Sistemi e Architetture per Big Data - A.A. 2020/21
Progetto 2: Analisi di dati marittimi geo-spaziali con Flink/Storm

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 relativo a dati provenienti da dispositivi *Automatic Identification System* (AIS), utilizzando il framework Apache Flink o, in alternativa, Apache Storm. I dispositivi di identificazione automatica AIS sono utilizzati per garantire la sicurezza delle navi in mare e nei porti, permettendo lo scambio di informazioni riguardo lo stato di navi in movimento [1]. Tra queste informazioni, ad oggi, il porto di destinazione è impostato manualmente dall'operatore. La frequenza di produzione di tali dati è in funzione dello stato di moto, con un periodo temporale variabile tra i 2 secondi in fase di manovra a 5 minuti in fase di navigazione ad alta velocità. Un dataset formato da questi dati geo-spaziali e relativo al mar Mediterraneo è stato indicato a lezione, in formato CSV (con i campi separati da ,). Nello specifico, ciascun evento AIS è una tupla con i seguenti attributi:

- SHIP_ID: stringa esadecimale che rappresenta l'identificativo della nave.
- SHIP_TYPE: numero intero che rappresenta la tipologia della nave [1].
- SPEED: numero in virgola mobile che rappresenta la velocità misurata in nodi a cui procede la nave all'istante di segnalazione dell'evento; il separatore decimale è il punto.
- LON: numero in virgola mobile che rappresenta la coordinata cartesiana in gradi decimali della longitudine data dal GPS; il separatore decimale è il punto.
- LAT: numero in virgola mobile che rappresenta la coordinata cartesiana in gradi decimali della latitudine data dal sistema GPS; il separatore decimale è il punto.
- COURSE: numero intero che rappresenta la direzione del movimento ed è espresso in gradi; è definito come l'angolo in senso orario tra il nord Vero e il punto di destinazione (rotta vera).
- HEADING: numero intero che rappresenta la direzione verso cui la nave è orientata ed è espresso in gradi; è definito come l'angolo in senso orario tra il nord Vero e l'asse longitudinale della barca (prua o prora vera).
- TIMESTAMP: rappresenta l'istante temporale della segnalazione dell'evento AIS; il timestamp è espresso con il formato GG-MM-YY hh:mm:ss (giorno, mese, anno, ore, minuti e secondi dell'evento).

[1] Link significato tipo nave.

- DEPARTURE_PORT_NAME: stringa che rappresenta l'identificativo del porto di partenza del viaggio in corso.
- REPORTED_DRAUGHT: numero intero che rappresenta la profondità della parte immersa della nave (in centimetri) tra la linea di galleggiamento e la chiglia.
- TRIP_ID: stringa alfanumerica che rappresenta l'identificativo del viaggio; è composta dai primi 7 caratteri (inclusi 0x) di SHIP_ID, concatenati con la data di partenza e di arrivo.

Ai fini del progetto, si limita l'area marittima d'interesse alle seguenti coordinate di latitudine e longitudine: LAT $\in [32.0, 45.0]$ e LON $\in [-6.0, 37.0]$. Si suppone inoltre che l'area considerata venga divisa in celle rettangolari di uguale dimensione. Le celle sono ottenute dividendo la LAT in 10 settori e la LON in 40 settori. I settori di LAT vengono identificati dalle lettere che vanno da A a J, mentre i settori di LON dai numeri interi che vanno da 1 a 40. Ad ogni cella è associato un *id* dato dalla combinazione della lettera del settore LAT e dal numero di settore LON. Ad esempio, il punto con LAT 32.0 e LON -6.0 è situato nella cella A1.

Il progetto è dimensionato per un gruppo composto da **2 studenti**; per gruppi composti da 1 oppure 3 studenti, si vedano le indicazioni specifiche. Per lo svolgimento della prova, si chiede di ordinare il dataset fornito sulla base del timestamp e di effettuarne il replay (accelerando opportunamente la scala temporale). Le query a cui rispondere in tempo reale sono:

