CORSO DI BIG DATA
Primo Progetto
26 aprile 2023

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 preparato opportunamente il dataset (per esempio eliminando dati errati o non significativi), progettare e realizzare in MapReduce, Hive, Spark almeno due delle seguenti applicazioni:

1. Un job che sia in grado di generare, per ciascun anno, i 10 prodotti che hanno ricevuto il maggior numero di recensioni e, per ciascuno di essi, le 5 parole con almeno 4 caratteri più frequentemente usate nelle recensioni (campo text), indicando, per ogni parola, il numero di occorrenze della parola.

2. Un job che sia in grado di generare una lista di utenti ordinata sulla base del loro apprezzamento, dove l'apprezzamento di ogni utente è ottenuto dalla media dell'utilità (rapporto tra HelpfulnessNumerator e HelpfulnessDenominator) delle recensioni che hanno scritto, indicando per ogni utente il loro apprezzamento.

3. Un job in grado di generare gruppi di utenti con gusti affini, dove gli utenti hanno gusti affini se hanno recensito con score superiore o uguale a 4 almeno 3 prodotti in comune, indicando, per ciascun gruppo, i prodotti condivisi. Il risultato deve essere ordinato in base allo UserId del primo elemento del gruppo e non devono essere presenti duplicati.

Per ciascun job bisogna illustrare e documentare in un rapporto finale:

- Le operazioni di preparazione dei dati che sono state eventualmente effettuate
- Una possibile implementazione dei job sopra indicati in MapReduce (pseudocodice), Hive, Spark core (pseudocodice) e SparkSQL.
- Le prime 10 righe dei risultati dei vari job.
- Tabella e grafici di confronto dei tempi di esecuzione in locale e su cluster dei vari job con dimensioni crescenti dell'input[2].
- 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 25 maggio 2023** in un unico file compresso di formato a piacere sul sito moodle del corso disponibile all'indirizzo: https://ingegneriacivileinformaticatecnologieaeronautiche.el.uniroma3.it/course/view.php?id=1338.

[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.