CORSO DI BIG DATA
Primo Progetto
13 maggio 2024

Si consideri il dataset **Daily Historical Stock Prices**, scaricabile da Kaggle (https://www.kaggle.com/) all'indirizzo https://www.kaggle.com/datasets/ehallmar/daily-historical-stock-prices-1970-2018. Esso contiene l'andamento giornaliero di una selezione di azioni sulla borsa di New York (NYSE) e sul NASDAQ dal 1970 al 2018. Il dataset è formato da due file CSV.

Il primo (**historical_stock_prices**) ha i seguenti campi:

- ticker: simbolo univoco dell'azione (https://en.wikipedia.org/wiki/Ticker_symbol)
- open: prezzo di apertura
- close: prezzo di chiusura
- adj_close: prezzo di chiusura "modificato" (potete trascurarlo)
- low: prezzo minimo
- high: prezzo massimo
- volume: numero di transazioni
- date: data nel formato aaaa-mm-gg

Il secondo (**historical_stocks**) ha invece questi campi:

- ticker: simbolo dell'azione
- exchange: NYSE o NASDAQ
- name: nome dell'azienda
- sector: settore dell'azienda (per esempio "technology")
- industry: industria di riferimento per l'azienda (per esempio "semiconductors")

Dopo avere preparato opportunamente il dataset (per esempio eliminando dati errati o non significativi), progettare e realizzare almeno due delle seguenti applicazioni in almeno tre tra le seguenti tecnologie: MapReduce, Hive, Spark core e Spark SQL:

1. Un job che sia in grado di generare le statistiche di ciascuna azione dall'anno in cui è entrata in borsa indicando, per ogni azione: (a) il simbolo, (b) il nome dell'azienda, (c) una lista con l'andamento dell'azione in ciascun anno della presenza dell'azione in borsa indicando, per ogni anno: (i) la variazione percentuale della quotazione nell'anno (differenza percentuale arrotondata tra il primo prezzo di chiusura e l'ultimo prezzo di chiusura dell'anno), (ii) il prezzo minimo nell'anno, (iii) quello massimo nell'anno e (iv) il volume medio dell'anno.
2. Un job che sia in grado di generare un report contenente, per ciascun'industria e per ciascun anno: (a) la variazione percentuale della quotazione dell'industria[1] nell'anno, (b) l'azione dell'industria che ha avuto il maggior incremento percentuale nell'anno (con indicazione dell'incremento) e (c) l'azione dell'industria che ha avuto il maggior volume di transazioni nell'anno (con indicazione del volume). Nel report le industrie devono essere raggruppate per settore e ordinate per ordine decrescente di variazione percentuale.
3. Un job in grado di generare gruppi di aziende le cui azioni hanno avuto lo stesso trend in termini di variazione annuale per almeno tre anni consecutivi a partire dal 2000, indicando le aziende e il trend comune (es. {Apple, Intel, Amazon}: 2011:-1%, 2012:+3%, 2013:+5%).

Per ciascun job bisogna illustrare e documentare in un rapporto finale:

- Le operazioni di preparazione dei dati che sono state eventualmente effettuate
- Una possibile implementazione MapReduce (pseudocodice), Hive, Spark core (pseudocodice) e SparkSQL.
- Le prime 10 righe dei risultati dei vari job.
- Tabella e grafici di confronto dei tempi di esecuzione in locale e su cluster dei vari job con dimensioni crescenti dell'input[2].
- Il relativo codice completo MapReduce e Spark (basta un link a un repository ad accesso libero).

Tutte le specifiche non definite in questo documento possono essere scelte liberamente.

Consegnare il rapporto **entro il 13 giugno 2024** in un unico file pdf sul sito moodle del corso disponibile all'indirizzo: https://ingegneriacivileinformaticatecnologieaeronautiche.el.uniroma3.it/course/view.php?id=1614.

[1] La quotazione di un'industria si ottiene sommando le quotazioni (prezzo di chiusura) di tutte le azioni dell'industria
[2] Per aumentare le dimensioni dell'input si suggerisce di generare copie del file dato, eventualmente alterando alcuni dati.