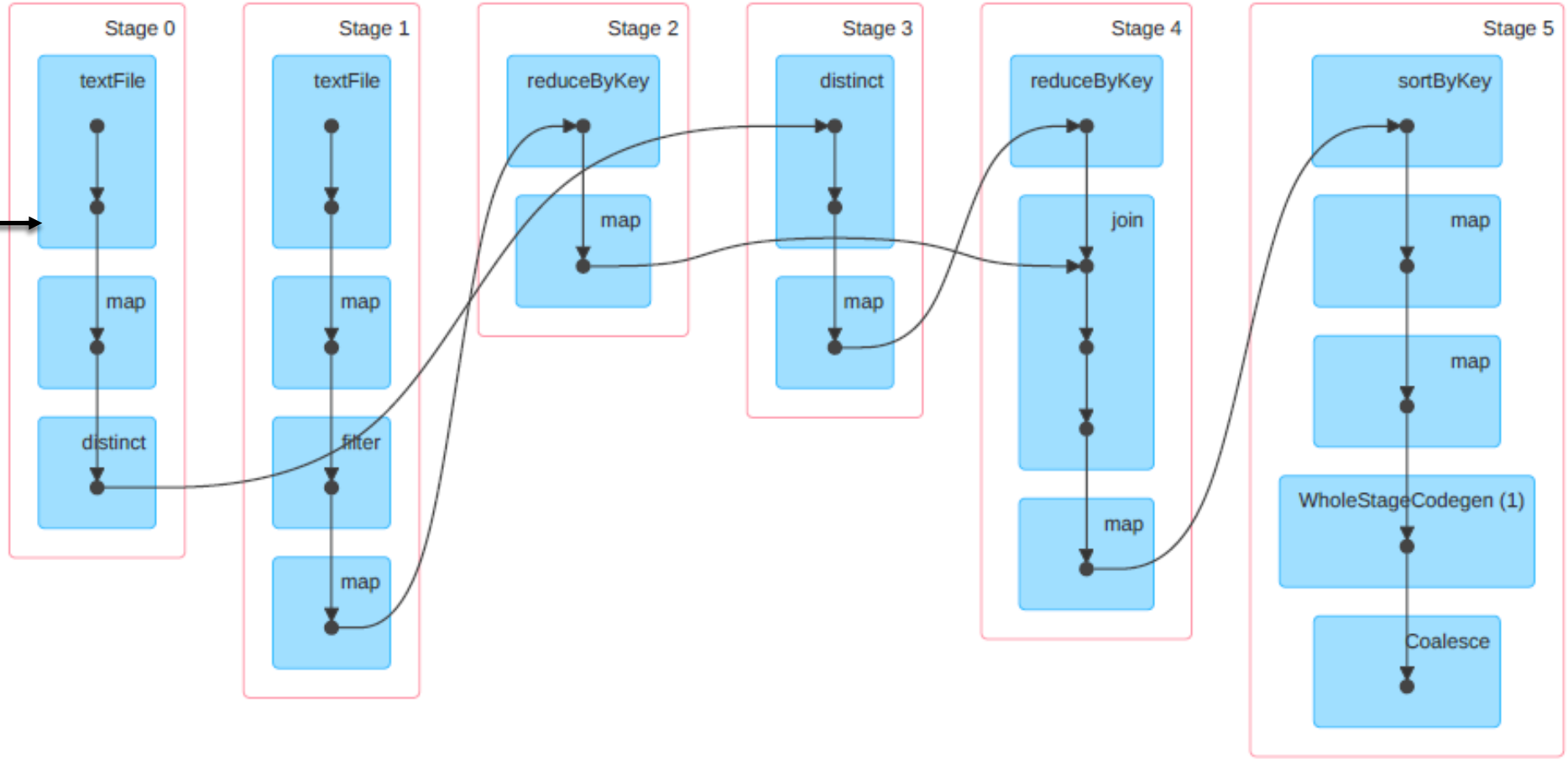
x_w *centri vaccinali nella regione* w
d_i*sono le vaccinazioni effettuate in un giorno* i *fissate la regione ed il mese*

