|  |
| --- |
| Immagine correlata  Corso di laurea magistrale in ingegneria informatica indirizzo Big Data  A.A. 2021/2022 |

**ANALISI DI DATI DI CONTABILITÀ AZIENDALE**



|  |  |
| --- | --- |
| Nome dello studente: | Francesco Patanè |
| Matricola | *530HHHINGINFOR* |
| Data Appello | *Gennaio 2022* |
| Esame | *Introduzione Big Data* |

Indice

|  |  |
| --- | --- |
| Introduzione al lavoro svolto | Pag 3 |
| Calcolo media e varianza di vendita | Pag 6 |
| Calcolo mese di ogni anno con maggiore e minore vendita | Pag 11 |
| Conclusioni | Pag 15 |
| Riferimento bibliografici, Sitografia | Pag 16 |
| Appendice | Pag 17 |

**Introduzione al lavoro svolto**

In questa trattazione si illustrerà il processo di analisi dei dati contabili di una ditta, alla quale ci si riferirà con il nome di “Steccaparapetutti s.r.l”, allo scopo di estrapolare dai dati delle informazioni generali sul suo andamento di vendita negli ultimi anni.

I dati messi a disposizione da Steccaparapetutti consistono in un registro degli ordini effettuati a partire dal gennaio 2016 fino all’agosto 2020, per un totale di 5600 ordini messi a disposizione tramite un apposito file csv.

Il file è strutturato nela forma: **Tipo-documento,data(aaaammgg),costo(€)**. Di seguito sono riportate alcune righe come esempio:

1. FATTURA, 20160104,139.8
2. FATTURA, 20160104,169.2
3. FATTURA, 20160104,65.3
4. …

Vengono trattati 8 tipi di ordini:

1. FATTURA
2. RICEVUTA
3. NOTA.DI.CREDITO
4. DDT
5. OFFERTA
6. PREVENTIVO
7. INVENTARIO
8. BUONO.PRELIEVO

Le informazioni che Steccaparapetutti desidera estrapolare sono le seguenti:

1. media di vendita per ogni mese di ogni anno
2. varianza di vendita per ogni mese di ogni anno
3. mese di ogni anno con maggiore vendita
4. mese di ogni anno con minore vendita

Allo scopo di soddisfare le richieste, i dati verranno analizzati tramite script realizzati in linguaggio R, un linguaggio di programmazione specifico per l'analisi statistica dei dati. Successivamente, per ottenere conferma della bontà e correttezza dei risultati ottenuti, la stessa analisi verrà svolta nuovamente utilizzando un approccio differente: il linguaggio di programmazione sarà di nuovo R, ma utilizzato secondo il paradigma MapReduce su piattaforma Hadoop File System, simulando un approccio di calcolo distribuito.

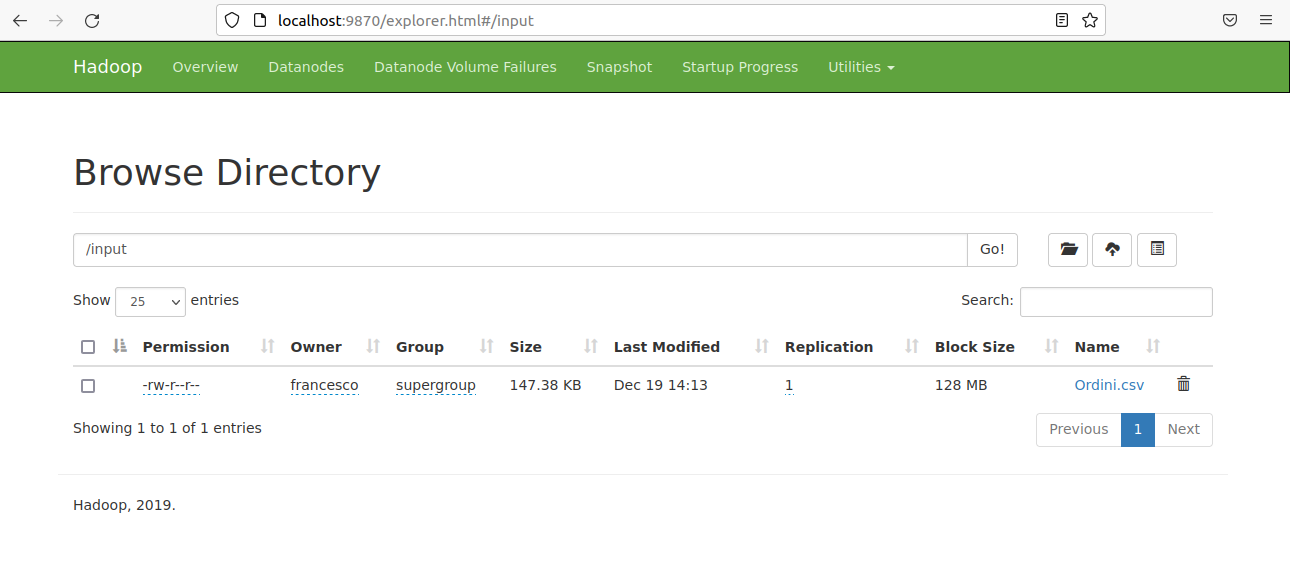
**Note sugli strumenti tecnici utilizzati**

