Università degli Studi di Roma “Tor Vergata”
Dipartimento di Ingegneria Civile e Ingegneria Informatica

Project 2

Corso di Sistemi e Architetture per Big Data
A.A. 2016/17

Valeria Cardellini, Matteo Nardelli
Project delivery

• Submission deadline:
  – July 14\textsuperscript{th}, 2017
  – After the deadline, the maximum achievable score will be decreased by 2 points for each week of delay

• What to deliver:
  – Link to cloud storage or repository containing the project code
  – Slides of your presentation (max. 15 \textbf{minutes} per group), to be delivered after the presentation

• Team
  – 2/3 students per team
You will use a real dataset from the DEBS Grand Challenge 2013: Soccer monitoring (http://debs.org/?p=41)

Data originates from a number of wireless sensors embedded in the shoes and a ball used during a soccer match and spans the whole duration of the game.

The real-time analytics includes the continuous computation of statistics of relevance to spectators as well as trainers and team managers.
DEBS GC 2013 dataset

- Available from http://lafayette.tosm.ttu.edu/debs2013/grandchallenge/full-game.gz

- Data:
  - collected on the Nuremberg Stadium in Germany
  - from sensors located near the players’ shoes (1 sensor per leg) and in the ball (1 sensor)
  - the goalkeeper is equipped with two additional sensors, one at each hand

- Sensors in the players’ shoes and hands produce data with 200Hz frequency
- Sensor in the ball produces data with 2000Hz frequency
- The total data rate reaches roughly 15,000 position events per second
DEBS 2013 dataset

The event schema is:

\[
\text{sid, ts, x, y, z, } |v|, |a|, vx, vy, vz, ax, ay, az
\]

where

- \text{sid} is a sensor id which produced the position event
- \text{ts} is a timestamp in picoseconds (start of the game: 10753295594424116; end of the game: 14879639146403495);
- \text{x, y, z} describe the position of the sensor in mm (the origin is the middle of a full size football field)
- \text{|v|} (in µm/s), \text{vx, vy, vz} describe speed magnitude and direction by a vector with size of 10,000. The speed of the object in x-direction in SI-units (m/s) is calculated by
  \[
  v_x = |v| \times vx \times 10^{-4} \times 10^{-6}
  \]
- \text{|a|} (in µm/s²), \text{ax, ay, az} describe the absolute acceleration and its constituents in 3 dimensions (the acceleration in m/s² is calculated similar to that of the velocity)
DEBS 2013 dataset

(0,33965)   (52477,33941)

(-50,-33960)   (52489,-33939)
DEBS 2013 dataset

In addition to sensor data:

• Separate data stream for referee events
  – Includes the time when a game was paused and the time when a game was resumed. Moreover, referee events contain the time and player_ids for substitutions.

• Mapping between player ids and team ids as well as between sensor id and player id is provided in the metadata file.

Other details:

• Game played on a half-size field with teams of 8 players each

• Game duration: two halves of 30 minutes each

• We assume that data arrives at the system under test without any delays, nor omissions
General requirements for the project

• Use a DSP framework of your choice among:
  – Apache Storm
  – Apache Spark Streaming
  – Apache Flink
  – Twitter Heron

• Include in your report the queries’ response time (latency) on your reference architecture
Query #1

• Goal: analyze the running performance of each of the players currently participating in the game

• Output: the aggregate running statistics
  ts_start, ts_stop, player_id, total distance, avg speed

• The aggregate running statistics must be calculated using three different time windows:
  – 1 minute
  – 5 minutes
  – entire match
Query #2

• Goal: analyze the running performance of every player participating in the game

• Output 2: top-5 players by average speed

```
ts_start, ts_stop, player_id_1, avg_speed_1, player_id_2, avg_speed_2, player_id_3, avg_speed_3, ...
```

• The aggregate running statistics must be calculated using three different time windows:
  – 1 minute
  – 5 minutes
  – entire match
Query #3

• Goal: calculate statistics for how long each of the players spent in which region of the field
• We define a grid with 8 rows along the x-axis and 13 columns along the y-axis of equal size (grid of 104 cells)
• Provide, for each player, the percentage of time that the player spent in each cell over two different time windows: 10 minutes and the whole game duration
• Output 3: Percentage spent in each cell of the field
  ts, player_id, cell_id1, percent_time_in_cell1,
  cell_id2, percent_time_in_cell2,
  cell_id3, percent_time_in_cell3, ...
Group composed of 1 student

- Solve queries #1 and #2
Optional part A

- **Compulsory** for group composed of 3 students

- Implement one of the queries (suggested query 2) using a different DSP framework

- Include in the report the query times achieved using the second DSP framework (running on your reference architecture) and compare the results achieved by the two DSP frameworks
Optional part B

• Compare the results achieved by your implementation with that achieved by another group that has used another DSP framework
  – For example, if your group uses Spark Streaming, compare your results with those achieved by another group that uses Storm
• Use the same reference architecture for comparing the frameworks
• Include in the report the comparison results