Sistemi e Architetture per Big Data - AA 2020/2021

Primo progetto

Giuseppe Lasco
Dipartimento di Ingegneria dell'Informazione
Università degli studi di Roma "Tor Vergata"
Roma, Italia
giuseppe.lasco17@gmail.com

Marco Marcucci
Dipartimento di Ingegneria dell'Informazione
Università degli studi di Roma "Tor Vergata"
Roma, Italia
marco.marcucci96@gmail.com

***Abstract*—Questo documento riporta i dettagli implementativi riguardanti l'analisi mediante *Spark* dei dataset contenenti informazioni relative all'andamento nazionale italiano dei vaccini effettuati. Viene, inoltre, descritta l'architettura a supporto dell'analisi e gli ulteriori *framework* utilizzati.**

I. Introduzione

L'analisi effettuata si pone lo scopo di valutare delle statistiche relative ai vaccini contro il COVID-19, su dati resi disponibili dal Commissario straordinario per l'semergenza Covid-19, Presidenza del Consiglio dei Ministri.

Dataset

Il primo file preso in considerazione è *punti-somministrazione-tipologia.csv*, il quale contiene dati sui punti di somministrazione per ciascuna Regione e Provincia Autonoma.

Il secondo file preso in considerazione è *somministrazioni-vaccini-latest.csv*, il quale contiene dati sulle somministrazioni giornaliere dei vaccini suddivisi per regioni, fasce detà e categorie di appartenenza dei soggetti vaccinati. Tale dataset risulta ordinato per data, inoltre è stata riscontrata l'assenza di numerose tuple relative a delle specifiche regioni, fasce d'etaà e mesi. Questo fenomeno ha reso necessario un intervento di preprocessamento utile a inserire date mancanti per rendere più accurato il lavoro di regressione sui dati, sotto l'assunzione che i dati mancanti fossero dovuti all'assenza di vaccinazioni in un determinato giorno.

Il terzo file preso in considerazione è *somministrazioni-vaccini-summary-latest.csv*, il quale contiene dati sul totale delle somministrazioni giornaliere per regioni e categorie di appartenenza dei soggetti vaccinati. Il dataset in questione risulta, invece, non ordinato, per cui si è reso necessario un effort di preprocessamento al fine di ordinarlo.

L'ultimo file preso in considerazione è *totale-popolazione.csv*, che tiene traccia della popolazione totale residente in una data Regione o Provincia Autonoma.

Query

L'obiettivo di questo progetto è quello di implementare ed eseguire tre query utilizzando *Spark*.

La prima query ha come scopo quello di calcolare il numero medio di vaccinazioni giornaliere in ciascun centro di ciascuna area.

La seconda consiste nel determinare le prime 5 aree per le quali previsto il maggior numero di vaccinazioni il primo giorno del mese successivo per le donne, per ogni fascia anagrafica e per ogni mese solare. A tale scopo si utilizza una retta di regressione, addestrata sui dati relativi al mese precedente a quello per cui viene fatta la predizione al primo giorno. I dati presi in considerazione partono dal 1 Febbraio 2021.

L'ultima query prevede di effettuare una previsione della percentuale totale delle somministrazioni dei vaccini al 1 Giugno 2021 per ogni regione, utilizzando tutti i dati relativi ai mesi precedenti, a partire dal 27 Dicembre 2020. Inoltre, vengono utilizzati due algoritmi di clustering in grado di raggruppare le Regioni in base alla previsione sopra citata.

Framework

Il progetto prevede l'utilizzo di alcuni *framework* che permettono di rendere la computazione parallela e distribuita. Come *framework* di processamento batch è stato utilizzato *Apache Spark* che comunica con lo storage distribuito *Hadoop Distributed File System*. Per la raccolta dei risultati è stato impiegato *HBase*, uno storage No-SQL column family. Infine, come *framework* di data ingestion è stato utilizzato *NiFi*.

II. Architettura

L'architettura si compone di un insieme di container *Docker*, su cui eseguono i servizi introdotti precedentemente. Inoltre, sempre sulla stessa macchina, una JVM ospita l'esecuzione di *Apache Spark*. I container comunicano attraverso la stessa rete, creata appositamente.