Come detto nel paragrafo precedente, gli strumenti utilizzati saranno il linguaggio di programmazione R e Hadoop file system per l’implementazione del paradigma MapReduce.

Per quanto riguarda R, la versione adoperata è stata la 3.3.1. Non è stato necessario adottare configurazioni particolari. Il codice è stato organizzato in script .r, da eseguire da linea di comando grazie al tool Rscript incluso nella distribuzione.

Per quanto riguarda Hadoop, la versione adoperata è stata la 3.2.1, con una installazione in ambiente locale. Per l’integrazione con R è stato utilizzato il tool hadoop-streaming della versione corrispondente.

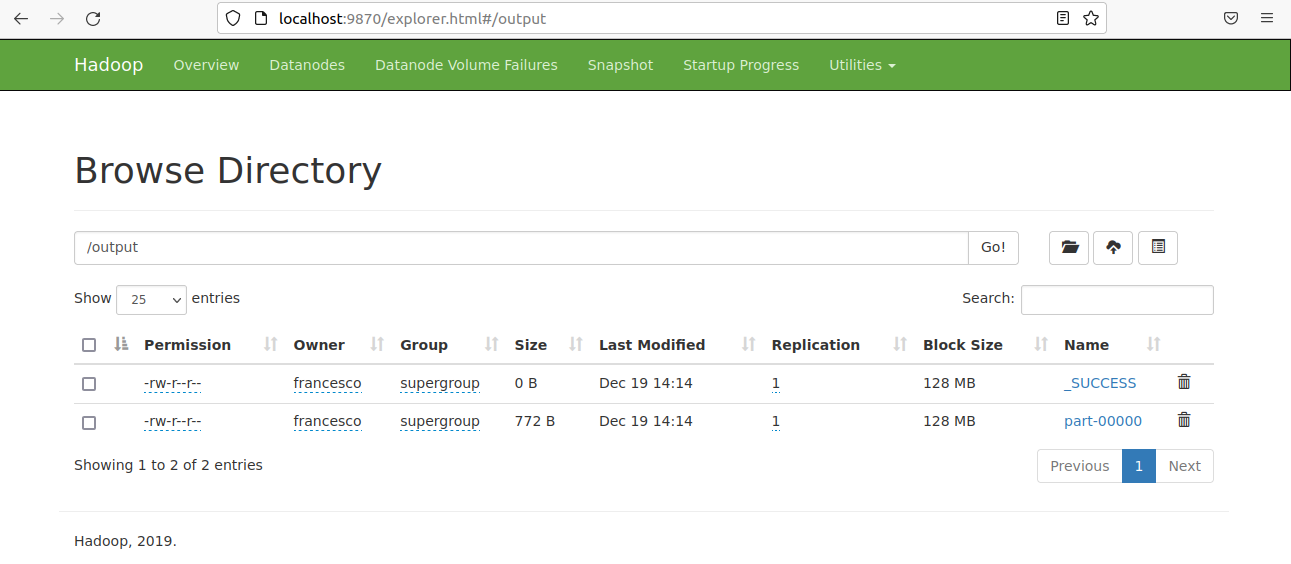
L’hdfs è stato organizzato creando una cartella input contenente il file csv contenente i dati aziendali.