QUERY 1

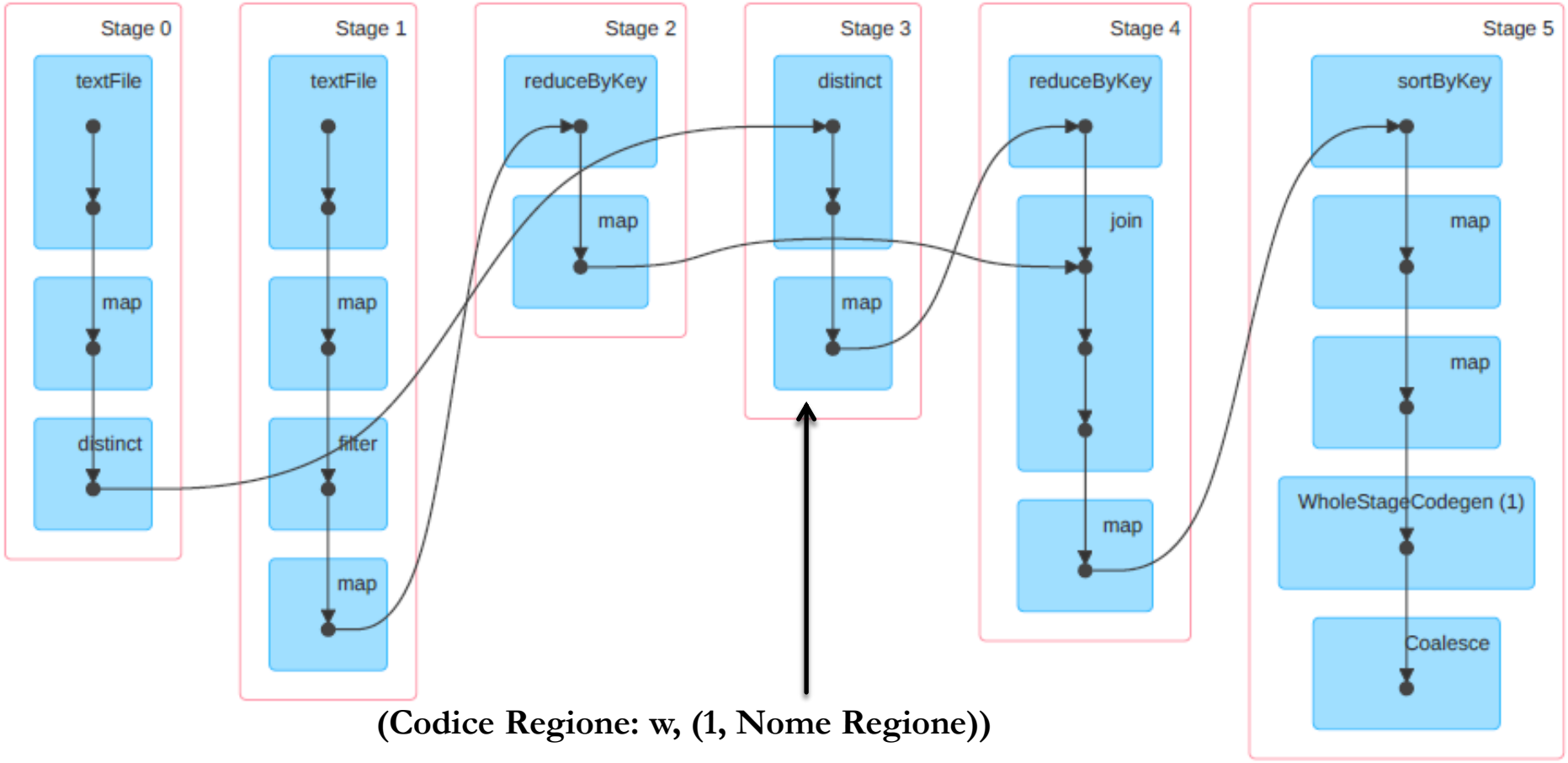

Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce

QUERY 1

Punti somministrazione

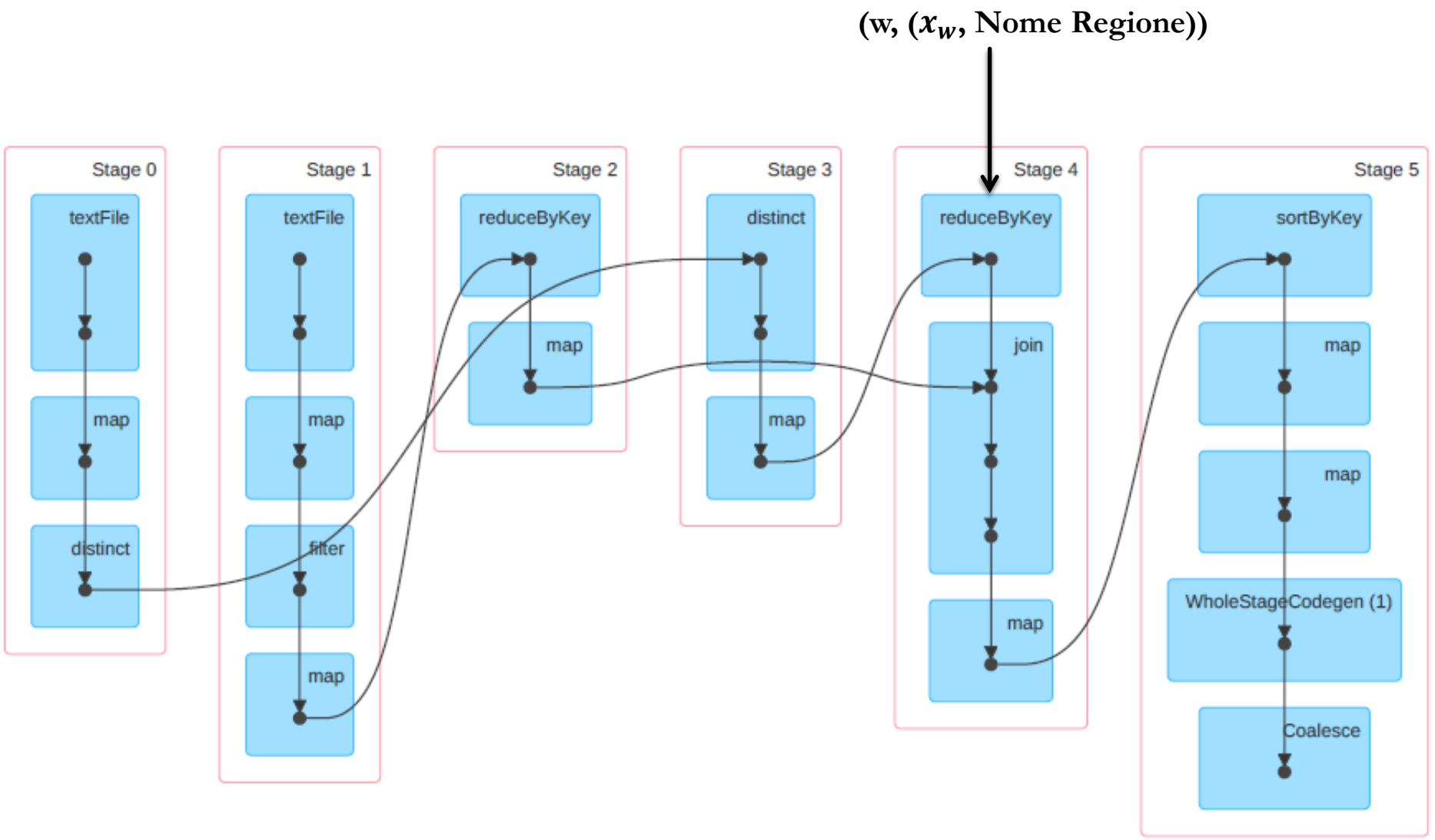


QUERY 1

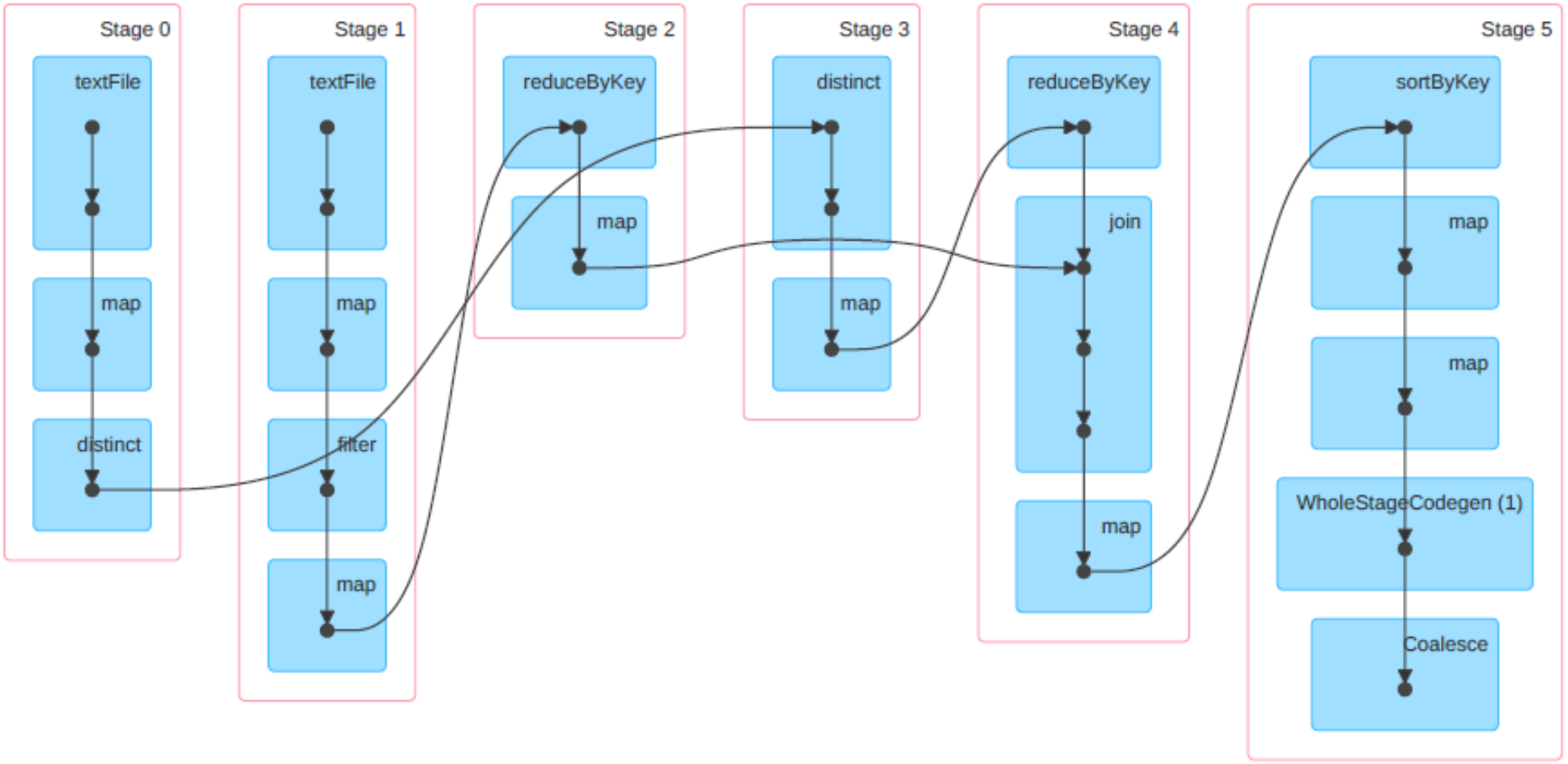


(Codice Regione: w, (1, Nome Regione))

