Progetto 1 Sistemi e Architetture per Big Data

Analisi di dataset Covid-19 con Framework Spark

Davide Salvadore
Ingegneria dell'Informazione
Università degli studi di Roma "Tor Vergata"
Roma,Italia
salvadore.davide@gmail.com

ABSTRACT

In questo documento sono riportati i dettagli implementativi e architetturali del progetto realizzato con l'utilizzo del framework "Spark" , per l'analisi di due dataset, uno nazionale italiano e uno globale, riguardo le dinamiche dei contagi relative al virus Covid-19.

KEYWORDS

Contagi, Covid-19, Tamponi, Statistiche, Spark, HDFS, AWS, EC2, Amazon EMR, Amazon S3

1 Introduzione

Lo scopo dell'analisi è la computazione di statistiche sulle dinamiche della diffusione del Covid-19 a partire da dataset contenenti informazioni relative alla pandemia del virus su scala nazionale italiana e su scala globale.

Il primo dataset, dpc-covid19-ita-andamento-nazionale, contiene informazioni dettagliate, aggiornate con granularità giornaliera alle ore 18:00 CET a partire dal 24 Febbraio 2020, sulle dinamiche relative al Covid-19 in Italia. In particolare vengono considerati per ogni giorno, espresso nel campo data, i ricoveri con sintomi, terapia intensiva, totale ospedalizzati, isolamento domiciliare, totale positivi, variazione totale positivi, nuovi positivi, dimessi guariti, deceduti, totale casi, tamponi, casi testati, e due campi note note it e note en. I dati relativi ai campi elencati in precedenza sono cumulativi, fatta eccezione per i campi variazione totale positivi e nuovi positivi.

Il secondo dataset time series covid19 confirmed global contiene informazioni sull'andamento del numero dei contagi di Covid-19 confermati ed è aggiornato con granularità giornaliera alle ore 23:59 UTC a partire dal 22 Gennaio 2020. Per ogni riga, nel dataset sono indicati: stato, nazione, latitudine e longitudine e ogni colonna contiene il numero di contagi riscontrati relativi ad una particolare data, espressa nel formato mm/dd/yy.

1.1 Formulazione Query 1

Si richiede, per ogni settimana, di calcolare il numero medio di pazienti dimessi guariti ed il numero di tamponi effettuati

1.2 Formulazione Query2

Si richiede, per ogni continente, di calcolare la media, la deviazione standard, il minimo e il massimo del numero di contagi confermati giornalmente per ogni settimana. Si richiede di considerare solamente i 100 stati più colpiti del dataset utilizzando come discriminante il **Trendline Coefficient**, ricavato effettuando il calcolo della retta di regressione che approssima la tendenza degli incrementi giornalieri

2 Architettura

Per l'architettura sono state proposte due soluzioni.

La prima prevede una JVM installata su un nodo standalone su cui viene eseguito il framework **Spark** il quale preleva i file di input da una directory locale ed salva i file di output in un'altra directory locale. Le

directory di output vengono successivamente copiate su un cluster **HDFS** creato a partire da 4 istanze **Amazon EC2** (1 Namenode, 3 Datanode).

La seconda prevede un cluster **Amazon EMR**, con istanze di tipo "**m5.xlarge**" su cui è avviata l'esecuzione del framework "Spark", il quale preleva i dati di input da un bucket **Amazon S3** e invia i dati di output all'HDFS del cluster EMR.

2.1 Spark

Spark viene eseguito sia su una JVM di un nodo Standalone che su un cluster Amazon EMR. È stato utilizzato solamente lo Spark Core per l'analisi dei dati. È stato utilizzato Spark nella versione 2.4.5 e la gestione delle dipendenze è eseguita attraverso Maven.

2.2 Cluster HDFS

Il cluster HDFS è stato generato a partire da 4 istanze Amazon EC2, un Namenode e tre Datanode, sulle quali è stato installato "Hadoop-2.7.3". Su tutti quanti i nodi sono state effettuate le configurazioni di "core-site.xml", "hdfs-site.xml" ed è stato reso possibile il binding dei Datanode in modo che il Namenode potesse contattare correttamente i Datanode per far partire il cluster all'esecuzione del comando "start-dfs". Il cluster ha il compito di accogliere i dati di output generati dall'esecuzione del framework "Spark". È stato inoltre creato uno script che all'avvio delle istanze del cluster effettua la rimozione delle cartelle di output e l'avvio dell'HDFS stesso.

2.3 Amazon S3

È stato creato un bucket di Amazon S3 chiamato "**sabdprojonebucket**" a cui è stato fornito l'accesso pubblico, così facendo l'applicazione può accedere ad i file di input.

