CORSO DI BIG DATA
Primo Progetto
22 aprile 2021

Si consideri il dataset **Daily Historical Stock Prices**, scaricabile dal sito Moodle del corso, che contiene l'andamento giornaliero di un'ampia selezione di azioni sulla borsa di New York (NYSE) e sul NASDAQ dal 1970 al 2018. Il dataset è formato da due file CSV.

Ogni riga del 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)
- lowThe: prezzo minimo
- highThe: 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
- industry: industria di riferimento per l'azienda

Dopo avere eventualmente eliminato dal dataset dati errati o non significativi, progettare e realizzare in: (a) MapReduce, (b) Hive e (c) Spark core (quindi senza usare Spark SQL):

1. Un job che sia in grado di generare un report contenente, per ciascuna azione: (a) la data della prima quotazione, (b) la data dell'ultima quotazione, (c) la variazione percentuale della quotazione (differenza percentuale tra il primo e l'ultimo prezzo di chiusura presente nell'archivio), (d) il prezzo massimo e quello minimo e (e) (facoltativo) il massimo numero di giorni consecutivi in cui l'azione è cresciuta (chiusura maggiore dell'apertura) con indicazione dell'anno in cui questo è avvenuto. Il report deve essere ordinato per valori decrescenti del punto b.
2. Un job che sia in grado di generare un report contenente, per ciascun settore e per ciascun anno del periodo 2009-2018: (a) la variazione percentuale della quotazione del settore[1] nell'anno, (b) l'azione del settore che ha avuto il maggior incremento percentuale nell'anno (con indicazione dell'incremento) e (c) l'azione del settore che ha avuto il maggior volume di transazioni nell'anno (con indicazione del volume). Il report deve essere ordinato per nome del settore.
3. Un job in grado di generare le coppie di aziende che si somigliano (sulla base di una soglia scelta a piacere) in termini di variazione percentuale mensile nell'anno 2017 mostrando l'andamento mensile delle due aziende (es. Soglia=1%, coppie: 1:{Apple, Intel}: GEN: Apple +2%, Intel +2,5%, FEB: Apple +3%, Intel +2,7%, MAR: Apple +0,5%, Intel +1,2%, …; 2:{Amazon, IBM}: GEN: Amazon +1%, IBM +0,5%, FEB: Amazon +0,7%, IBM +0,5%, MAR: Amazon +1,4%, IBM +0,7%, ..)

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

- Una possibile implementazione MapReduce (pseudocodice), Hive e Spark (pseudocodice).
- Le prime 10 righe dei risultati dei vari job.
- Tabella e grafici che confrontano i tempi di esecuzione in locale e su cluster dei vari job con dimensioni crescenti dell'input[2].
- Il relativo codice completo MapReduce e Spark (da allegare al documento)

Tutte le specifiche non definite in questo documento possono essere scelte liberamente. Consegnare il rapporto **entro il 22 maggio 2021** in un unico file compresso di formato a piacere sul sito moodle del corso.

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