Sistemi e Architetture per Big Data - A.A. 2022/23
Progetto 2: Analisi in tempo reale di dati finanziari con Apache Flink

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 di dati finanziari fornito dall'azienda fintech Infront Financial Technology [3, 5], utilizzando l'approccio di processamento a *stream* con Apache Flink. Per gli scopi di questo progetto, viene fornita una versione ridotta del dataset indicato in [4] e descritto in [1], che è disponibile all'URL[1]:

http://www.ce.uniroma2.it/courses/sabd2223/project/out600_combined+header.csv.

Il dataset riguarda lo scambio di strumenti finanziari su tre principali borse europee nel corso di una settimana. I dati si basano su eventi reali acquisiti da Infront Financial Technology per la settimana dall'8 al 14 novembre 2021 (cinque giorni lavorativi seguiti da sabato e domenica). Il dataset ridotto contiene circa 4 milioni di eventi (a fronte dei 289 milioni del dataset originario) che coprono 600 azioni (equities) e indici (indices) sulle borse europee: Parigi (FR), Amsterdam (NL) e Francoforte/Xetra (ETR). Gli eventi sono registrati così come sono stati acquisiti; al momento della registrazione i dati vengono etichettati con data ed ora di ricezione dell'aggiornamento (campi: *date* e time). Alcuni eventi sembrano essere privi di payload. **Attenzione:** il dataset ridotto è *diverso* rispetto alla versione usata per il progetto 1 del corso.

La Tabella 1 descrive i campi di ogni record; i campi rilevanti per questo progetto sono evidenziati nella tabella con un asterisco (*). All'interno del dataset, ogni strumento finanziario è associato ad un identificatore (`ID`), che è costituito da una stringa univoca che indica il nome dello specifico strumento ed il codice di scambio della borsa su cui tale strumento è negoziato (ad esempio, I2GS.FR, in cui FR indica la borsa di Parigi). I timestamp orari sono nel formato HH:MM:SS.ssss, le date in DD-MM-YYYY e i prezzi in 12.3456 (sei cifre).

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 del progetto, si chiede di effettuare il replay del dataset, accelerando opportunamente la scala temporale. La riproduzione accelerata deve preservare la coerenza tra intervalli temporali (ad esempio, accelerando con un fattore 3600, 1 ora in event time viene riprodotta in 1 secondo).

Le query a cui rispondere in modalità *streaming* sono:

Q1 Per le azioni (campo `SecType` con valore pari a `E`) scambiate sui mercati di Parigi (`FR`) che iniziano per "G", calcolare il numero di eventi ed il valor medio del prezzo di vendita (campo `Last`).

L'output della query ha il seguente schema:

[1] sha1sum di out600_combined+header.csv: dbeaf8405c7de661a4de8e0312586af596cfa995

Tabella 1: Formato dei dati forniti da Infront Financial Technology

Campo	Descrizione	Rilevante
ID	Unique ID	*
SecType	Security Type (E)quity/(I)ndex	*
Date	System date for last received update	*
Time	System time for last received update	*
Ask	Price of best ask order	
Ask volume	Volume of best ask order	
Bid	Price of best bid order	
Bid volume	Volume of best bid order	
Ask time	Time of last ask	
Day's high ask	Day's high ask	
Close	Closing price	
Currency	Currency (ISO 4217)	
Day's high ask time	Day's high ask time	
Day's high	Day's high	
ISIN	International Securities Identification Number	
Auction price	Price at midday's auction	
Day's low ask	Lowest ask price of the current day	
Day's low	Lowest price of the current day	
Day's low ask time	Time of lowest ask price of the current day	
Open	First price of current trading day	
Nominal value	Nominal Value	
Last	Last trade price	*
Last volume	Last trade volume	
Trading time	Time of last update (bid/ask/trade)	*
Total volume	Cumulative volume for current trading day	
Mid price	Mid price (between bid and ask)	
Trading date	Date of last trade	*
Profit	Profit	
Current price	Current price	
Related indices	Related indices	
Day high bid time	Days high bid time	
Day low bid time	Days low bid time	
Open Time	Time of open price	
Last trade time	Time of last trade	
Close Time	Time of closing price	
Day high Time	Time of days high	
Day low Time	Time of days low	
Bid time	Time of last bid update	
Auction Time	Time when last auction price was made	

```
ts, ID, count, avg_last_value
```

dove:

- `ts`: timestamp relativo all'inizio del periodo su cui è stata calcolata la media;
- `ID`: identificativo dell'azione;
- `count`: numero di misurazioni;
- `avg_last_value`: valor medio del prezzo di vendita nel periodo considerato.

Calcolare la query sulle finestre temporali:

- 1 ora (event time)
- 1 giorno (event time);
- dall'inizio del dataset.

Si faccia attenzione agli eventi privi di payload.

Q2 Calcolare la classifica aggiornata in tempo reale delle 5 azioni (di qualsiasi mercato) che registrano la più alta variazione del prezzo di vendita calcolato nella finestra temporale indicata di seguito e delle 5 azioni (di qualsiasi mercato) con la variazione del prezzo di vendita più bassa.

