Sistemi e Architetture per Big Data - A.A. 2020/21
Progetto 1: Analisi del dataset delle vaccinazioni anti Covid-19 con Hadoop/Spark

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

Requisiti del progetto

Lo scopo del progetto è rispondere ad alcune query riguardanti le vaccinazioni anti Covid-19 in Italia, utilizzando il framework di data processing Apache Hadoop oppure Apache Spark.

Per gli scopi di questo progetto viene utilizzato il dataset disponibile all'URL https://github.com/italia/covid19-opendata-vaccini/tree/master/dati e fornito dal Commissario straordinario per l'emergenza Covid-19, Presidenza del Consiglio dei Ministri. Il dataset viene aggiornato ogni giorno e sono possibili correzioni dei dati da parte del proprietario in caso di inaccuratezze. In particolare, per il progetto si utilizzano i seguenti file in formato CSV (Comma-separated values):

- punti-somministrazione-tipologia.csv,
- somministrazioni-vaccini-latest.csv,
- somministrazioni-vaccini-summary-latest.csv.

Nello specifico, il file `punti-somministrazione-tipologia.csv` contiene dati sui punti di somministrazione per ciascuna Regione e Provincia Autonoma. La Tabella 1 descrive i campi in ogni riga.

Tabella 1: punti-somministrazione-tipologia.csv

Campo	Tipo di dati	Descrizione
area	string	Regione
denominazione_struttura	string	Presidio di somministrazione
tipologia	string	Tipologia del presidio di somministrazione: ospedaliero o territoriale
codice_NUTS1	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 1
codice_NUTS2	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 2
codice_regione_ISTAT	integer	Codice ISTAT della Regione
nome_regione	string	Denominazione standard dell'area (dove necessario denominazione bilingue)

Il file `somministrazioni-vaccini-latest.csv` contiene dati sulle somministrazioni giornaliere dei vaccini suddivisi per regioni, fasce d'età e categorie di appartenenza dei soggetti vaccinati. La Tabella 2 descrive i campi in ogni riga; il numero di somministrazioni effettuate è indicato per giorno e area.

Tabella 2: somministrazioni-vaccini-latest.csv

Campo	Tipo di dati	Descrizione
data_somministrazione	datetime	Data di somministrazione
fornitore	string	Nome completo del fornitore del vaccino
area	string	Sigla della regione di consegna
fascia_anagrafica	string	Fascia anagrafica a cui appartengono i soggetti a cui è stato somministrato il vaccino
sesso_maschile	integer	Totale dei soggetti di sesso maschile a cui è stato somministrato il vaccino
sesso_femminile	integer	Totale dei soggetti di sesso femminile a cui è stato somministrato il vaccino
categoria_operatori_sanitari_sociosanitari	integer	Numero di somministrazioni effettuate agli operatori sanitari e sociosanitari
categoria_personale_non_sanitario	integer	Numero di somministrazioni effettuate al personale non sanitario impiegato in strutture sanitarie e in attività lavorativa a rischio
categoria_ospiti_rsa	integer	Numero di somministrazioni effettuate ai soggetti ospiti di comunità residenziali indicate per giorno, regione e fascia d'età
categoria_60_69	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica compresa tra 60 e 69 anni, non appartenenti ad altre categorie prioritarie
categoria_70_79	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica compresa tra 70 e 79 anni, non appartenenti ad altre categorie prioritarie
categoria_over80	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica maggiore o uguale a 80 anni, non appartenenti ad altre categorie prioritarie
categoria_forze_armate	integer	Numero di somministrazioni effettuate al personale del comparto difesa e sicurezza
categoria_personale_scolastico	integer	Numero di somministrazioni effettuate al personale scolastico
categoria_soggetti_fragili	integer	Numero di somministrazioni effettuate ai soggetti fragili e loro caregiver
categoria_altro	integer	Numero di somministrazioni effettuate ai soggetti non riconducibili alle precedenti categorie
prima_dose	integer	Numero prime somministrazioni
seconda_dose	integer	Numero seconde somministrazioni
codice_NUTS1	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 1
codice_NUTS2	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 2
codice_regione_ISTAT	integer	Codice ISTAT della Regione
nome_area	string	Denominazione standard dell'area (dove necessario denominazione bilingue)

Il file `somministrazioni-vaccini-summary-latest.csv` contiene dati sul totale delle somministrazioni giornaliere per regioni e categorie di appartenenza dei soggetti vaccinati. La Tabella 3 descrive i campi in ogni riga; il numero di somministrazione effettuate è indicato per giorno e area.

