TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Analisi di dati energetici con Apache Spark

Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Agenda

01 Obiettivo

02 Architettura

03 Implementazione

04 Risultati

05 Analisi prestazioni

Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

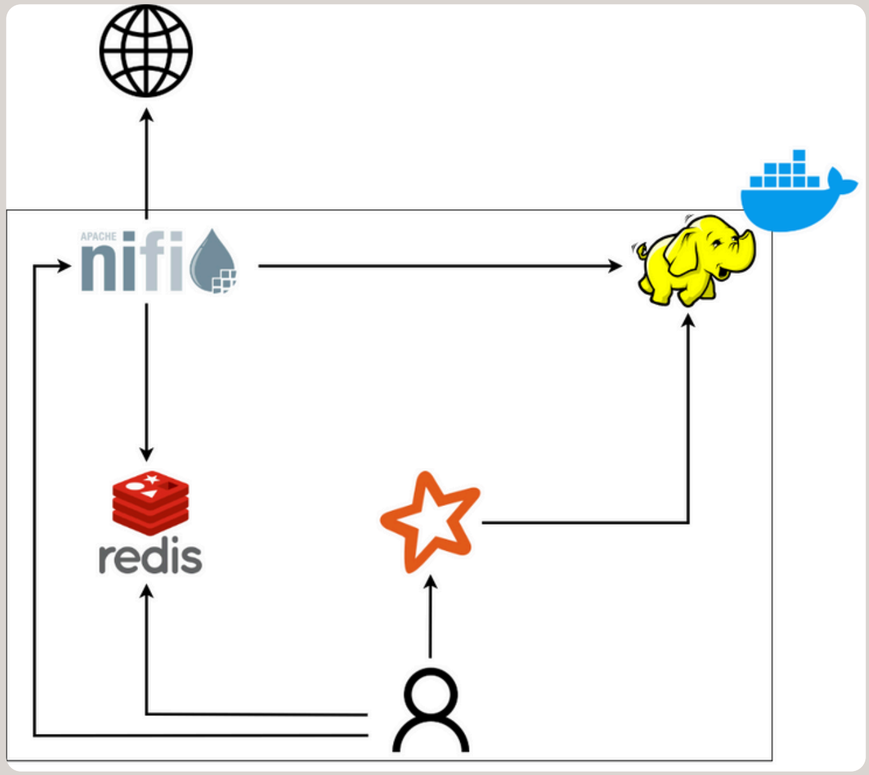
Obiettivo

Analizzare i dati storici orari forniti da Electricity Maps relativi alla Carbon Intensity e alla Carbon-Free Energy Percentage (CFE%) di Italia e Svezia, rispondendo a specifiche query con Apache Spark.

Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

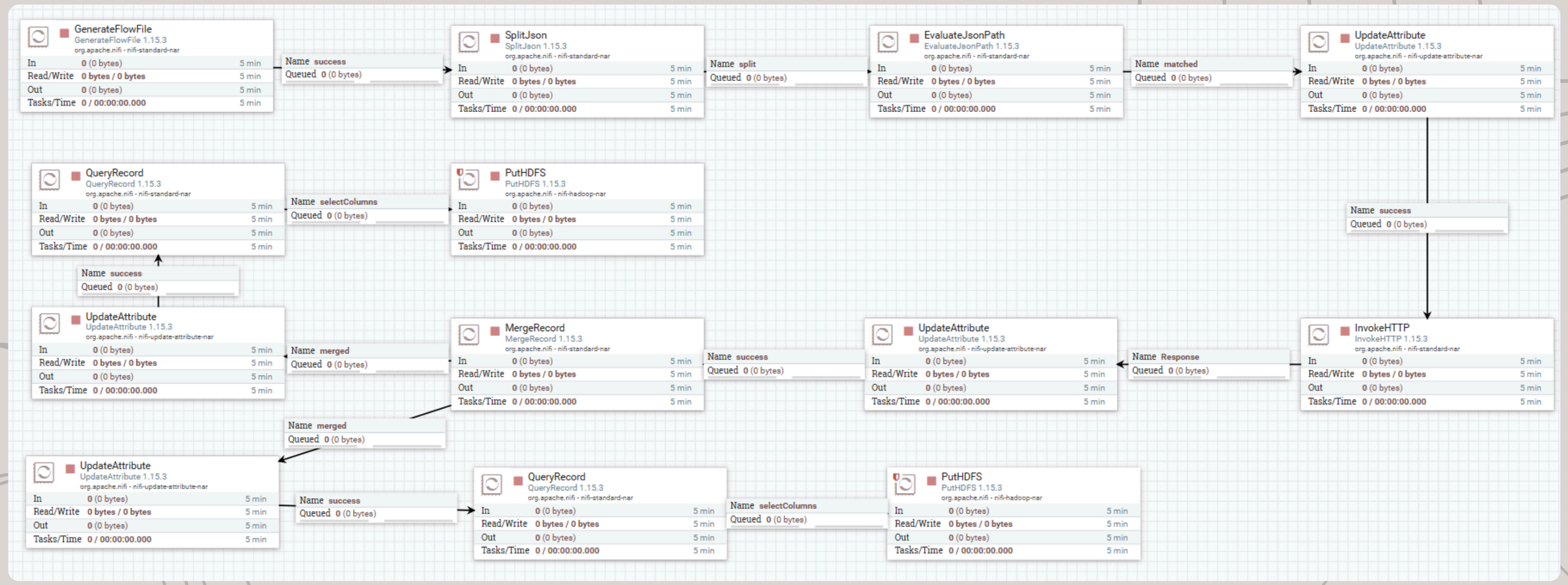
Architettura



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Data Ingestion



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Formati di Memorizzazione

CSV

- Testo semplice
- Nessuna compressione
- No metadata

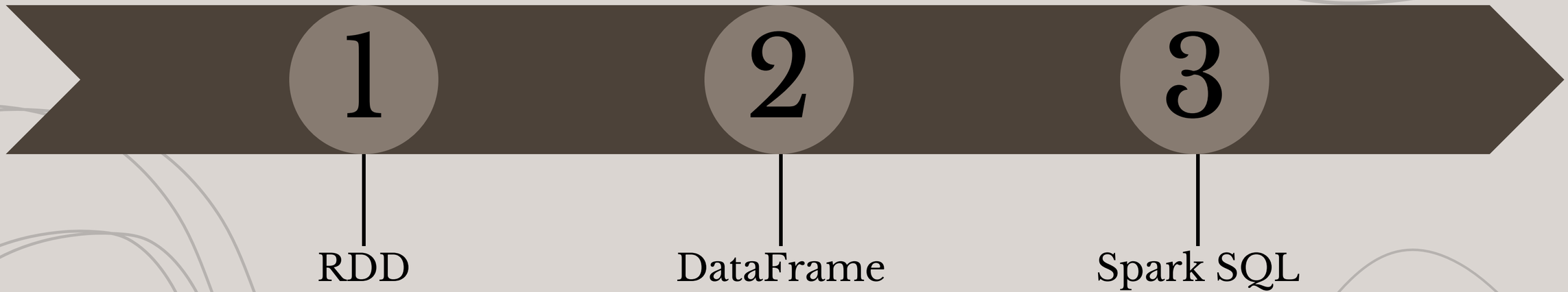
Parquet

- Colonnare, Binario
- Compressione Integrata
- Metadata

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

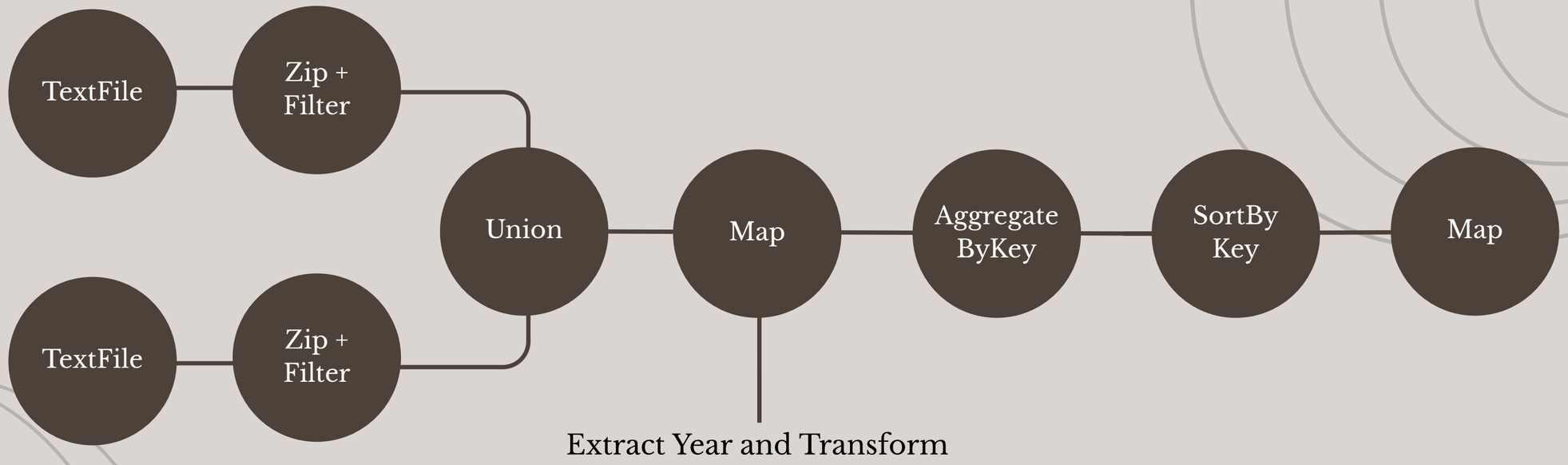
Query 1

Facendo riferimento al dataset dei valori energetici dell'**Italia** e della **Svezia**, **aggregare i dati su base annua.** Calcolare la **media, il minimo ed il massimo** di "**Carbon intensity gCO2eq/kWh (direct)**" e "**Carbon-free energy percentage (CFE%)**" per ciascun anno dal **2021** al **2024**.
Inoltre, considerando il valor medio di "Carbon intensity gCO2eq/kWh (direct)" e "Carbon-free energy percentage (CFE%)" aggregati su base annua, generare due grafici che consentano di confrontare visivamente l'andamento per Italia e Svezia.

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query 1



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

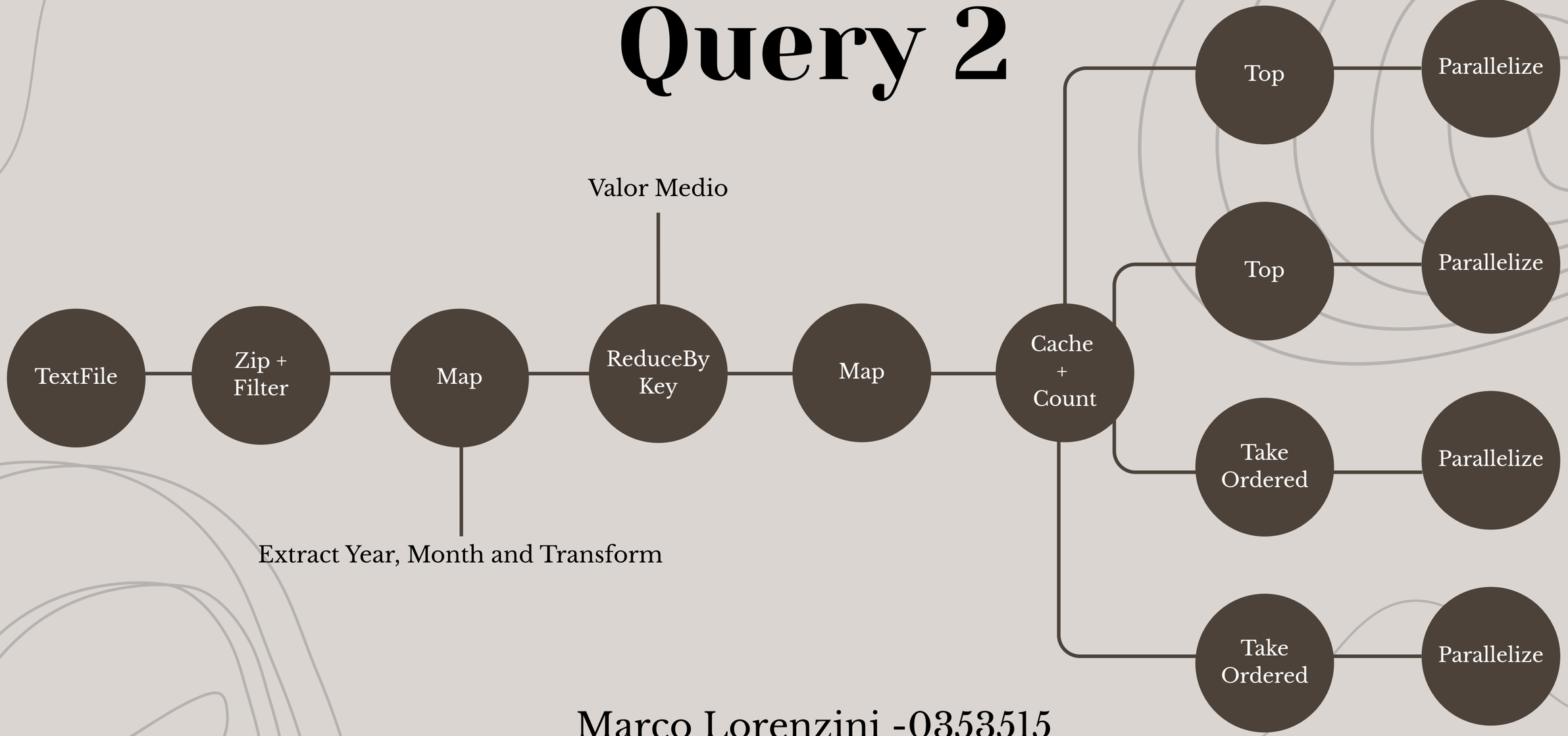
Query 2

Considerando il solo **dataset italiano**, aggregare i dati sulla coppia (**anno, mese**), calcolando il **valor**
medio di “**Carbon intensity gCO2eq/kWh (direct)**” e “**Carbon-free energy percentage (CFE%)**”. Calcolare
la **classifica delle prime 5 coppie** (anno, mese) ordinando per “**Carbon intensity gCO2eq/kWh (direct)**” decrescente, crescente e “**Carbon-free energy percentage (CFE%)**” decrescente, crescente.
In totale sono attesi 20 valori. Inoltre, considerando il valor medio di “Carbon intensity gCO2eq/kWh
(direct)” e “Carbon-free energy percentage (CFE%)” aggregati sulla coppia (anno, mese) per l’Italia,
generare due grafici che consentano di valutare visivamente l’andamento delle due metriche.