NiFi

NiFi è il servizio che permette di recuperare i dataset in formato *comma separated value* da *GitHub*, trasformarli in formato *parquet* e inviarli al servizio di storage distribuito *HDFS*. Tale *framework* è stato istanziato su container *Docker* utilizzando l'immagine *apache/nifi*. L'uso di *parquet* ha permesso di comprimere i dati migliorando le prestazioni in termini di occupazione di memoria. ***colonne***

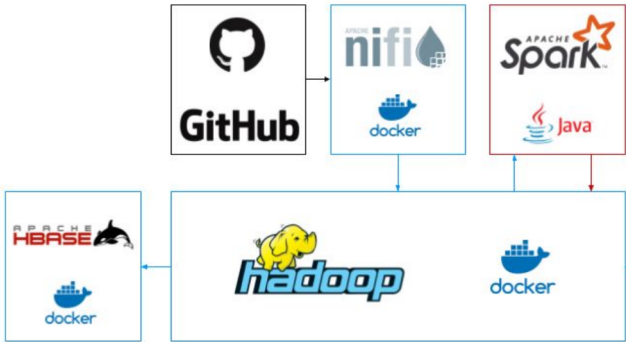

Fig. 1. Schema dell'architettura

Al fine di eseguire le operazioni elencate, sono stati impiegati due *processori*, uno che permette di collegarsi al servizio di hosting *GitHub* e scaricare i dati e uno che permette la trasformazione in *parquet* di questi ultimi e l'upload su *HDFS*. La struttura è definita mediante il template in figura 2.

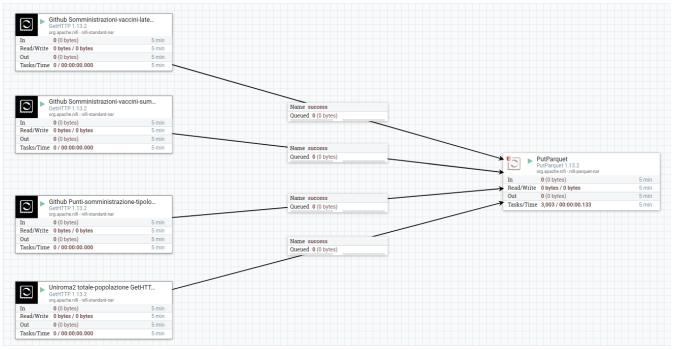

Fig. 2. NiFi template

HDFS

HDFS rappresenta il mezzo che permette l'archiviazione dei dati in maniera distribuita. Il servizio si compone di un nodo *master* e tre nodi *worker* con un livello di replicazione pari a 2. Tale servizio permette di rendere disponibili i dati su cui *Spark* esegue la computazione e memorizza gli *output* dell'analisi, che vengono, in seguito, esportati su *HBase*, per eventuali analisi, manipolazione e rappresentazione dei dati. Il deployment del framework avviene attraverso l'utilizzo dell'immagine *Docker effeerre/hadoop*, istanziata su container. In seguito all'avvio del servizio, uno script permette di eseguire lo sturtup del *Namenode* e dei *Datanode*, e crea le directory `/data`, dove *NiFi* inserisce i dati, e `/output`, in cui risiedono i risultati dell'analisi, concedendo i permessi di lettura, scrittura ed esecuzione.

Spark

Al fine di preocessare i dati ed eseguire le *query*, viene utilizzato *Apache Spark* in locale, tramite lo script `$SPARK_HOME/bin/spark-submit`. Oltre allo *Spark Core*, che espone un set di API di *trasformazioni* ed *azioni*, è stata impiegata la libreria di *Machine Learning MLlib*, utile per effettuare *clustering* sui risultati della terza *query*.

HBase

Hbase è stato utilizzato come datastore *NoSQL* sul quale importare i risultati delle query, anche questo servizio è stato istanziato utilizzando un container *Docker* realizzato, questa volta, a partire dall'immagine *harisekhon/hbase*. Affinchè fosse possibile l'esportazione dei risultati da *HDFS* a *HBase*, è stata creata la classe `HBaseQueries.java` che permette la creazione delle tabelle e l'inserimento dei dati, sfruttando la classe `HBaseClient.java`. Quest'ultima contiene informazioni riguardo la configurazione di *HBase* e *Zookeeper*, e le principali operazioni di gestione del datastore.

Lista:

0 ewf
0 ef

III. QUERY

Query 1

Al fine di soddisfare la seguente *query*, si è reso necessario l'utilizzo di due file, *somministrazioni-vaccini-summary-latest.parquet* e *punti-somministrazione-tipologia.parquet*.

Tali file sono stati caricati in `Dataset` e trasformati in `JavaPairRDD`, considerando le sole colonne di interesse: *data_somministrazione*, *area* e *totale* per il primo e *area* per il secondo.

