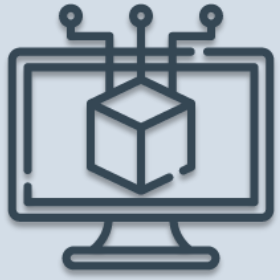
Sistemi e architetture per big data

Progetto 1 - Analisi dei dati inerenti alle vaccinazioni anti-Covid-19 in Italia

Giuseppe Lasco

Marco Marcucci

Sommario



Architettura

- NiFi
- HDFS
- Spark
- HBase

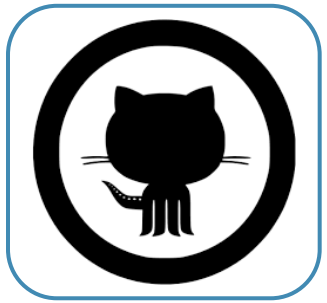
Query

- Query 1
- Query 2
- Query 3

Benchmark

- Prestazioni Query e Clustering
- Qualità clustering

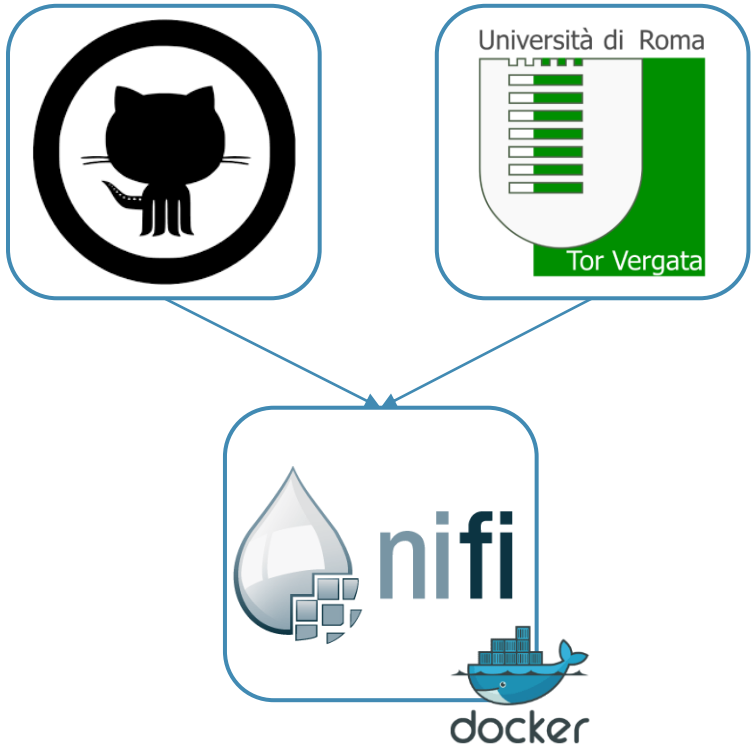
Architettura



- somministrazioni-vaccini-summary-latest.csv
- somministrazioni-vaccini-latest.csv
- punti-somministrazione-tipologia.csv
- totale-popolazione.csv