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query 2



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

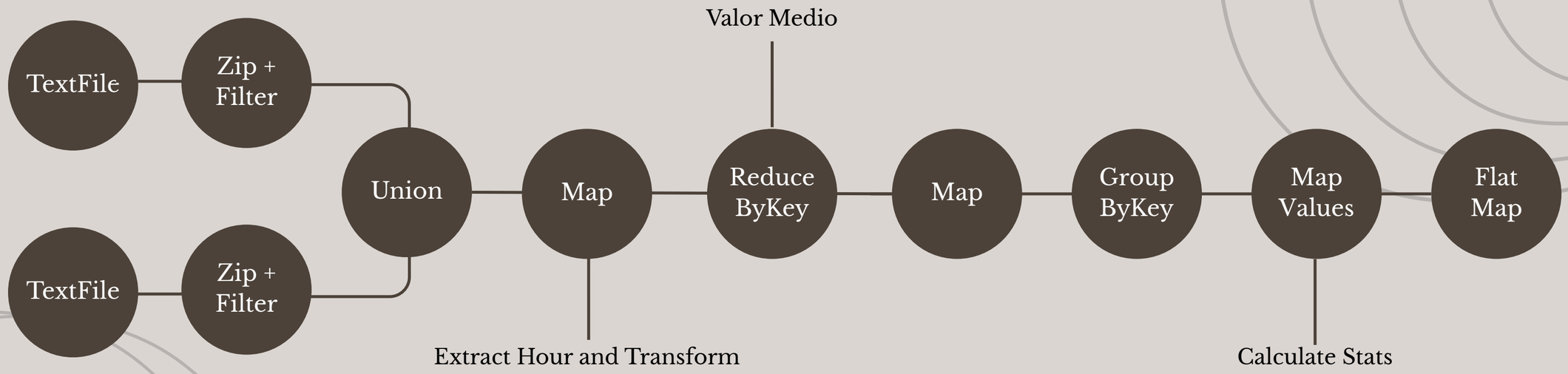
Query 3

Facendo riferimento al dataset dei valori energetici dell'**Italia** e della **Svezia**, aggregare i dati di ciascun
paese sulle **24 ore** della giornata, calcolando il **valor medio** di "**Carbon intensity gCO2eq/kWh (direct)**" e "**Carbon-free energy percentage (CFE%)**". Calcolare il **minimo, 25-esimo, 50-esimo, 75-**
esimo percentile e massimo del valor medio di "**Carbon intensity gCO2eq/kWh (direct)**" e "**Carbonfree**
energy percentage (CFE%)". Inoltre, considerando il valor medio di "Carbon intensity gCO2eq/kWh
(direct)" e "Carbon-free energy percentage (CFE%)" aggregati sulle 24 fasce orarie giornaliere, generare
due grafici che consentano di confrontare visivamente l'andamento per Italia e Svezia.

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query 3



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

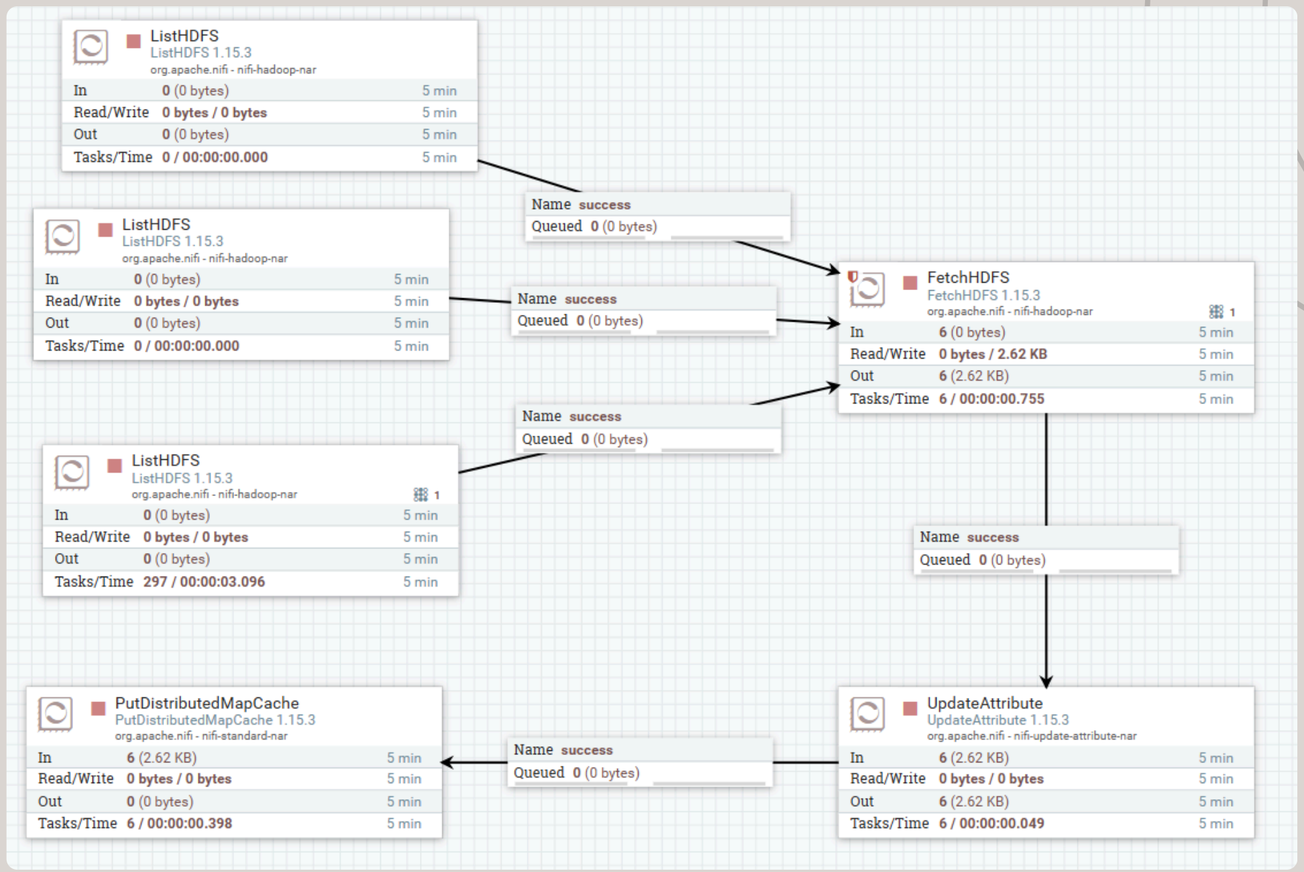
Query 3 – Calculate Stats

Input: dati_ordinati, $p \in [0, 1]$
  $n = \text{len}(\text{dati\_ordinati})$
  if $n = 0$ **then**
    return 0.0
  $\text{indice} = p \cdot (n - 1)$
  $\text{indice\_basso} = \lfloor \text{indice} \rfloor$
  $\text{indice\_alto} = \min(\text{indice\_basso} + 1,\ n - 1)$
  if $\text{indice\_basso} = \text{indice\_alto}$ **then**
    return $\text{dati\_ordinati}[\text{indice\_basso}]$
  $\text{peso} = \text{indice} - \text{indice\_basso}$
  $\text{valore\_basso} = \text{dati\_ordinati}[\text{indice\_basso}]$
  $\text{valore\_alto} = \text{dati\_ordinati}[\text{indice\_alto}]$
  return $\text{valore\_basso} \cdot (1 - \text{peso}) + \text{valore\_alto} \cdot \text{peso}$

