INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH AND INNOVATIVE STUDIES
ISSN: 2820-7157
www.ijsrisjournal.com

October 2025 Volume 4 Number 5 178-183

Big Data and decision-making processes in the 5 V dimension

Prof. Dr AHMED ERRHAMANI [1], Prof. Dr GHAZI AZIZ [2], Doctorant Blaise KAPALALA KAPENDA [3], Doctorant MASUDI DEOGRACIAS DEO [4]

Dr Ahmed Errhamani – Sup 'management Maroc, Email:aerrhmani@hotmail.com

Dr. Ghazi Aziz – Sup 'management Maroc, Email: aziz.ghazi40@gmail.com

Mr KAPALALA KAPENDA BLAISE, PhD student – ISP – Gombe Kinshasa, Email:blaisekapalala94@gmail.com Mr MASUDI DEOGRACIAS DEO, PhD Student – ISP – Gombe Kinshasa, Email:graciasmasudi12@gmail.com

Abstract

The high use of computers has created large volumes of data that cannot be managed by traditional software and hardware. Take the case of large companies like Microsoft and Google, which must store, manage and manipulate billions of pieces of data. This perplexity in managing these large volumes of data gave birth to the term Big Data. The potentially infinite amounts of data, as well as the constraints that derive from it, pose many problems in terms of storage and processing of these very large data sets in terms of time and calculation using dedicated platforms such as Hadoop, which is one of the best platforms for Big Data and which is based on the Map Reduce paradigm.

Technological evolution in the 21st century; from day to day the data evolves in time and space, from space to space we talk about Big data and the cluster in the decision-making process in the 5 V dimension, this is the very important research topic that interests many researchers in the unsupervised classification of data, and it is a research axis no less important than the 5 V itself, several indices of data validation have been proposed in the literature review either for external acceptance or internal.

External acceptance requires a priori knowledge of the optimal real partition and is based on the comparison of the big data obtained by a supervised learning algorithm for any KNN classification with the known optimal partition. The most well-known error indices are: purity, measure, entropy, synaptic coefficient, etc.

In this work, we proposed a parallel and distributed model to solve the problem of scaling external validation indices of Big Data, and more precisely the "synaptic coefficient" index. We used the Hadoop platform, and more precisely the Map Reduce paradigm, for the implementation of the proposed model. The results obtained show the validity of this model.

Keywords: Big Data, supervised learning, decision process, clustering, Hadoop, Map Reduce, 5 V dimension, cluster and KNN classification, synaptic coefficient.

INTRODUCTION

In the literature, a multitude of clustering techniques have been developed. These techniques may differ in their principles, properties, parameters and general forms of the partitioning generated. The four main categories of clustering techniques available in the literature are: partitioning techniques, hierarchical techniques, density-based techniques and grid-based techniques.

The search for useful information among enormous amounts of data is known as data mining. This interdisciplinary field draws its techniques and tasks from several other fields. **One task in particular, big data**, **is an exploratory analysis process that divides data into a set of clusters**. These discovered clusters can be used to explain characteristics of the underlying data distribution. **Clustering reduces the data to a smaller set of representatives, allowing for a simplified representation of the initial data.**

The clustering problem has been addressed in many contexts and by many researchers in many disciplines, reflecting its broad appeal and usefulness as one of the steps in exploratory data analysis. Its applications are numerous, including statistics, image processing, artificial intelligence, pattern recognition, web analysis, marketing, medical diagnosis, biology, surgery and many others.

The result of clustering algorithms is measured by validation indices, which can be done in two ways depending on whether or not a known optimal partition is used. In the first case, we refer to external validation, and in the second case, we refer to internal validation. On the other hand, Big Data is a new revolution in the field of computing and relates to data sets that are becoming so large and difficult to manage with traditional database management tools that they require the use of dedicated platforms and tools for managing this data, including the Hadoop platform.

Hadoop consists of two essential components: MapReduce, which is a new programming paradigm on which parallel and distributed calculations of large amounts of data are performed, and HDFS, which is a distributed file management system.

