Sistemi e Architetture per Big Data - A.A. 2024/25
Progetto 1: Analisi di dati energetici con Apache Spark

Docenti: Valeria Cardellini, Matteo Nardelli
Dipartimento di Ingegneria Civile e Ingegneria Informatica
Università degli Studi di Roma "Tor Vergata"

Requisiti del progetto

Lo scopo del progetto è usare il framework di data processing Apache Spark per rispondere ad alcune query su dati storici forniti da Electricity Maps [1] sull'elettricità e sulle emissioni di CO_2 per produrla.

Electricity Maps fornisce un dataset per ogni paese e per diverse granularità temporali (ora, giorno, mese, anno)[1]. Il dataset di ogni paese contiene 35065 eventi, che descrivono la produzione di elettricità dal 1 gennaio 2021 al 31 dicembre 2024. Ogni evento è caratterizzato dalle seguenti informazioni:

- **Intensità di carbonio**: quantità di gas serra emessi per unità di elettricità, espressa in grammi di CO_2 equivalenti per kilowattora (gCO_2eq/kWh); Sono forniti sia i fattori di emissione diretti che quelli relativi al ciclo di vita.
- **Percentuale di energia senza emissioni di carbonio**: percentuale di elettricità disponibile sulla rete da fonti a basse o nulle emissioni di CO_2. Sono esclusi i combustibili fossili e sono inclusi l'energia solare, eolica, idroelettrica, geotermica, da biomassa e nucleare.
- **Percentuale di energia rinnovabile**: percentuale di elettricità consumata da fonti rinnovabili. Sono inclusi l'energia da biomassa, geotermica, idroelettrica, solare ed eolica.

Dataset

Recuperare dal sito Electricity Maps[1] il dataset relativo all'Italia ed alla Svezia con granularità oraria, per gli anni dal 2021 al 2024. Durante la fase di data ingestion, valutare il modo più opportuno di memorizzare i dati, se tramite file separati o tramite un unico file per il periodo dal 2021 al 2024. Inoltre, per gli scopi del progetto, sono di interesse solo i dati relativi all'*intensità di carbonio diretta* e la *percentuale di energia senza emissioni di carbonio*.

Query

Considerando il dataset indicato, le query a cui rispondere sono:

Q1 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 gCO_2eq/kWh (direct)" e "Carbon-free energy percentage (CFE%)" per ciascun anno dal 2021 al 2024. Inoltre, considerando il

[1] `https://portal.electricitymaps.com/datasets`

valor medio di “Carbon intensity gCO_2eq/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.

Esempio di output:
 # date, country, carbon-mean, carbon-min, carbon-max, cfe-mean, cfe-min, cfe-max
2021, IT, 280.08, 121.24, 439.06, 46.305932, 15.41, 77.02
2022, IT, 321.617976, 121.38, 447.33, 41.244127, 13.93, 77.44
...
2021, SE, 5.946325, 1.50, 55.07, 98.962411, 92.80, 99.65
2022, SE, 3.875823, 0.54, 50.58, 99.551723, 94.16, 99.97
...

Q2 Considerando il solo dataset italiano, aggregare i dati sulla coppia (anno, mese), calcolando il valor medio di “Carbon intensity gCO_2eq/kWh (direct)” e “Carbon-free energy percentage (CFE%)”. Calcolare la classifica delle prime 5 coppie (anno, mese) ordinando per“Carbon intensity gCO_2eq/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 gCO_2eq/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.

Esempio di output:
 # date, carbon-intensity, cfe
2022_12, 360.520000, 35.838320
2022_3, 347.359073, 35.822218
2021_11, 346.728514, 33.076681
2022_10, 335.784745, 39.167164
2022_2, 330.489896, 38.980595

2024_5, 158.240887, 68.989731
2024_4, 170.670889, 66.253958
2024_6, 171.978792, 65.487792
2024_3, 192.853871, 60.919556
2024_7, 200.595995, 57.939099

2024_5, 158.240887, 68.989731
2024_4, 170.670889, 66.253958
2024_6, 171.978792, 65.487792
2024_3, 192.853871, 60.919556
2023_5, 203.494489, 59.877003

2021_11, 346.728514, 33.076681
2022_3, 347.359073, 35.822218
2022_12, 360.520000, 35.838320
2022_1, 326.947876, 36.603683
2021_12, 329.303508, 37.868817

Q3 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 gCO_2eq/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 gCO_2eq/kWh (direct)" e "Carbon-free energy percentage (CFE%)". Inoltre, considerando il valor medio di "Carbon intensity gCO_2eq/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.

Esempio di output:
 # county, data, min, 25-perc, 50-perc, 75-perc, max
IT, carbon-intensity, 219.029329, 241.060318, 279.202916, 285.008504, 296.746208
IT, cfe, 42.203176, 45.728436, 47.600110, 53.149180, 57.423648
SE, carbon-intensity, 3.150062, 3.765761, 4.293638, 4.876138, 5.947180
SE, cfe, 99.213936, 99.338007, 99.411328, 99.472495, 99.540979

Il risultato di ciascuna query deve essere consegnato in formato CSV.