GitHub & Uniroma2

Architettura

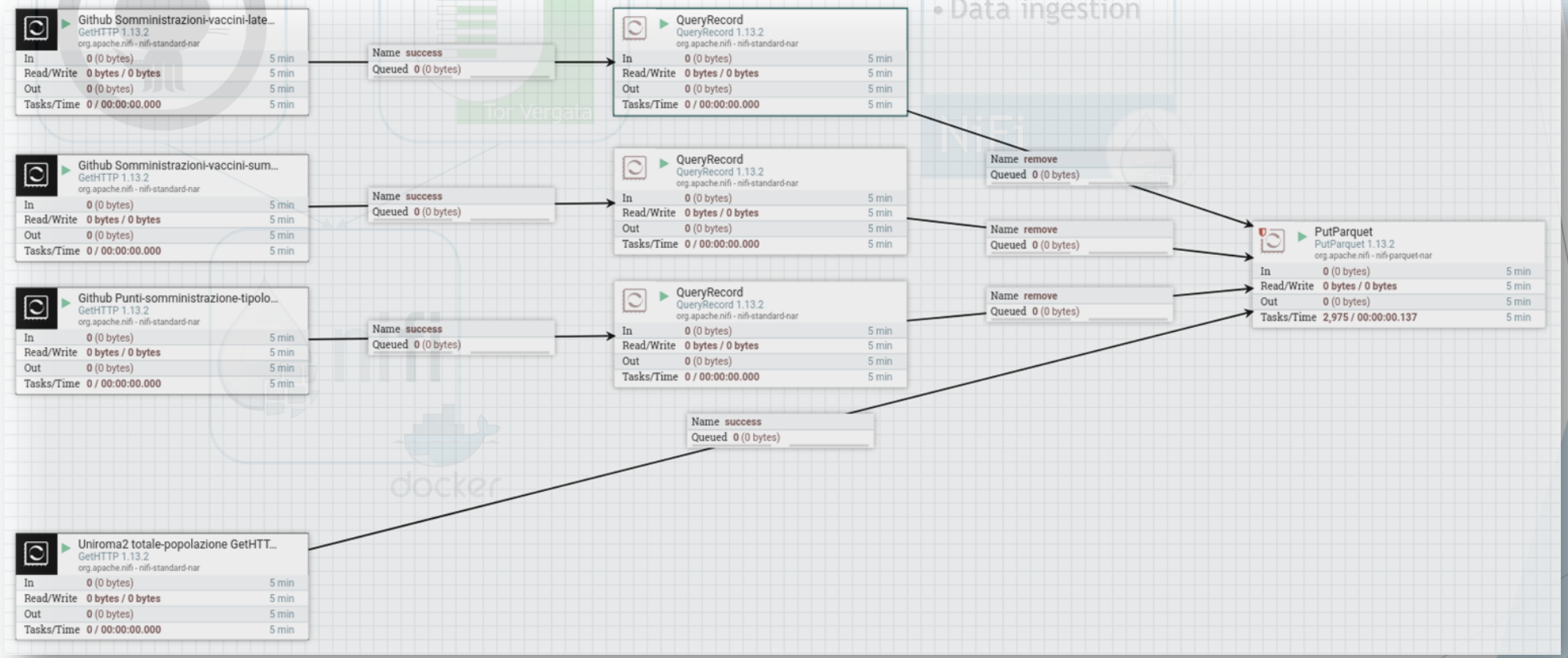


- Data acquisition
- Data parsing
- Data ingestion

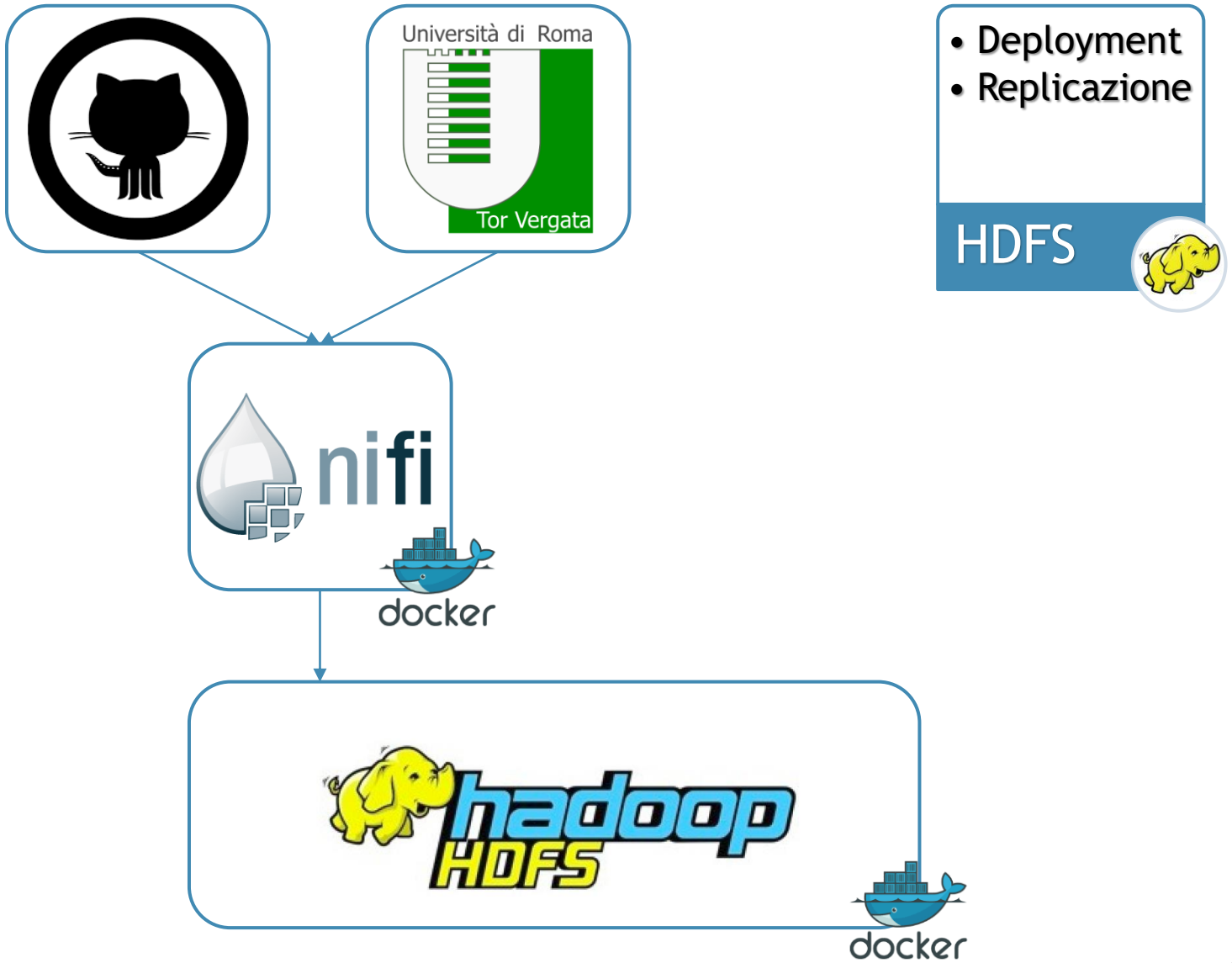
NiFi

Architettura

- Data acquisition
- Data parsing
- Data ingestion



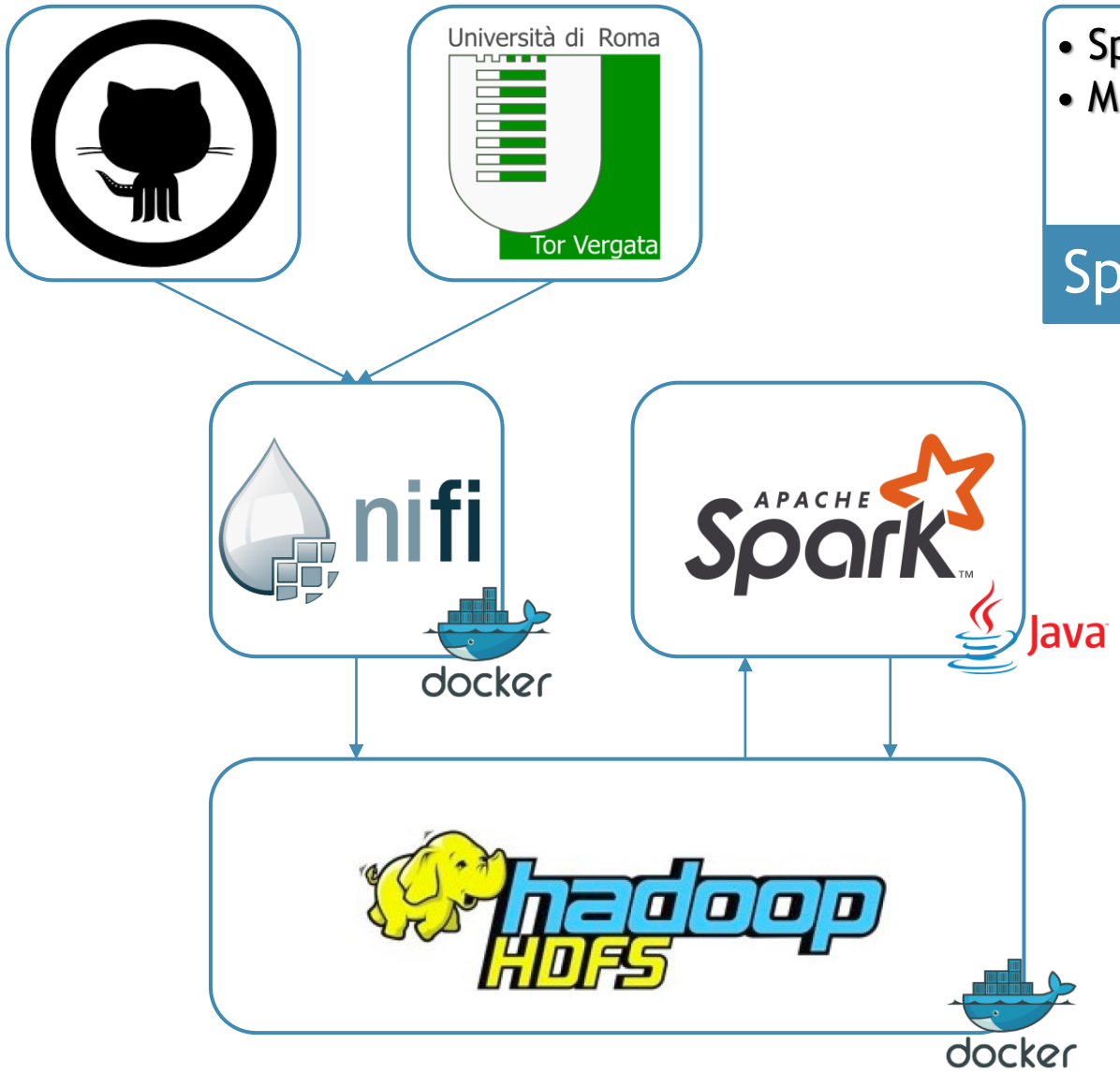
Architettura



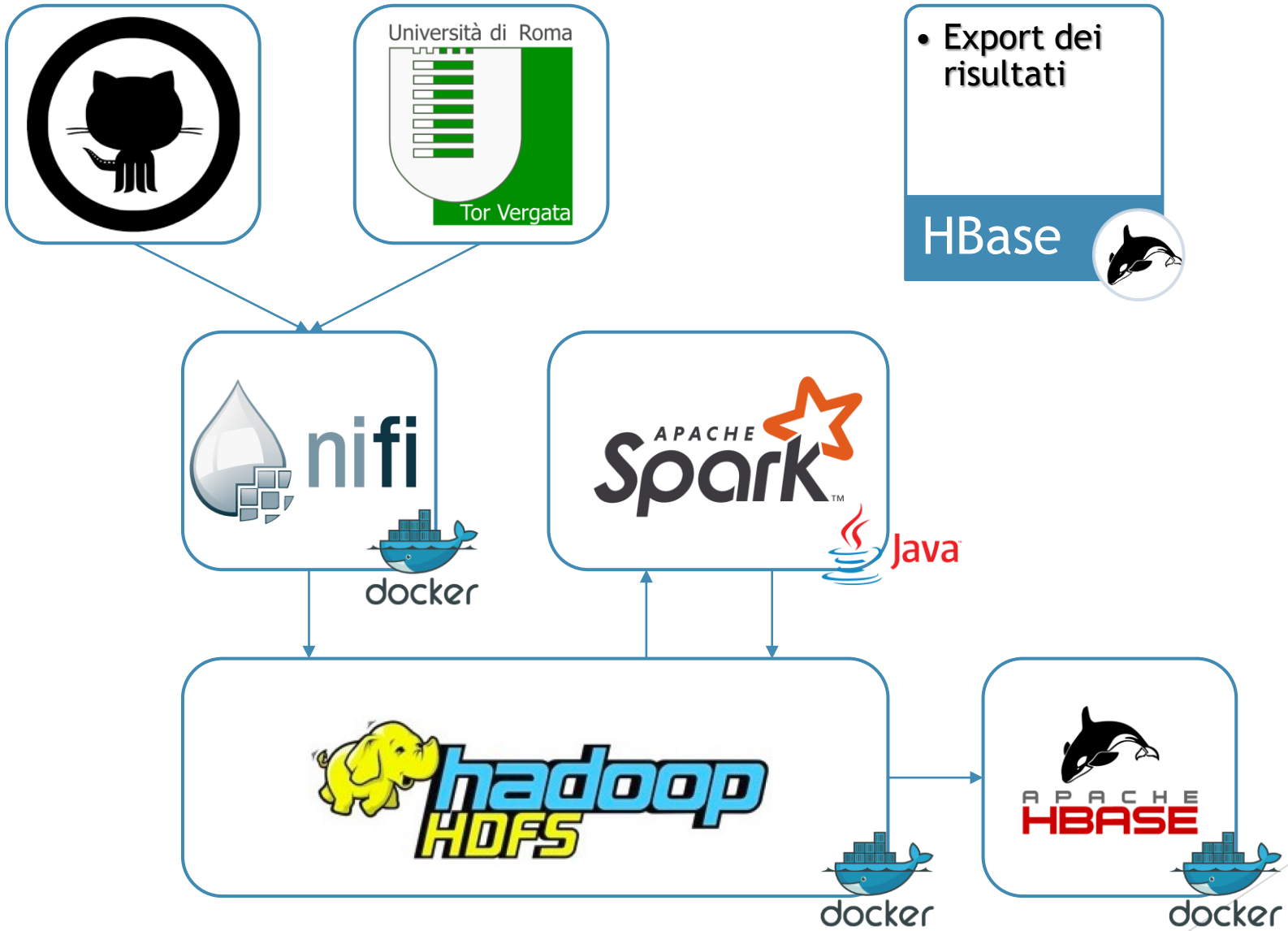
- Deployment
- Replicazione

HDFS

Architettura

