Sistemi e Architetture per Big Data - A.A. 2016/17
Progetto 1: Analisi del dataset MovieLens con Hadoop

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 di rispondere, utilizzando il framework Apache Hadoop, ad alcune query riguardanti il dataset MovieLens (`http://movielens.org`), in particolare la versione 20M [1]. Tale dataset contiene 20 milioni di valutazioni (su scala da 0 a 5 stelle) effettuate da 138000 utenti e riguardanti 27000 film, appartenenti a diversi generi cinematografici.

Il dataset è memorizzato in 6 file di testo, nel formato comma-separated value (csv). In particolare, i file di interesse per il progetto sono: `movies.csv` e `ratings.csv`. Il primo contiene informazioni sui film; ogni riga del file (eccetto la prima di intestazione) ha il formato:

```
movieId,title,genres
```

dove:

- `movieId` è l'ID del film;
- `title` è il titolo del film;
- `genres` è una lista (i cui elementi sono separati da |) dei generi attribuiti al film; i valori possibili sono: `Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, (no genres listed)`.

Il file `ratings.csv` contiene le valutazioni dei film; ogni riga del file (eccetto la prima di intestazione) ha il formato:

```
userId,movieId,rating,timestamp
```

dove:

- `userId` è l'ID dell'utente che ha inserito la valutazione del film;
- `movieId` è l'ID del film;
- `rating` è la valutazione del film, su una scala a 5 stelle, con incremento di mezza stella (da 0.5 a 5.0 stelle).
- `timestamp` è la data della valutazione, rappresentata in secondi a partire dalla mezzanotte UTC del 1 Gennaio 1970.

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. Individuare i film con una valutazione maggiore o uguale a 4.0 e valutati a partire dal 1 Gennaio 2000.
2. Calcolare la valutazione media e la sua deviazione standard per ciascun genere di film.
3. Trovare i 10 film che hanno ottenuto la più alta valutazione nell'ultimo anno del dataset (dal 1 Aprile 2014 al 31 Marzo 2015) e confrontare, laddove possibile, la loro posizione nella classifica rispetto a quella conseguita nell'anno precedente (dal 1 Aprile 2013 al 31 Marzo 2014).

Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query sulla piattaforma di riferimento usata per la realizzazione del progetto. Tale piattaforma può essere un nodo standalone oppure in alternativa è possibile utilizzare un servizio Cloud per Hadoop (ad es. Amazon EMR o Google Dataproc), utilizzando i 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 a scelta (e.g., Flume, Kite, ...);
- esportare i dati di output da HDFS ad un sistema di storage a scelta (e.g., HBase, ...).

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 oppure Pig) per rispondere alle 3 query. Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query ottenuti con Hive o Pig e di confrontarli con quelli ottenuti usando il solo framework Hadoop.

Svolgimento e consegna del progetto

Comunicare ai docenti la composizione del gruppo entro venerdì 26 maggio 2017.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2016/17 e deve essere consegnato **entro lunedì 12 giugno 2017** 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 (**durata massima di 15 minuti per gruppo**), da inviare via email ai docenti *dopo* lo svolgimento della presentazione.

Valutazione del progetto

I principali criteri di valutazione del progetto saranno:

1. rispondenza ai requisiti;

2. originalità;
3. organizzazione del codice;
4. efficienza;
5. organizzazione, chiarezza e rispetto dei tempi della presentazione orale.

Riferimenti bibliografici

[1] GroupLens Research. MovieLens 20M Dataset. `https://grouplens.org/datasets/movielens/20m/`, 2016.