Sistemi e Architetture per Big Data - A.A. 2018/19
Progetto 1: Analisi del dataset sulle condizioni meteorologiche 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 un dataset sulle condizioni meteorologiche di alcune città degli Stati Uniti e dello stato di Israele, utilizzando il framework di data processing Apache Hadoop oppure Apache Spark.

Le misurazioni orarie riportate nel dataset ricoprono un periodo di 5 anni, a partire dal 1 novembre 2012, ore 12:00:00, al 30 novembre 2017, ore 00:00:00. Nello specifico, per gli scopi di questo progetto vengono forniti i seguenti file in formato CSV, disponibili all'indirizzo `http://www.ce.uniroma2.it/courses/sabd1819/projects/prj1_dataset.tgz`:

- `city_attributes.csv`: per ogni città, viene riportata la latitudine e la longitudine;
- `humidity.csv`: per ogni città, viene riportata l'umidità (in %) registrata in una specifica ora di uno specifico giorno di un dato anno del dataset;
- `pressure.csv`: per ogni città, viene riportata la pressione (in hPa) registrata in una specifica ora di uno specifico giorno di un dato anno del dataset;
- `temperature.csv`: per ogni città, viene riportata la temperatura (in gradi Kelvin) registrata in una specifica ora di uno specifico giorno di un dato anno del dataset;
- `weather_description.csv`: per ogni città viene riportata la descrizione (espressa in formato `String`) delle condizioni meteo in una specifica ora, in uno specifico giorno di un dato anno del dataset. Ad esempio, una giornata serena viene descritta con la stringa `sky is clear`.

Nei file `humidity.csv`, `pressure.csv` e `temperature.csv` l'ora è espressa usando l'orario UTC (quindi ad es. UTC-7 per ottenere l'ora locale di San Francisco, UTC-4 per ottenere l'ora locale di New York, UTC+3 per ottenere l'ora locale di Haifa).

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 anno del dataset individuare le città che hanno almeno 15 giorni al mese di tempo sereno nei mesi di marzo, aprile e maggio.
   Nota: tempo sereno nei mesi di marzo, aprile e maggio da intendersi in AND. Determinare un criterio per decidere se il giorno è sereno, considerando l'informazione oraria a disposizione.

2. Individuare, per ogni nazione, la media, la deviazione standard, il minimo, il massimo della temperatura, della pressione e dell'umidità registrata in ogni mese di ogni anno.
   Nota: la nazione a cui appartiene ogni città non viene indicata in modo esplicito nel dataset, ma deve essere ricavata.

3. Individuare, per ogni nazione, le 3 città che hanno registrato nel 2017 la massima differenza di temperature medie nella fascia oraria locale 12:00-15:00 nei mesi di giugno, luglio, agosto e settembre rispetto ai mesi di gennaio, febbraio, marzo e aprile. Confrontare la posizione delle città nella classifica dell'anno precedente (2016).
   Nota: le medie delle temperature devono essere effettuate sulla base dei valori registrati nei mesi di giugno, luglio, agosto e settembre (e, allo stesso modo, gennaio, febbraio, marzo e aprile); non deve essere quindi considerato ogni mese singolarmente.

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 nell'eventuale 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 3; 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 3 query. 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).

Svolgimento e consegna del progetto

Comunicare ai docenti la composizione del gruppo entro **venerdì 10 maggio 2019**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2018/19 e deve essere consegnato **entro venerdì 24 maggio 2019** 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;

2. lucidi della presentazione orale, da inviare via email ai docenti *dopo* lo svolgimento della presentazione.

3. *opzionale*: relazione di lunghezza compresa tra le 4 e le 6 pagine, usando il formato ACM proceedings (`https://www.acm.org/publications/proceedings-template`) oppure il formato IEEE proceedings (`https://www.ieee.org/conferences_events/conferences/publishing/templates.html`).

La presentazione si terrà **giovedì 30 maggio 2019**; 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.