Sistemi e Architetture per Big Data - A.A. 2022/23
Progetto 1: Analisi di dati finanziari con Apache 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 di dati finanziari fornito dall'azienda fintech Infront Financial Technology [2, 5], utilizzando il framework di data processing Apache Spark. Per gli scopi di questo progetto, viene fornita una versione ridotta del dataset indicato in [3] e descritto in [1], che è disponibile all'URL[1]:

http://www.ce.uniroma2.it/courses/sabd2223/project/out500_combined+header.csv.

Il dataset riguardo 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 500 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; alcuni eventi sembrano essere privi di payload.

La Tabella 1 descrive i campi di ogni record. Gli attributi rilevanti per questo progetto sono evidenziati nella tabella. All'interno del dataset, qualsiasi strumento finanziario è identificato da 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 es. 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.

Considerando il dataset indicato, le query a cui rispondere sono:

Q1 Per ogni ora, calcolare il valore minimo, medio e massimo del prezzo di vendita (campo `Last`) per le sole azioni (campo `SecType` con valore pari a `E`) scambiate sui mercati di Parigi (`FR`). Nell'output indicare anche il numero totale di eventi utilizzati per il calcolo delle statistiche. Si faccia attenzione agli eventi privi di payload.

Esempio di output (*i valori indicati sono solo a titolo esemplificativo*):
 # DD-MM-YYYY, *HH:MM*, *ID*, *min*, *mean*, *max*, *count*
08-11-2021, 08:00, I2GS.FR, 10.000, 12.356, 13.000, 241

[1] sha1sum di out500_combined+header.csv: 8c12da295ded47596d1849d758030a88b1bd2695

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	

Q2 Per ogni giorno, per ogni azione scambiata su qualsiasi mercato, calcolare il valor medio e la deviazione standard della loro variazione del prezzo di vendita calcolata su una finestra temporale di un'ora. Ad esempio, se l'azione `I2GS.FR` ha valore 12.356 alle ore 11:00 e valore 13.000 alle ore 12:00, la sua variazione è di 644; variazioni negative sono ammissibili, se il titolo ha perso valore nel corso di quell'ora. Dopo aver calcolato gli indici statistici su base giornaliera, determinare la classifica delle migliori 5 azioni che nella giornata hanno registrato la migliore variazione di prezzo e delle 5 peggiori azioni che hanno registrato la peggiore variazione di prezzo. Nell'output indicare anche il numero totale di eventi utilizzati per il calcolo delle statistiche.

Esempio di output:
 # DD-MM-YYYY, *ID*, *mean*, *std dev*, *count*
08-11-2021, I2GS.FR, 685, 336.83, 120
08-11-2021, SBF80.FR, 369, 126.98, 131
08-11-2021, FR20N.FR, 210, 893.12, 119
...
08-11-2021, CAEWC.FR, -725, -26.83, 149
08-11-2021, IPSRW.FR, -525, -26.83, 149
...
09-11-2021, CLEWJ.FR, 854, 1506.33, 256
...

Q3 Per ogni giorno, calcolare il 25-esimo, 50-esimo, 75-esimo percentile della variazione del prezzo di vendita delle azioni scambiate sui singoli mercati. La statistica deve essere calcolata considerando tutte e sole le azioni appartenenti ad ogni specifico mercato: Parigi (FR), Amsterdam (NL) e Francoforte/Xetra (ETR). Nell'output indicare anche il numero totale di eventi utilizzati per il calcolo delle statistiche. Si faccia attenzione agli eventi privi di payload. Esempio di output:
 # DD-MM-YYYY, *XID*, *25perc*, *50perc*, *75perc*, *count*
08-11-2021, FR, 125, 236, 595, 1346
08-11-2021, NL, 658, 986, 1256, 3986
08-11-2021, ETR, 365, 1266, 2504, 4982
09-11-2021, FR, 294, 759, 1985, 9346
09-11-2021, NL, 338, 786, 956, 2986
09-11-2021, ETR, 125, 556, 985, 3982

Il risultato di ciascuna query deve essere consegnato in formato CSV.

Inoltre, si chiede di valutare sperimentalmente i tempi di processamento delle query sulla piattaforma di riferimento usata per la realizzazione del progetto e di riportare tali tempi nella relazione e nella presentazione del progetto. Tale piattaforma può essere un nodo standalone, oppure è possibile utilizzare un servizio cloud per il processamento di Big Data (e.g., Amazon EMR) avvalendosi del grant a disposizione.

Infine, si chiede di realizzare la fase di data ingestion per:

- importare i dati di input in HDFS, eventualmente gestendo la conversione del formato dei dati, usando un framework di data ingestion a scelta (e.g., Apache Flume, Apache NiFi, Apache Kafka, Apache Pulsar);
- esportare i risultati 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 2; inoltre, la gestione del data ingestion è opzionale, ma il dataset deve comunque essere letto da HDFS ed i risultati di output scritti su HDFS.

Per gruppi composti da 3 studenti: in aggiunta ai requisiti sopra elencati, si richiede di utilizzare SparkSQL per rispondere alle tre query utilizzando SQL. Si chiede inoltre di valutare sperimentalmente i tempi di processamento delle 3 query ottenuti con Spark SQL e di confrontarli con quelli ottenuti usando il solo framework Spark, riportando l'analisi del confronto nella relazione e nella presentazione.

Opzionale: Fornire una rappresentazione grafica dei risultati delle query utilizzando un framework di visualizzazione (e.g., Grafana [4]).

Svolgimento e consegna del progetto

Comunicare la composizione del gruppo ai docenti entro **venerdì 26 maggio 2023**.

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 lunedì 12 giugno 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 lunedì 12 giugno 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 **giovedì 15 giugno 2023**; 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. 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.

[3] 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.

[4] Grafana. https://grafana.com/.

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