Sistemi e Architetture per Big Data - A.A. 2016/17
Progetto 2: Analisi di dati in tempo reale di una partita di calcio

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 analizzare in tempo reale, tramite un framework open-source di data stream processing (a scelta tra Apache Flink, Apache Spark Streaming, Apache Storm, Twitter Heron) il dataset del DEBS 2013 Grand Challenge [1, 2] riguardante una partita di calcio, rispondendo ad alcune query rilevanti per gli allenatori delle due squadre e per gli spettatori della partita.

Tale dataset riguarda i data acquisiti tramite sensori wireless durante una partita di calcio tra 2 squadre da 8 giocatori ciascuna svoltasi al Nuremberg Stadium in Germania. La partita è stata giocata su un campo di calcio di dimensione pari alla metà di quella standard, in due tempi della durata di 30 minuti ciascuno. Ciascun giocatore e l'arbitro avevano due sensori nei parastinchi, i due portieri avevano due sensori aggiuntivi nei guanti. Il pallone aveva un sensore localizzato nel centro. I sensori nei parastinchi e nei guanti producono dati ad una frequenza di 200Hz, quello nel pallone ad una frequenza di 2000Hz. Il tasso di dati totale è di circa 15000 eventi di posizione al secondo.

Il dataset ha il seguente schema:

```
sid, ts, x, y, z, |v|, |a|, vx, vy, vz, ax, ay, az
```

dove

```
sid, // sensor id
ts, // time stamp
x, y, z, // sensor coordinates
|v|, // velocity
|a|, // acceleration
vx, vy, vz, // direction vector
ax, ay, az // acceleration vector
```

La descrizione del dataset è fornita in [2].

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

Assumendo che i dati arrivino al sistema di data stream processing senza ritardi e omissioni e supponendo di effettuarne il replay, si chiede di rispondere alla seguenti query in tempo reale:

1. Analizzare le prestazioni nella corsa di ogni giocatore che partecipa alla partita. L'output della statistica aggregata sulla corsa ha il seguente schema:

   ```
   ts_start, ts_stop, player_id, total distance, avg speed
   ```

dove

```
ts_start , // start of the run
ts_stop , // end of the run
player_id , // identifier of a player for which the measurement is made
total distance , // total length of the run performed by the player
avg speed // average speed of the run
```

Tali statistiche aggregate dovranno essere calcolate per diverse finestre temporali, in modo tale da permettere di confrontare le prestazioni di ciascun giocatore durante lo svolgimento della partita. Le finestre temporali richieste hanno durata di:

- 1 minuto;
- 5 minuti;
- l'intera partita.

2. A complemento della query precedente, si richiede di fornire la classifica aggiornata in tempo reale dei 5 giocatori più veloci. L'output della classifica ha il seguente schema:

   ```
   ts_start , ts_stop , player_id_1 , avg_speed_1 , player_id_2 , avg_speed_2 ,
   player_id_3 , avg_speed_3 , player_id_4 , avg_speed_4 , player_id_5 ,
   avg_speed_5
   ```

   dove

   ```
   ts_start , // start of the run
   ts_stop , // end of the run
   player_id_1 , // identifier of the fastest player
   avg_speed_1 // average speed of the run of the fastest player
   player_id_2 , // identifier of the second fastest player
   avg_speed_2 // average speed of the run of the second fastest player
   ...
   ```

   Tali statistiche aggregate dovranno essere calcolate per diverse finestre temporali, in modo tale da permettere di confrontare le prestazioni dei giocatori più veloci durante lo svolgimento della partita. Le finestre temporali richieste hanno durata di:

   - 1 minuto;
   - 5 minuti;
   - l'intera partita.

3. L'obiettivo della terza query è di calcolare le statistiche relative a quanto tempo ciascun giocatore trascorre nelle diverse zone del campo di gioco, quindi una sorta di heat map. A tale scopo, si suddivide il campo di gioco in una griglia di celle di uguale dimensione, con 8 celle lungo l'asse x e 13 celle lungo l'asse y (quindi una griglia composta da 104 celle).

   Si chiede di fornire per ciascun giocatore la percentuale di tempo che il giocatore trascorre in ciascuna cella usando due differenti finestre temporali:

- 1 minuto;
- l'intera partita.

L'output della query 3 ha il seguente schema:

```
ts, player_id, cell_id1, percent_time_in_cell1, cell_id2,
percent_time_in_cell2, cell_id3, percent_time_in_cell3, ...
```

dove

```
ts, // timestamp di inizio statistica
player_id, // identifier of the player
cell_id1, // identifier of the cell #1
percent_time_in_cell1, // percentage of time that given player
                       // spent in the cell #1 during the time window
...
```

Si chiede inoltre di valutare sperimentalmente i tempi di latenza ed il throughput delle tre query durante il processamento sulla piattaforma di riferimento usata per la realizzazione del progetto.

Opzionale: Insieme ad un gruppo che ha utilizzato un altro framework di data stream processing, confrontare, sulla stessa piattaforma di riferimento, le prestazioni in termini di tempo di latenza e throughput delle query ottenute dai due framework.

Per gruppi composti da 1 studente: si richiede di rispondere alle query 1 e 2.

Per gruppi composti da 3 studenti: in aggiunta ai requisiti sopra elencati, si richiede di utilizzare un altro framework di data stream processing per rispondere ad una delle tre query (si suggerisce la query 2) e di confrontare le prestazioni in termini di latenza e throughput con quelle ottenute dal primo framework scelto.

Svolgimento e consegna del progetto

Comunicare ai docenti la composizione del gruppo entro venerdì 30 giugno 2017; la composizione del gruppo può essere diversa rispetto al primo mini-progetto.

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 venerdì 14 luglio 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] DEBS 2013 Grand Challenge: Soccer Monitoring. `http://debs.org/?p=41`, 2016.

[2] C. Mutschler, H. Ziekow, and Z. Jerzak. The DEBS 2013 Grand Challenge. In *Proc. of 7th ACM Int'l Conf. on Distributed Event-based Systems*, DEBS '13, pages 289–294, 2013.