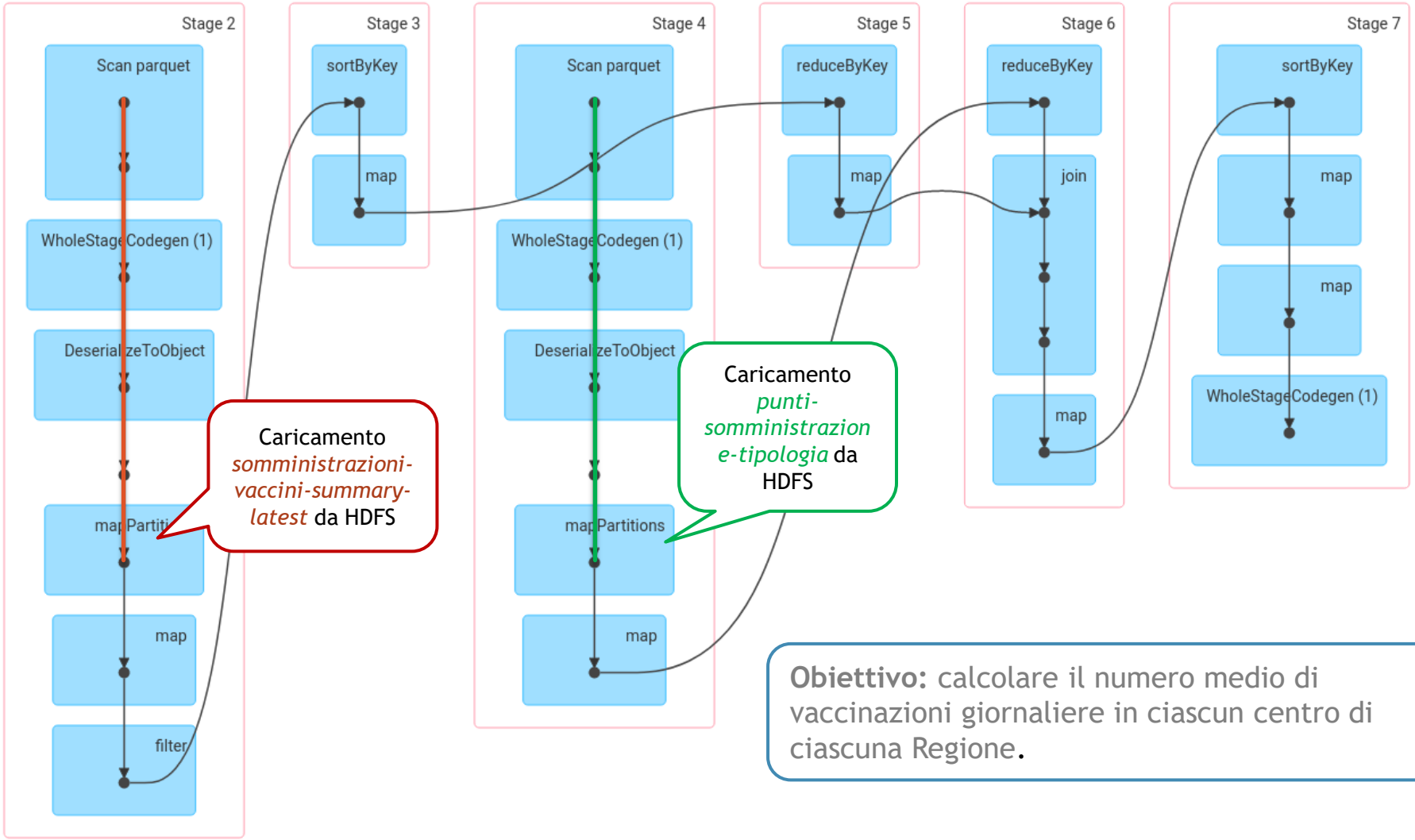


Architettura



Query

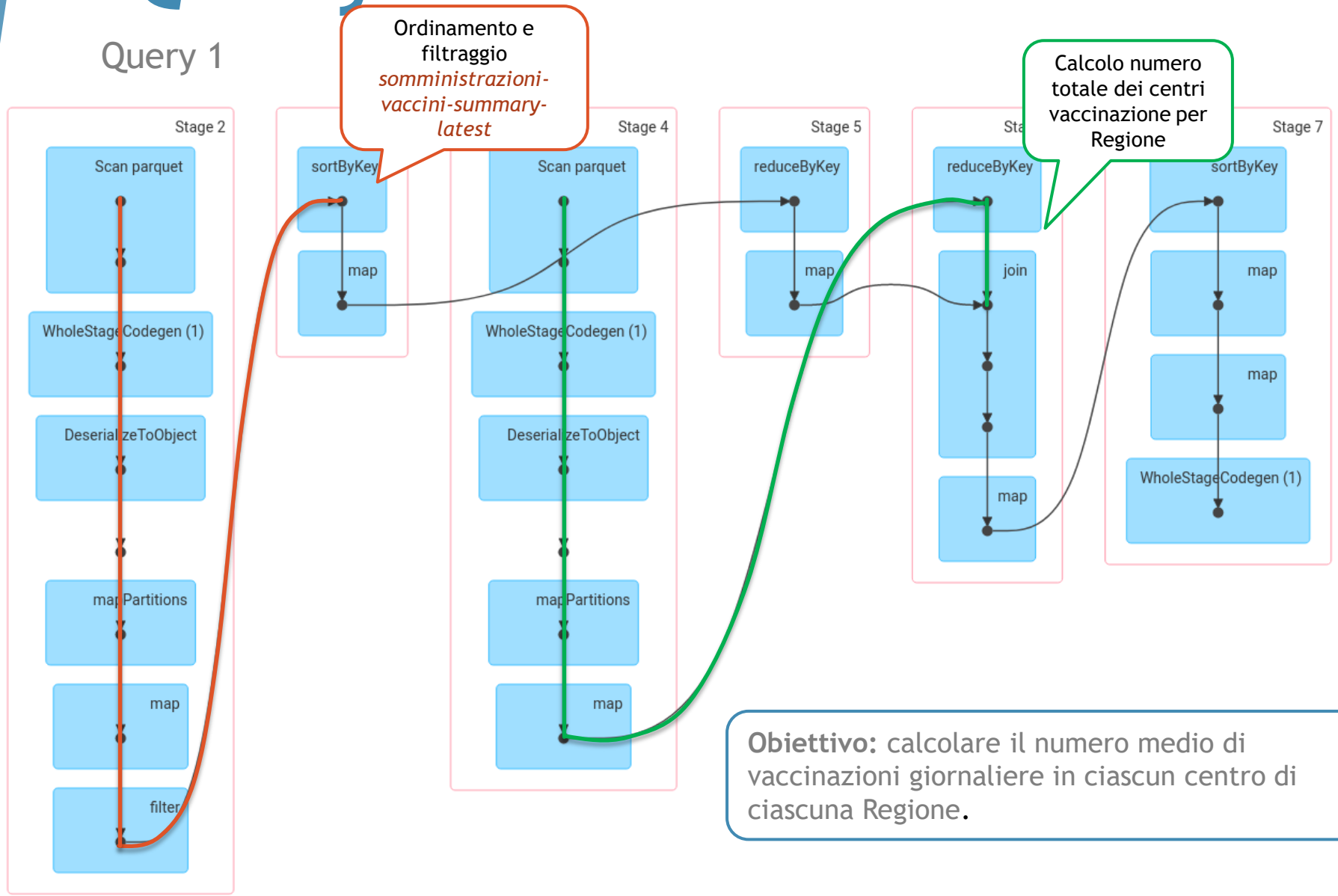
Query 1



Obiettivo: calcolare il numero medio di vaccinazioni giornaliere in ciascun centro di ciascuna Regione.

Query

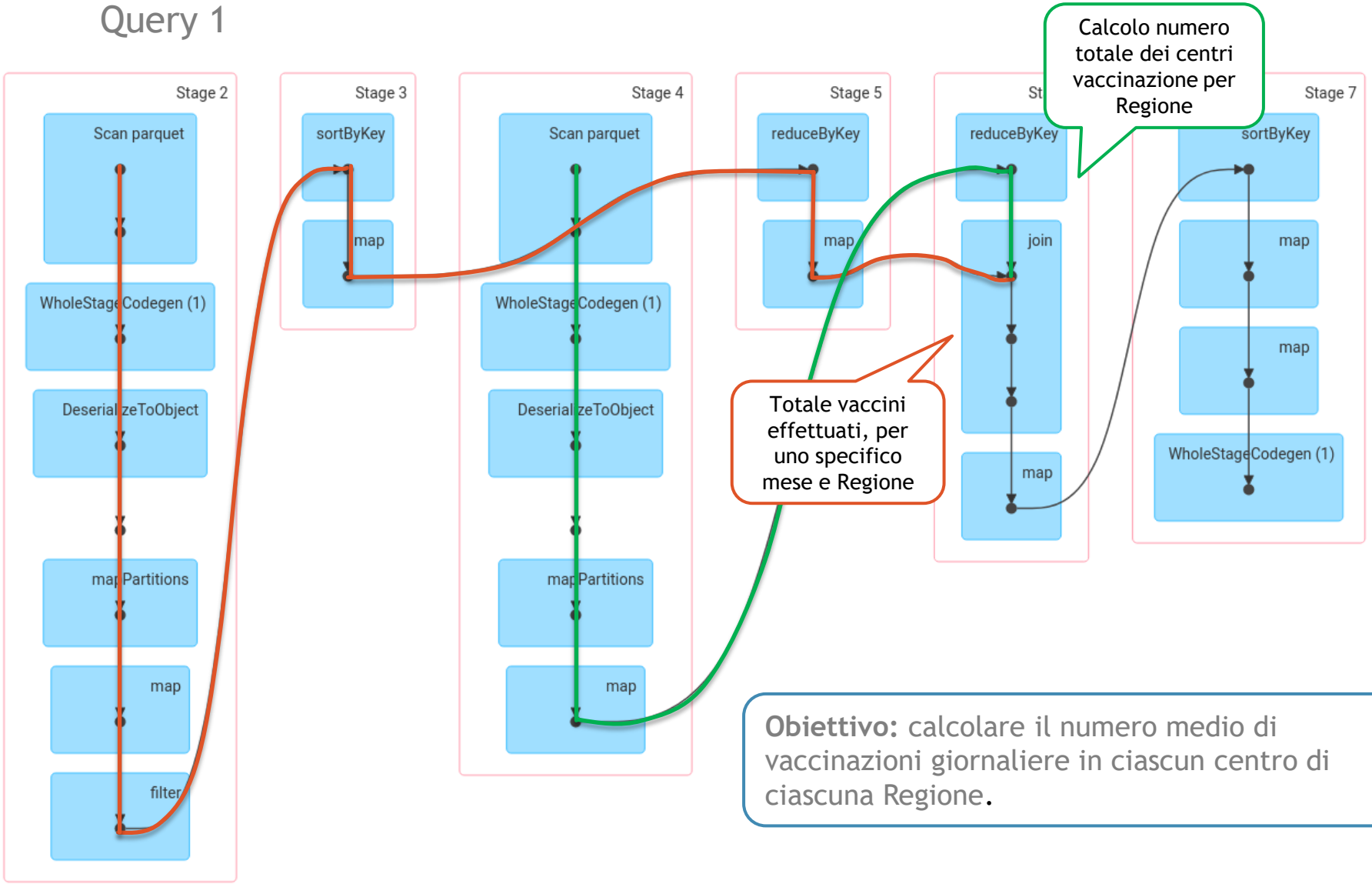
Query 1



Obiettivo: calcolare il numero medio di vaccinazioni giornaliere in ciascun centro di ciascuna Regione.