3 Query1

3.1 Pre-processamento del Dataset

In questa fase vengono selezionate solamente le colonne che contengono dati significativi per l'esecuzione della query, ovvero i campi "dismessi guariti" e "tamponi", per creare due RDD uno contenente i dati giornalieri non modificati e uno contenente i dati giornalieri ai quali è stata omessa la prima riga e le date hanno subito uno shift all'indietro di un giorno.

3.2 Esecuzione della query

Dagli RDD ottenuti nella fase di pre-processamento vengono ottenuti i dati relativi agli incrementi giornalieri, sottraendo al dato cumulativo registrato per una certa data quello del giorno precedente. Dall'RDD così ottenuto vengono eseguiti i seguenti step:

- Creazione di un PairRDD tramite associazione delle date ad una determinata "YEAR_WEEK" (ottenuta tramite la classe "Locale" di java), utilizzata come chiave a cui sono stati associati i valori giornalieri registrati per quella specifica settimana.
- I dati sono seguentemente stati raggruppati per settimana così ottenere dei vettori corrispondenti a tutti i dati settimanali registrati.
- I valori così raggruppati vengono utilizzati per computare le statistiche richieste

L'RDD generato al termine del processamento viene salvato con il metodo "saveAsTextFile" generando la cartella di output "query1_results".

4 Query2

4.1 Pre-processamento del Dataset

In questa fase, per sfruttare le funzionalità della classe "Locale" di java, sono state effettuate alcune modifiche al dataset sostituendo al nome della nazione/Stato registrato, il quale per alcune entries non veniva correttamente riconosciuto, il nome esteso oppure il codice ISO a due caratteri (ad esempio "Czechia" è stato sostituito con "Czech Republic" e "North Macedonia" con "MK")

4.2 Esecuzione della query

Vengono eseguiti i seguenti step:

- Il dataset viene ridotto calcolando il "Trendline Coefficient" e ordinandolo in maniera discendente in base ad esso, successivamente vengono presi i primi 100 elementi del dataset per costruire un secondo RDD.
- La computazione viene divisa in due parti generando due RDD:
    - Identificazione del continente, caricando come ulteriore RDD un dataset "country_continent.csv" contenente i Codici ISO a due caratteri di ogni Stato associato al codice del relativo continente.
    - Raggruppamento dei dati per settimana.
- I dati ottenuti nei due RDD vengono messi insieme tramite una join e viene generato un PairRDD con chiave una Tupla contenente i campi < YEAR_WEEK, Continente > e viene effettuato un raggruppamento in base alla chiave.
- I dati così raggruppati vengono utilizzati per computare le statistiche richieste.

L'RDD generato al termine del processamento viene salvato con il metodo "saveAsTextFile" generando la cartella di output "query2_results".

5 Analisi delle prestazioni

L'analisi delle prestazioni è stata effettuata sia su un nodo locale (con 12 GB di RAM e una CPU Intel Core i7 quad-core con frequenza 2,7 GHz) sia su Cluster EMR eseguendo un multi-run per catturare le prestazioni prima e dopo l'istanziazione della JVM da parte di Spark. Di seguito vengono riportare le tabelle relative alle velocità di esecuzione della prima e della seconda query sia su Cluster EMR che su nodo locale.

5.1 Esecuzione locale

Nella tabella sono indicati i tempi di esecuzione in secondi e nel formato "media $\pm$ deviazione standard" delle query 1 e 2 eseguite in locale.

	Cold Start	**Hot Start**
Query1	$1.1826 \pm \mathbf{0.09679}$	$0.0609 \pm \mathbf{0.06976}$
Query2	$1.9248 \pm \mathbf{0.10097}$	$0.1450 \pm \mathbf{0.05865}$

5.2 Esecuzione su cluster EMR

Nella tabella sono indicati i tempi di esecuzione in secondi e nel formato "media $\pm$ deviazione standard" delle query 1 e 2 eseguite sul cluster EMR.

	Cold Start	**Hot Start**
Query1	$6.9618 \pm \mathbf{0.81543}$	$0.1922 \pm \mathbf{0.08573}$
Query2	$9.0306 \pm \mathbf{0.38549}$	$0.5067 \pm \mathbf{0.72773}$

REFERENCES

[1] https://github.com/pcm-dpc/COVID-19/blob/master/dati-andamento-nazionale/dpc-covid19-ita-andamento-nazionale.csv.
[2] https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
[3] https://dev.maxmind.com/geoip/legacy/codes/country_continent/