Sistemi e Architetture per Big Data - A.A. 2019/2020
Progetto 1: Analisi del dataset 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 la pandemia Covid-19, utilizzando il framework di data processing Apache Hadoop oppure Apache Spark.

Per gli scopi di questo progetto vengono forniti i seguenti file in formato CSV:

- `dpc-covid19-ita-andamento-nazionale`, disponibile all'URL https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv;
- `time_series_covid19_confirmed_global` disponibile all'URL https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Nello specifico, il file `dpc-covid19-ita-andamento-nazionale` contiene, in ogni riga, la *data* (formato *yyyy-mm-dd*T*hh:mm:ss*), lo *stato* (ITA), il numero di *ricoverati con sintomi*, *ricoverati in terapia intensiva*, *totale pazienti ospedalizzati*, il numero di persone in *isolamento domiciliare*, il *totale dei positivi*, la *variazione del totale dei positivi*, il numero dei *nuovi positivi*, il numero dei pazienti *guariti e dimessi*, il numero dei pazienti *deceduti*, il *totale dei casi registrati*, il numero di *tamponi* effettuati. Il dataset viene aggiornato su base nazionale con granularità giornaliera alle ore 18:00 CET dal 24 Febbraio 2020. Quasi tutti i dati riportati nelle colonne del dataset sono cumulativi (ovvero il dato di un determinato giorno è pari al dato del giorno precedente incrementato del valore di quel giorno). Fanno eccezione le colonne *variazione totale positivi* e *nuovi positivi* in cui i dati sono puntuali (ovvero fanno riferimento ad un singolo giorno).

Il file `time_series_covid19_confirmed_global` contiene i dati sull'andamento mondiale dei casi confermati di Covid-19. In particolare, in ogni riga sono indicati: lo *stato*, la *nazione*, la *latitudine* e *longitudine*. Inoltre, per ogni giorno dal 22 Gennaio 2020, ogni colonna con la data *mm/dd/yy* riporta il totale dei casi confermati fino a quel giorno (dato cumulativo)). Il dataset viene aggiornato ogni giorno con una nuova riga alle ore 23:59 UTC e sono possibili correzioni nel caso in cui vengano riscontrate delle inaccuratezze.

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

Le query a cui rispondere sono:

1. Per ogni settimana, calcolare il numero medio di guariti e dei tamponi effettuati in Italia in quella settimana.

2. Per ogni continente, calcolare la media, la deviazione standard, il minimo e il massimo del numero di casi confermati per ogni settimana. Nel calcolo delle statistiche, considerare solo i 100 stati più colpiti dalla pandemia. Qualora lo stato non fosse indicato, considerare la nazione. Per determinare gli stati più colpiti nell'intero dataset, si consideri l'andamento degli incrementi giornalieri dei casi confermati attraverso il *trendline coefficient*. Per stimare il *trendline coefficient*, si calcoli la pendenza della retta di regressione che approssima la tendenza degli incrementi giornalieri.
   Nota: il continente a cui appartiene ogni nazione non viene indicato in modo esplicito nel dataset, ma deve essere ricavato. Si considerino 6 continenti: Africa, America, Antartide, Asia, Europa, Oceania.

3. Considerando i 50 stati più colpiti calcolati su base mensile secondo il trendline coefficient, per ogni mese nel dataset applicare l'algoritmo di clustering K-means [2, 3] con $K = 4$. Determinare gli stati (o nazioni) che fanno parte di ogni cluster. Ogni cluster dovrebbe raggruppare gli stati che hanno un simile andamento degli incrementi giornalieri dei casi confermati.

   Per l'algoritmo di clustering K-means, si effettui un confronto tra le prestazioni di un'implementazione näive dell'algoritmo e l'implementazione fornita dalla libreria Spark MLlib [5] o Apache Mahout [4].

Al momento della consegna, includere anche il risultato prodotto da ciascuna query in formato CSV, specificando la data dei dataset utilizzati.

Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query sulla piattaforma di riferimento usata per la realizzazione del progetto e di riportare tali tempi nella presentazione e nella relazione. Tale piattaforma può essere un nodo standalone, oppure è possibile utilizzare un servizio Cloud per il processamento di Big Data (Amazon EMR o Google Dataproc) avvalendosi dei rispettivi 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 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 presentazione (e nell'eventuale relazione).

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 **venerdì 8 maggio 2020**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2019/20 ed il codice deve essere consegnato **entro venerdì 29 maggio 2020** 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 29 maggio 2020**; 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 1 giugno 2020**; 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 4 giugno 2020**; 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] k-means clustering. `https://en.wikipedia.org/wiki/K-means_clustering`.

[3] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*, chapter Clustering. Cambridge University Press, USA, 3rd edition, 2020. `http://infolab.stanford.edu/~ullman/mmds/ch7.pdf`.

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

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