Per eseguire il codice R su hadoop, è necessario porre il jar di streaming nella stessa directori dei file di map e reduce. Avviando l’esecuzione con

*hadoop jar hadoop-streaming-3.2.1.jar -input /input -output /output -mapper map.R -reducer reduce.R*

viene creata una cartella output sul dfs contenente l’output del programma sotto forma di file di testo.



**Calcolo media e varianza di vendita**

**Calcolo media**

Il primo problema affrontato è quello del calcolo della media di vendita per ogni mese di ogni anno.

In sostanza, bisognerà calcolare la media per ogni mese di ognuno degli anni coperti dal registro contabile fornito. Per il mese di gennaio riporteremo la media delle vendite relative a gennaio 2016, 2017, 2018, 2019 e 2020, e così via per ogni mese.

In questo primo approccio all’analisi si pone una domanda, che sarà rilevante anche per tutte le analisi successive: cosa definire come vendita.

Come accennato nella introduzione, i dati forniti riguardano diverse tipologie di ordini: fattura, ricevuta, nota di credito, ddt, offerta, preventivo, inventario e buono prelievo. Esaminando le varie tipologie, possiamo concludere che le occorrenze relative a vendite sono, ovviamente, le fatture, a cui possiamo aggiungere i buoni prelievo, ovvero buoni che il cliente acquista e che può riutilizzare per effettuare ulteriori compere in un secondo momento.

Esaminiamo la struttura dei dati. E’ evidente come per ogni occorrenza, la prima sezione, contenente il tipo di ordine, è il discriminante per le occorrenze da considerare o meno nell’analisi.

**FATTURA**, 20160104,139.8

**OFFERTA**,20180518,712.36

**BUONO.PRELIEVO**,20190103,10.39

L’informazione utile al raggruppamento temporale, ovvero l’identificativo del mese relativamente all’anno, è invece contenuto nella seconda sezione, precisamente nei suoi primi 6 caratteri.

FATTURA, **20160104**,139.8

La terza sezione contiene il valore monetario relativo alla vendita, è il valore sul quale effettuare i nostri calcoli.

Identificate tutte le informazioni necessarie, possiamo passare allo svolgimento pratico dell’analisi, come già accennato nell’introduzione svolta prima programmaticamente tramite script R e poi svolta nuovamente secondo paradigma Map-Reduce in modo da avere una controprova dei risultati ottenuti.

**Svolgimento in R**

Per lo svolgimento pratico dell’analisi in R, è stato implementato un algoritmo secondo i passaggi elencati di seguito:

1. Importazione del file di ordini in memoria come oggetto data frame.
2. Filtraggio del data frame mantenendo solo le righe corrispondenti a fatture o buoni prelievo.
3. Ridurre i valori data alla sola componente aaaMM, in base alla quale poi aggregare.
4. Calcolare la media di vendita come aggregazione in base alla componente aaaMM.
5. Esportazione dei risultati ottenuti in forma csv.
6. Generare una rappresentazione grafica dei risultati.

Lo script nella sua interezza è riportato in appendice.

Di seguito i risultati ottenuti, in forma testuale e grafica:

Mese Media vendite (€) Mese Media vendite(€)