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Data Export



Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

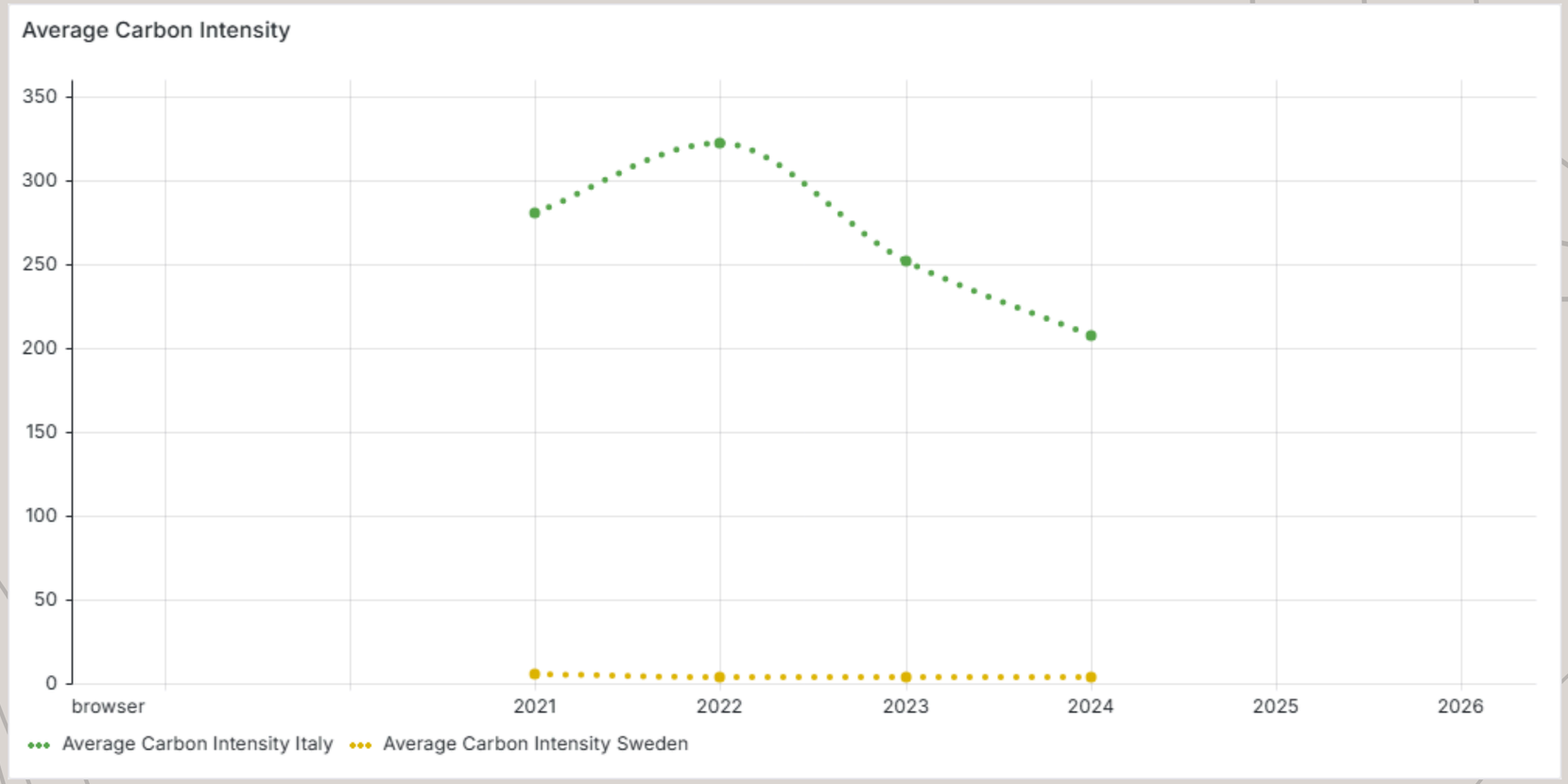
Redis

```
/data # redis-cli
127.0.0.1:6379> KEYS *
1) "query1rdd.csv"
2) "CfeBottom.csv"
3) "CfeTop.csv"
4) "CarbonDirectBottom.csv"
5) "query3rdd.csv"
6) "CarbonDirectTop.csv"
127.0.0.1:6379>
```