QUERY 1

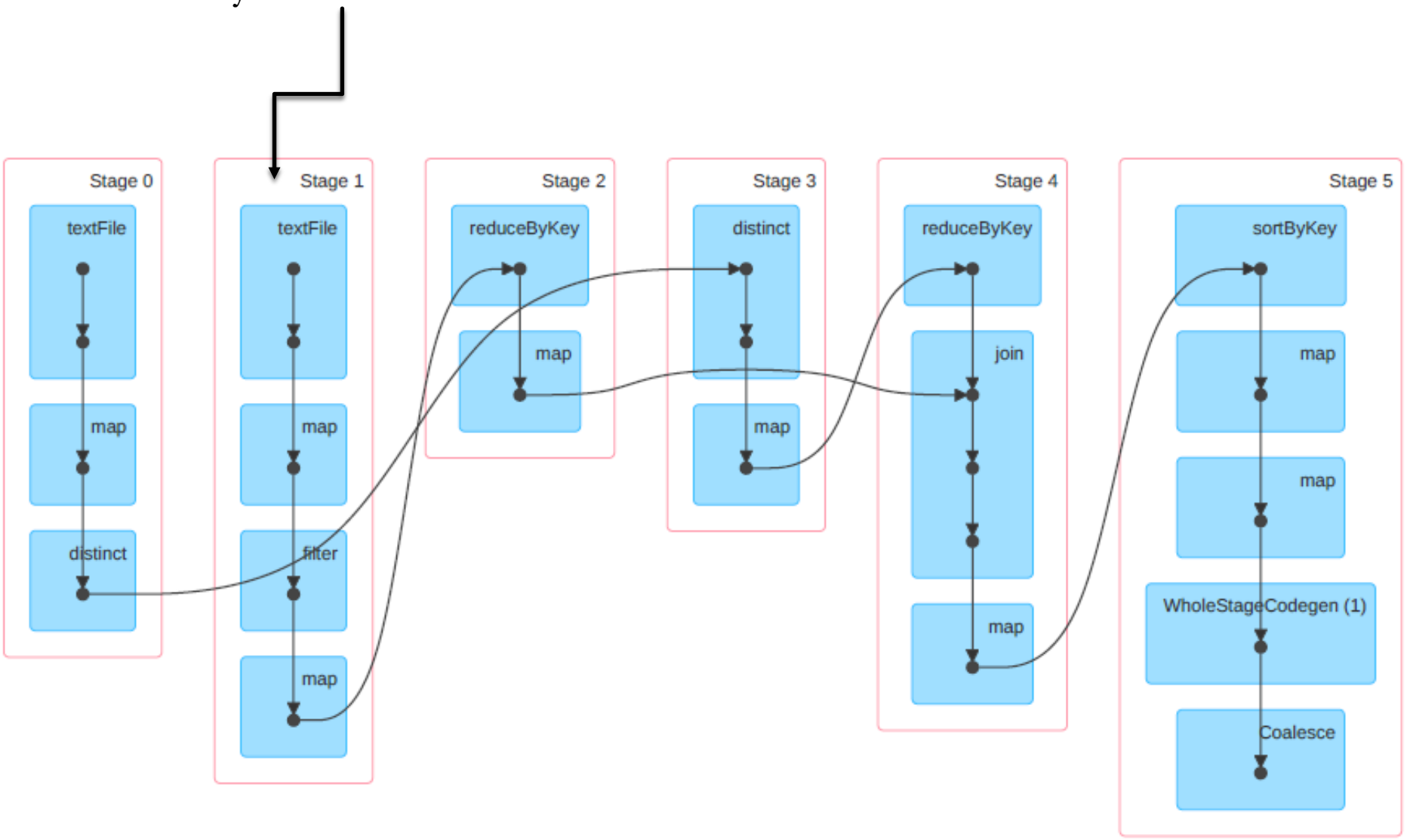

(w, (x_w, Nome Regione))
Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce

QUERY 1


Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce

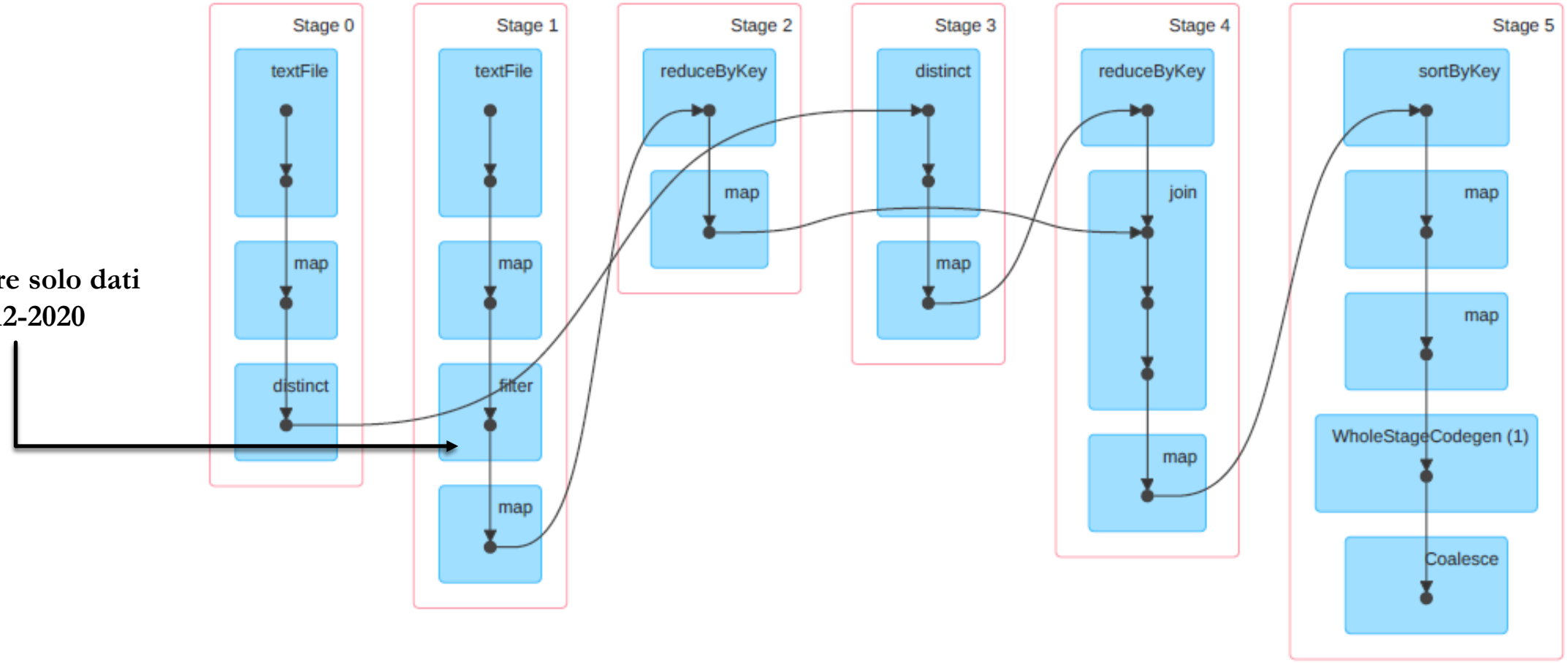
QUERY 1

Somministrazione Vaccini Summary

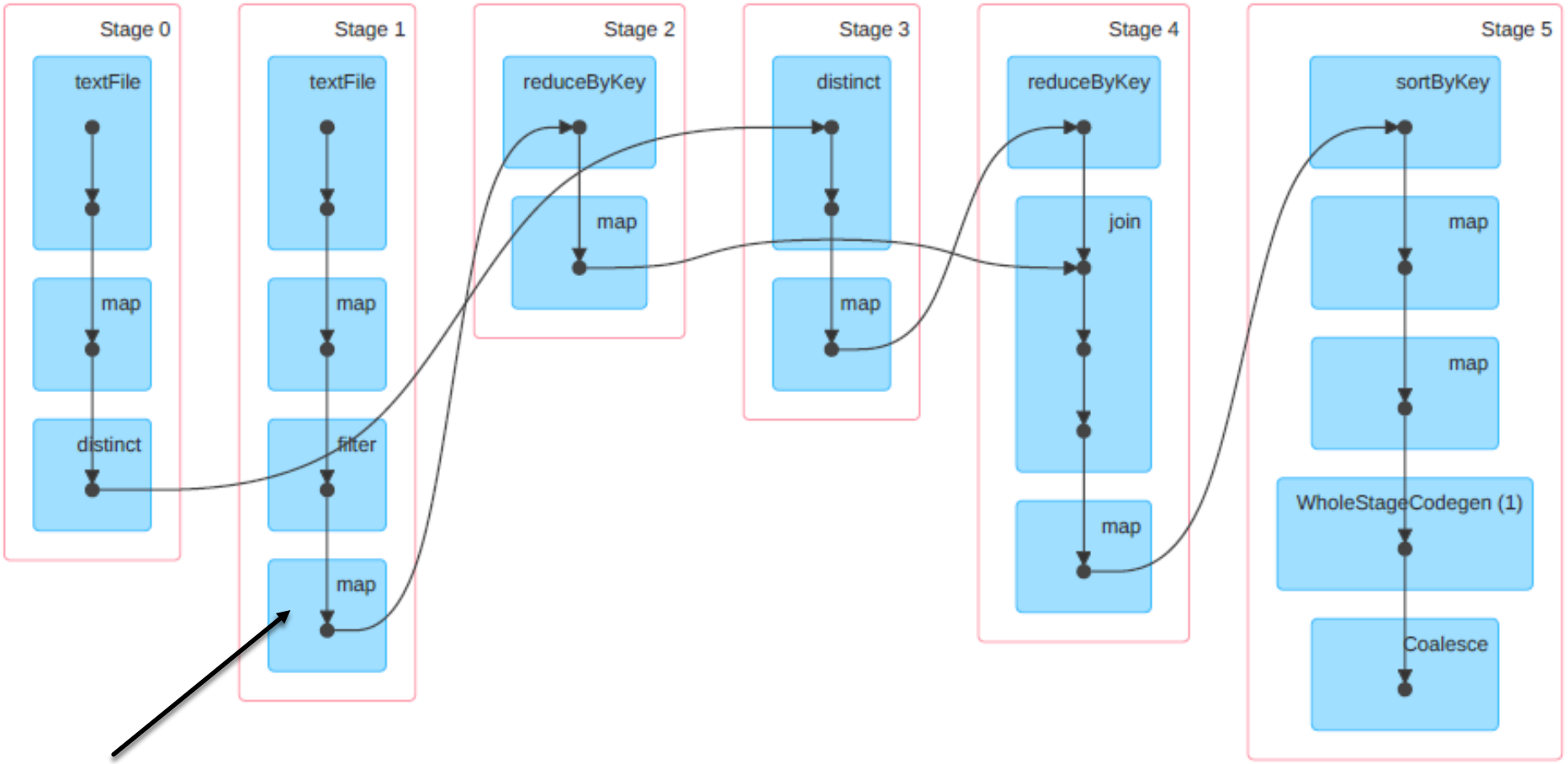


QUERY 1

Mantenere solo dati dopo 31-12-2020

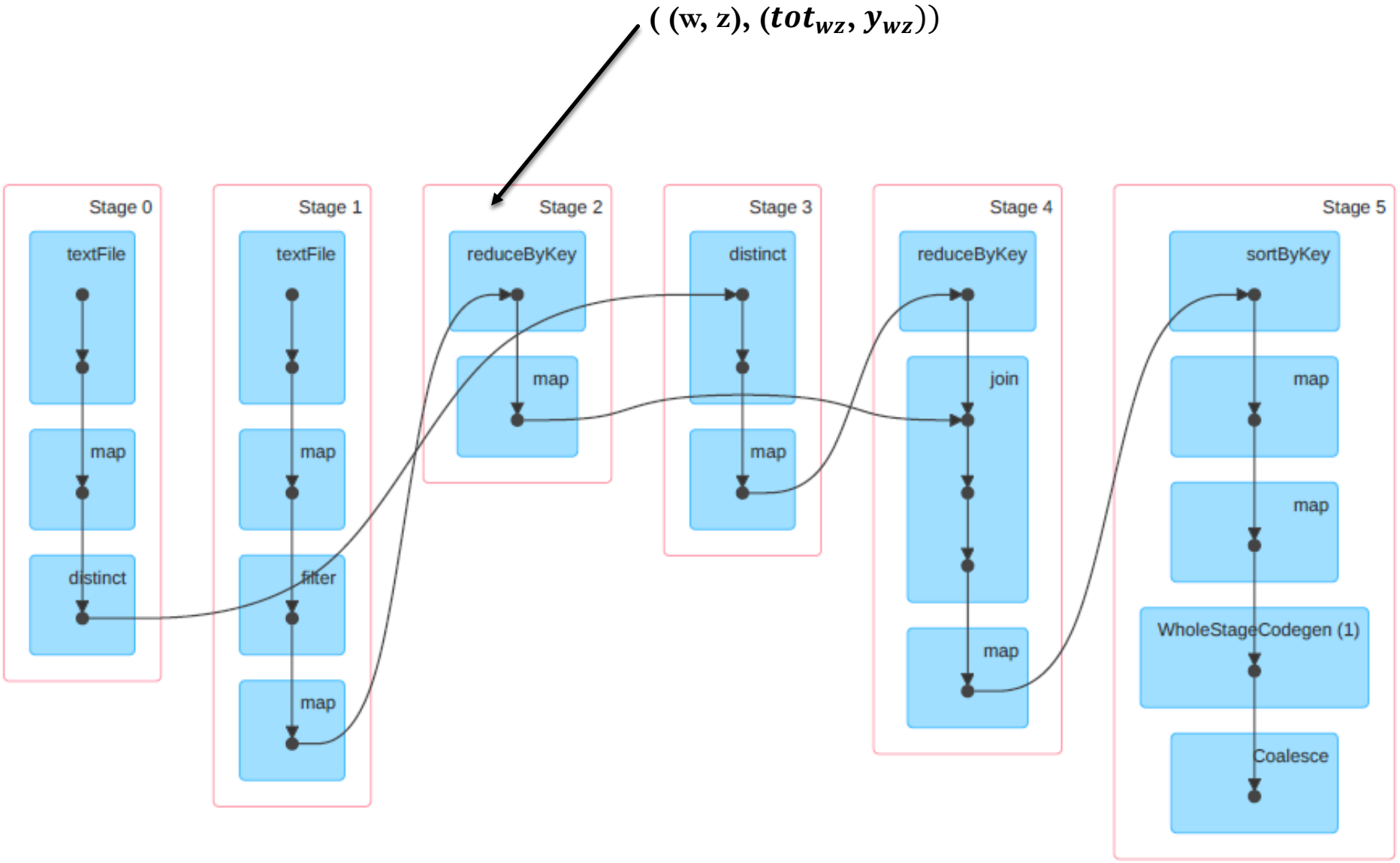


QUERY 1

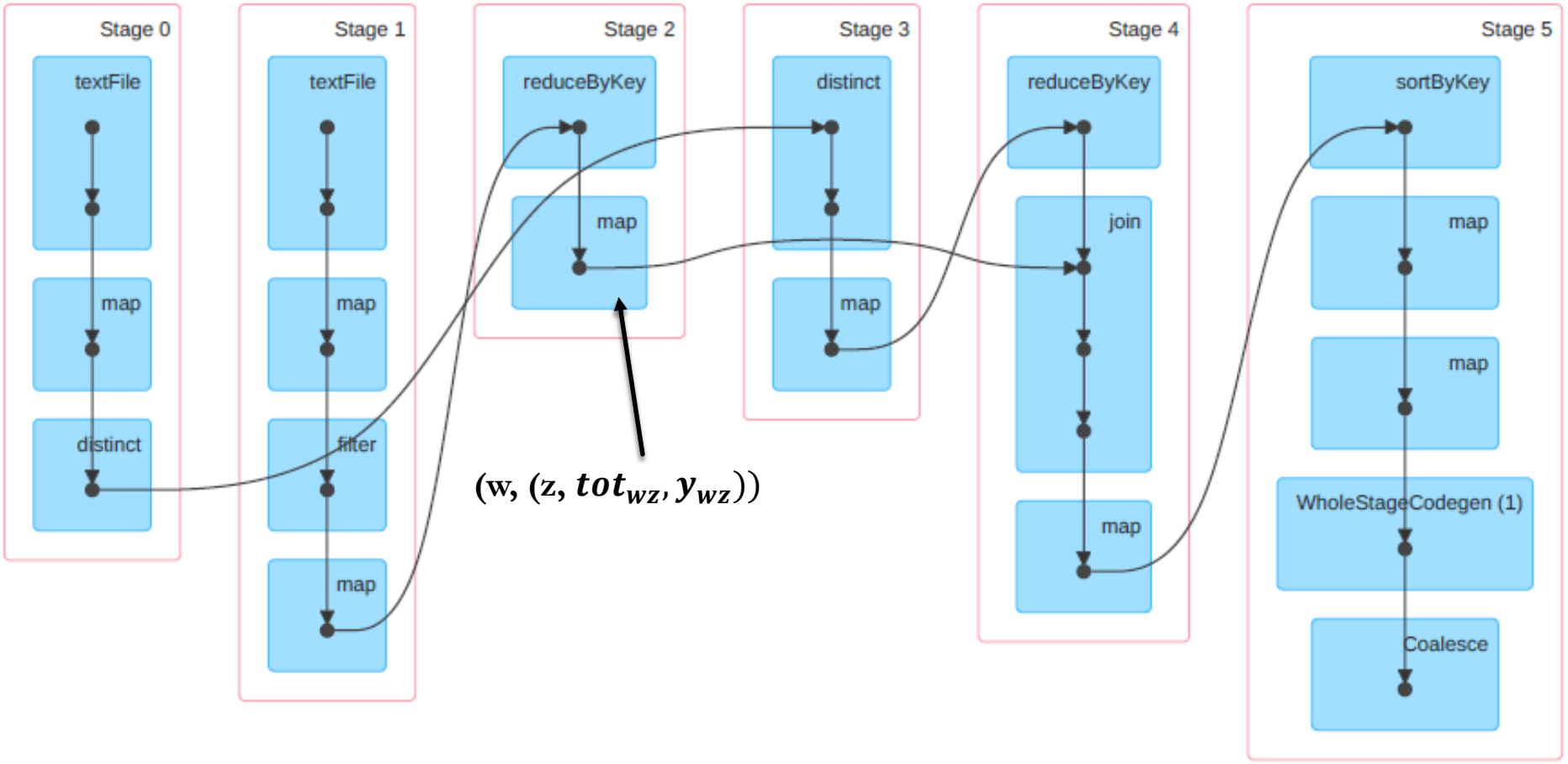


((Codice Regione: w, mese: z), (totale giornaliero: d_i, 1))

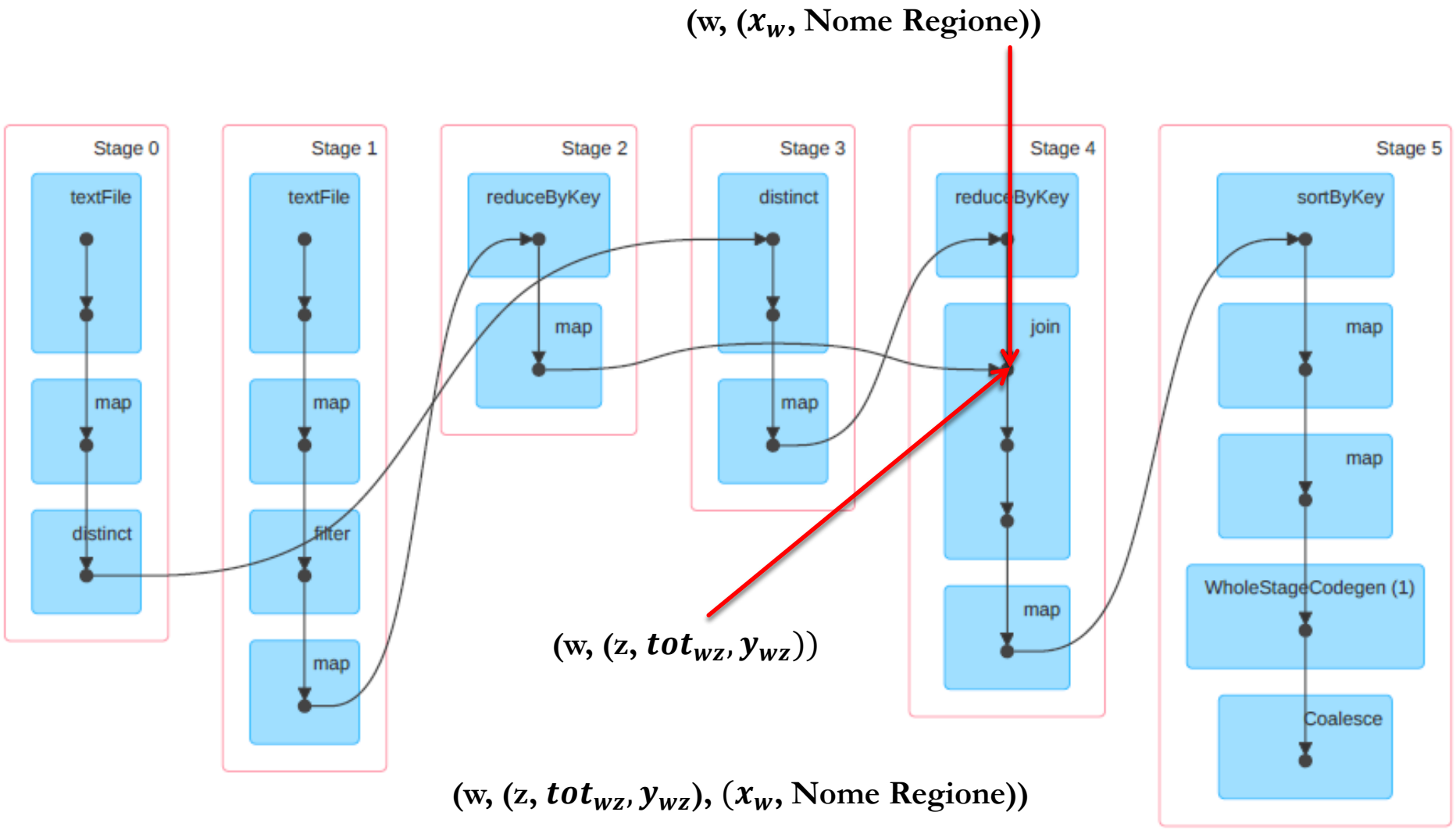
QUERY 1


((w, z), (tot_{wz}, y_{wz}))
Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce

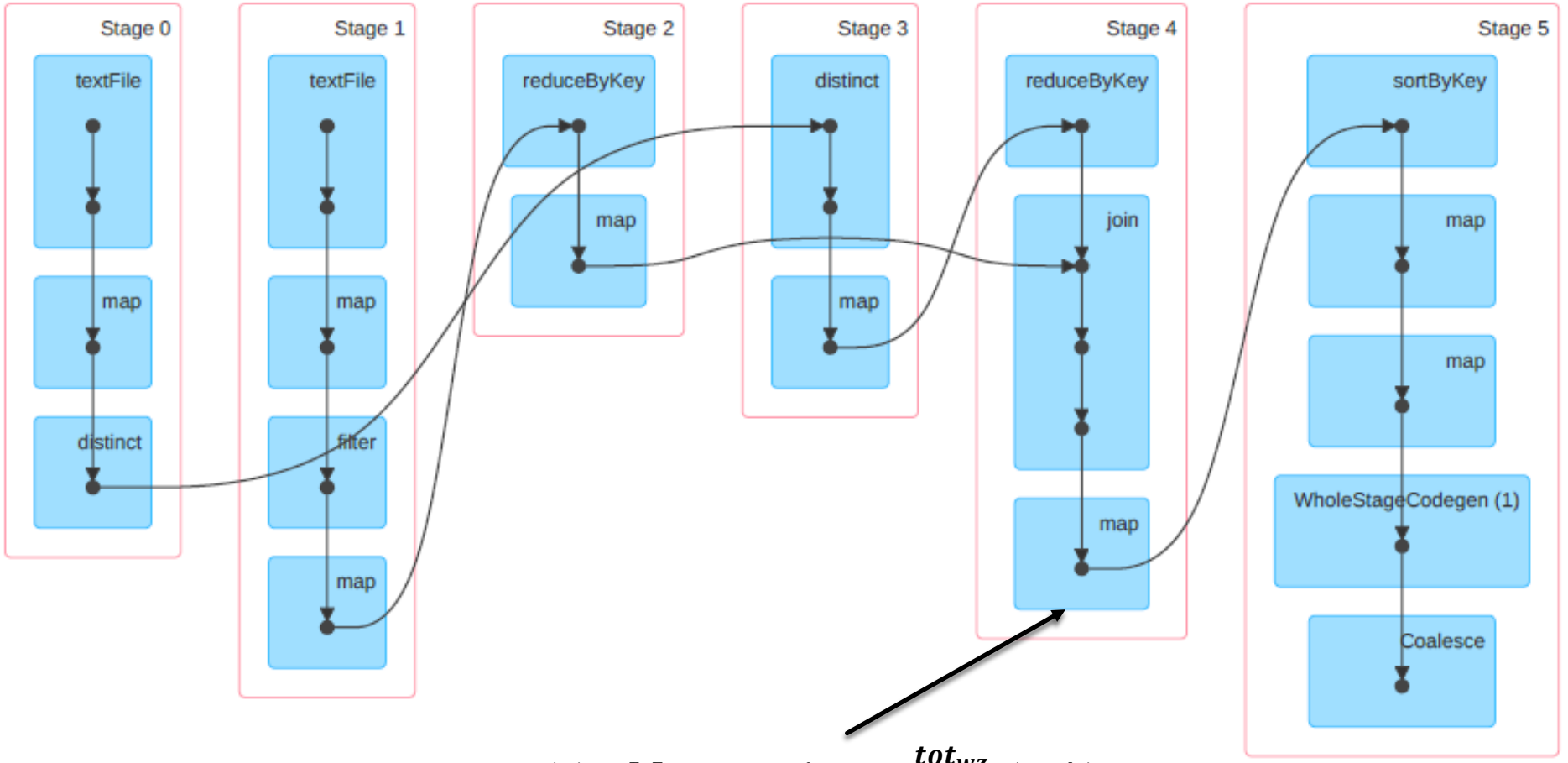
QUERY 1


Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce
(w, (z, tot_{wz}, y_{wz}))

QUERY 1

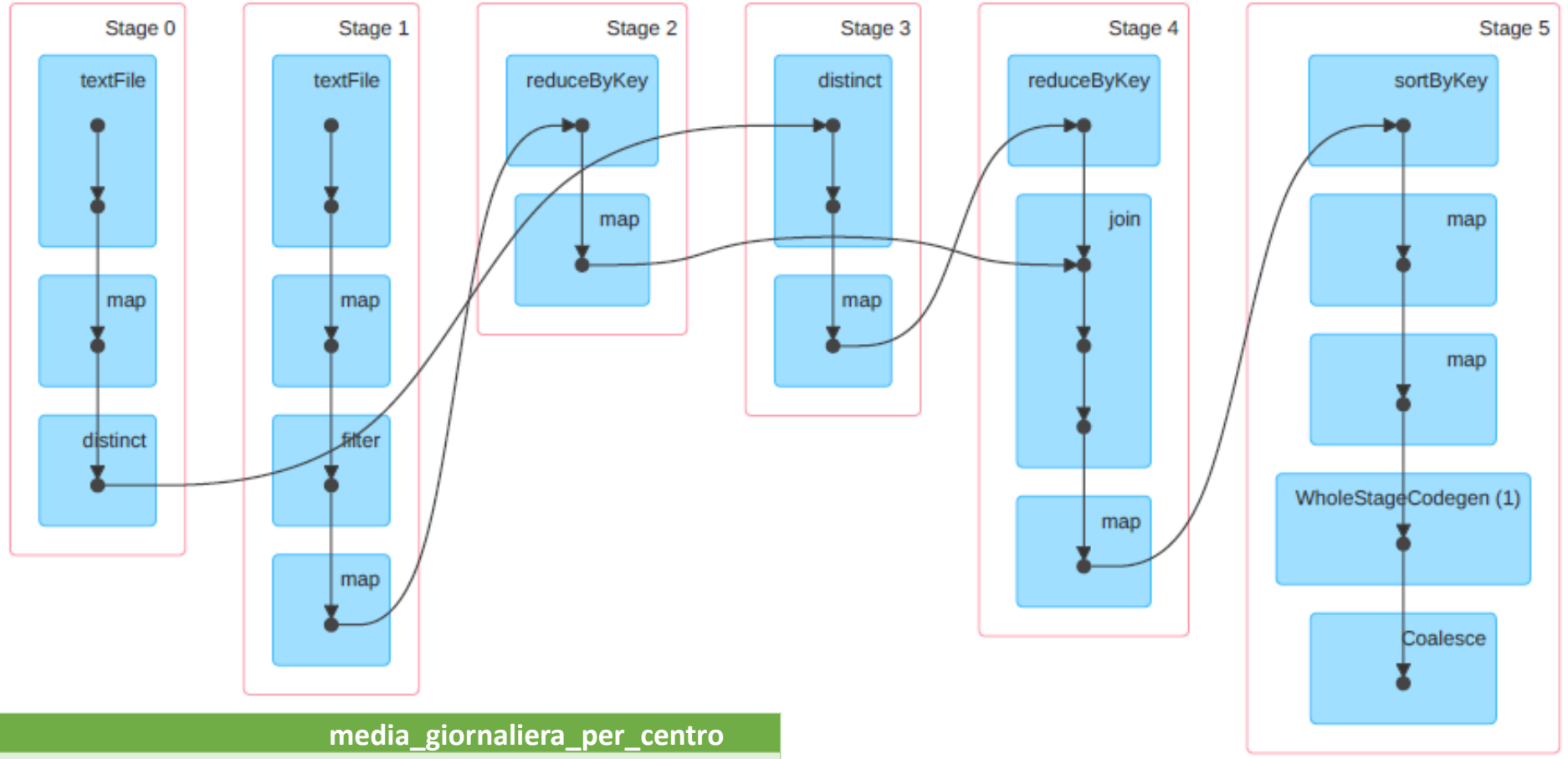

(w, (x_w, Nome Regione))
Stage 0
textFile
map
distinct
Stage 1
textFile
map
filter
map
Stage 2
reduceByKey
map
Stage 3
distinct
map
Stage 4
reduceByKey
join
map
Stage 5
sortByKey
map
map
WholeStageCodegen (1)
Coalesce
(w, (z, tot_{wz}, y_{wz}))
(w, (z, tot_{wz}, y_{wz}), (x_w, Nome Regione))

QUERY 1



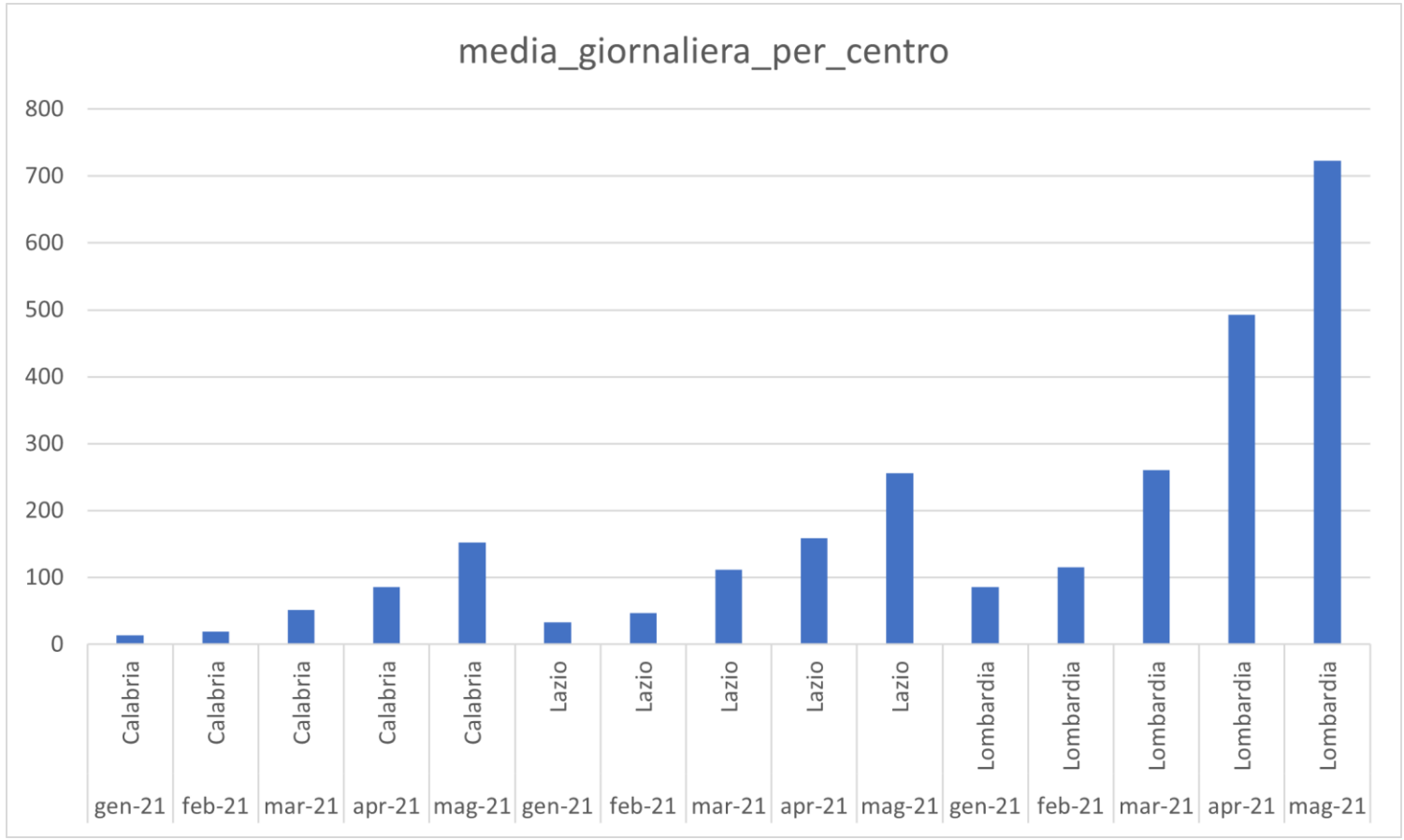
$$\left((z,\ \text{Nome regione},\ \frac{tot_{wz}}{y_{wz} * x_w}),\ 1 \right)$$