1. Objectives

With the enormous growth in data, its heterogeneity and the frequency with which it is generated, captured and shared, traditional clustering validation indices have become incapable of managing this data and need to be revisited in order to properly manage these large data sets. A number of clustering validation indices have been proposed in the literature. These indices differ in whether they use a known optimal partition or not.
In the first case, we refer to external validation, and in the second case, we refer to internal validation. In our work, we are interested in scaling external clustering validation indices. To this end, ***we propose a parallel and distributed model for one of the clustering validation indices***, the "synaptic coefficient"**,** on a platform dedicated to big data processing, in this case **Hadoop.** This choice is mainly justified by the fact that it allows complex processing and calculations to be performed on very large data sets. We began by proposing a model architecture, then we determined the model's Map and Reduce functions, and finally we implemented it on the Hadoop platform using the Map Reduce Framework in order to validate the proposed model.

2. Problem

With the enormous growth in data, its heterogeneity and the frequency with which it is generated, captured and shared, traditional clustering validation indices have become incapable of managing this data and need to be revisited in order to properly manage these large data sets. A number of clustering validation indices have been proposed in the literature. These indices differ in that they use an optimal partition that is either known a priori or not. In the first case, we refer to external validation, and in the second case, we refer to internal validation.

3. Proposed solutions

In this work, our actions and investigations focus on scaling external clustering validation indices. To this end, we propose a parallel and distributed model for one of the clustering validation indices, the "synaptic coefficient", on a platform dedicated to big data processing, in this case Hadoop. This choice is mainly justified by the fact that it allows complex processing and calculations to be performed on very large data sets.

We began by proposing a model architecture, then determined the model's Map Reduce functions, and finally implemented it on the Hadoop platform.

4. Methodologys

To carry out this work, we used three methodologies as data collection tools, among others:

- ***The Delphi Method:*** This method was useful for organising expert consultations on a specific topic with a significant forward-looking aspect.
- ***The questionnaire method:*** This method allows us to interact directly with the target audience and offers specific advantages in terms of research accuracy. It enabled us to obtain results in real time for quick and easy analysis.
- ***The offline questionnaire method:*** Offline data collection tools work exactly like online tools, with all the features designed for an Internet audience, but with the
- added advantage of being able to collect data in the field. The most interesting feature is that no internet connection is required when you visit the survey site to collect data

with offline tools. This eliminates the use of paper while removing dependence on the internet. We can collect and store data even when you are not connected to the internet.

5. Sample

Virtualisation relies on a system called a hypervisor. There are two types of hypervisors: type 1 hypervisors and type 2 hypervisors.

In either case, to perform virtualisation, your machine must have a hardware configuration that supports **Intel-VT** or **AMD-V** technology. Rest assured, this is a standard feature in machines that have been on the market for several years now. I propose to address this concept of "hypervisor type" in this article. Also available in video format.

Since the data to be injected into our system are partitions of data sets somewhere for which the number of clusters is 15 and the actual optimal result is known.

- Step 1: represents the results of job1 and job2;
- Step 2: concerns the results of job3 and job4;

- Step 3: represents the results of job5 and job6;
- Step 4: finally, the result of job7, which is the result of synaptic coefficients.

The main objective of our experiments was to ensure the feasibility and validity of the model developed. We considered an example and went further with the case of the *synaptic coefficient* index. The first scenario: the result of phase 1 (job1 and job2) differs completely from the actual clustering. The second scenario: the result of phase 2 (job3 and job4) differs partially from the actual clustering.

R_s	C_i
1	1
1	1
1	2
1	2
1	2
1	2
2	1
2	1
2	1
2	2
2	2
2	2
2	2

The result corresponding to this scenario is:

$$R_s \quad C_i$$

R_s : Number of data points belonging to each clusterR_j of the actual clustering.C_i : Number of data points belonging to each clusterC_i of the obtained clustering. We obtained this result:

6. Analysis

We used three containers representing a master node (Namenode) and two slave nodes (Datanodes) respectively. The use of containers guarantees consistency between development environments and will significantly reduce the complexity of machine configuration (in the case of native access) as well as the heavy execution load associated with the use of a virtual machine, which we opted for.