201601 179.93 201812 220.09

201602 131.92 201901 173.09

201603 222.56 201902 196.26

201604 182.44 201903 122.9

201605 200.41 201904 230.5

201606 143.58 201905 230.66

201607 185.63 201906 217.89

201608 66.67 201907 108.2

201609 119.33 201908 170.93

201610 174.44 201909 137

201611 256.15 201910 114.12

201612 250.21 201911 140.53

201701 194.52 201912 97.02

201702 137.06 202001 108.09

201703 174.73 202002 91.74

201704 274.47 202003 87.97

201705 157.32 202004 108.19

201706 370.18 202005 116.23

201707 274.78 202006 114.34

201708 120.59 202007 148.93

201709 157.89 202008 171.31

201710 206.53

201711 96.63

201712 282.65

201801 144.47

201802 152.43

201803 190.62

201804 248.88

201805 158.39

201806 176.2

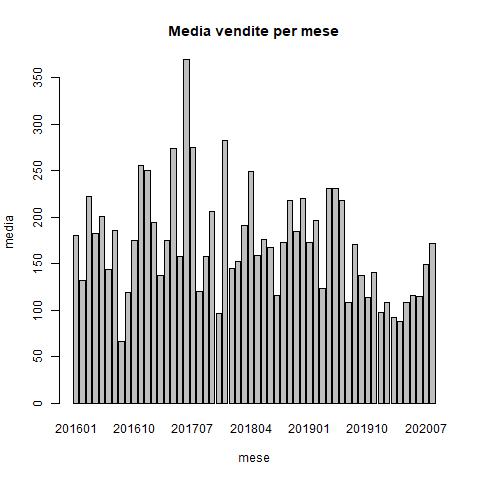
201807 167.15

201808 115.29

201809 173.22

201810 217.67

201811 184.74



Tramite la rappresentazione grafica si può estrapolare in modo immediato un’idea dell’andamento delle vendite nel periodo temporale in esame.

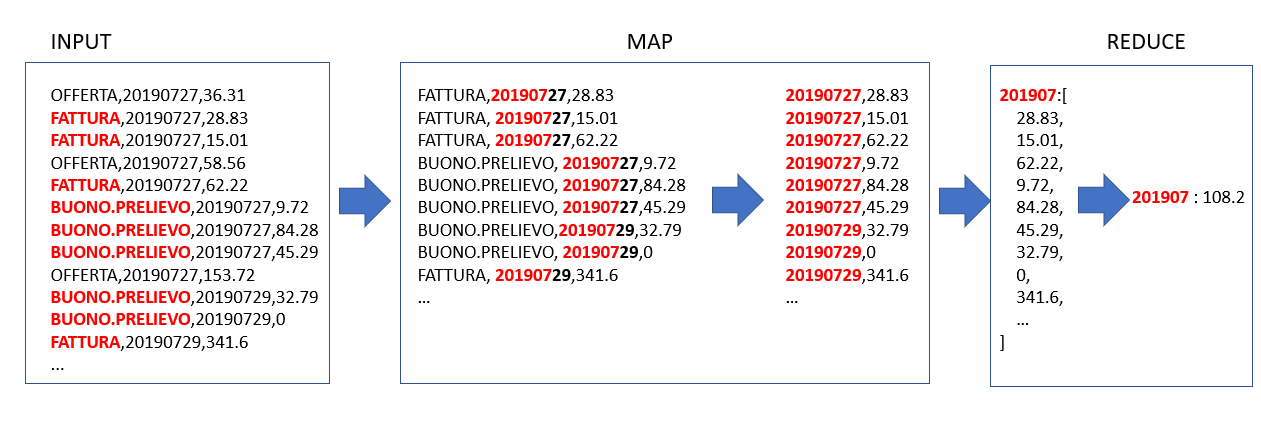
Nel prossimo paragrafo illustreremo lo svolgimento della stessa analisi mediante approccio MapReduce, al fine di valutare la bontà dei risultati ottenuti.

**Svolgimento in MapReduce**

Per l’analisi della problematica secondo il paradigma MapReduce, la questione più immediata è quella di identificare un’opportuna ripartizione del calcolo tra le fasi di Map e di Reduce. L’approccio adottato è stato il seguente:

1. Map:
   1. Lettura file.
   2. Filtraggio delle righe includendo solo fatture e buoni prelievo.
   3. Strutturazione dei dati nella forma chiave-valore data-costo.
2. Reduce:
   1. Riduzione della chiave nella struttura aaaaMM.
   2. Aggregazione dei valori di vendita per chiave comune.
   3. Calcolo della media per chiave.

Di seguito una rappresentazione visuale della modalità di operazione sui dati:



Anche in questo caso il codice e i risultati ottenuti sono riportati in appendice.

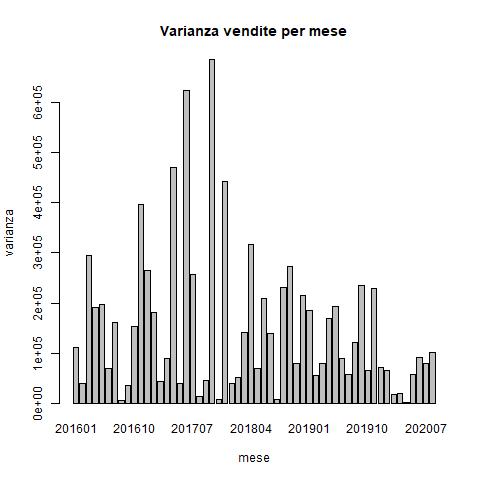
Possiamo constatare che i risultati ottenuti coincidono quelli della precedente analisi. È lecito quindi assumere i risultati ottenuti come veritieri.