Query

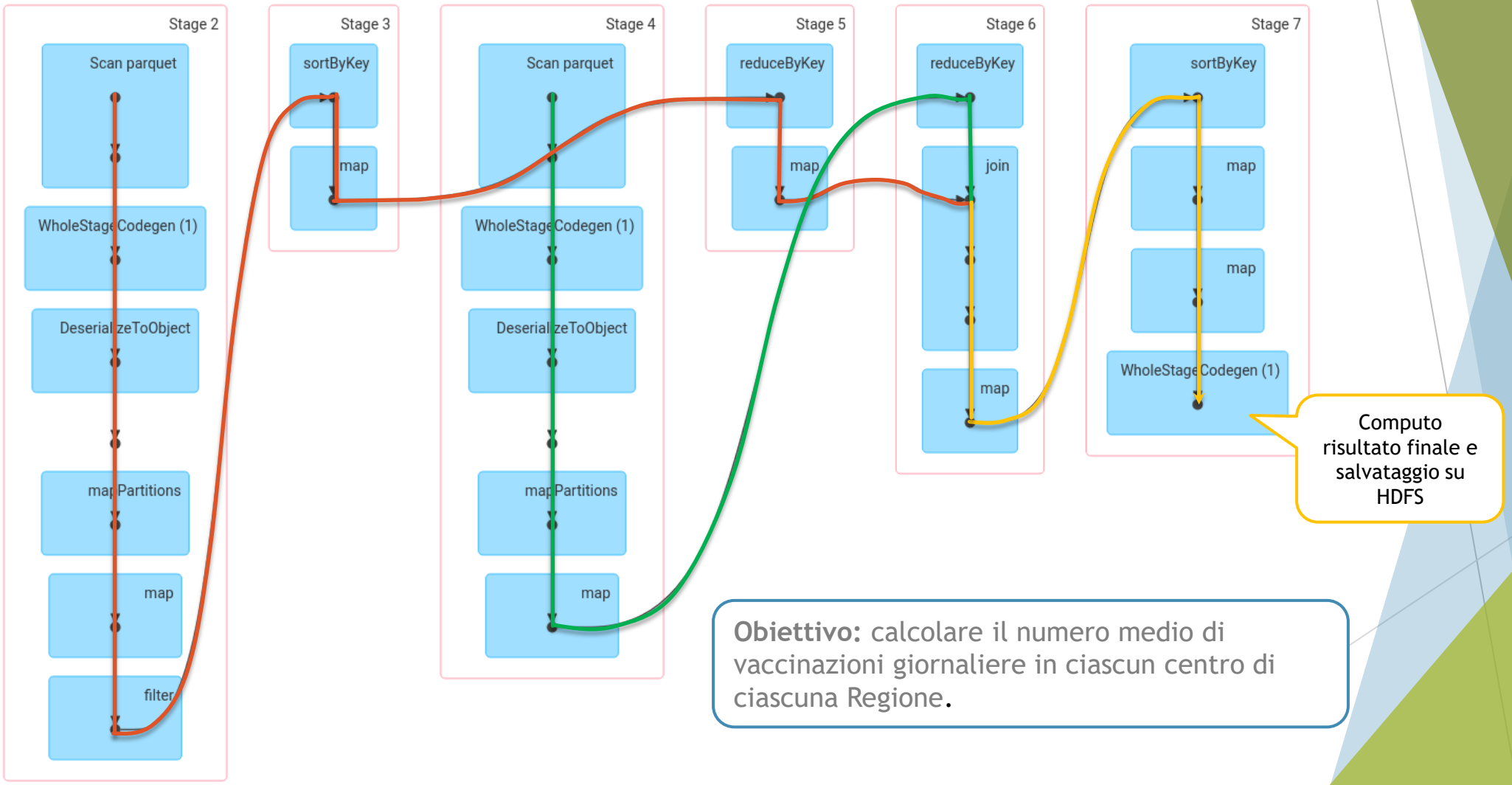
Query 1



Obiettivo: calcolare il numero medio di vaccinazioni giornaliere in ciascun centro di ciascuna Regione.

Query

Query 1

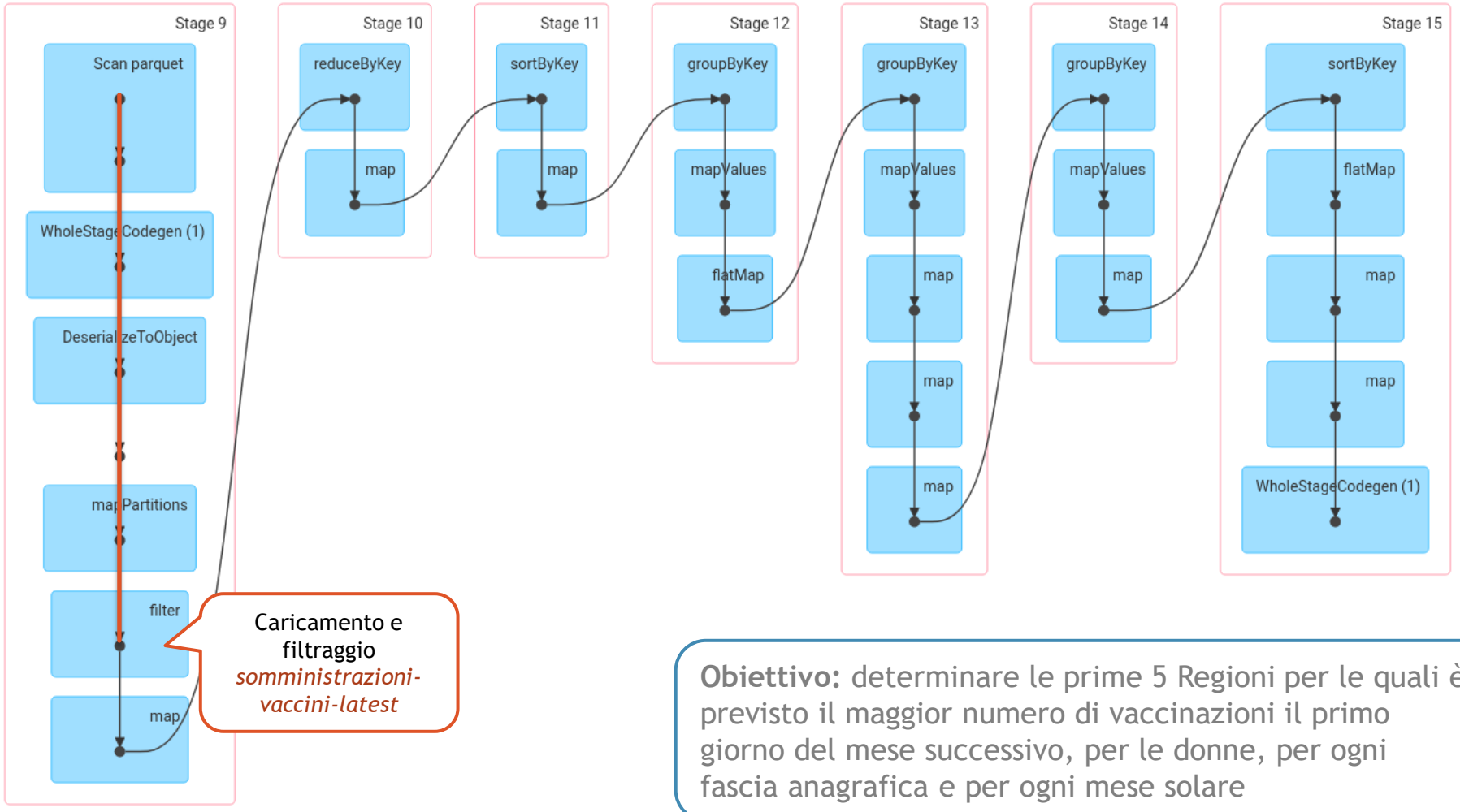


Computo risultato finale e salvataggio su HDFS

Obiettivo: calcolare il numero medio di vaccinazioni giornaliere in ciascun centro di ciascuna Regione.