Marco Lorenzini -0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

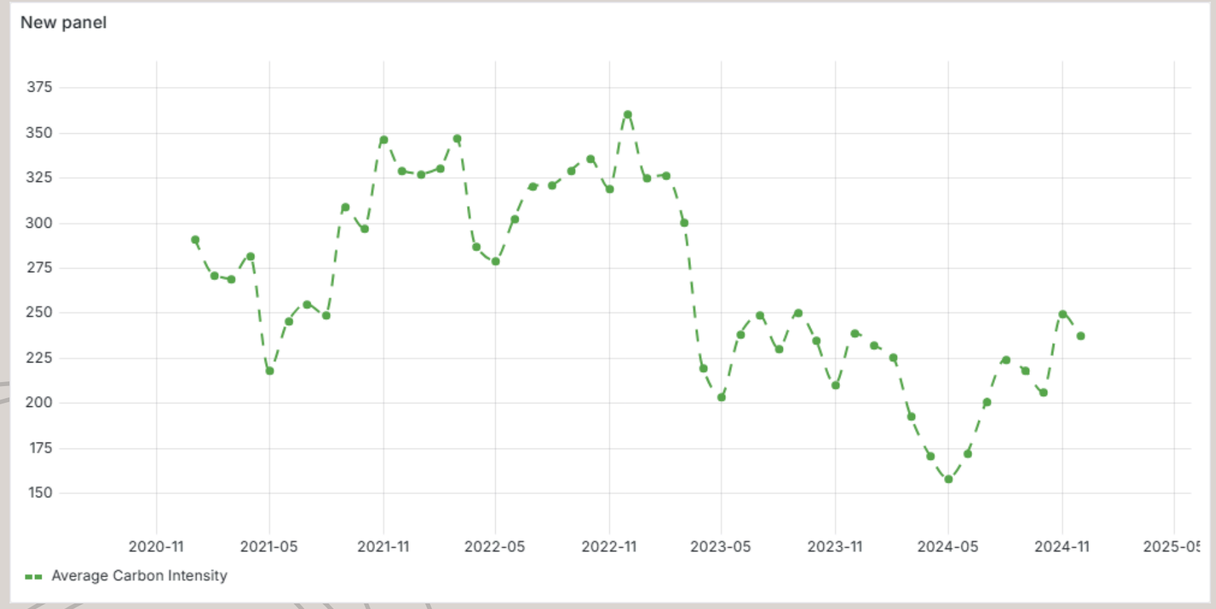
Query 1 Result

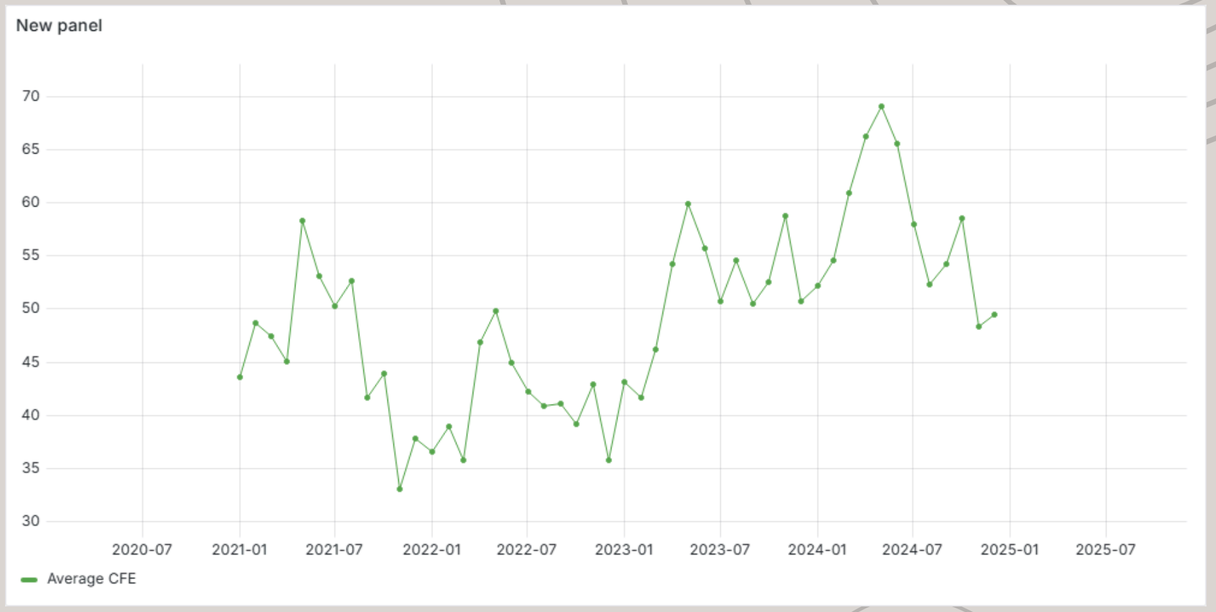


Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query 2 Result

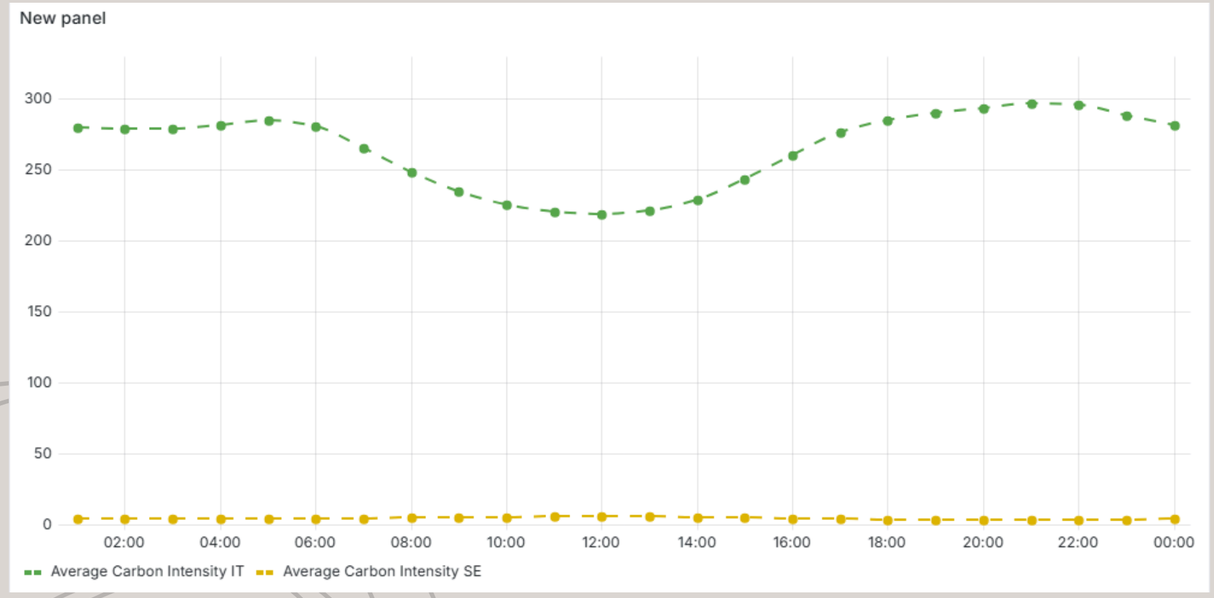


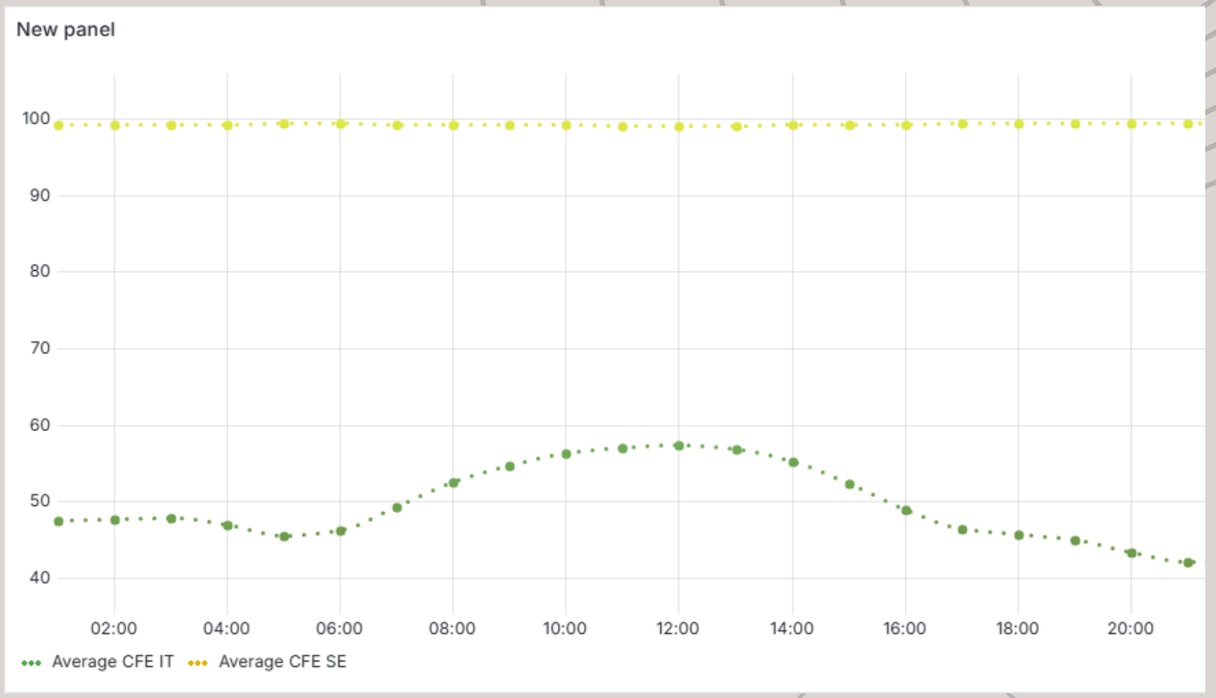


Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Query 3 Result

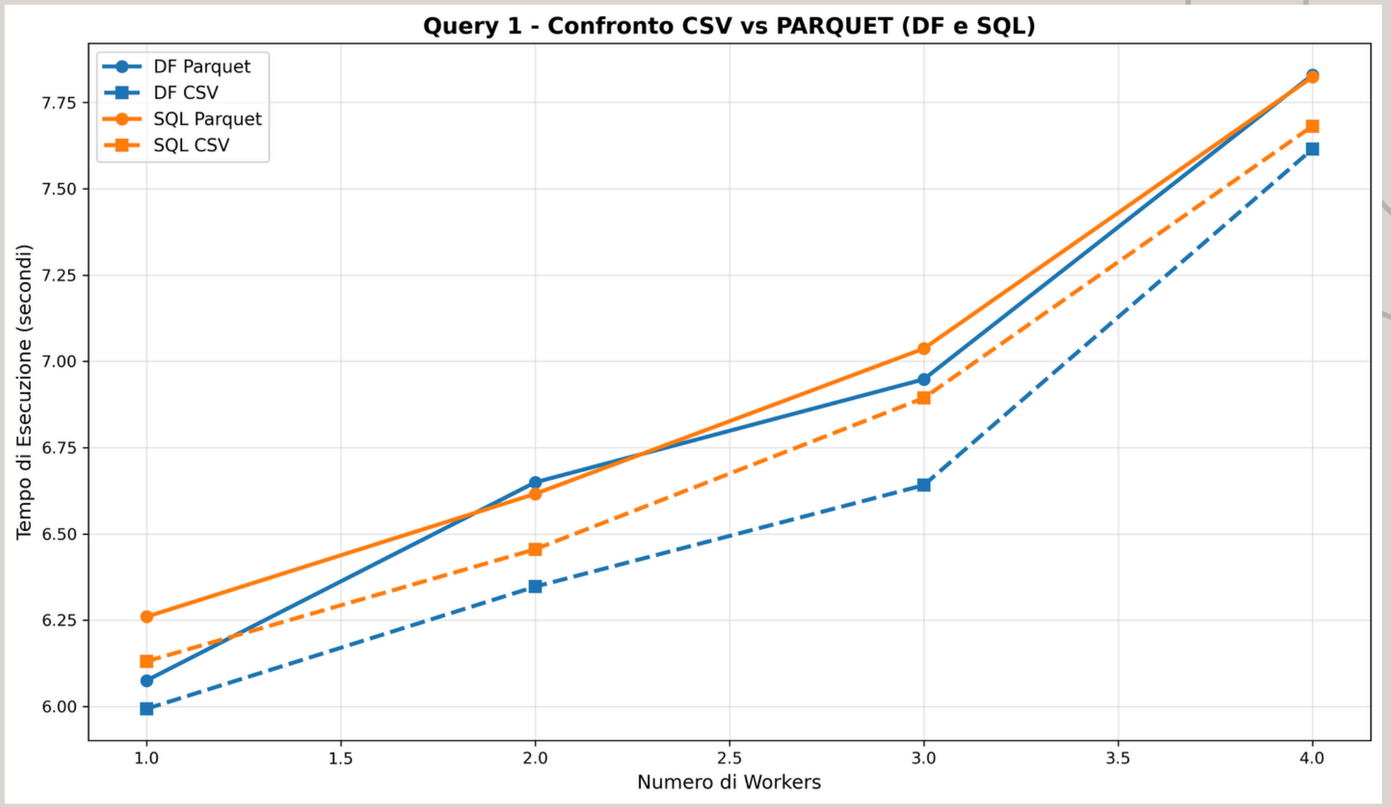




Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

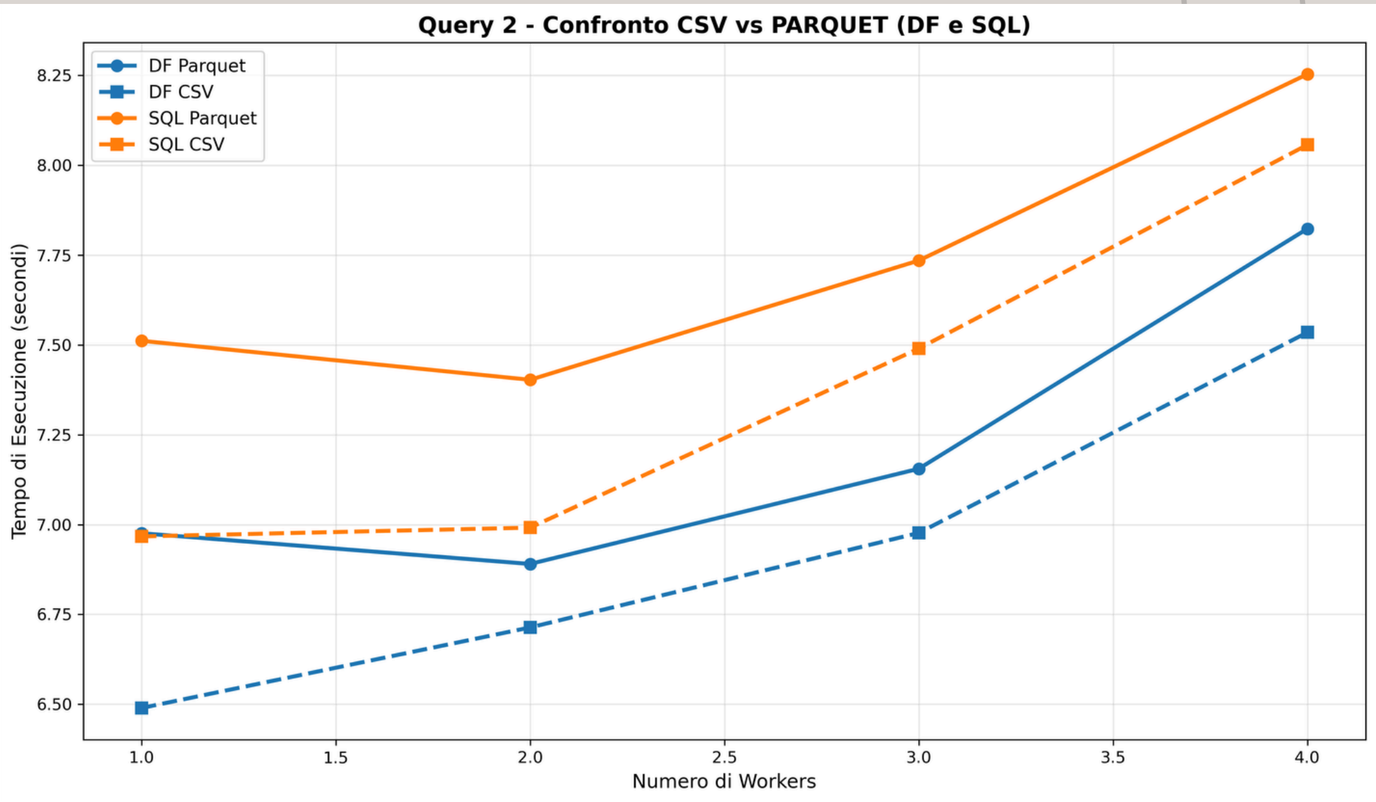
Store Format Result – Q1



Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

