Sistemi e Architetture per Big Data - A.A. 2017/18
Progetto 1: Analisi del dataset ACM DEBS Grand Challenge 2014 con Hadoop/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 rispondere ad alcune query riguardanti il dataset DEBS 2014 Grand Challenge [2, 3], utilizzando il framework Apache Hadoop o, in alternativa, Apache Spark.

Il dataset Grand Challenge 2014 contiene registrazioni di dati provenienti da prese di corrente intelligenti, installate in case private. Una presa di corrente intelligente è un dispositivo che si interpone tra la presa a muro ed il dispositivo ad essa collegato. È equipaggiata con diversi sensori di consumo energetico, che catturano ed emettono periodicamente informazioni di consumo (ad esempio, potenza istantanea, consumo energetico totale). Ogni sensore lavora in modo indipendente ed emette le sue misurazioni. Si osserva che i dati sono stati collezionati in un ambiente reale ed in modo non controllato, pertanto questi possono contenere valori mancanti nonché dati malformati.

Il dataset assume che i sensori siano identificati univocamente per mezzo di una gerarchia di identificatori. Nello specifico, l'entità di più alto livello è la casa, identificata da un `house_id` univoco. Ogni casa può contenere una o più famiglie, identificate da un `household_id` univoco all'interno della casa. Ogni famiglia contiene una o più prese intelligenti, identificate da un `plug_id` univoco all'interno della famiglia. Ogni presa intelligente contiene esattamente due sensori: (1) un sensore che misura la potenza istantanea consumata, espressa in Watt; (2) un sensore che misura la quantità totale di energia consumata dall'avvio (o dal riavvio) del sensore, espressa in kWh.

Per gli scopi di questo progetto, viene fornita una versione ridotta del dataset originale [1]; esso contiene registrazioni di al più 3 prese intelligenti per un massimo di 10 case, dove ogni sensore emette misurazioni ogni 20 secondi. Si considera un'unica famiglia per casa. Ogni riga del file ha il formato:

```
id,timestamp,value,property,plug_id,household_id,house_id
```

dove:

- `id` è un identificatore univoco della misura (intero senza segno a 32 bit);
- `timestamp` è il timestamp (secondi dal 1 gennaio 1970, 00:00:00 GMT; intero senza segno a 32 bit);
- `value` è la misura (numero in virgola mobile a 32 bit);
- `property` identifica la tipologia di misura: 0 per l'energia totale consumata, 1 per la potenza istantanea consumata (booleano);

- `plug_id` è l'identificatore, univoco per famiglia, della presa intelligente (intero senza segno a 32 bit);
- `household_id` è l'identificatore, univoco per casa, della famiglia dove la presa intelligente è localizzata (intero senza segno a 32 bit);
- `house_id` è l'identificatore univoco della casa dove la presa intelligente è installata (intero senza segno a 32 bit).

Le misure riportate nel dataset ricoprono un periodo di un mese, con il primo timestamp pari a 1377986401 (1 settembre 2013, ore 00:00:00) e l'ultimo timestamp pari a 1380578399 (30 settembre 2013, ore 23:59:59). Tutti gli eventi sono ordinati in base al proprio timestamp; gli eventi aventi lo stesso timestamp sono ordinati in modo casuale. I valori di energia sono cumulativi, con granularità di 1 kWh, per cui i cambiamenti sono visibili solo quando sono state registrate diverse misurazioni. Pertanto, i valori di potenza si prestano meglio per calcolare piccole quantità di consumi energetici.

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 le case con consumo di potenza istantaneo maggiore o uguale a 350 Watt.
2. Per ogni casa, calcolare il consumo energetico medio e la sua deviazione standard nelle quattro fasce orarie: notte, dalle ore 00:00 alle ore 05:59; mattino, dalle 06:00 alle 11:59; pomeriggio, dalle 12:00 alle 17:59; e sera, dalle 18:00 alle 23:59. Il consumo energetico di una casa è dato dalla somma dei consumi energetici di ogni presa intelligente collocata nella casa.
3. Si considerino le seguenti fasce di consumo dell'energia elettrica, differenziate in base all'ora e al giorno della settimana, ed a cui corrispondono differenti tariffe: *fascia di punta*, che si applica negli orari diurni dal lunedì al venerdì (dalle 06:00 alle 17:59) ed a cui corrisponde una tariffa più alta, e *fascia fuori punta*, che si applica negli orari notturni dal lunedì al venerdì (dalle 18:00 alle 05:59), nel fine settimana (sabato e domenica) e nei giorni festivi. Calcolare la classifica delle prese intelligenti in base alla differenza dei consumi energetici medi mensili tra la fascia di punta e la fascia fuori punta. Nella classifica le prese sono ordinate in modo decrescente, riportando, come primi elementi, le prese che non sfruttano la fascia fuori punta.

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. 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. Gli studenti interessati ad usare Google Dataproc o comunque un servizio Cloud di Google per il deployment dei framework usati sono invitati a richiedere l'attivazione del grant (che non richiede un numero di carta di credito) ai docenti, segnalando lo username del proprio account Google.

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.

Svolgimento e consegna del progetto

Comunicare ai docenti la composizione del gruppo entro **lunedì 21 maggio 2018**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2017/18 e deve essere consegnato **entro venerdì 1 giugno 2018** 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 7 giugno 2018**; 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. organizzazione del codice;
4. efficienza;
5. organizzazione, chiarezza e rispetto dei tempi della presentazione orale.

Riferimenti bibliografici

[1] DEBS 2014 Grand Challenge. Lightweight dataset. `http://www.ce.uniroma2.it/courses/sabd1718/resources/debs14_reduced.tar.gz`, 2018.

[2] DEBS 2014 Grand Challenge: Smart homes. http://debs.org/debs-2014-smart-homes/, 2018.

[3] Z. Jerzak and H. Ziekow. The DEBS 2014 grand challenge. In *Proc. of 8th ACM Int'l Conf. on Distributed Event-Based Systems*, DEBS '14, pages 266–269, 2014.