QUERY 1

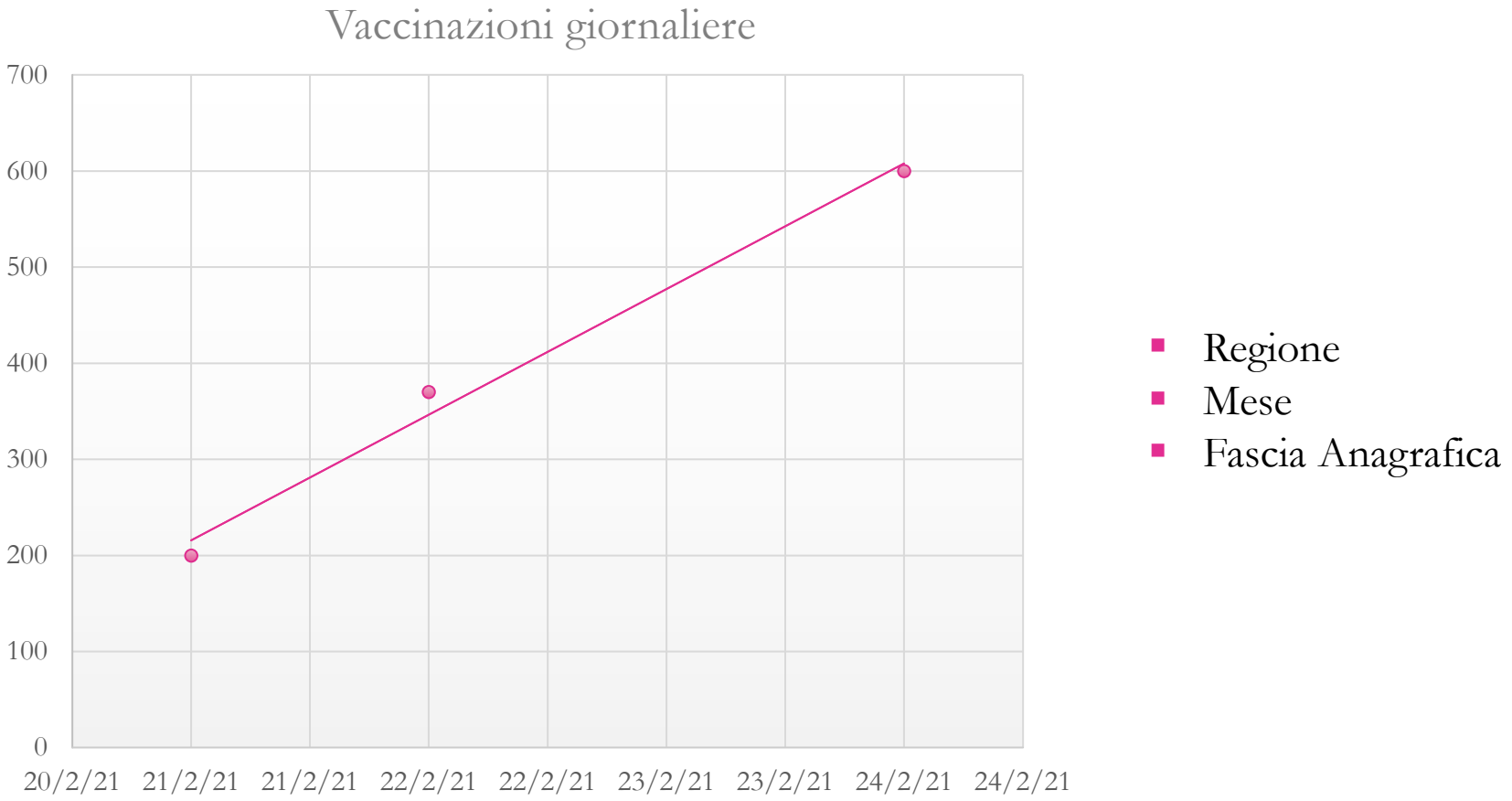


mese	regione	media_giornaliera_per_centro
01/01/2021	Abruzzo	17,157
01/01/2021	Basilicata	17,847
01/01/2021	Calabria	13,41
01/01/2021	Campania	104,247
01/01/2021	Emilia-Romagna	41,961
01/01/2021	Friuli-Venezia Giulia	26,736
01/01/2021	Lazio	33,469
01/01/2021	Liguria	11,673

ANALISI RISULTATI QUERY 1



QUERY 2



- Regione
- Mese
- Fascia Anagrafica

```
org.apache.commons.math3.stat.regression.SimpleRegression
```

QUERY 2: AVOID GROUPBYKEY

- Da una serie di puti $(\text{key}, \text{val}) \rightarrow (key, [val_1, val_2, \ldots, val_n])$
- Map($f: (key, val_i) \rightarrow (key, [val_i])$
- ReduceByKey(f: $(\text{key}, \text{val}_\text{i}), (key, val_j) \rightarrow (key, [val_i, val_j])$
- Per i test sotto riportati si usa un Oggetto Custom (src/main/java/utils/MyIterable.java) serializzabile e contenente una lista.

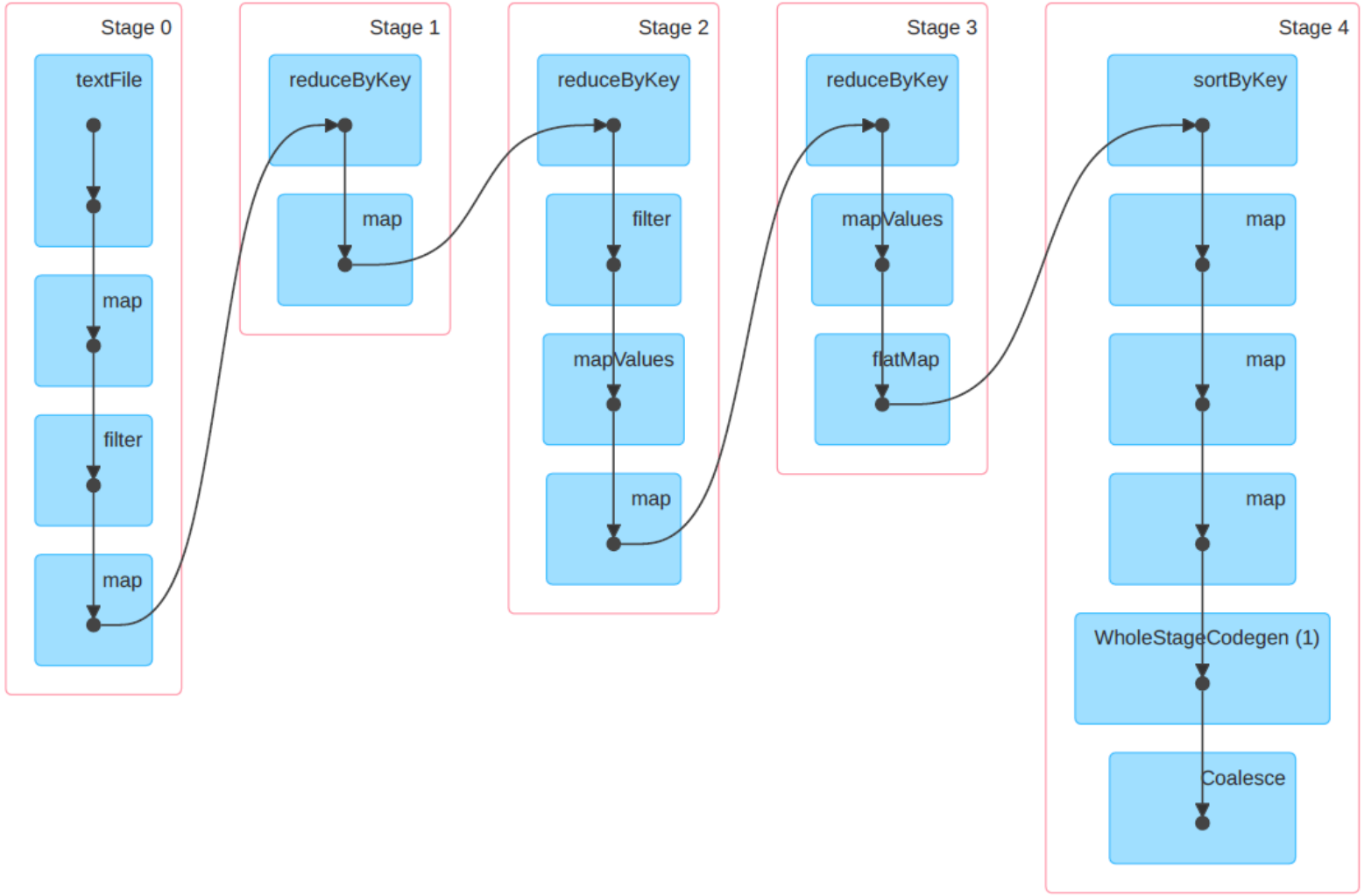
Performance GroupByKey

Metric	Min	25th percentile	Median
Duration	0.1 s	0.1s	0.1s
GC Time	0.0	0.0	0.0
Shuffle Size/Record	371.5Kib/18650	371.5Kib/18650	371.5Kib/18650

Performance map & reduceByKey

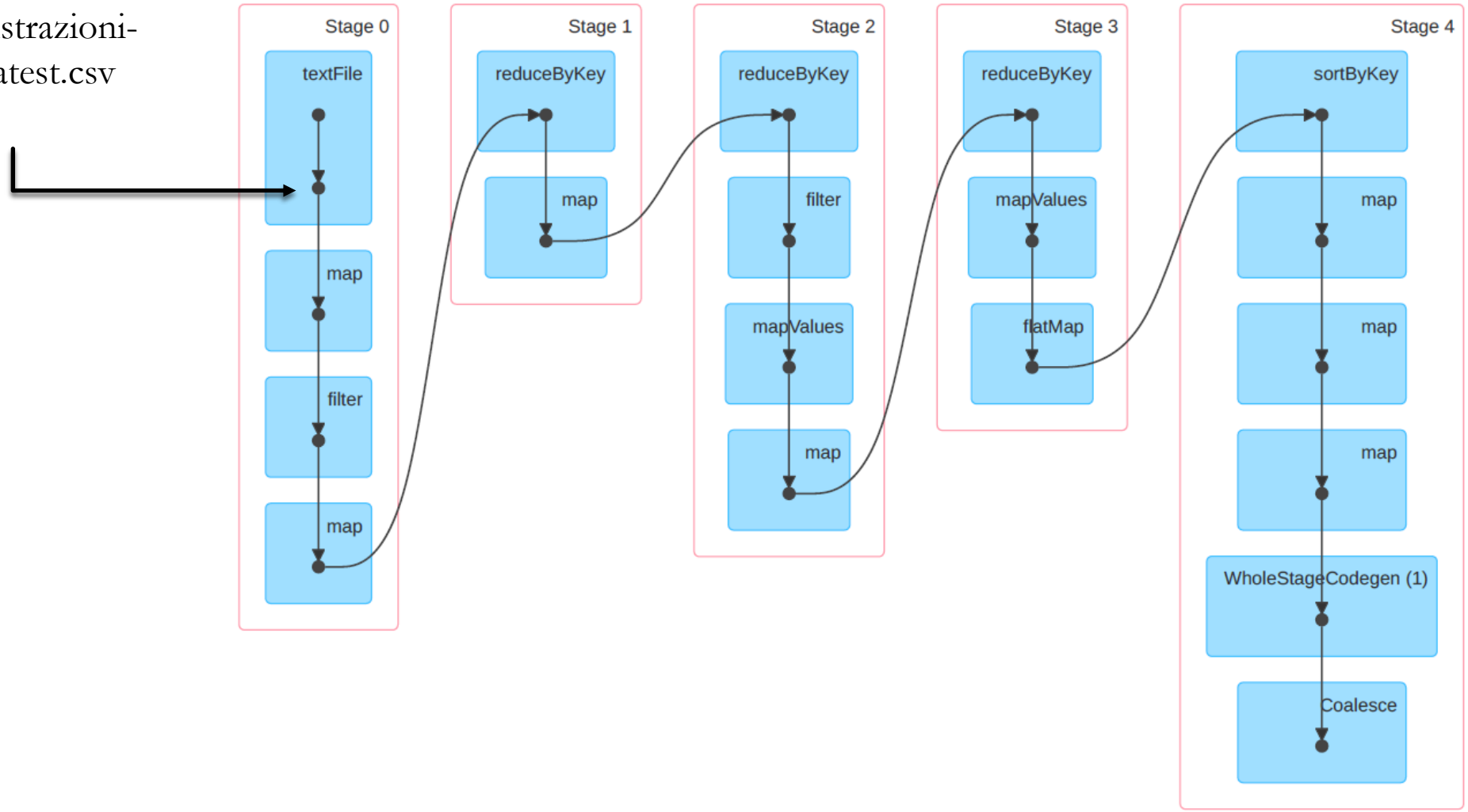
Metric	Min	25th percentile	Median
Duration	0.1 s	0.1s	0.1s
GC Time	6.0 ms	6.0 ms	6.0 ms
Shuffle Size/Record	169.6 Kib/755	169.6 Kib/755	169.6 Kib/755

QUERY 2


Stage 0
textFile
map
filter
map
Stage 1
reduceByKey
map
Stage 2
reduceByKey
filter
mapValues
map
Stage 3
reduceByKey
mapValues
flatMap
Stage 4
sortByKey
map
map
map
WholeStageCodegen (1)
Coalesce