**Calcolo varianza**

Il problema del calcolo della varianza per ogni mese di ogni anno è analogo al calcolo della media. Lo svolgimento è speculare, sia per la parte R che MapReduce, con la sola differenza di usare la funzione di calcolo della varianza al posto di quella per il calcolo della media.

Di seguito i risultati ottenuti e una rappresentazione grafica:

Mese Varianza vendite

201601 110972.691204257 

201602 38351.0435100466

201603 295872.781138519

201604 191570.684792203

201605 196337.278810343

201606 69653.8695388889

201607 160827.842888087

201608 5731.73667285714

201609 35471.3577466949

201610 152986.941006066

201611 397335.011873956

201612 264896.325849901

201701 181979.970995667

201702 43393.8605261279

201703 88570.205400847

201704 471101.399361996

201705 39083.5122050641

201706 624458.34956256

201707 256962.885998192

201708 13170.9308356725

201709 44541.9327838972

201710 687600.467134253

201711 7355.77151949495

201712 442718.918099512

201801 38217.6561699608

201802 50629.9956360324

201803 142062.611345507

201804 317879.844685161

201805 69215.7165818027

201806 208360.387216836

201807 139695.276144637

201808 7884.58927205882

201809 231756.571460918

201810 272246.003412605

201811 78327.0505054299

201812 214416.386952273

201901 184759.874082656

201902 54997.569054878

201903 79830.8775473987

201904 169999.102631523

201905 192411.130895038

201906 89277.4086640112

201907 57077.3639738796

201908 120844.375429994

201909 234843.343474732

201910 64392.9493331564

201911 229925.526507311

201912 70895.0544836076

202001 65327.3108314733

202002 18086.2925111623

202003 18976.1629754778

202004 825.39845

202005 56468.2611723219

202006 90620.1698420444

202007 80054.7082105621

202008 101551.251912643

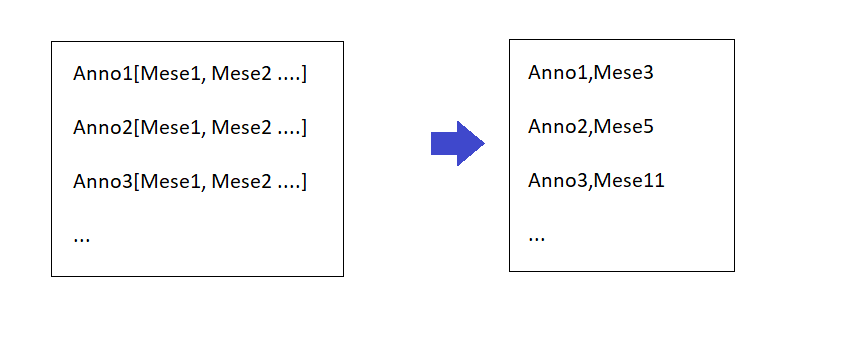
**Calcolo mese di ogni anno con maggiore e minore vendita**

Il secondo problema affrontato è quello dell’individuazione dei mesi con maggiore e minore vendita per ognuno degli anni presi in esame.

Rispetto all’analisi precedente, oltre a filtrare i dati includendo le sole vendite, come definite in precedenza, questa presenta maggiore complessità nel raggruppamento dei dati in base al campo temporale. Non sarà più sufficiente, infatti, isolare semplicemente il mese nella forma aaaaMM, ma bisognerà, dopo aver calcolato il totale delle vendite di ogni mese per ogni anno, raggruppare i vari mesi in base all’anno:

FATTURA, **201601**04,139.8

Una volta effettuato il raggruppamento, bisognerà identificare il mese con vendite totali maggiori o minori per l’anno di riferimento.