In fact, we have obviously created three containers, but also a network that will allow us to connect and then map the ports of the host machine to those of the container. However, we can launch Hadoop and Yarn. Using the command: ***/start-hadoop.sh,*** we will find ourselves in the namenode shell, and we will be able to manipulate the cluster as we wish.

```
root@hadoop-master:~# ./start-hadoop.sh


Starting namenodes on [hadoop-master]
hadoop-master: Warning: Permanently added 'hadoop-master,172.22.0.2' (ECDSA) to the list of known hosts.
hadoop-master: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-hadoop-master.out
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.22.0.3' (ECDSA) to the list of known hosts.
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.22.0.4' (ECDSA) to the list of known hosts.
hadoop-slave2: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave2.out
hadoop-slave1: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-hadoop-slave1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts.
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-hadoop-master.out


starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-hadoop-master.out
hadoop-slave2: Warning: Permanently added 'hadoop-slave2,172.22.0.4' (ECDSA) to the list of known hosts.
hadoop-slave1: Warning: Permanently added 'hadoop-slave1,172.22.0.3' (ECDSA) to the list of known hosts.
hadoop-slave2: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave2.out
hadoop-slave1: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-hadoop-slave1.out
```

Figure 1 : root@hadoop-master:~#

7. Hadoop

All commands interacting with the Hadoop system begin with **hadoop fs.** The options added are largely inspired by standard Unix commands.

- We created a directory in the HDFS file system. As it is a distributed system, this was very useful for distributing files across the cluster machines.

The command that allowed us to create an "input" directory in HDFS

hadoop fs –mkdir -p input

Error

If for some reason you are unable to create the *input* directory, with a message similar to this: ls: `.': No such file or directory, please build the main user (root) tree as follows:

hadoop fs -mkdir -p /user/root

- We will use the purchases.txt file as input for MapReduce processing. This file is already located in the main directory of our master machine.
- Load the purchases file into the input directory we created: hadoop fs –put purchases.txt input.
- To display the contents of the *input* directory, the command is: hadoop fs –ls input

To display the last lines of the purchases file: hadoop fs -tail input/purchases.txt

We get the following result:

```
root@hadoop-master:~# hadoop fs -tail input/purchases.txt
31      17:59   Norfolk Toys    164.34  MasterCard
2012-12-31      17:59   Chula Vista     Music   380.67  Visa
2012-12-31      17:59   Hialeah Toys    115.21  MasterCard
2012-12-31      17:59   Indianapolis    Men's Clothing  158.28  MasterCard
2012-12-31      17:59   Norfolk Garden  414.09  MasterCard
2012-12-31      17:59   Baltimore       DVDs    467.3   Visa
2012-12-31      17:59   Santa Ana       Video Games     144.73  Visa
2012-12-31      17:59   Gilbert Consumer Electronics    354.66  Discover
2012-12-31      17:59   Memphis Sporting Goods  124.79  Amex
2012-12-31      17:59   Chicago Men's Clothing  386.54  MasterCard
2012-12-31      17:59   Birmingham      CDs     118.04  Cash
2012-12-31      17:59   Las Vegas       Health and Beauty       420.46  Amex
2012-12-31      17:59   Wichita Toys    383.9   Cash
2012-12-31      17:59   Tucson  Pet Supplies    268.39  MasterCard
2012-12-31      17:59   Glendale        Women's Clothing        68.05   Amex
2012-12-31      17:59   Albuquerque     Toys    345.7   MasterCard
2012-12-31      17:59   Rochester       DVDs    399.57  Amex
2012-12-31      17:59   Greensboro      Baby    277.27  Discover
2012-12-31      17:59   Arlington       Women's Clothing        134.95  MasterCard
2012-12-31      17:59   Corpus Christi  DVDs    441.61  Discover
root@hadoop-master:~#
```

Figure 2: ***root@hadoop-master:~#***

8. Launching the job

We launched the job on the ***purchases.txt*** file previously loaded into the HDFS ***input*** directory. Once the job was complete, an *output* directory was created. We obtained the following display:

Figure 3: the ***purchases.txt*** file in the HDFS ***input*** directory

Results

By going to the page: http://localhost:8088, we can monitor our Map Reduce jobs.