Query

Query 2

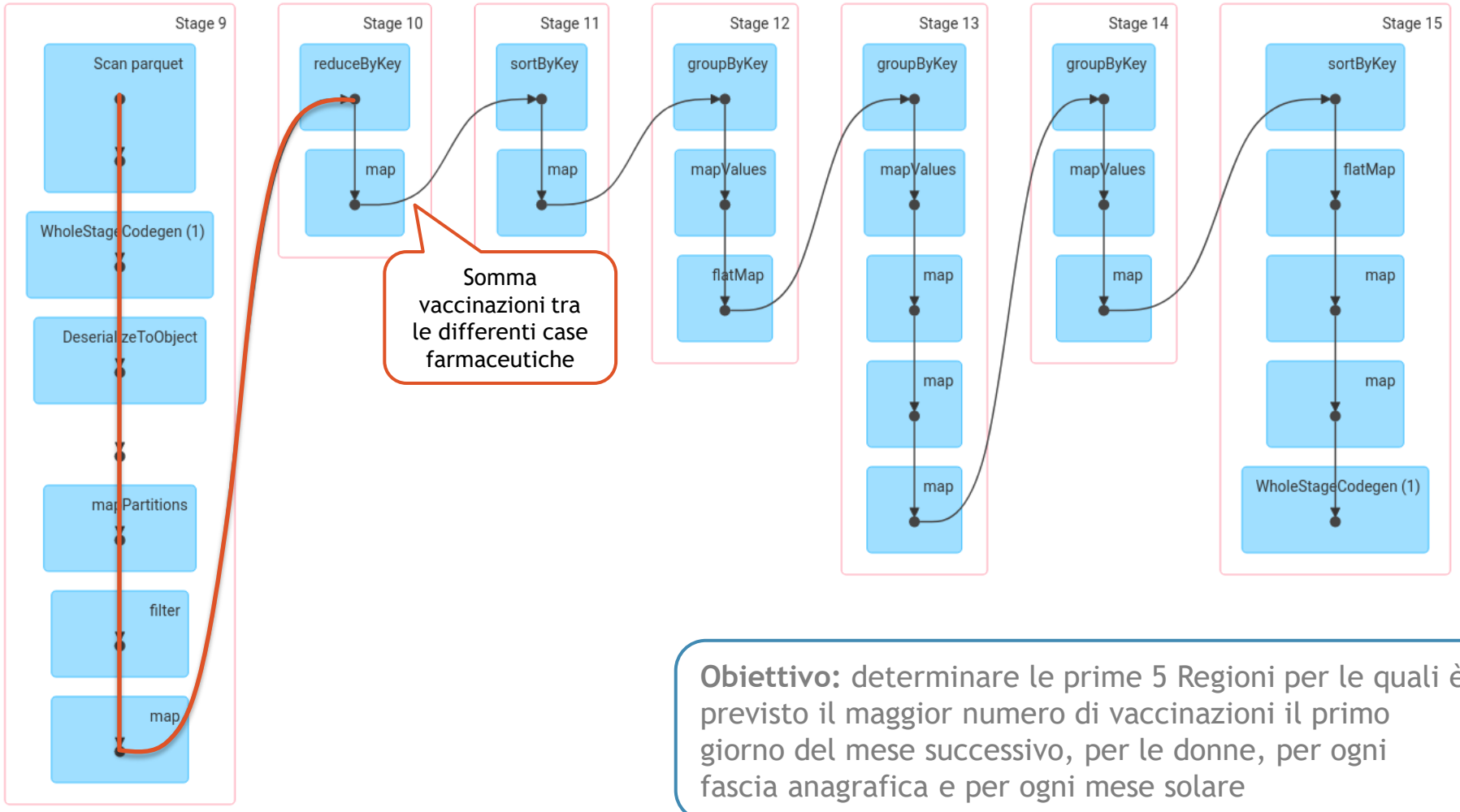


Caricamento e filtraggio *somministrazioni-vaccini-latest*

Obiettivo: determinare le prime 5 Regioni per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo, per le donne, per ogni fascia anagrafica e per ogni mese solare

Query

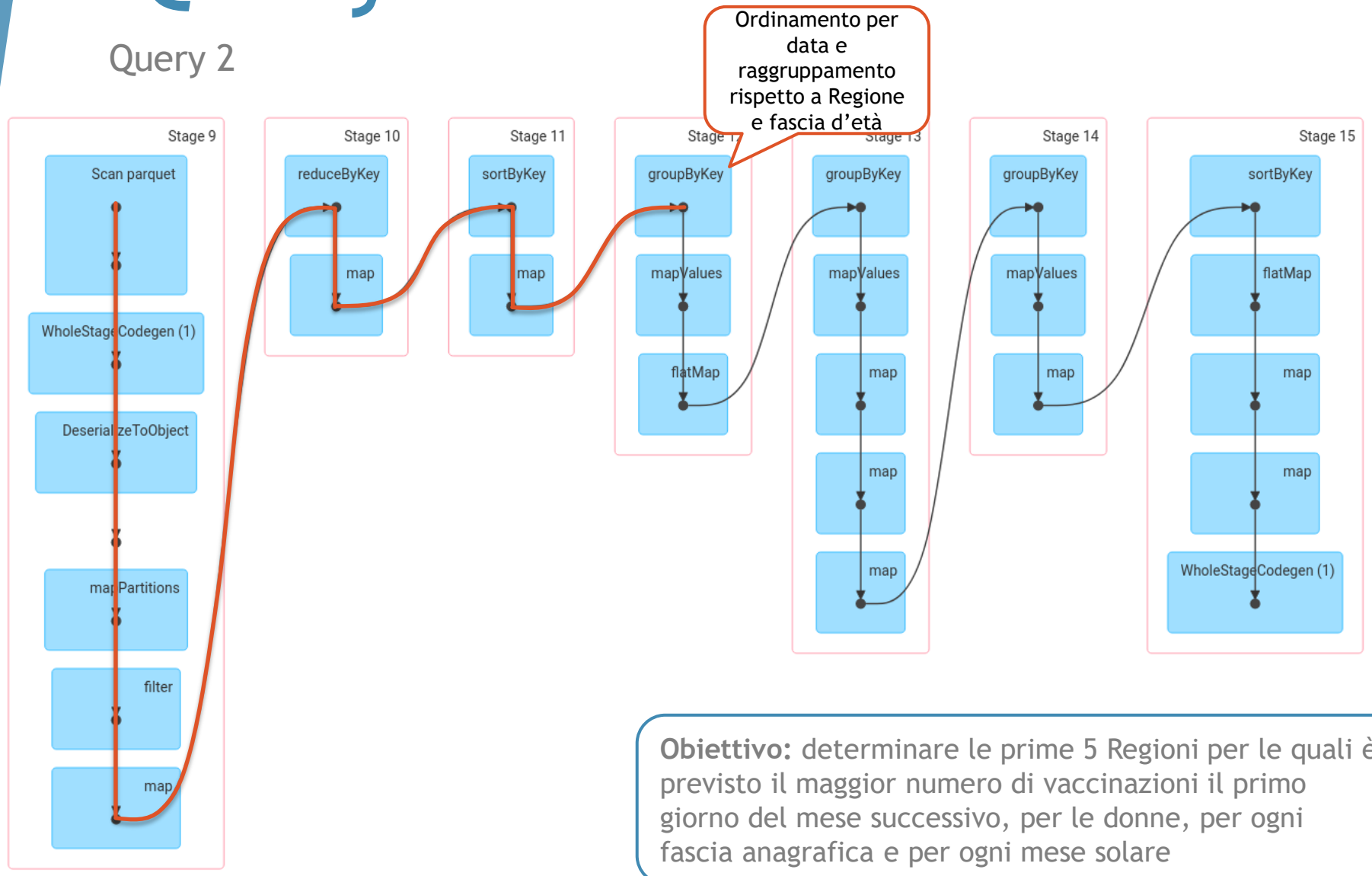
Query 2



Obiettivo: determinare le prime 5 Regioni per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo, per le donne, per ogni fascia anagrafica e per ogni mese solare

Query

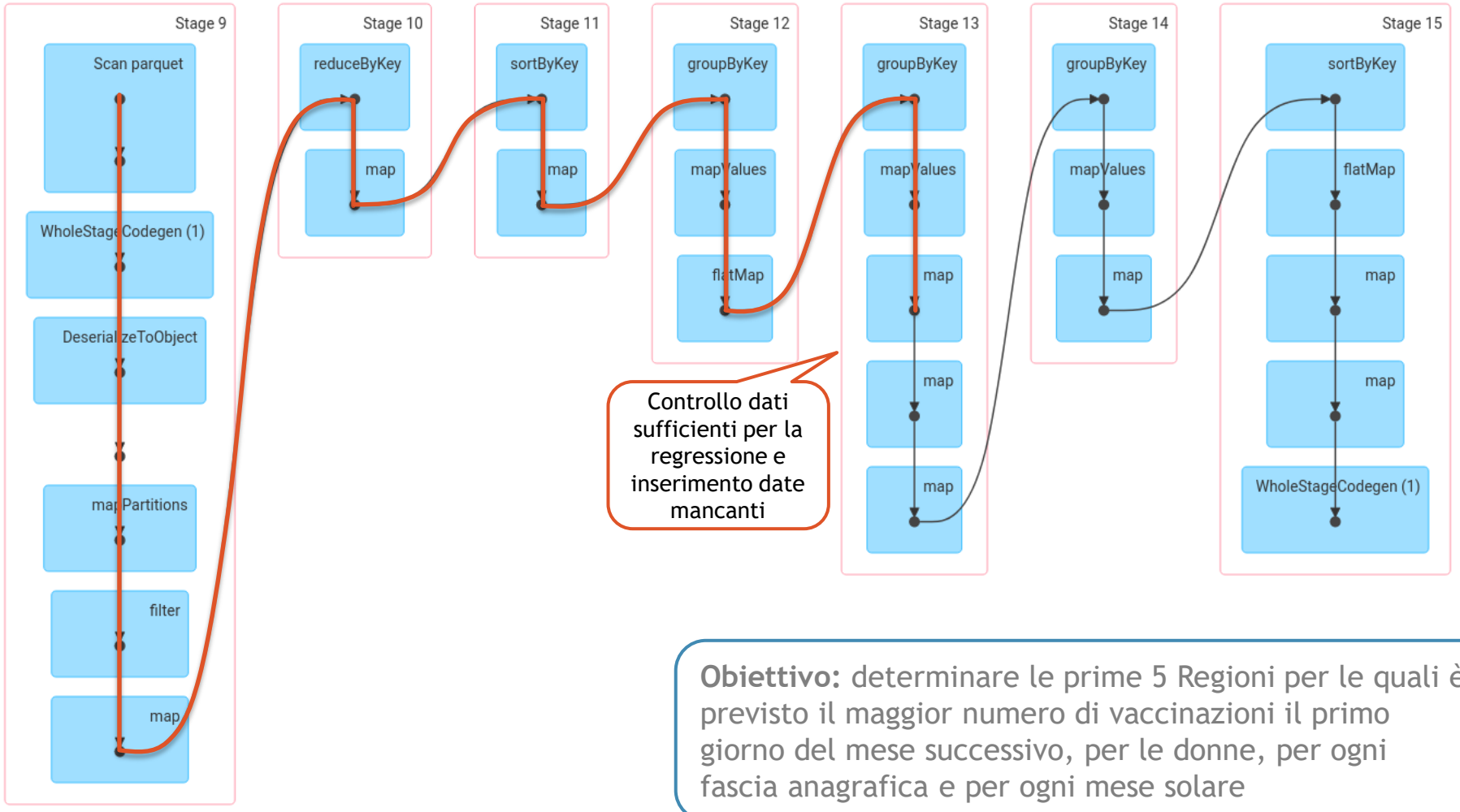
Query 2



Obiettivo: determinare le prime 5 Regioni per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo, per le donne, per ogni fascia anagrafica e per ogni mese solare