Il progetto è dimensionato per un gruppo composto da **2 studenti**; per gruppi composti da 1 oppure 3 studenti, si vedano le indicazioni specifiche.

Considerando i dati fino al 31 Maggio 2021, le query a cui rispondere sono:

Q1 Utilizzando `somministrazioni-vaccini-summary-latest.csv` e `punti-somministrazione-tipologia.csv`, per ogni mese solare e per ciascuna area, calcolare il numero medio di somministrazioni che è stato effettuato giornalmente in un centro vaccinale generico in quell'area e durante quel mese. Considerare i dati a partire dall'1 Gennaio 2021.
Poiché il file `somministrazioni-vaccini-summary-latest.csv` non è ordinato in base alla data di somministrazione, per i gruppi composti da 2 o 3 persone si richiede di ordinare il file all'inizio del processamento.

Esempio di output:

Tabella 3: somministrazioni-vaccini-summary-latest.csv

Campo	Tipo di dati	Descrizione
data_somministrazione	datetime	Data di somministrazione
area	string	Sigla della regione di consegna
totale	integer	Numero totale di dosi di vaccino somministrate
sesso_maschile	integer	Totale dei soggetti di sesso maschile a cui è stato somministrato il vaccino
sesso_femminile	integer	Totale dei soggetti di sesso femminile a cui è stato somministrato il vaccino
categoria_operatori_sanitari_sociosanitari	integer	Numero di somministrazioni effettuate agli operatori sanitari e sociosanitari
categoria_personale_non_sanitario	integer	Numero di somministrazioni effettuate al personale non sanitario impiegato in strutture sanitarie e in attività lavorativa a rischio
categoria_ospiti_rsa	integer	Numero di somministrazioni effettuate ai soggetti ospiti di comunità residenziali
categoria_personale_scolastico	integer	Numero di somministrazioni effettuate al personale scolastico
categoria_60_69	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica compresa tra 60 e 69 anni, non appartenenti ad altre categorie prioritarie
categoria_70_79	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica compresa tra 70 e 79 anni, non appartenenti ad altre categorie prioritarie
categoria_over80	integer	Numero somministrazioni effettuate ai soggetti con età anagrafica maggiore o uguale a 80 anni, non appartenenti ad altre categorie prioritarie
categoria_soggetti_fragili	integer	Numero di somministrazioni effettuate ai soggetti fragili e loro caregiver
categoria_forze_armate	integer	Numero di somministrazioni effettuate al personale del comparto difesa e sicurezza
categoria_altro	integer	Numero di somministrazioni effettuate ai soggetti non riconducibili alle precedenti categorie
prima_dose	integer	Numero prime somministrazioni
seconda_dose	integer	Numero seconde somministrazioni
codice_NUTS1	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 1
codice_NUTS2	string	Classificazione europea delle unità territoriali NUTS: livello NUTS 2
codice_regione_ISTAT	integer	Codice ISTAT della Regione
nome_area	string	Denominazione standard dell'area (dove necessario denominazione bilingue)

Gennaio, Abruzzo, num. medio di vaccini per centro
...
Gennaio, Veneto, num. medio di vaccini per centro
Febbraio, Abruzzo, num. medio di vaccini per centro
...