Il target dell’analisi sarà ottenere un risultato nella forma:

* Anno1 meseMax/Min
* Anno2 meseMax/Min
* Anno3 meseMax/Min
* Anno4 meseMax/Min
* Anno5 meseMax/Min

Passiamo ora all’approccio pratico al problema.

**Svolgimento in R**

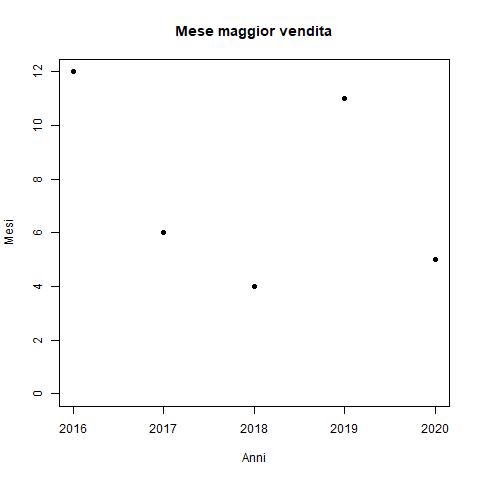
Per lo svolgimento pratico dell’analisi in R, è stato implementato un algoritmo secondo i passaggi elencati di seguito:

1. Importazione del file di ordini in memoria come oggetto data frame.
2. Filtraggio del data frame mantenendo solo le righe corrispondenti a fatture o buoni prelievo.
3. Ridurre i valori data alla sola componente aaaaMM, in base alla quale poi aggregare.
4. Calcolo vendite totali per ogni mese come aggregazione in base alla componente aaaaMM.
5. Divisione del campo temporale da aaaaMM in 2 colonne: aaaa e MM.
6. Calcolo del mese con maggiori/minori vendite per ogni anno.
7. Esportazione dei risultati ottenuti in file txt.
8. Generare una rappresentazione grafica dei risultati.

Lo script nella sua interezza è riportato in appendice.

Di seguito i risultati ottenuti, in forma testuale e grafica.

**Mese con maggiore vendita:**

Anno Mese con maggior vendita

2016 12

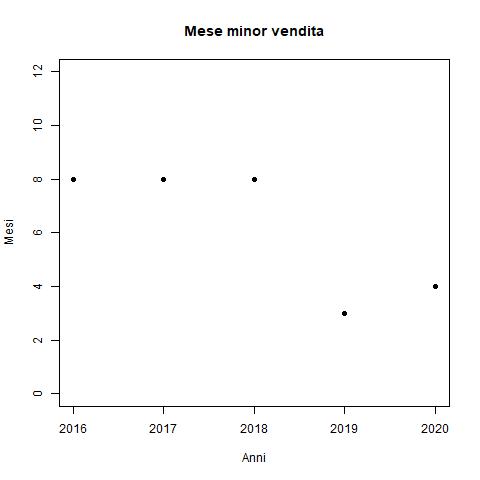
2017 06

2018 04

2019 11

2020 05

**Mese con minore vendita:**

Anno Mese con minor vendita

2016 08

2017 08

2018 08

2019 03

2020 04

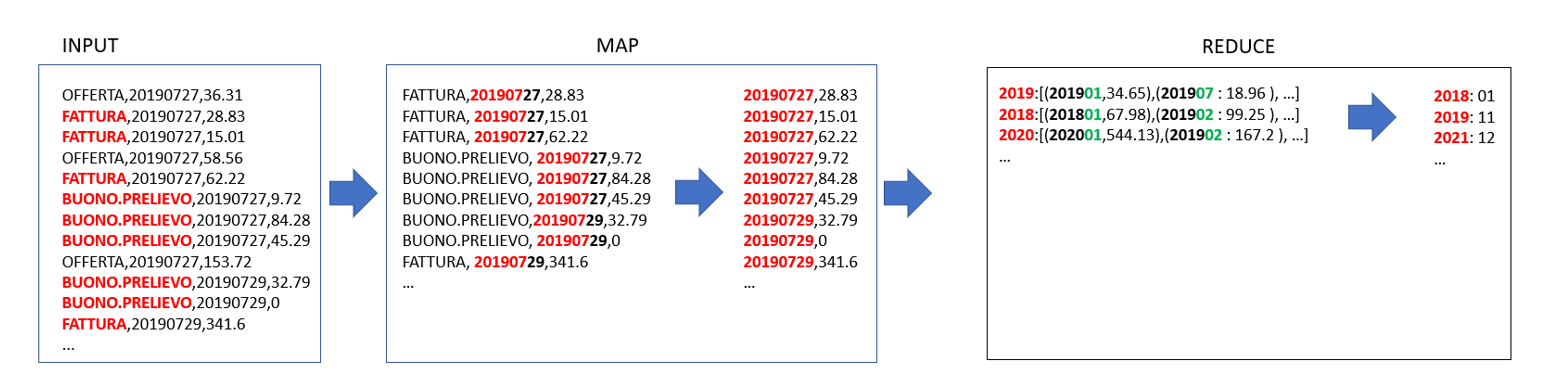
Nel prossimo paragrafo illustreremo lo svolgimento della stessa analisi mediante approccio MapReduce, al fine di valutare la bontà dei risultati ottenuti.