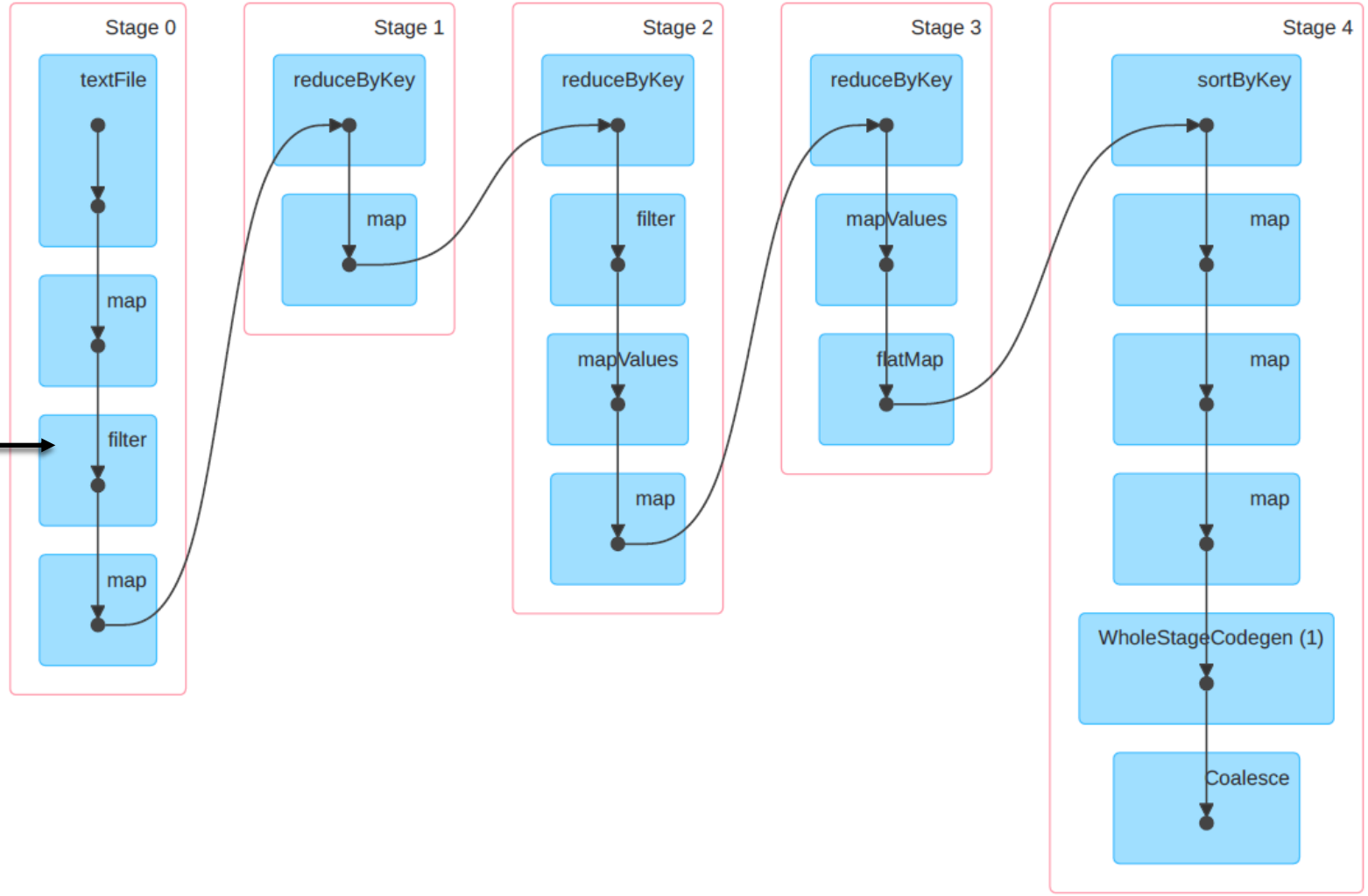
QUERY 2

somministrazioni-vaccini-latest.csv

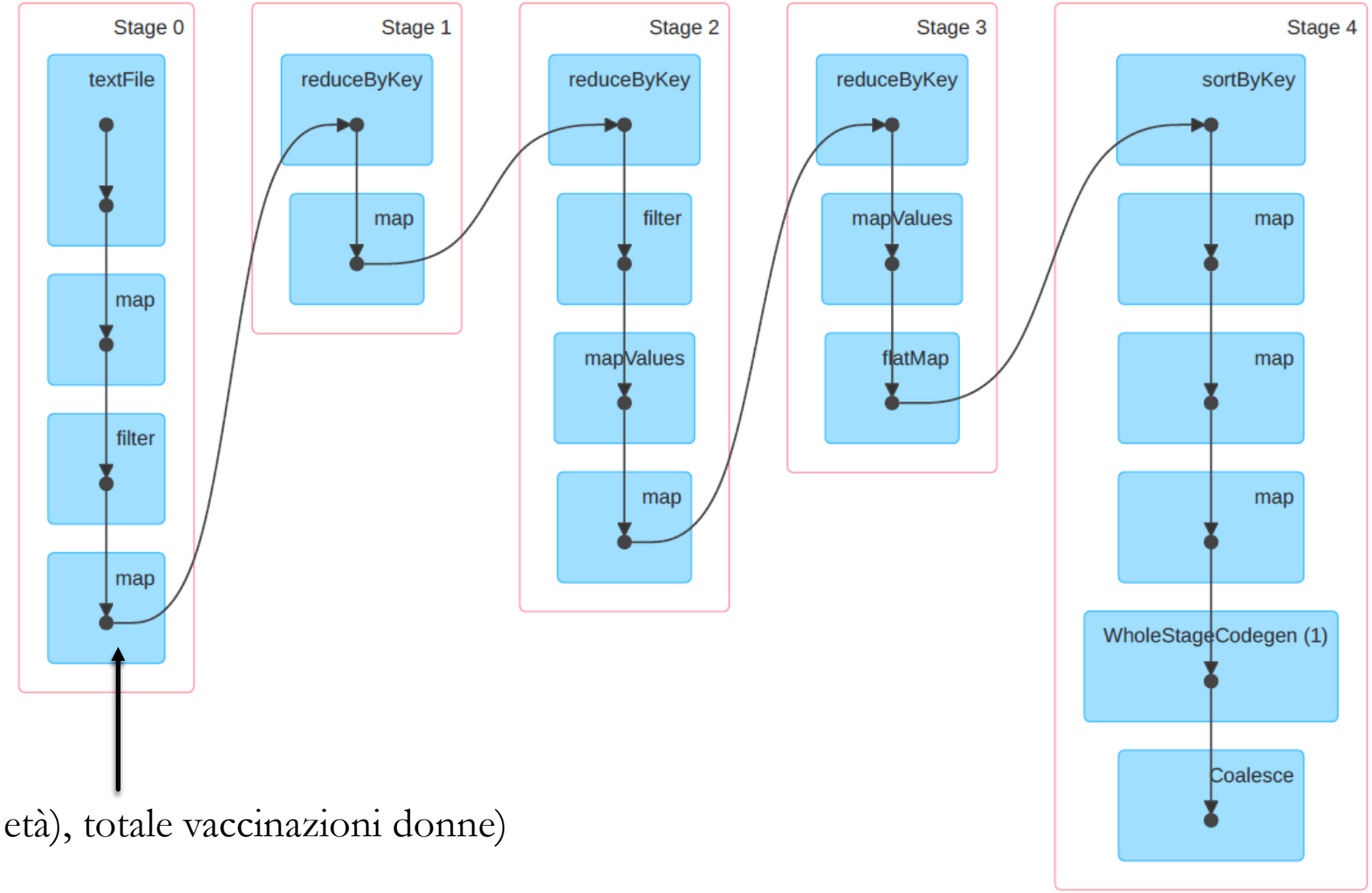


QUERY 2

giorni t.c. vaccinazioni donne > 0 && data dopo 31/01/2021



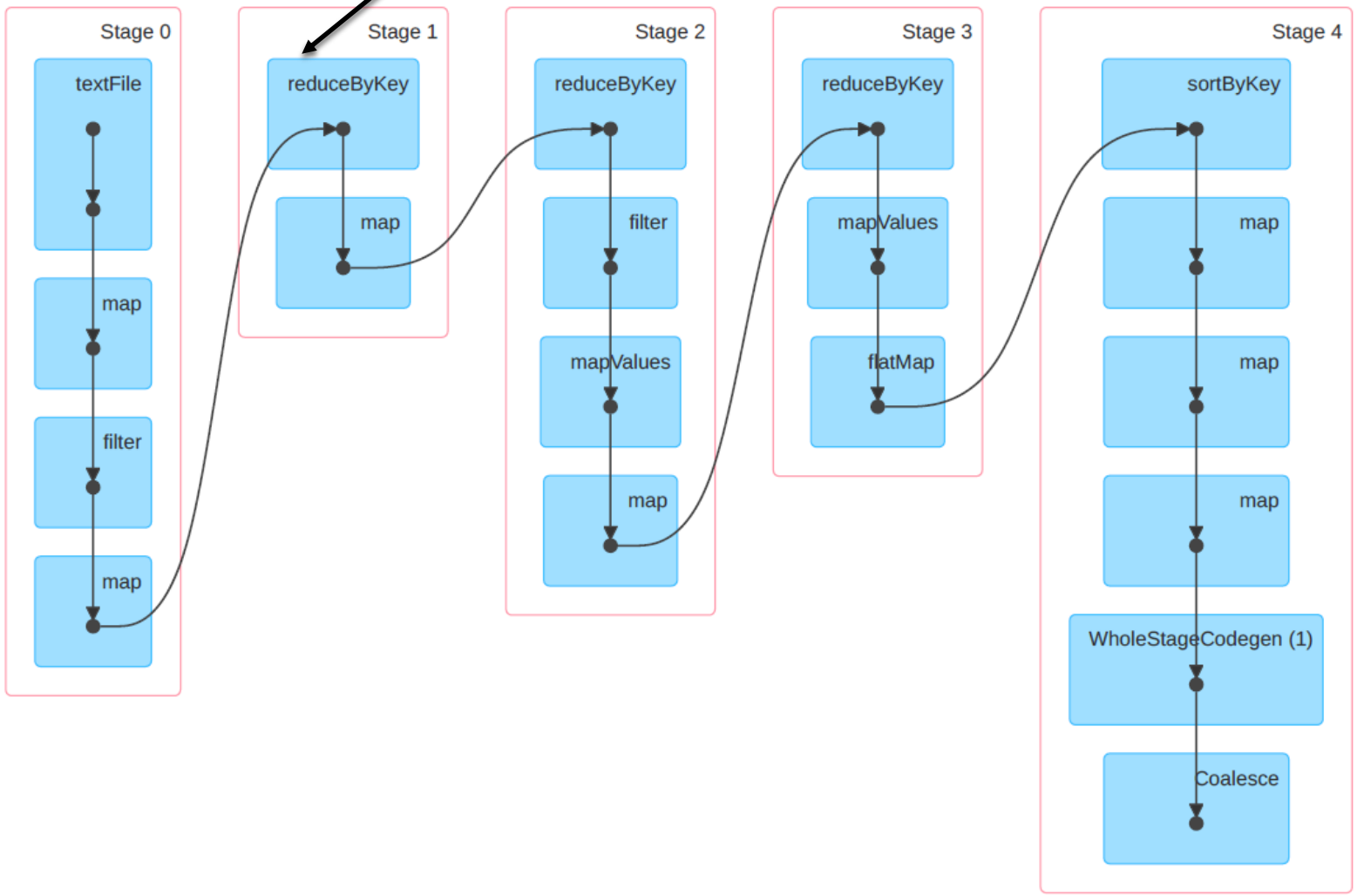
QUERY 2



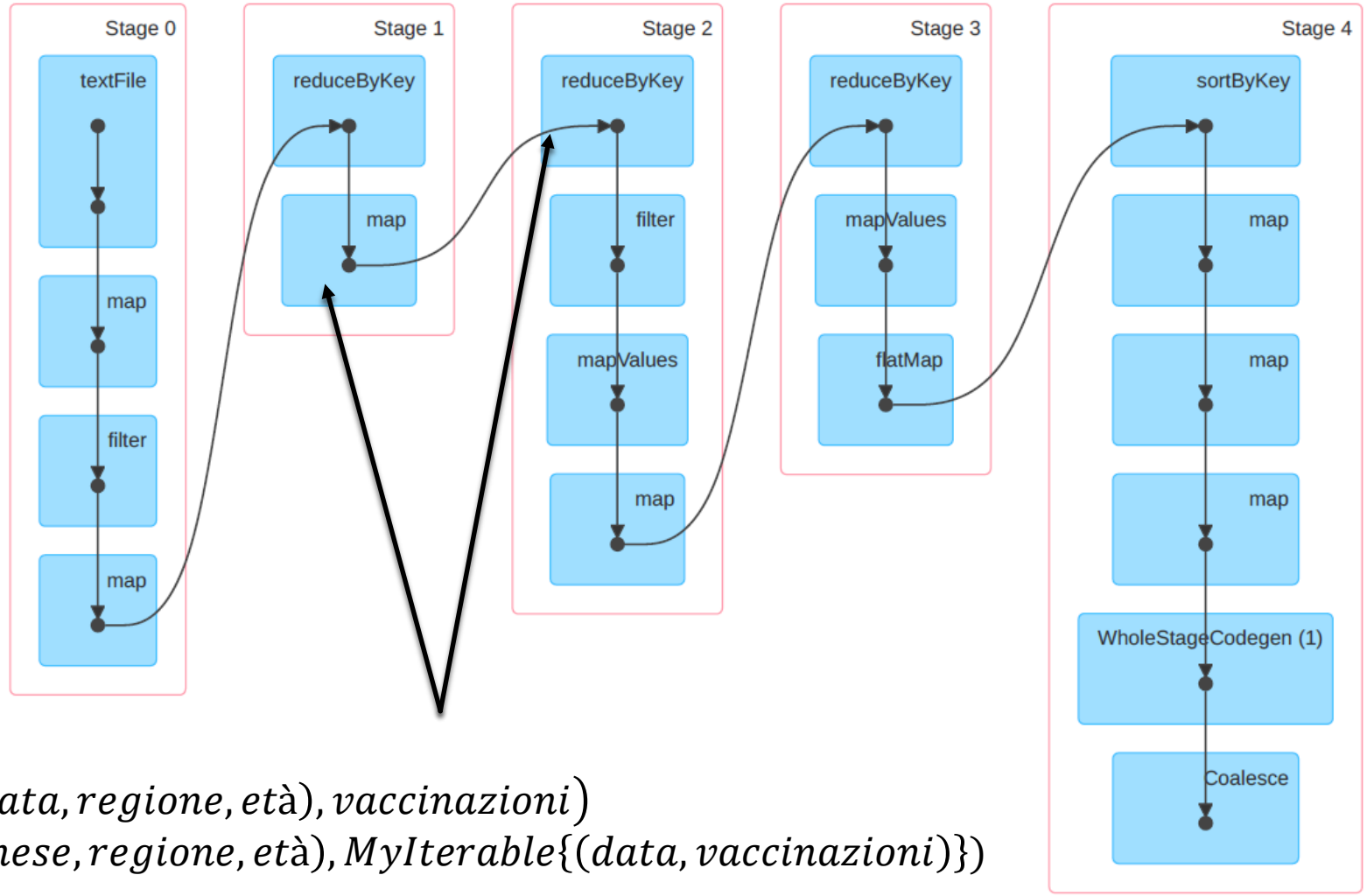
((data, regione, fascia età), totale vaccinazioni donne)

QUERY 2

Aggregare dati relativi a tipologie di vaccini differenti facendo la somma.