Query

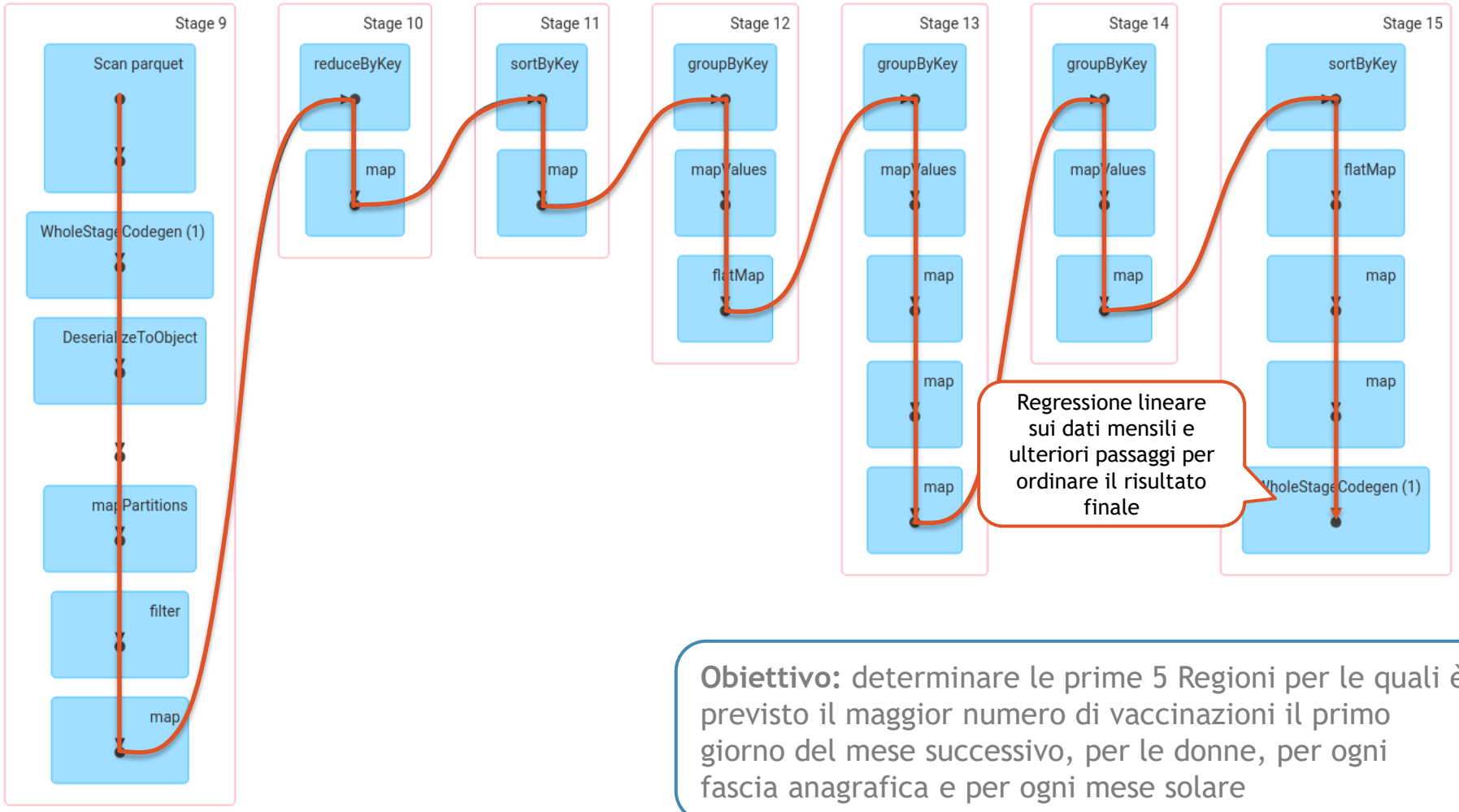
Query 2



Obiettivo: determinare le prime 5 Regioni per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo, per le donne, per ogni fascia anagrafica e per ogni mese solare

Query

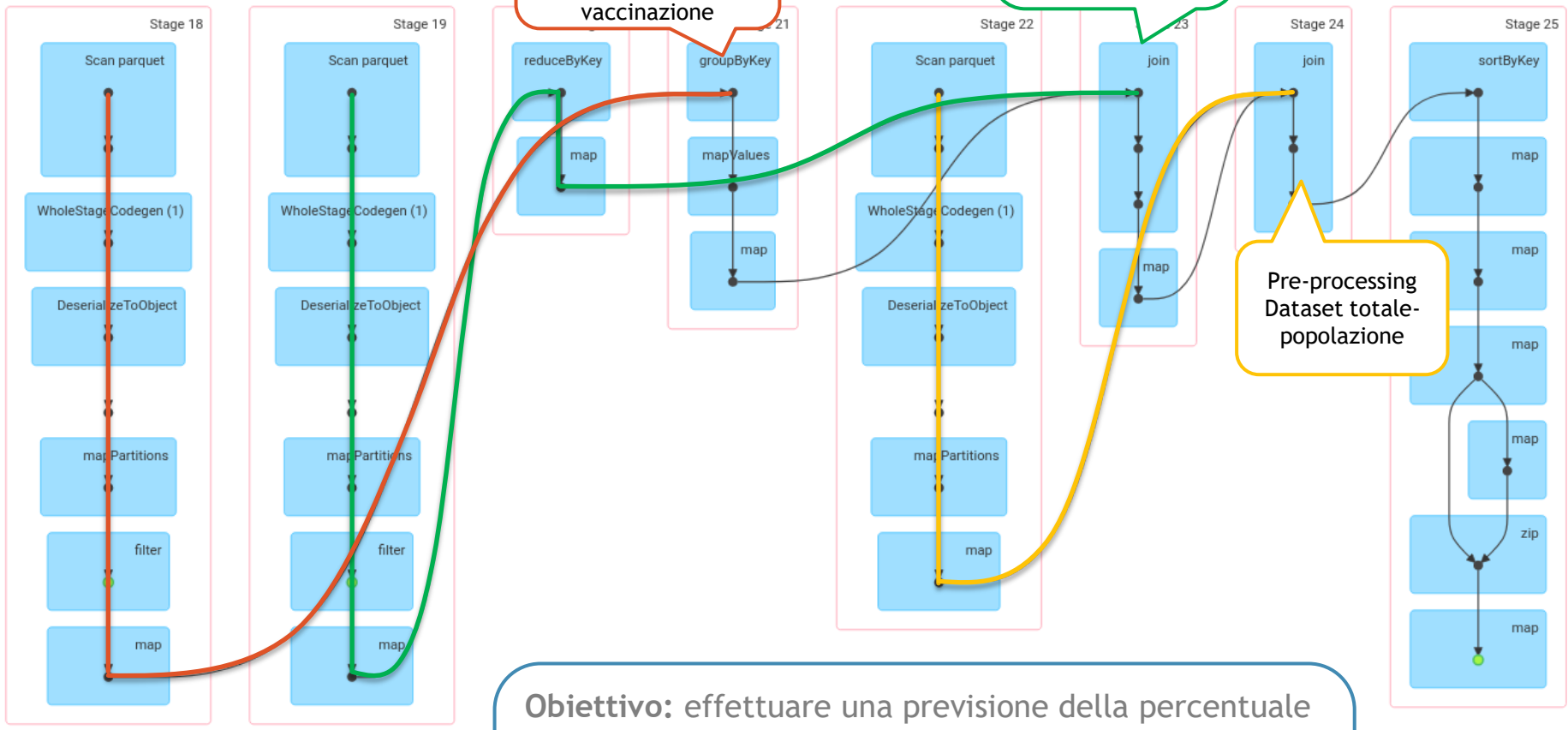
Query 2



Obiettivo: determinare le prime 5 Regioni per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo, per le donne, per ogni fascia anagrafica e per ogni mese solare

