CORSO DI BIG DATA
Primo Progetto
21 aprile 2017

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 suddiviso per anni in 4 file di dimensione crescente (1999-2006, 2007-2008, 2009-2010 e 2011-2012), è in formato CSV e ogni riga ha i seguenti attributi:

- 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 mese, i cinque prodotti che hanno ricevuto lo score medio più alto, indicando ProductId e score medio e ordinando il risultato temporalmente.
2. Un job che sia in grado di generare, per ciascun utente, i 10 prodotti preferiti (ovvero quelli che ha recensito con il punteggio più alto), indicando ProductId e Score. Il risultato deve essere ordinato in base allo UserId.
3. [Facoltativo] Un job in grado di generare coppie di utenti con gusti affini, dove due utenti hanno gusti affini se hanno recensito con score superiore o uguale a 4 almeno tre prodotti in comune, indicando le coppie di utenti e i prodotti recensiti che condividono Il risultato deve essere ordinato in base allo UserId 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.
- Il relativo codice completo MapReduce e Spark (da allegare al documento)
- Un test di uso con logs e file di output (da allegare)

Tutte le specifiche non definite in questo documento possono essere scelte liberamente. Consegnare il rapporto **entro il 18 maggio 2017** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: http://moodle3.ing.uniroma3.it/.