QUERY 2



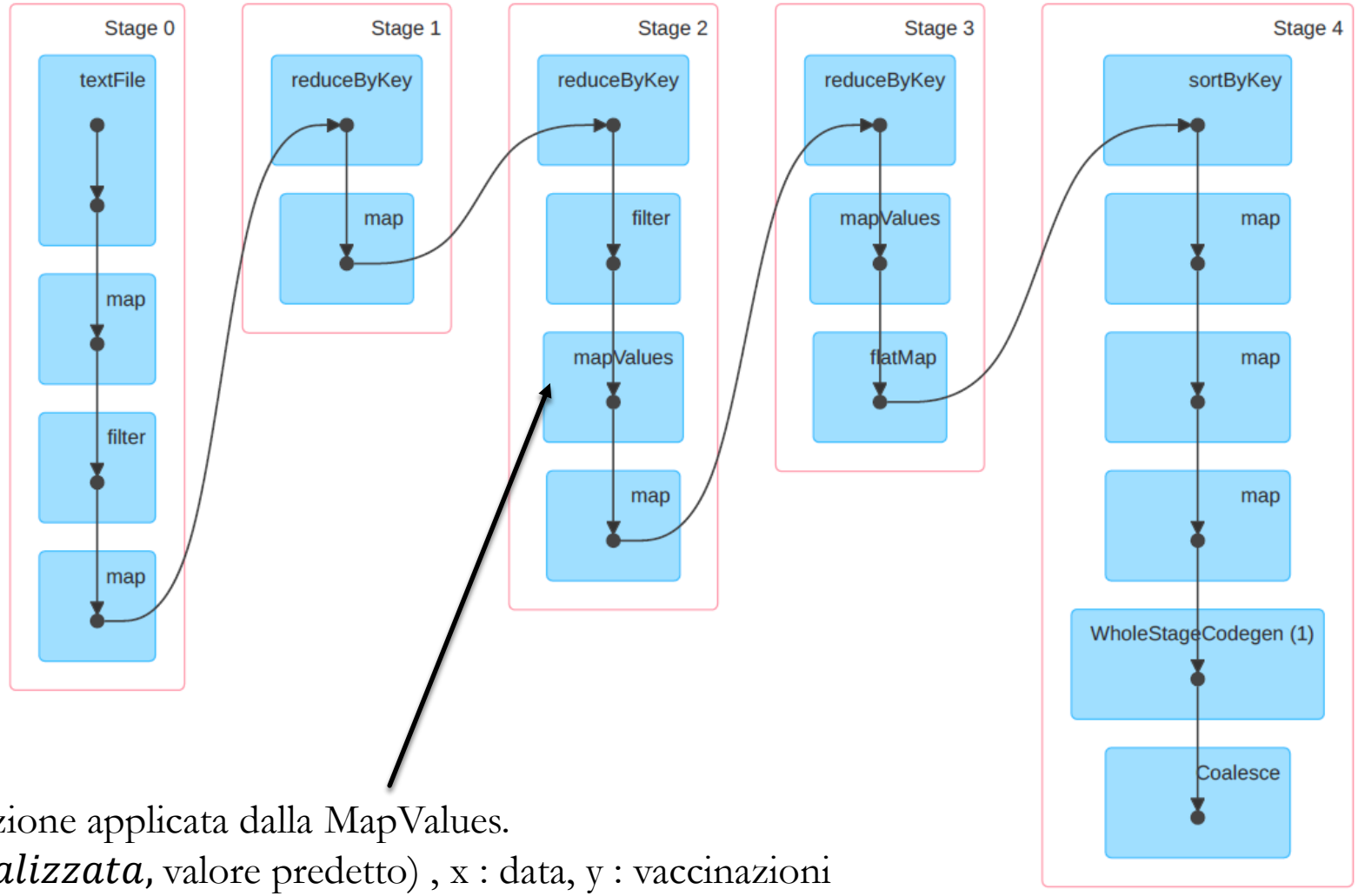
MapToPair

$$f: \big((data, regione, età), vaccinazioni\big)$$
$$\rightarrow ((mese, regione, età), MyIterable\{(data, vaccinazioni)\})$$

ReduceByKey:

$$f : ((mese, regione, età), MyIterable\{(data, vaccinazioni)\})$$
$$\rightarrow (\text{mese, regione, età}), \text{MyIterable}\{(\text{data}_1, \text{vaccinazioni}_1), \dots, (\text{data}_n, \text{vaccinazioni}_n)\}$$

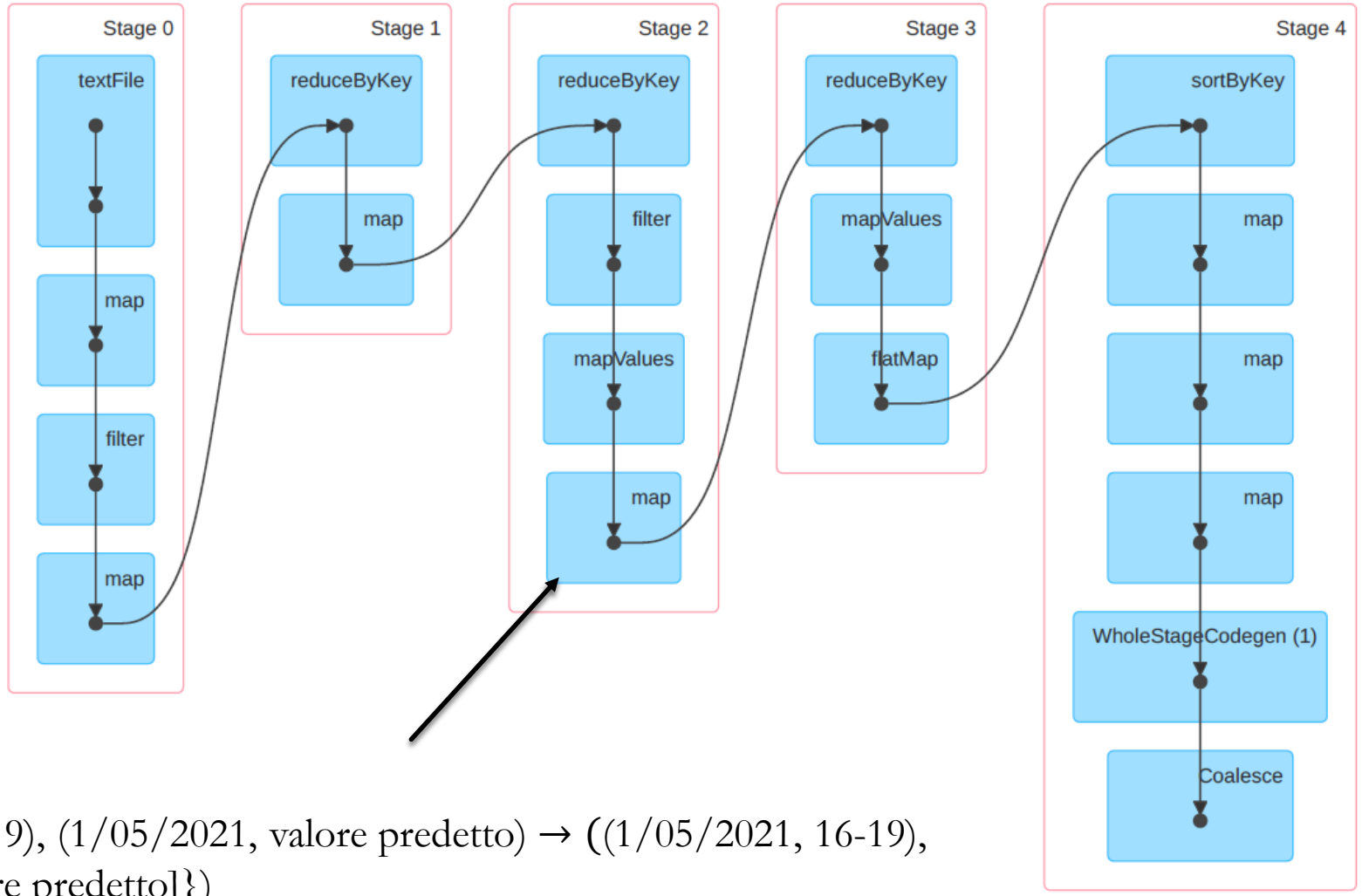
QUERY 2



Predittore: implementa funzione applicata dalla MapValues.
$f: (x, y) \rightarrow$ $(data\_normalizzata,$ valore predetto) , x : data, y : vaccinazioni

((04-2021, Abruzzo, 16-19), (1-05-2021, valore predetto))

QUERY 2



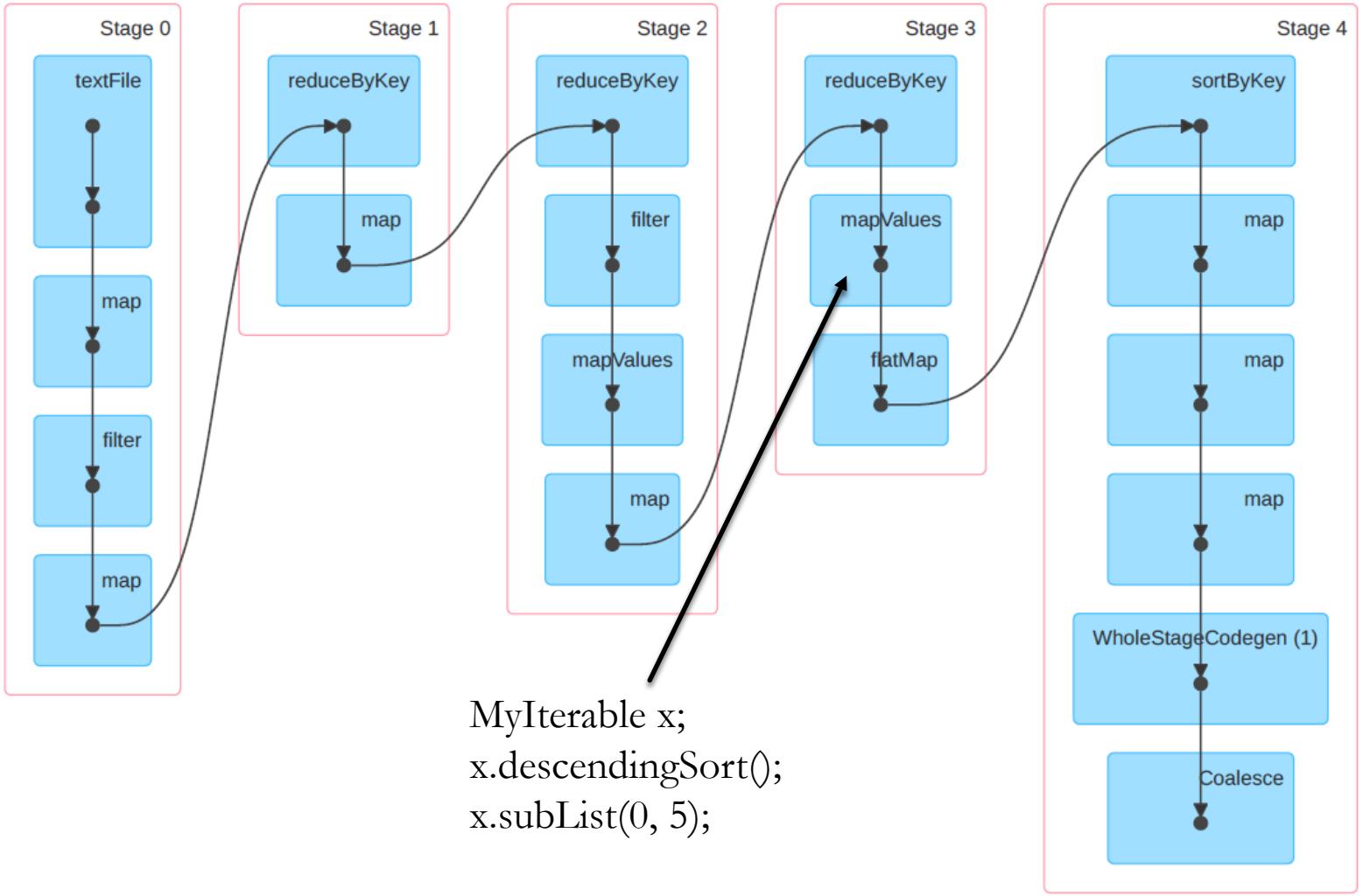
MapToPair

f: (04/2021, Abruzzo, 16-19), (1/05/2021, valore predetto) → ((1/05/2021, 16-19), MyIterable{[Abruzzo, valore predetto]})