Per la rappresentazione grafica dei risultati delle query, utilizzare un framework di visualizzazione (e.g., Grafana [2]).

Inoltre, si chiede di valutare sperimentalmente i tempi di processamento delle query sulla piattaforma di riferimento usata per la realizzazione del progetto e di riportare tali tempi nella relazione e nella presentazione del progetto. Tale piattaforma può essere un nodo standalone (si suggerisce di utilizzare Docker Compose), oppure è possibile utilizzare un servizio cloud per il processamento di Big Data (e.g., Amazon EMR) avvalendosi del grant a disposizione.

Infine, si chiede di realizzare la fase di data ingestion per:

- importare i dati di input in HDFS, eventualmente gestendo la conversione del formato dei dati, usando un framework di data ingestion a scelta (e.g., Apache NiFi, Apache Kafka, Apache Pulsar);
- esportare i risultati di output da HDFS ad un sistema di storage a scelta (e.g., HBase, Redis).

Composizione dei gruppi

Il progetto è dimensionato per un gruppo composto da **2 studenti**.

Per gruppi composti da 1 studente: si richiede di rispondere alle query 1 e 2; inoltre, la gestione del data ingestion è opzionale, ma il dataset deve comunque essere scaricato dal sito, letto da HDFS ed i risultati di output scritti su HDFS.

Per gruppi composti da 3 studenti: in aggiunta ai requisiti sopra elencati, eseguire un'analisi di clustering sui dati relativi al "Carbon intensity gCO_2eq/kWh (direct)", aggregati su base annua e per l'anno 2024. I dati fanno riferimenti ai valori medi annui per ciascun paese. L'obiettivo è individuare insieme di nazioni con comportamenti simili in termini di emissioni di carbonio, usando l'algoritmo di clustering *k-means*. Oltre ad applicare l'algoritmo di clustering sui dati, determinare un valore ottimale di *k* (numero di cluster) mediante l'uso di metriche appropriate, come ad esempio *elbow method*[2] o l'*indice di silhouette*[3]. Si considerino i seguenti 15 paesi europei: Austria, Belgio, Francia, Finlandia, Germania, Gran Bretagna, Irlanda, Italia, Norvegia, Polonia, Repubblica Ceca, Slovenia, Spagna, Svezia e Svizzera. Si scelgano inoltre 15 paesi extra-europei, tra cui Stati Uniti, Emirati Arabi, Cina, India, per un totale di 30 paesi a livello mondiale.

[2] `https://en.wikipedia.org/wiki/Elbow_method_(clustering)`
[3] `https://en.wikipedia.org/wiki/Silhouette_(clustering)`

Generare un grafico che consenta di visualizzare il risultato del clustering.

Opzionale: Si richiede di utilizzare Spark SQL per rispondere alle tre query utilizzando SQL. Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query ottenuti con Spark SQL e di confrontarli con quelli ottenuti usando il solo framework Spark, riportando l'analisi del confronto nella relazione e nella presentazione.

Svolgimento e consegna del progetto

Comunicare la composizione del gruppo ai docenti entro **venerdì 23 maggio 2025**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2024/25 ed il codice deve essere consegnato **entro lunedì 9 giugno 2025**.

La consegna del progetto consiste in:

1. link a spazio di Cloud storage o repository contenente il codice del progetto da comunicare via email ai docenti **entro lunedì 9 giugno 2025**; inserire i risultati delle query in formato CSV in una cartella denominata `Results`.
2. relazione di lunghezza compresa tra le 4 e le 8 pagine, da inserire all'interno della cartella denominata `Report`; per la relazione si consiglia di usare il formato ACM proceedings (https://www.acm.org) oppure il formato IEEE proceedings (https://www.ieee.org); includere nella relazione lo schema dell'architettura di sistema utilizzata.
3. slide della presentazione orale, da inviare via email ai docenti **dopo** lo svolgimento della presentazione.

La presentazione si terrà **lunedì 16 giugno 2025** (da confermare); ciascun gruppo avrà a disposizione **massimo 15 minuti** per presentare la propria soluzione (20 minuti per gruppi composti da 3 persone).

Valutazione del progetto

I principali criteri di valutazione del progetto saranno:

1. rispondenza ai requisiti;
2. originalità;
3. architettura del sistema e deployment;
4. organizzazione del codice;
5. efficienza;
6. organizzazione, chiarezza e rispetto dei tempi della presentazione orale.

Riferimenti bibliografici

[1] Electricity Maps. Carbon Intensity Data. `https://www.electricitymaps.com/`, 2025.

[2] Grafana. `https://grafana.com/grafana/`.