CORSO DI BIG DATA
Primo Progetto
24 aprile 2020

Si consideri il dataset **Daily Historical Stock Prices**, scaricabile dal sito 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:

1. Un job che sia in grado di generare le statistiche di ciascuna azione tra il 2008 e il 2018 indicando, per ogni azione: (a) il simbolo, (b) la variazione della quotazione (differenza percentuale arrotondata tra i prezzi di chiusura iniziale e finale dell'intervallo temporale), (c) il prezzo minimo, (e) quello massimo e (f) il volume medio nell'intervallo, ordinando l'elenco in ordine decrescente di variazione della quotazione.

2. Un job che sia in grado di generare, per ciascun settore, il relativo "trend" nel periodo 2008-2018 ovvero un elenco contenete, per ciascun anno nell'intervallo: (a) il volume annuale[1] medio delle azioni del settore, (b) la variazione annuale[2] media delle aziende del settore e (c) la quotazione giornaliera media delle aziende del settore.

3. Un job in grado di generare gruppi di aziende le cui azioni hanno avuto lo stesso trend in termini di variazione annuale nell'ultimo triennio disponibile, indicando le aziende e il trend comune (es. {Apple, Intel, Amazon}: 2016:-1%, 2017:+3%, 2018:+5%).

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[3].
- 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 24 maggio 2020** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: http://moodle3.ing.uniroma3.it/.

[1] Per volume annuale di un'azione si intende la somma dei volumi dell'azione in tutti i giorni dell'anno in cui la borsa è aperta.
[2] Per variazione annuale di un'azione si intende la differenza percentuale tra la quotazione di fine anno e quella di inizio anno
[3] Per aumentare le dimensioni dell'input si suggerisce di generare copie del file dato, eventualmente alterando alcuni dati.