ReduceByKey:

f : ((1/05/2021, 16−19), MyIterable{[Abruzzo, valore predetto]}) → ((1/05/2021, 16−19), $\text{MyIterable}\{(\text{Abruzzo}, \text{vaccinazioni}_{\text{ABR}}), \ldots, (\text{Lazio}, \text{vaccinazioni}_{\text{Laz}})\}$)

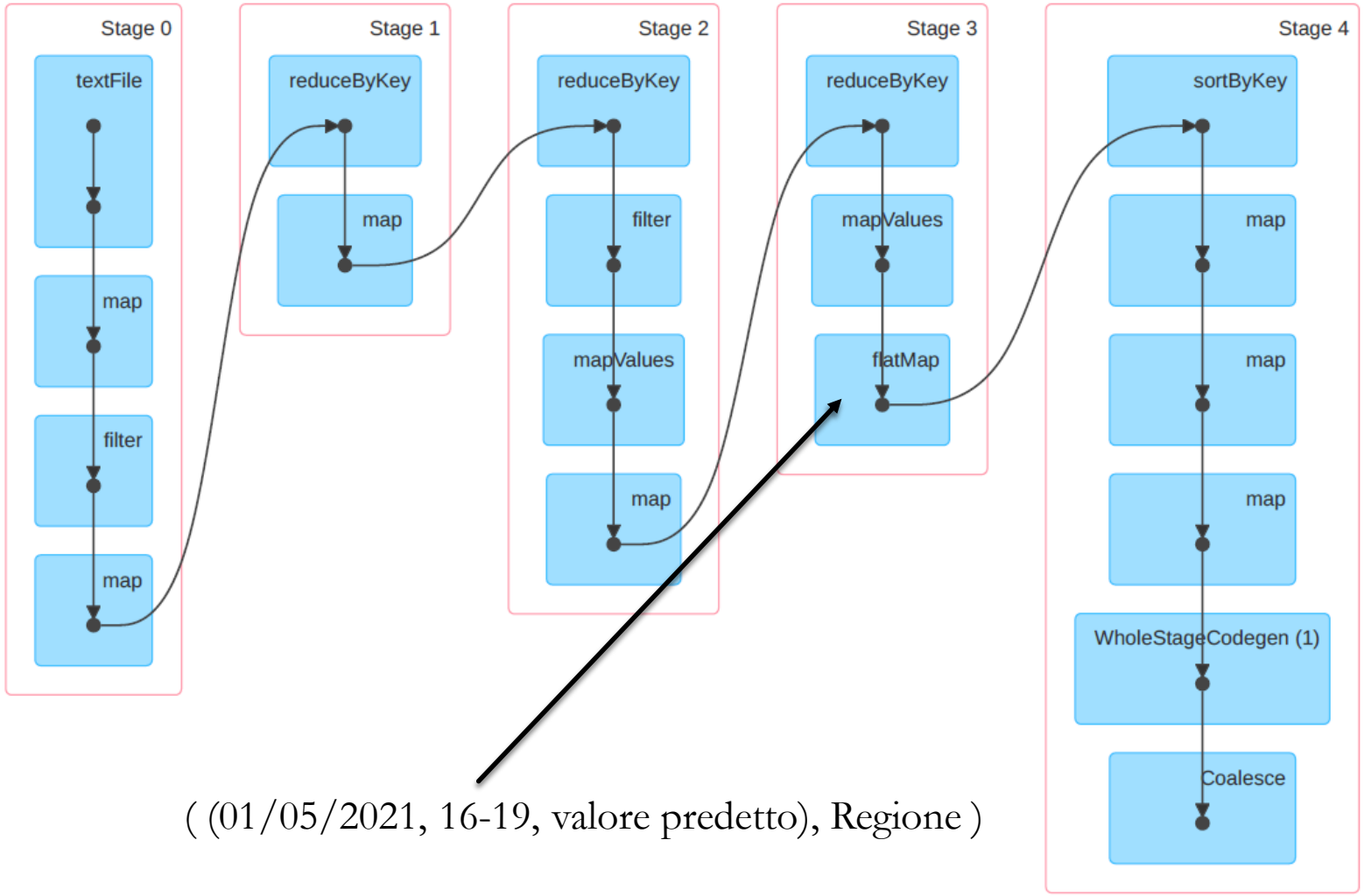
QUERY 2



MyIterable x;
x.descendingSort();
x.subList(0, 5);

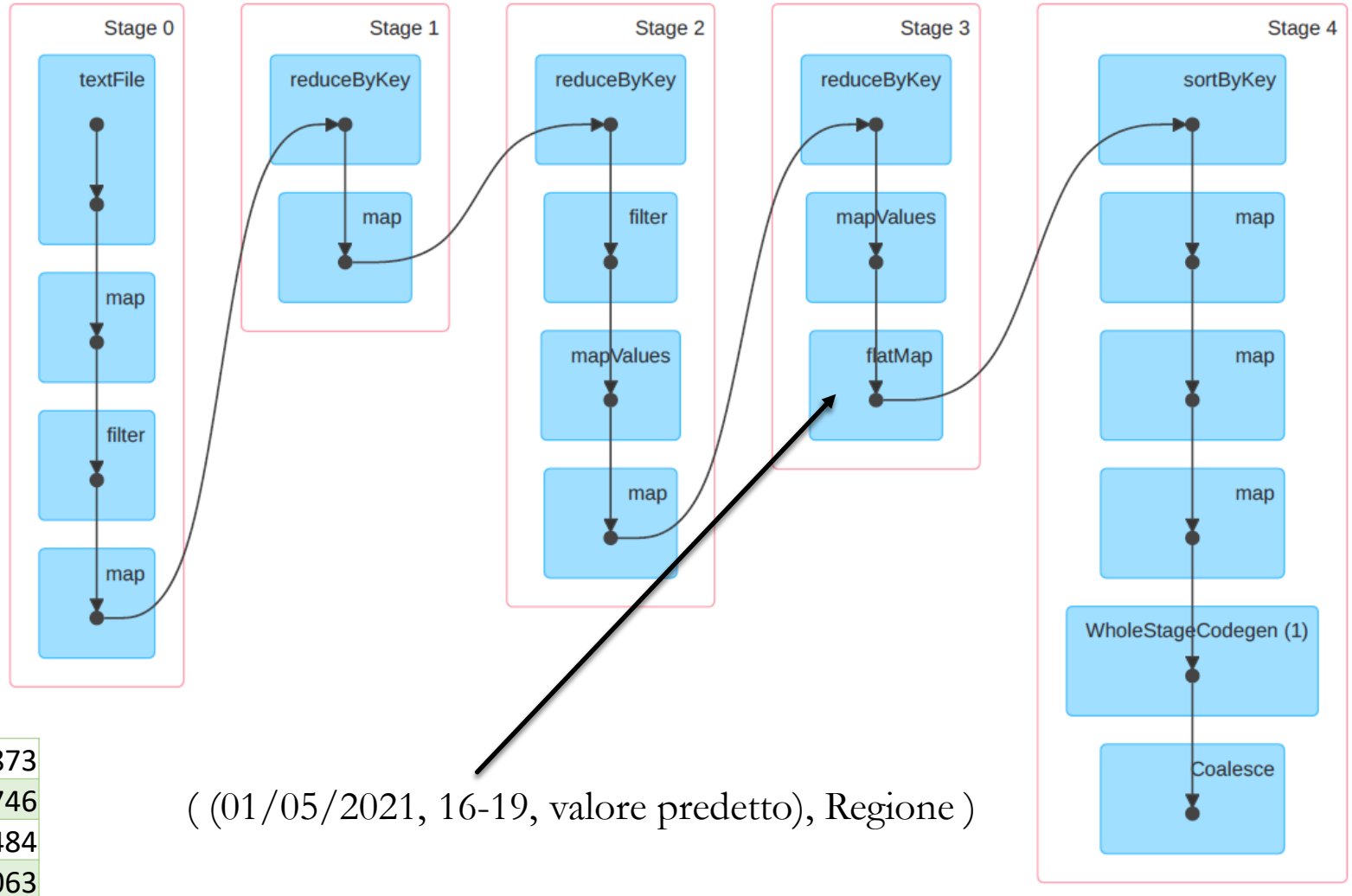
Dati sono stati aggrgati in base al primo giorno del mese e alla fascia anagrafica.
Tutte le liste ottenute hanno una lunghezza limitata (< 22) al numero delle regioni d'Italia.

QUERY 2



((01/05/2021, 16-19, valore predetto), Regione)

QUERY 2



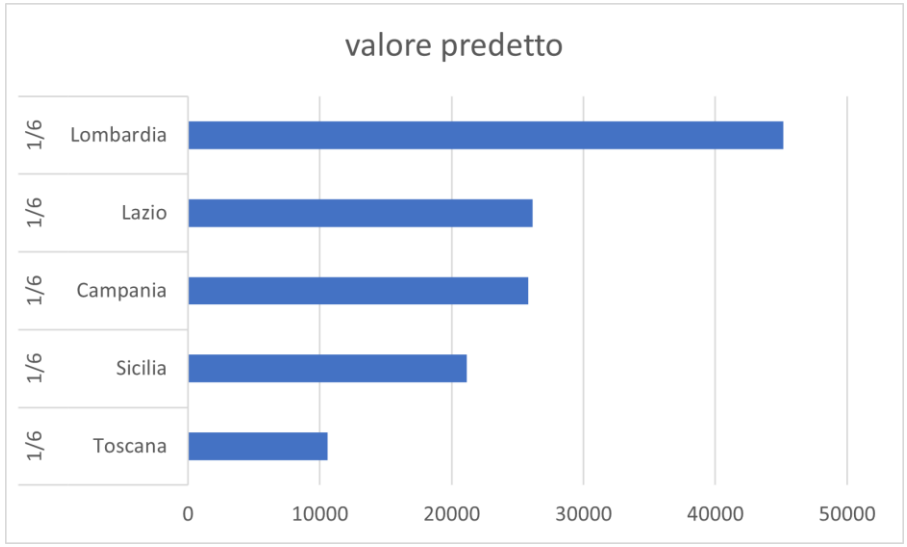
((01/05/2021, 16-19, valore predetto), Regione)

01/03/2021	20-29	Lombardia	271,873
01/03/2021	20-29	Piemonte	283,746
01/03/2021	20-29	Veneto	334,484
01/03/2021	20-29	Puglia	363,063
01/03/2021	20-29	Toscana	444,738
01/03/2021	30-39	Lombardia	387,81
01/03/2021	30-39	Piemonte	394,119
01/03/2021	30-39	Puglia	545,508
01/03/2021	30-39	Campania	564,786
01/03/2021	30-39	Toscana	911,492

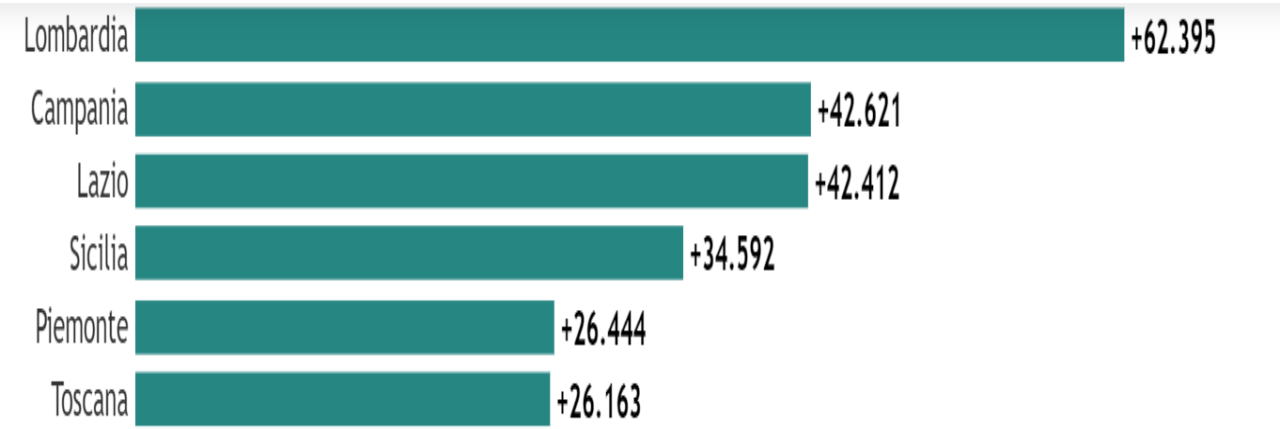
ANALISI RISULTATI QUERY 2

VS

VALORI PREDETTI



CLASSIFICA UFFICIALE



ANALISI TEMPI DI ESECUZIONE

- *Tempi di esecuzione query 1*

Stage	0	1	2	3	4	5
Durata	79ms	0.2 s	43ms	41ms	27ms	0.2 s

- *Tempi di esecuzione query 2*

Stage	0	1	2	3	4
Durata	0.6 s	0.2 s	0.2 s	31ms	0.2s

- *Tempi esecuzione programma*

Programma	query1	query2	query1&2
Durata Media	2.83 s	3.55 s	3.76 s
Query1	2.83 s	-	2.84 s
Query2	-	3.55 s	0.92 s

Processore Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz 3.60 GHz
RAM installata 16,0 GB
Tipo sistema Sistema operativo a 64 bit, processore basato su x64

GRAZIE PER L'ATTENZIONE