L'output della query ha il seguente schema:

```
ts, ID1, delta_last1,..., ID5, delta_last5, ID6, delta_last6, ...,
ID10, delta_last10
```

dove:

- `ts`: il timestamp relativo all'inizio del periodo su cui è stata calcolata la statistica;
- `ID[1-5]`: identificativo dell' azione in posizione [1-5] nella top-5 delle azioni con variazione del prezzo di vendita più alta;
- `delta_last[1-5]`: variazione del prezzo di vendita dell'azione `ID[1-5]` nella finestra considerata;
- `ID[6-10]`: identificativo dell'azione in posizione [1-5] nella top-5 delle azioni con variazione del prezzo di vendita più bassa;
- `delta_last[6-10]`: variazione del prezzo di vendita dell'azione `ID[6-10]` nella finestra considerata.

Calcolare la query sulle finestre temporali:

- 30 minuti (event time);
- 1 ora (event time);
- 1 giorno (event time).

Si faccia attenzione agli eventi privi di payload.

Q3 Dopo aver calcolato la variazione di prezzo di vendita delle azioni scambiate sui diversi mercati per la finestra temporale di seguito indicata, raggruppare le azioni per mercato (`ETR`, `FR` e `NL`) e calcolare il 25-esimo, 50-esimo, 75-esimo percentile della variazione del prezzo di vendita per ciascun mercato.

L'output della query ha il seguente schema:

```
ts, gID1, 25p_gID1, 50p_gID1, 75p_gID1, ..., gID3, 25p_gID3, 50p_gID3,
75p_gID3
```

dove:

- `ts`: il timestamp relativo all'inizio del periodo su cui è stata calcolata la statistica;

- `gID[1-3]`: identificativo del mercato azionario (ad es., `FR`);
- `25p_gID[1-3]`: 25-esimo percentile della variazione del prezzo di vendita delle azioni del mercato `gID[1-3]`;
- `50p_gID[1-3]`: 50-esimo percentile della variazione del prezzo di vendita delle azioni del mercato `gID[1-3]`;
- `75p_gID[1-3]`: 75-esimo percentile della variazione del prezzo di vendita delle azioni del mercato `gID[1-3]`.

Si suggerisce di calcolare il valore dei percentili in tempo reale, senza ordinare tutti i valori e possibilmente senza accumularli; si veda ad esempio [2, 6].

Calcolare la query sulle finestre temporali:

- 30 minuti (event time);
- 1 ora (event time);
- 1 giorno (event time).

Non è necessario emettere le statistiche per gruppi privi di eventi.

Il risultato di ciascuna query deve essere consegnato in formato CSV (Flink supporta la scrittura di file in tale formato). Si chiede inoltre di valutare sperimentalmente i tempi di latenza ed il throughput delle query durante il processamento sulla piattaforma usata per la realizzazione del progetto. Riportare l'analisi del confronto nella relazione e nella presentazione del progetto.

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 query 1 e 2 e di confrontare, sulla stessa piattaforma di riferimento, le prestazioni in termini di tempo di latenza e throughput con quelle ottenute usando Flink. Riportare l'analisi del confronto nella relazione e nella presentazione del progetto.

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 con quelle ottenute usando Flink.

Svolgimento e consegna del progetto

Comunicare la composizione del gruppo ai docenti entro **venerdì 23 giugno 2023** (sole se diversa rispetto al progetto 1).

Per ogni comunicazione via email è necessario specificare *[SABD]* nell'oggetto (subject) dell'email. Il progetto è valido **solo** per l'A.A. 2022/23 ed il codice deve essere consegnato **entro mercoledì 12 luglio 2023**.

La consegna del progetto consiste in:

1. link a spazio di Cloud storage o repository contenente il codice del progetto da comunicare via email ai docenti **entro mercoledì 12 luglio 2023**; 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 inserire all'interno della cartella denominata `Report`; per la relazione 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 ai docenti **dopo** lo svolgimento della presentazione.

La presentazione si terrà **venerdì 14 luglio 2023** nel pomeriggio; ciascun gruppo avrà a disposizione **massimo 15 minuti** per presentare la propria soluzione.

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] DEBS 2022. Call for Grand Challenge Solutions. `https://2022.debs.org/call-for-grand-challenge-solutions/`, 2022.

[2] S. Engelhardt. Calculating Percentiles on Streaming Data Part 1: Introduction. `https://www.stevenengelhardt.com/2018/03/06/calculating-percentiles-on-streaming-data-part-1-introduction/`.

[3] S. Frischbier, M. Paic, A. Echler, and C. Roth. Managing the complexity of processing financial data at scale - an experience report. In *Complex Systems Design & Management*, pages 14–26. Springer International Publishing, Cham, Switzerland, Nov. 2019. `https://link.springer.com/chapter/10.1007/978-3-030-34843-4_2`.

[4] S. Frischbier, J. Tahir, C. Doblander, A. Hormann, R. Mayer, and H.-A. Jacobsen. DEBS 2022 Grand Challenge Data Set: Trading Data. `https://doi.org/10.5281/zenodo.6382482`, 2022.

[5] Infront Financial Technology. `https://www.infrontfinance.com/`, 2023.

[6] R. Jain and I. Chlamtac. The p^2 algorithm for dynamic calculation of quantiles and histograms without storing observations. *Communications of the ACM*, 28(10):1076–1085, 1985. `https://www.cse.wustl.edu/~jain/papers/ftp/psqr.pdf`.