È stato effettuato un'ordinamento dell'`RDD` *somministrazioni-vaccini-summary-latest* in base alla data, scartando i dati precedenti al 1 Gennaio 2021 e successivi al 31 Maggio 2021, utilizzando la *trasformazione* di *filter*. Un'*azione* di *reduceByKey* ha permesso di ottenere il totale di vaccinazioni per ogni mese. Utilizzando un approccio simile al *word count*, sono stati contati i centri riferiti ad una determinata Regione relativi all'`RDD` *punti-somministrazione-tipologia*. La *join* ha permesso di unire i due `RDD`, utilizzando come chiave la Regione. Il risultato finale è stato ottenuto dividendo il totale per il numero di giorni del mese di riferimento e per il numero di centri della regione di riferimento, ordinando, infine, il risultato in termini di mese e regione.

Query 2

Relativamente alla seconda *query* è stato utilizzato il file *somministrazioni-vaccini-latest.parquet*, il quale, in seguito al caricamento da *HDFS*, è stato trasformato in `JavaPairRDD`. Durante questa fase sono state scartate le colonne irrilevanti ai fini della richiesta. La *trasformazione "filter"* ha permesso l'eliminazione delle entry relative a date precedenti al 1 Febbraio 2021 e successivi al 1 Giugno 2021. Considerando come chiave la tupla *area, data e fascia anagrafica* sono stati sommati i vaccini relativi ad aziende farmaceutiche differenti. La *trasformazione "groupByKey"* è stata applicata al fine di raggruppare tutte le tuple *data, numero somministrazioni giornaliere* relative ad una certa regione e fascia anagrafica. Per ogni mese è stato eseguito una operazione di inserimento

di date e valori macanti ed è stata effettuata *regressione lineare* per ogni mese, in modo da prevedere il numero di donne vaccinate al primo giorno del mese successivo. Il modello di regressione lineare é stato addestrato attraverso l'implementazione fornita dalla libreria di regressione di `Apache Commons`.

Query 3

L'ultima *query* fa uso dei dati presenti nei file *somministrazioni-vaccini-summary-latest.parquet* e *totale-popolazione.parquet*. In seguito si è passati al caricamento dei file e alla trasformazione in `JavaPairRDD`. Sui dati relativi a *somministrazioni-vaccini-summary-latest* si è proceduto al raggruppamento delle tuple *data, numero somministrazioni giornaliere* per ogni regione, questa operazione ha permesso di svolgere regressione lineare su tutti i giorni dal 27 Dicembre 2020 al 31 Maggio 2021, in modo da prevedere il numero di vaccini effettuati in data 1 Giugno 2021. Una *reduceByKey* sulle regioni ha, invece, permesso di calcolare il totale di vaccini effettuati dal 27 Dicembre 2020 al 31 Maggio 2021. Infine, l'operazione di somma tra le proiezioni e il totale calcolato, ha decretato il numero totale previsto di vaccinati per regione al 1 Giugno 2021. Il *join* tra l'*RDD* in questione e quello contenente il numero totale di abitanti residenti in ciacuna regione, he reso possibile clcolare la percentuale prevista di vaccinati al 1 Giugno 2021. Utilizzando due algoritmi presenti in *MLLib*, è stato effettuato *clustering* utilizzando i risultati precedenti come dataset, con numero di cluster variabile da 2 a 5. Gli algoritmi utilizzati sono *K-means* e *Bisecting K-means*, mentre per la regressione è stata utilizzata l'implementazione fornita dalla libreria di regressione di `Apache Commons`.

IV. BENCHMARK

L'esecuzione del progetto e la valutazione delle prestazioni sono state eseguite su Linux Mint 19.3 Cinnamon, Intel Core i7-8750H CPU, 6 core, 12 thread e 32 GB di RAM, con archiviazione su HDD.

In tabella I sono riportati i tempi di processamento delle query. Sono stati considerate le performance di puro processamento, ovvero non considerando i tempi di sturtup della *Java Virtual Machine* su cui *Spark* opera, il caricamento dei file dall'*HDFS* e la scrittura dei risultati. Come si può notare, la query 3 risulta molto più lenta delle altre due, che, invece, mostrano risultati comparabili. Tale evidenza è causata dall'inclusione, nel totale, dei tempi di addestramento degli algoritmi di *clustering*, che possono essere osservati in tabella II.

TABLE I
TEMPI ESECUZIONE QUERY

Query	Media	Varianza
Query 1	11C	22C
Query 2	9C	19C
Query 3	10C	21C

TABLE II
TEMPI ESECUZIONE CLUSTERING

Numero cluster	Modello			
	K-means		Bisecting K-means	
	Media	Varianza	Media	Varianza

TABLE III
COSTO CLUSTERING

Numero cluster	Modello	
	K-means	Bisecting K-means

V. PREPARE YOUR PAPER BEFORE STYLING

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections V-A–V-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert

TABLE IV
TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	*Subhead*	*Subhead*
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 3. Example of a figure caption.

figures and tables after they are cited in the text. Use the abbreviation "Fig. 3", even at the beginning of a sentence.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.