Query

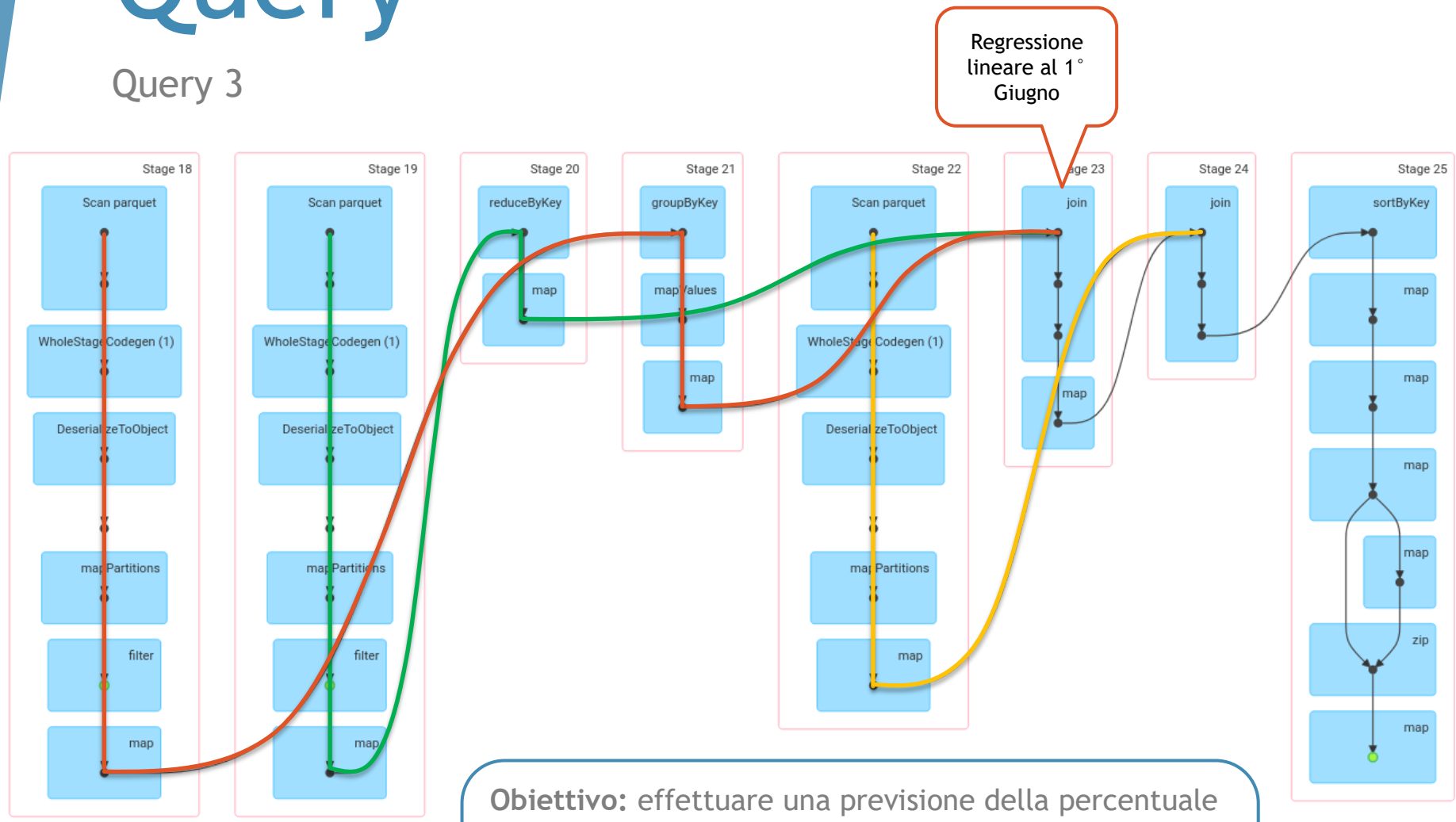
Query 3



Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Query

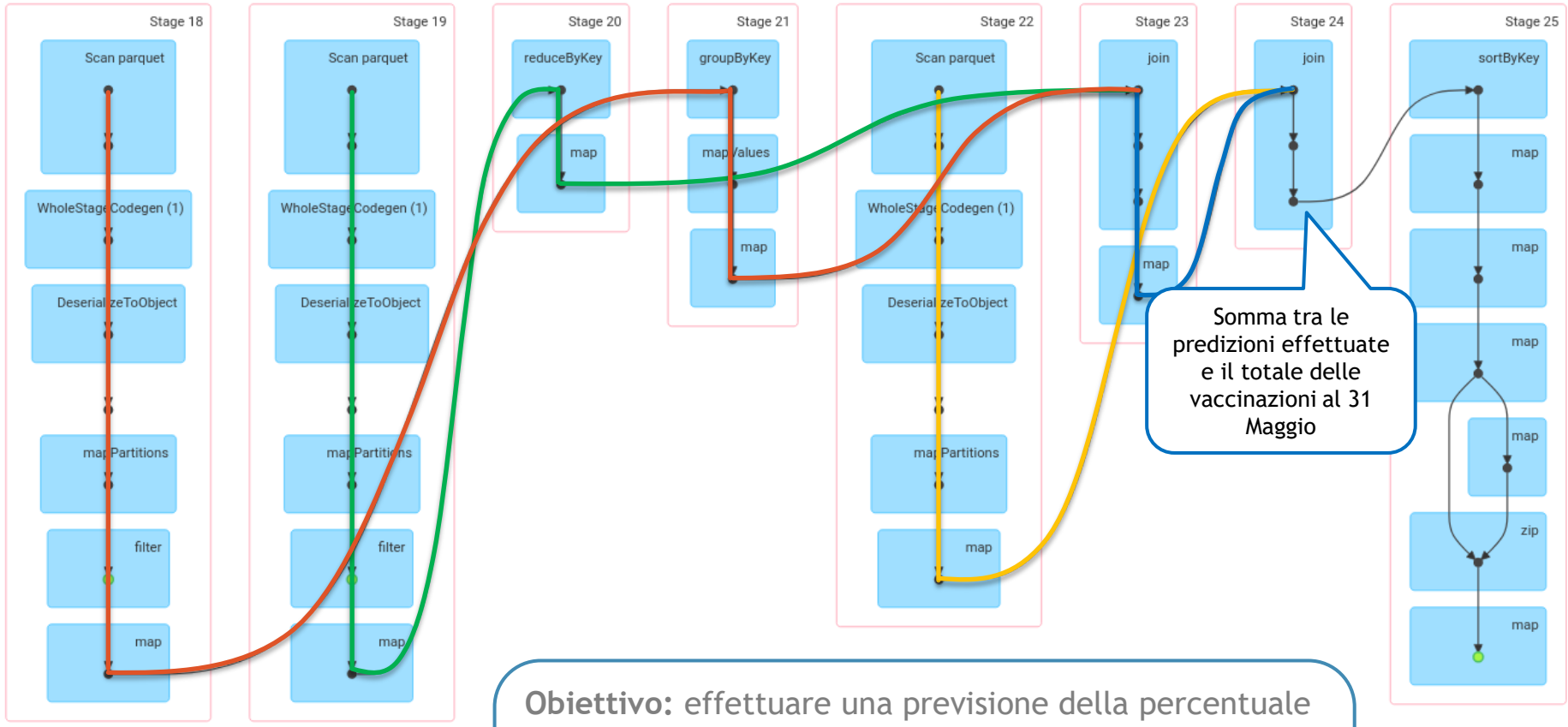
Query 3



Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Query

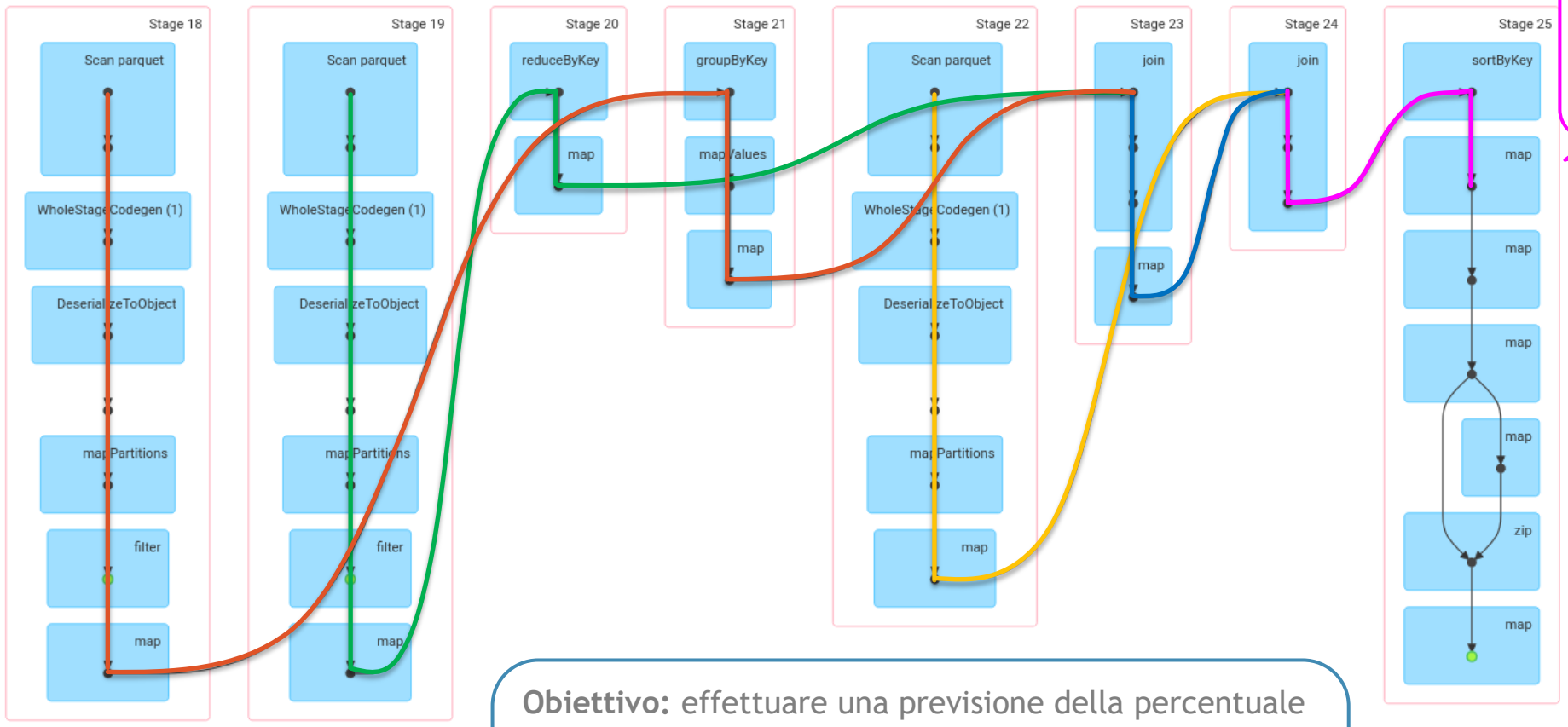
Query 3



Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Query

Query 3

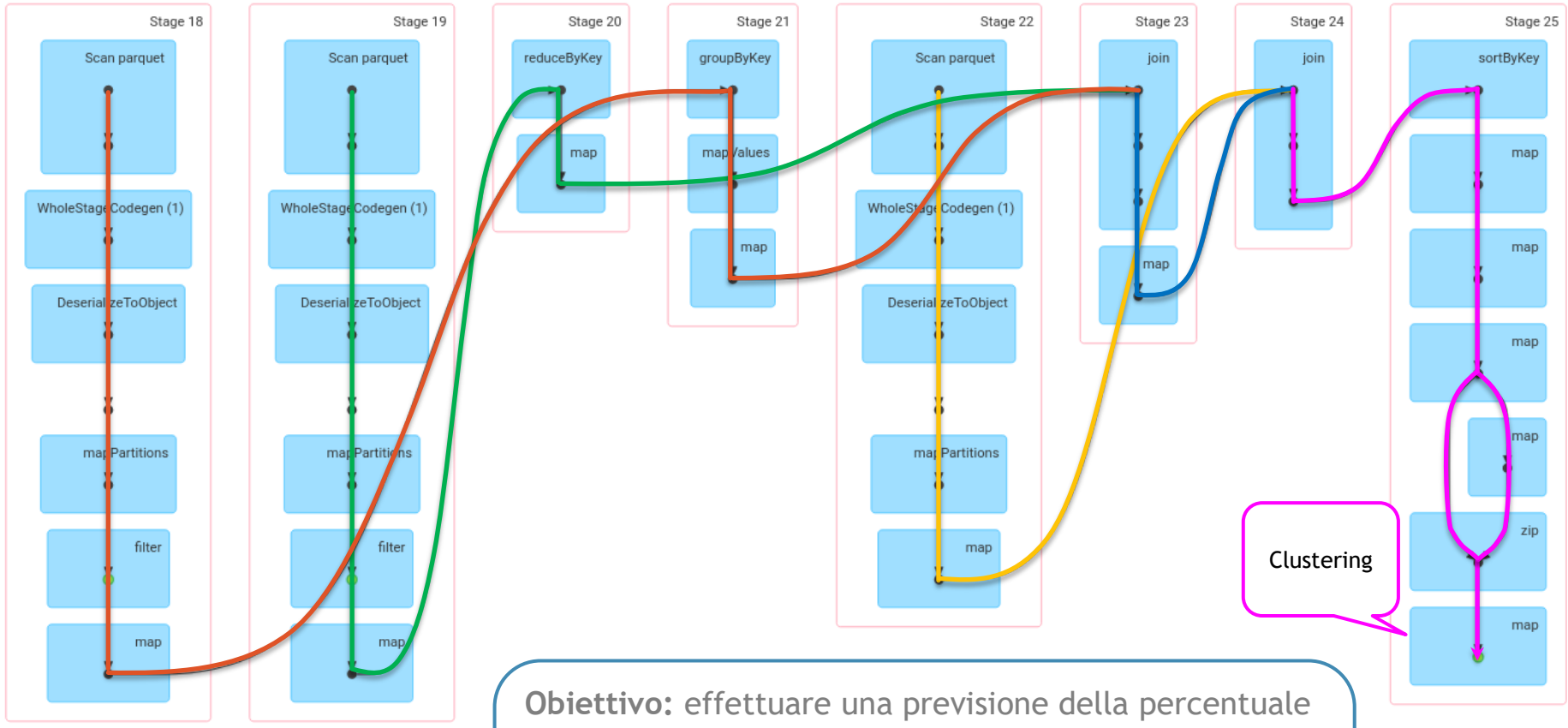


Calcolo percentuali vaccinati totali al 1° Giugno

Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Query

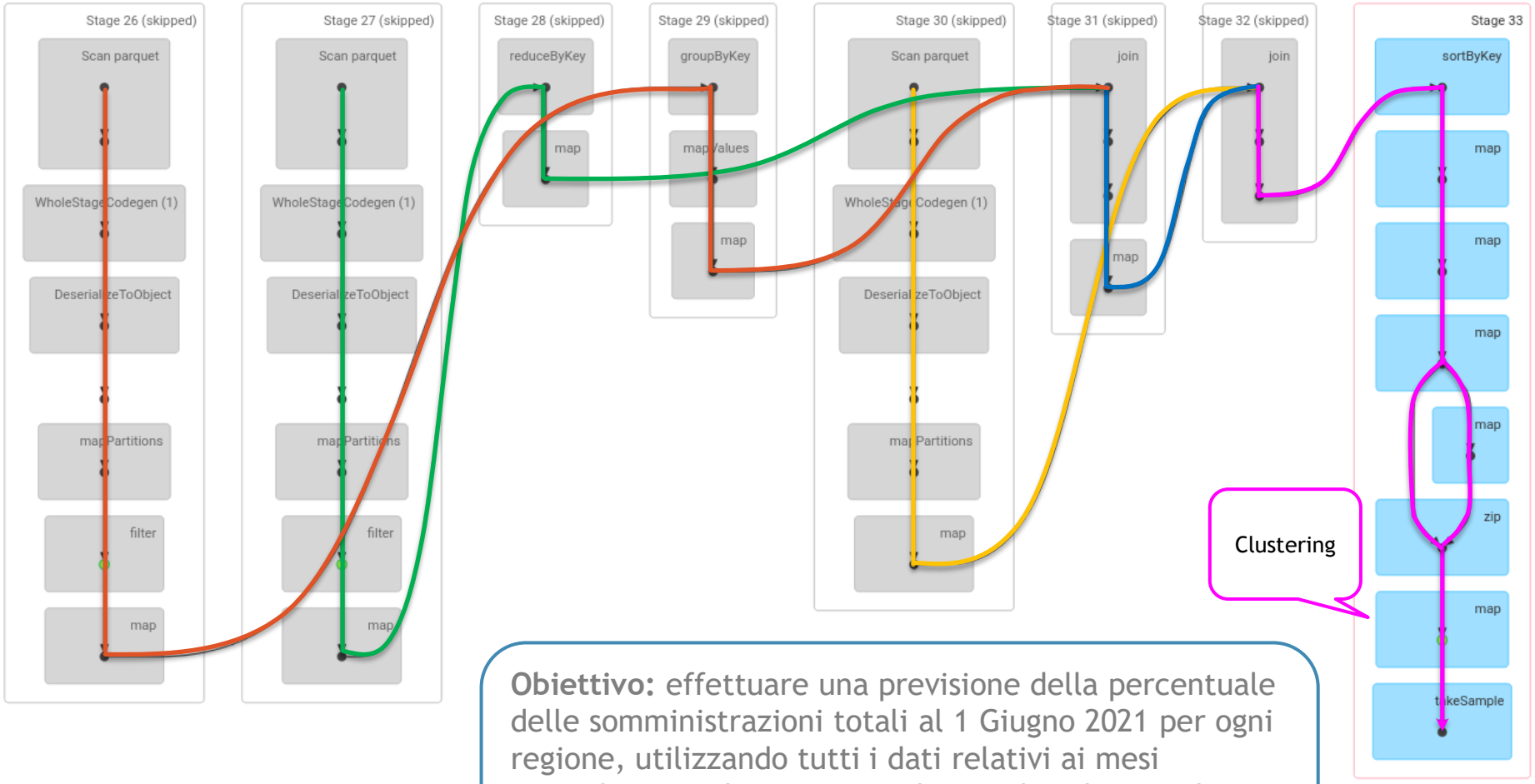
Query 3



Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Query

Query 3



Obiettivo: effettuare una previsione della percentuale delle somministrazioni totali al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti. Inoltre, viene richiesto di utilizzare due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Benchmark

Prestazioni Query e Clustering

Query	Media	Std Dev
Query1	221.1	41.1
Query2	925.2	34.2
Query3	2778.4	260.3

*i tempi sono espressi in millisecondi

Specifiche HW: Linux Xubuntu basato su Ubuntu 20.04 LTS virtualizzato tramite VBox, CPU AMD Ryzen 5 3600, 6 core, 12 thread (di cui 10 assegnati alla VM) e 16 GB di RAM(di cui 9 assegnati alla VM), con archiviazione SSD.

	Modello			
Numero Cluster	K-Means		Bisecting K-Means	
	Media	Std Dev	Media	Std Dev
2	347.3	43.6	203.2	62.4
3	136.2	21.6	127.1	24.8
4	153.1	39.5	148.2	25.5
5	148.9	51.9	144.3	42.1

*i tempi sono espressi in millisecondi

Benchmark

Qualità Clustering (WSSSE)

Numero Cluster	Modello	
	K-Means	Bisecting K-Means
2	0.005392	0.005639
3	0.003129	0.002325
4	0.001469	0.001585
5	0.000751	0.000751

*Libreria utilizzata: Apache Spark MLlib

Grazie per l’attenzione