CORSO DI BIG DATA
Primo Progetto
26 aprile 2019

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 dell'azione
- 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

Progettare e realizzare in: (a) MapReduce, (b) Hive e (c) Spark:

1. Un job che sia in grado di generare, in ordine, le dieci azioni la cui quotazione (prezzo di chiusura) è cresciuta maggiormente dal 1998 al 2018, indicando, per ogni azione: (a) il simbolo, (b) l'incremento percentuale, (c) il prezzo minimo raggiunto, (e) quello massimo e (f) il volume medio giornaliero in quell'intervallo temporale.
2. Un job che sia in grado di generare, per ciascun settore, il relativo "trend" nel periodo 2004-2018 ovvero un elenco contenete, per ciascun anno nell'intervallo: (a) il volume complessivo del settore, (b) la percentuale di variazione annuale (differenza percentuale arrotondata tra la quotazione di fine anno e quella di inizio anno) e (c) la quotazione giornaliera media. N.B.: volume e quotazione di un settore si ottengono sommando i relativi valori di tutte le azioni del settore.
3. Un job in grado di generare coppie di aziende di settori diversi le cui azioni che, negli ultimi 3 anni, hanno avuto lo stesso trend in termini di variazione annuale indicando le aziende e il trend comune (es. Apple, Fiat, 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 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 variabili dell'input[1].
- 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 28 maggio 2019** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: http://moodle3.ing.uniroma3.it/.

[1] Se si desidera aumentare le dimensioni dell'input si suggerisce di generare più copie del file dato, eventualmente alterando alcuni dati.