Q2 Utilizzando `somministrazioni-vaccini-latest.csv`, per le donne, per ogni categoria e per ogni mese solare, determinare le prime 5 aree per le quali è previsto il maggior numero di vaccinazioni il primo giorno del mese successivo. Per determinare la classifica mensile e prevedere il numero di vaccinazioni, considerare la retta di regressione che approssima l'andamento delle vaccinazioni giornaliere. Per la risoluzione della query, considerare le sole categorie per cui nel mese solare in esame vengono registrati almeno due giorni di campagna vaccinale. Viene inoltre richiesto di calcolare la classifica per ogni mese e categoria a partire dai dati raccolti dall'1 Febbraio 2021.

Esempio di output:
1 Marzo, 20-29, Lazio, 40 (assumendo che 40 sia il valore predetto per l'1 Marzo usando i dati del

mese di Febbraio)
... altre 4 regioni top nella fascia 20-29
1 Marzo, 30-39, Lombardia, 120
... altre 4 regioni top nella fascia 30-39
...

Q3 Utilizzando `somministrazioni-vaccini-summary-latest.csv`, stimare il numero totale di somministrazioni effettuate al 1 giugno 2021 a partire dal 27 dicembre 2020 considerando tutte le categorie e, usando un algoritmo di clustering, classificare le aree in K cluster considerando per ogni area la stima della percentuale di popolazione vaccinata. Il totale della popolazione stimata nel 2021 per ogni regione è disponibile all'URL
http://www.ce.uniroma2.it/courses/sabd2021/projects/totale-popolazione.csv.
Come algoritmi di clustering nel caso di Spark si considerino K-means e Bisecting K-means, implementati nella libreria Spark MLlib [3]; nel caso di Hadoop MapReduce si considerino K-means e Fuzzy K-means, implementati in Apache Mahout [2]. Confrontare la qualità del clustering e le prestazioni misurate in termini di tempo di processamento per i due algoritmi di clustering al variare di K da 2 a 5.

Si chiede di consegnare anche il risultato prodotto da ciascuna query in formato CSV, specificando la data del dataset utilizzato.

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, 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 trasformando la rappresentazione dei dati in un altro formato (e.g., Avro, Parquet, ...), usando un framework di data ingestion a scelta (e.g., Apache Kafka, Apache Pulsar, Apache Flume, Apache NiFi, ...);
- esportare i dati di output da HDFS ad un sistema di storage a scelta (e.g., HBase, Redis, ...).

Per gruppi composti da 1 studente: si richiede di rispondere alle query 1 e 2; inoltre, la gestione del data ingestion è opzionale.

Per gruppi composti da 3 studenti: in aggiunta ai requisiti sopra elencati, si richiede di utilizzare un framework di alto livello (Hive, Pig oppure SparkSQL) per rispondere alle query 1 e 2. Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query ottenuti con Hive, Pig o SparkSQL e di confrontarli con quelli ottenuti usando il solo framework Hadoop o Spark, riportando il confronto nella relazione e nella presentazione.

Opzionale: Fornire una rappresentazione grafica dei risultati delle query utilizzando un framework di visualizzazione (e.g., Grafana [1]).

Svolgimento e consegna del progetto

Comunicare alle docenti la composizione del gruppo entro **martedì 18 maggio 2021**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2020/21 ed il codice deve essere consegnato **entro venerdì 4 giugno 2021** per poter raggiungere il punteggio massimo.

La consegna del progetto consiste in:

1. link a spazio di Cloud storage o repository contenente il codice del progetto da comunicare via email **entro venerdì 4 giugno 2021**; inserire i risultati delle query in formato CSV in una cartella denominata `Results`.
2. relazione di lunghezza compresa tra le 3 e le 6 pagine, da inviare via email **entro lunedì 7 giugno 2021**; per la redazione si consiglia di usare il formato ACM proceedings (https://www.acm.org) oppure il formato IEEE proceedings (https://www.ieee.org);
3. slide della presentazione orale, da inviare via email alle docenti **dopo** lo svolgimento della presentazione.

La presentazione si terrà **giovedì 10 giugno 2021**; ciascun gruppo avrà a disposizione **massimo 15 minuti**.

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

[2] Apache Mahout. `https://mahout.apache.org/`.

[3] Apache Spark MLlib. `https://spark.apache.org/mllib/`.