CORSO DI BIG DATA
Primo Progetto
26 aprile 2018

Si consideri il dataset **Amazon Fine Food Reviews**, scaricabile dal sito del corso, che contiene circa 600.000 recensioni di prodotti gastronomici rilasciati su Amazon dal 1999 al 2012. Il dataset è in formato CSV e ogni riga ha i seguenti campi:

- Id,
- ProductId (unique identifier for the product),
- UserId (unique identifier for the user),
- ProfileName,
- HelpfulnessNumerator (number of users who found the review helpful),
- HelpfulnessDenominator (number of users who graded the review),
- Score (rating between 1 and 5),
- Time (timestamp of the review expressed in Unix time),
- Summary (summary of the review),
- Text (text of the review).

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

1. Un job che sia in grado di generare, per ciascun anno, le dieci parole che sono state più usate nelle recensioni (campo summary) in ordine di frequenza, indicando, per ogni parola, la sua frequenza, ovvero il numero di occorrenze della parola nelle recensioni di quell'anno.
2. Un job che sia in grado di generare, per ciascun prodotto, lo score medio ottenuto in ciascuno degli anni compresi tra il 2003 e il 2012, indicando ProductId seguito da tutti gli score medi ottenuti negli anni dell'intervallo. Il risultato deve essere ordinato in base al ProductId.
3. Un job in grado di generare coppie di prodotti che hanno almeno un utente in comune, ovvero che sono stati recensiti da uno stesso utente, indicando, per ciascuna coppia, il numero di utenti in comune. Il risultato deve essere ordinato in base allo ProductId del primo elemento della coppia e, possibilmente, non deve presentare duplicati.

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 22 maggio 2018** 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.