RICERCA DI INFORMAZIONI CHIAVE UTILIZZANDO MAPREDUCE

AUTORE: ANGELO ALEX CAMMAROTA (336494)

INTRODUZIONE

Il dataset utilizzato contiene una collezione di pazienti con relative caratteristiche cliniche (5111 pazienti).

Il dataset, in formato csv, è scaricabile al seguente idirizzo:
https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Nel progetto, per simulare un approccio classico ai big data, non vengono prefiltrati i dati ma viene caricato l'intero dataset e se si dovessero riscontrare dei problemi, i dati responsabili verranno scartati. In seguito saranno effettuate delle operazioni di analisi sui pazienti presenti nel dataset.

Per il progetto è stata utilizzata una macchina virtuale Ubuntu con le seguenti caratteristiche: 4096 MB di memoria primaria, 32.97GB di storage e 2 processori. Le specifiche riportate sono più che abbondanti, vista la ridotta dimensione del dataset.

DATAFLOW E TECNOLOGIE UTILIZZATE

Per eseguire le operazioni di analisi è stato utilizzato Hadoop MapReduce. Il modello di programmazione MapReduce ci consente di modellare in maniera semplice ed efficace le nostre richieste.

Il dataflow è il seguente:

- Il file csv in input viene copiato dal file system locale all'hadoop distributed file system (hdfs);

- L'input viene diviso in split da massimo 128 MB (nel nostro caso abbiamo un solo split, visto che le dimensioni dell'input sono circa 310kB);

- Utilizzando apposite interfacce il csv viene diviso per righe che vengono assegnate al mapper nella forma (chiave, valore);

- Il Map Task carica questi dati in RAM e saranno processati tramite il codice del mapper ottenendo coppie (chiave, valore) intermedie. In questo momento le coppie si troveranno sul disco locale della macchina;

- Fase di Shuffle e Sort;

- Il Reduce Task carica le coppie intermedie in RAM dal disco locale e saranno processate tramite il codice del reducer ottenendo l'output che verrà stampato come testo in un file in hdfs;

- Il file di output viene, quindi, copiato dall'hdfs al file system locale.

Le operazioni possono richiedere job multipli, quindi avverrà il passaggio per file di output intermedi che vengono copiati nel file sistem locale in apposite cartelle. Nel nostro caso questo succederà eseguendo lo script "**run-app2 sh**" che richiederà l'esecuzione di 2 round.

CASI D'USO

Per utilizzare il sistema sviluppato è necessario copiare nella cartella "**hadoop-2.10.0**" il contenuto della cartella "**Cammarota_Angelo_Alex_336494**", composto da:

- app-input

- app-cammarota.jar

- app2-cammarota.jar

- run-app.sh

- run-app2.sh

- start-hadoop.sh

- stop-hadoop.sh

Assicurarsi di avere nella cartella "**hadoop-2.10.0**" una cartella denominata "**output**".

Aprire la shell di Linux e spostarsi sulla cartella "**hadoop-2.10.0**" tramite il comando

"**cd hadoop-2.10.0**".

Eseguire successivamente lo script "**sh start-hadoop.sh**".

Eseguire poi il programma MapReduce tramite il comando "**sh run-app.sh**" se si vuole aggregare simultaneamente per sesso del paziente ed età del paziente il numero di persone soggette ad ictus.

Se si vuole invece aggregare simultanemente per sesso del paziente ed età del paziente il

numero di persone soggette ad ictus (primo step) e successivamente raggruppare per numero (secondo step) eseguire il comando "**sh run-app2.sh**".

I file di output saranno consultabili nella cartella "**output**". Inoltre le coppie chiave-valore finali saranno stampate a schermo.

Eseguire infine il comando "**sh stop-hadoop.sh**".

LIMITI E POSSIBILI ESTENSIONI

L'insieme delle funzioni di analisi messe a disposizione è limitato e potrebbe essere ampliato, accompagnato inoltre da una interfaccia grafica per migliorare il programma.

Inoltre, bisogna considerare, che nonostate tramite il paradigma MapReduce possiamo modellare agevolmente le richieste, a causa delle dimensioni ridotte del dataset non si stanno sfruttando a pieno le funzionalità del modello e dei suoi principali punti di forza come la parallelizzazione dei task.

Le applicazioni sviluppate possono essere già utilizzate con dataset relativi ad altri pazienti, a patto che venga rispettato il formato dei dati in input.

In questi script, l'estrazione dei dati avviene in modalità hard coded, se si volesse rendere il codice parametrico si potrrebbe utilizzare un file di configurazione.

Inoltre, c'è da considerare che nella macchina utilizzata, vi è una installazione Psudocluster di Hadoop che essendo sicuramente più realistica di una installazione locale, non presenta chiaramente la complessità di una installazione distribuita vera e propria con più macchine dove ogni macchina viene configurata per avere uno specifico ruolo.