**CORSO DI BIG DATA**
**Primo Progetto**
21 aprile 2022

Si consideri il dataset **Amazon Fine Food Reviews** di Kaggle[1], che contiene circa 500.000 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).

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, per ciascun anno, le dieci parole che sono state più usate nelle recensioni (campo text) in ordine di frequenza, indicando, per ogni parola, il numero di occorrenze della parola nell'anno.
2. Un job che sia in grado di generare, per ciascun utente, i prodotti preferiti (ovvero quelli che ha recensito con il punteggio più alto) fino a un massimo di 5, indicando ProductId e Score. Il risultato deve essere ordinato in base allo UserId.
3. 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 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 dei vari job con dimensioni variabili dell'input[2].
- Il relativo codice completo MapReduce e Spark (da allegare al documento)
- Un test di uso con logs e file di output (da allegare)
- [Facoltativo] Eseguire i vari job su un cluster a propria scelta (per esempio su dataproc di Google (https://cloud.google.com/dataproc) che offre 300$ di credito gratuito per 90 giorni) e confrontare i tempi con l'esecuzione in locale di uno o più dei job realizzati.

Tutte le specifiche non definite in questo documento possono essere scelte liberamente. Consegnare il rapporto **entro il 20 maggio 2022** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: https://ingegneria.el.uniroma3.it/course/view.php?id=386.

[1] https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
[2] Per aumentare le dimensioni dell'input si suggerisce di generare copie del file dato, eventualmente alterando alcuni dati.