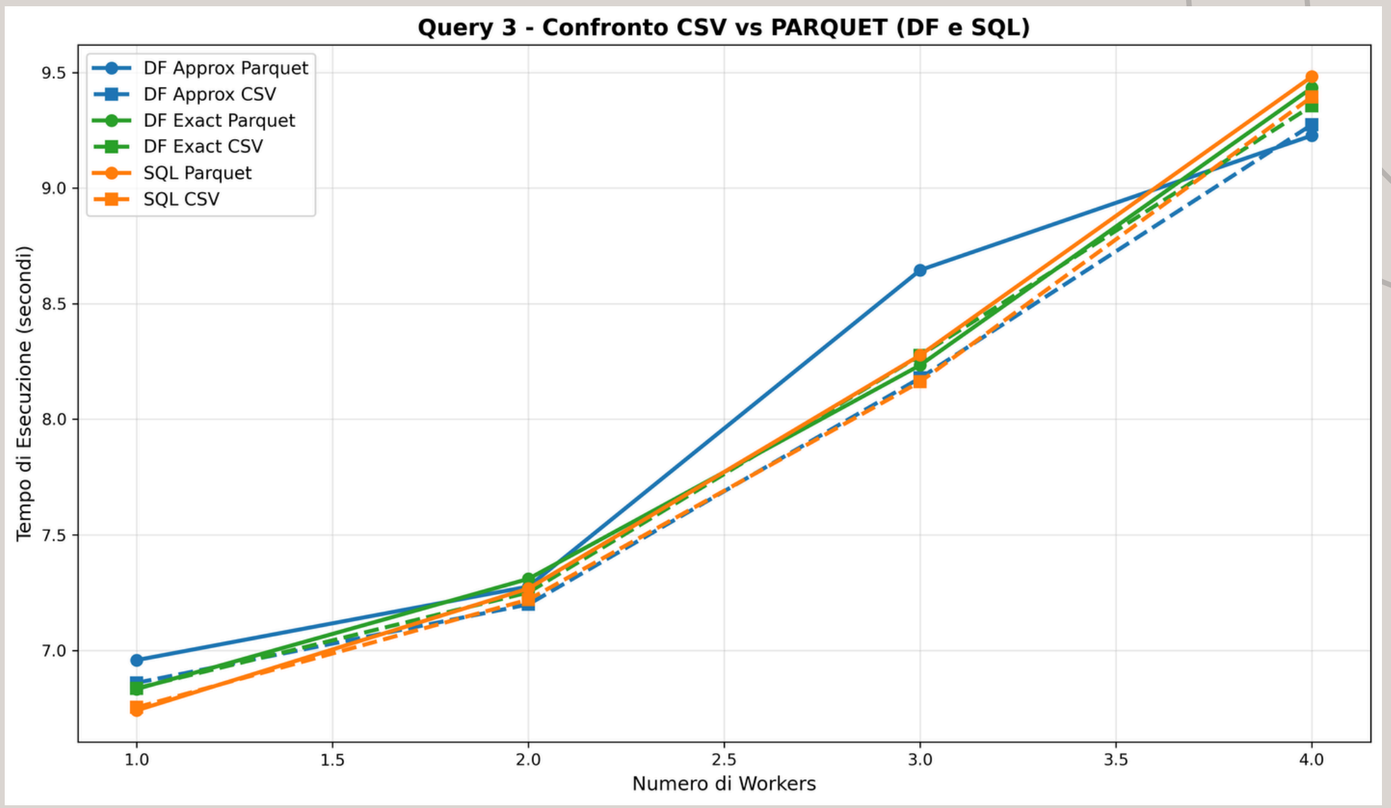
Store Format Result – Q2



Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

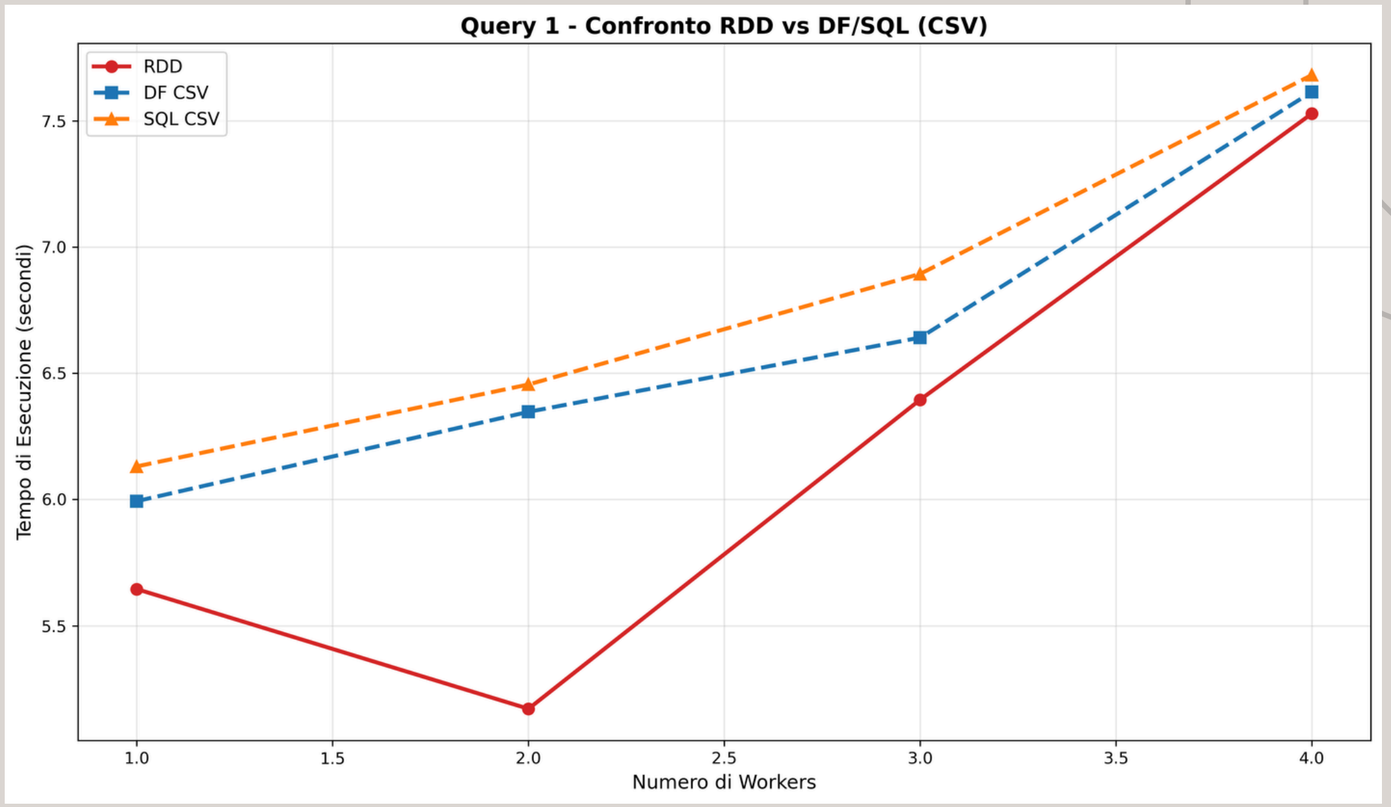
Store Format Result – Q3



Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

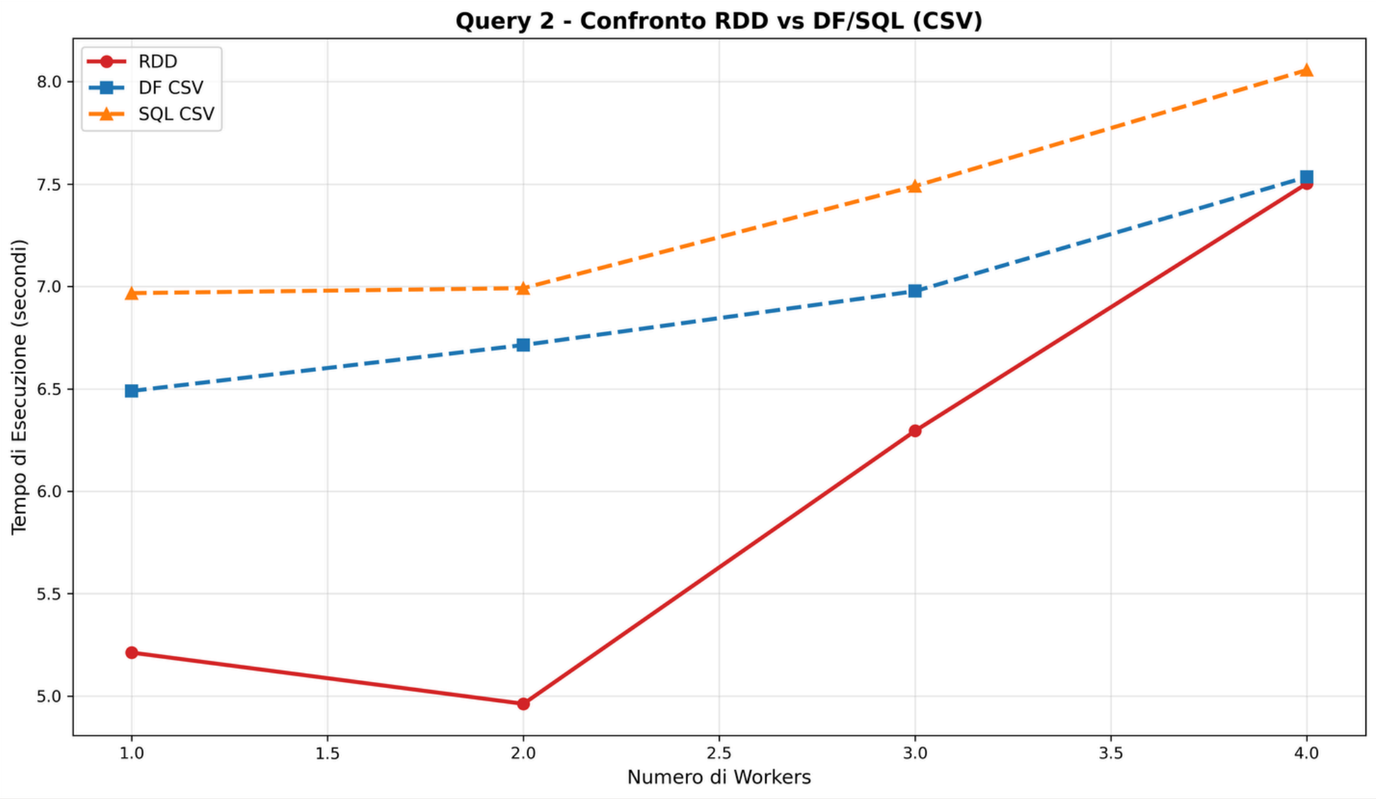
RDD vs DF vs SQL – Q1



Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

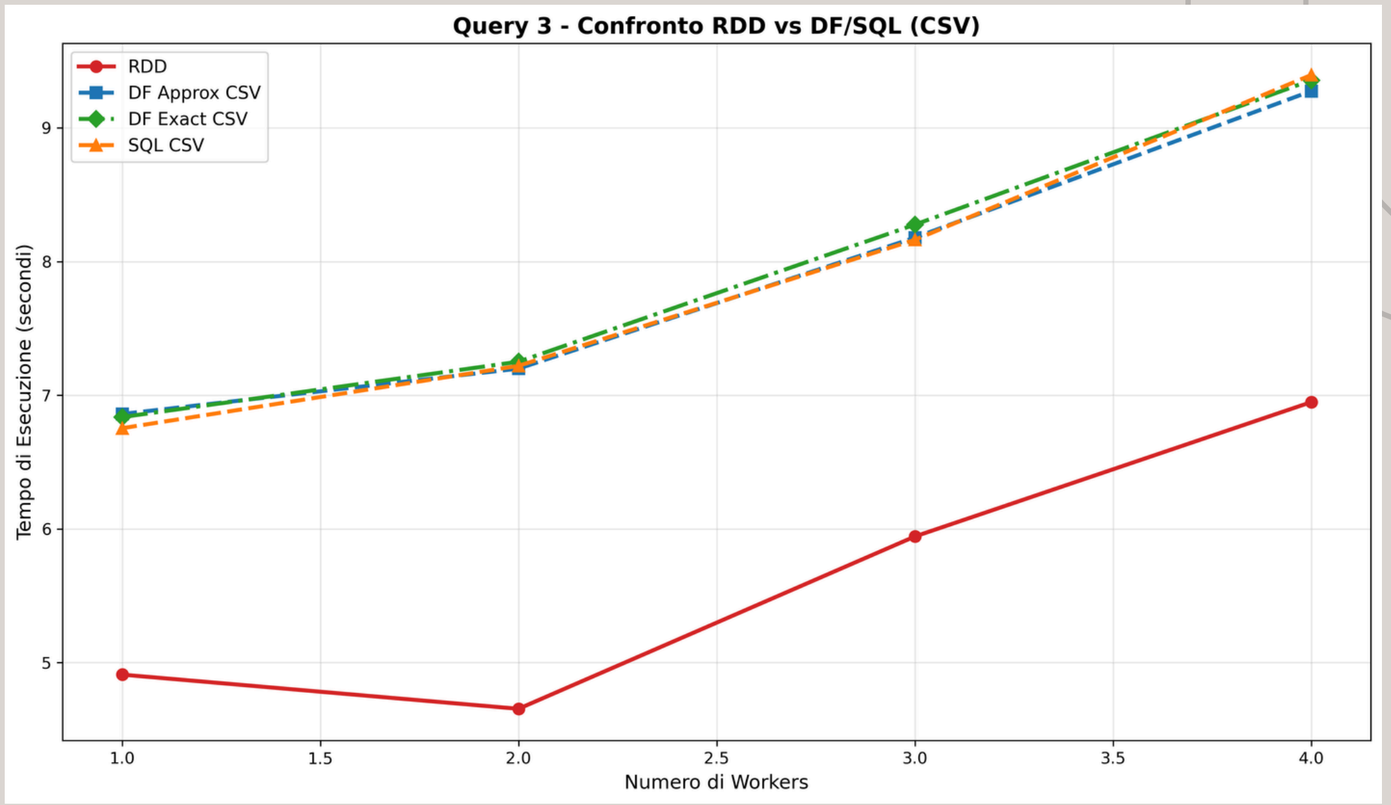
RDD vs DF vs SQL – Q2



Marco Lorenzini - 0353515

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

RDD vs DF vs SQL – Q3



Marco Lorenzini - 0353515

Thank You

https://github.com/MarcoLor01/BatchProcessingSABD/