**Svolgimento in MapReduce**

L’approccio adottato per la risoluzione del problema dell’individuazione dei mesi di massima minima vendita di ogni anno secondo paradigma MapReduce è stato il seguente:

1. Map:
   1. Lettura file.
   2. Filtraggio delle righe includendo solo fatture e buoni prelievo.
   3. Strutturazione dei dati nella forma chiave-valore data-costo.
2. Reduce:
   1. Riduzione della chiave nella struttura aaaaMM.
   2. Somma dei valori di vendita per mese comune, nella forma aaaaMM.
   3. Estrapolazione degli anni inclusi nei dati dall’insieme di dati.
   4. Aggregazione insiemistica delle coppie aaaaMM/valore per anno aaa. comune.
   5. Individuazione mese con valore di vendita massimo/minimo per ogni anno.

Di seguito una rappresentazione visuale della modalità di operazione sui dati:



Il codice e i risultati ottenuti sono riportati in appendice.

Possiamo constatare che i risultati ottenuti coincidono quelli della precedente analisi. È lecito quindi assumere i risultati ottenuti come veritieri.

**Conclusioni**

Con questa trattazione è stato illustrato il lavoro svolto, tanto gli approcci e le metodologie adottati quanto i risultati ottenuti.

È stato mostrato come dei dati di contabilità aziendale possano essere analizzati tramite strumenti informatici.

Il linguaggio adottato, R, ha mostrato una grande convenienza grazie alle molte funzioni fornite per immagazzinare e operare agevolmente sui dati, incluse le operazioni di carattere matematico e statistico e non ultima la capacità di creare rapidamente report visuali dei risultati ottenuti, permettendone una comprensione immediata.

Sebbene la quantità di dati da gestire non fosse tale da richiedere una vera elaborazione distribuita, analizzare una seconda volta i dati secondo MapReduce, con il suo particolare focus verso le manipolazioni successive dei dati in forma chiave-valore, si è rivelato un utile processo di verifica dei risultati ottenuti. A livello tecnico, Hadoop si è dimostrato uno strumenti valido, in particolare grazie al tool Hadoop-Streaming che ha permesso un’integrazione del linguaggio R, con grandi vantaggi per le elaborazioni statistiche.

Con i risultati ottenuti, l’azienda può formare un’idea chiara sull’andamento delle vendite nei vari periodi dell’anno, e pianificare le sue strategie di marketing di conseguenza.

**Riferimento bibliografici, Sitografia**

[1] <https://image.freepik.com/free-vector/business-accounting-profit-growth-calculation-data-analysis-analytics-statistics-accountant-bookkeeper-with-laptop-cartoon-character_335657-2973.jpg>

[2] <http://torlone.dia.uniroma3.it/bigdata/materiale.html>

[3] <https://hadoop.apache.org/docs/current/>

[4] <https://cran.r-project.org/doc/manuals/r-release/R-intro.html>

**Appendice**

|  |
| --- |
| Risultati calcolo media |
|  |
| Codice calcolo media |
|  |
| Risultati calcolo varianza |
|  |
| Codice calcolo varianza |
|  |
| Risultati calcolo massimo |
|  |
| Codice calcolo massimo |
|  |
| Risultati calcolo minimo |
|  |
| Codice calcolo minimo |
|  |