Q1 Calcolare per ogni cella del Mar Mediterraneo Occidentale [2], il numero medio di navi militari (SHIP_TYPE = 35), navi per trasporto passeggeri (SHIP_TYPE = 60-69), navi cargo (SHIP_TYPE = 70-79) e *others* (tutte le navi che non hanno uno SHIP_TYPE che rientri nei casi precedenti) negli ultimi 7 giorni (di event time) e 1 mese (di event time). L'output della query ha il seguente schema:

```
ts, id_cella, ship_t35, avg_t35, ..., ship_to, avg_to
```

dove

```
ts     // timestamp relativo all'inizio del periodo su cui e'
   ↪ calcolata la media
id_cella // id cella
ship_t35 // SHIP_TYPE = 35
avg_t35  // numero medio di navi con SHIP_TYPE = 35
...
ship_to // navi di tipo others
avg_to  // numero medio di navi di tipo others
```

Q2 Per il Mar Mediterraneo Occidentale ed Orientale [3] fornire la classifica delle tre celle più frequentate nelle due fasce orarie di servizio 00:00-11:59 e 12:00-23:59. In una determinata fascia oraria, il grado di frequentazione di una cella viene calcolato come il numero di navi diverse che attraversano la cella nella fascia oraria in esame. L'output della query ha il seguente schema:

```
ts, sea, slot_a, rank_a, slot_p, rank_p
```

dove

[2] https://it.wikipedia.org/wiki/Mediterraneo_occidentale
[3] https://it.wikipedia.org/wiki/Mediterraneo_orientale

```
ts     // timestamp di inizio classifica
sea  // area del Mar Mediterraneo in esame
slot_a // fascia oraria del mattino
rank_a // classifica delle 3 celle più frequentate nella fascia
   ↪ oraria del mattino
slot_p // fascia oraria del pomeriggio
rank_p // classifica delle 3 celle più frequentate nella fascia
   ↪ oraria del pomeriggio
```

La classifica dovrà essere calcolata sulle finestre temporali:

- 7 giorni (di event time),
- 1 mese (di event time).

Q3 Fornire la classifica in tempo reale dei 5 viaggi che hanno il punteggio di percorrenza più alto. Il punteggio di percorrenza viene calcolato come pari alla distanza percorsa fino a quel momento del viaggio. Per il calcolo della distanza, considerare la distanza euclidea. Come punto di partenza, considerare la prima coppia (LAT, LON) registrata per il viaggio in esame.

L'output della classifica ha il seguente schema:

```
ts, ship_id_1, rating_1, ship_id_2, rating_2, ..., ship_id_5,
   ↪ rating_5
```

dove

```
ts       // timestamp di inizio classifica
trip_1 // id del viaggio classificato primo
rating_1 // punteggio complessivo del viaggio classificato primo
...
trip_5 // id del viaggio classificato quinto
rating_5 // punteggio complessivo del viaggio classificato quinto
```

La classifica dovrà essere calcolata sulle finestre temporali:

- 1 ora (di event time),
- 2 ore (di event time).

Gli output delle query devono anche essere memorizzati in file CSV e consegnati.

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, riportando tali tempi nella presentazione e nella relazione. La piattaforma di data stream processing può essere un nodo standalone con Apache Flink o Apache Storm oppure in alternativa è possibile utilizzare un servizio Cloud per stream processing (ad es. Amazon EMR con Flink), avvalendosi del grant a disposizione.

Opzionale: Rispondere ad una query a scelta tra le tre sopra descritte usando Kafka Streams oppure Spark Streaming e 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 Kafka Streams oppure Spark Streaming per rispondere alle tre query e di confrontare, sulla stessa piattaforma di riferimento, le prestazioni in termini di latenza e throughput con quelle ottenute dal primo framework scelto.

Svolgimento e consegna del progetto

Comunicare alle docenti la composizione del gruppo entro **lunedì 21 giugno 2021**.

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2020/21 ed il codice deve essere consegnato **entro venerdì 9 luglio 2021** 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 da comunicare via email **entro venerdì 9 luglio 2021**; inserire i risultati delle query in formato CSV in una cartella denominata `Results`.
2. relazione di lunghezza compresa tra le 3 e le 6 pagine, da inviare via email **entro domenica 11 luglio 2021**; per la redazione si consiglia di usare il formato ACM proceedings (https://www.acm.org) oppure il formato IEEE proceedings (https://www.ieee.org);
3. slide della presentazione orale, da inviare via email alle docenti **dopo** lo svolgimento della presentazione.

La presentazione si terrà **lunedìì 12 luglio 2021**; 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.

Riferimenti bibliografici

[1] Automatic identification system. `https://en.wikipedia.org/wiki/Automatic_identification_system`.