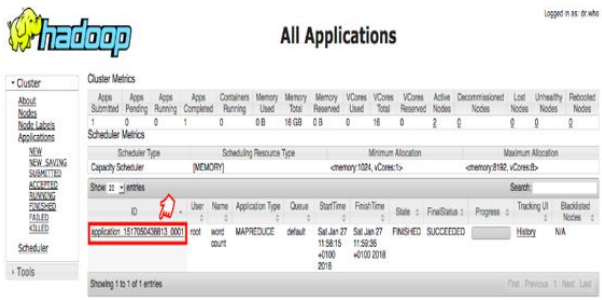

Figure 4: http://localhost:8088

It is also possible to see the behaviour of the slave nodes by going to the address: http://localhost:8041 for *slave1*, and http://localhost:8042 for *slave2*. We obtained the following:

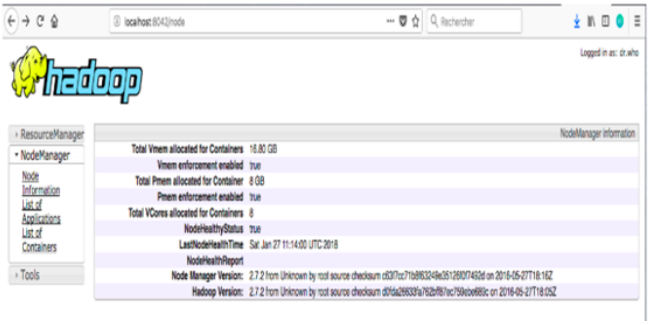

Figure 5: http://localhost:80 42

9. Map Reduce

Map Reduce enabled us to extract the necessary data in key/value form, so that we could then sort it according to the key, which was useful for data processing. We tested our model using an open source example, Wordcount, for data processing applications. This allows us to calculate the number of words in a given file, breaking down the calculation into two steps:

- Split the text into words and deliver a text stream as output, where each line contains the word found, followed by the value 1 (to indicate that the word was found once), *Mapping;*
- Adding up the 1s for each word to find the total number of occurrences of that word in the text, known as *reducing.*

CONCLUSION

We then moved on to similarity measures, and finally, we finished with the many problems and limitations of clustering. Digital data is growing exponentially, yet 90% of the data that currently exists was created in the last two years. Today, we no longer talk about gigabytes, but rather terabytes, petabytes, zettabytes and yottabytes.

This exponential growth can be explained mainly by the increase in internet users worldwide, from 56% in 1990 to 32.7% in 2011. Also, mobile phone usage has increased from 0.21%to 85.5%. Furthermore, the evolution of applications and social networks has played a major role in the creation of this enormous amount of data. However, this huge quantity of data exceeds the capabilities of traditional storage and analysis solutions.

Big Data represents the large amount of data produced and processed by companies. But the most interesting part of Big Data is analytics. We have detailed this concept in the following section of our article. It has given rise to new storage and processing systems and countless technologies. Data clustering is a task whose objective is to find groups within a data set. In this article, we have looked at the main concepts of clustering and the main techniques used in clustering. We have also presented other clustering techniques.

On the other hand, big data analysis involves risks related to privacy, confidentiality and free will, which need to be considered now. In the following section, we will discuss and detail an area that is beginning to revolutionise the world of computing: data clustering. As data becomes increasingly voluminous and complex, our traditional databases are limited in their ability to analyse and process it. With a view to saving time, new technology has emerged to facilitate, relieve and support companies that generate large amounts of data. Big data analysis is undoubtedly set to grow in importance, with some even talking about a technological revolution.

References

Aziz Ghazi, Machine Learning, 4th Year, Fez, Morocco, 2021

Aziz Ghazi, Big Data and Mass Data Processing, Sup 'Management, Fez, Morocco, 2021

Aziz Ghazi, Neural Networks, 4th Year, Fez, Morocco, 2021

Aziz Ghazi, Internet of Things, 4th Year, Fez, Morocco, 2021

Aziz Ghazi, Introduction to Embedded Systems, 4th Year, Fez, Morocco, 2020

Aziz Ghazi, "Deep Learning," 5th Year Fez, Morocco, 2022

Berkhin P "Survey of Clustering Data Mining Techniques" Technical Report, Accrue Software, San Jose, CA, 2002

Bruchez Rudi, "NoSQL Databases and Big Data," Eyrolles: Paris, 2015

Candillier L, Contextualisation, visualisation and evaluation in unsupervised learning, PhD thesis, 2006.

Cleuziou G, "An unsupervised method for rule learning and information retrieval," December 2004.

Constantine, "Model and architecture for validating data clustering in MapReduce," Master's thesis in computer science, 2016.

Emmanuel Faug Director of Business Intelligence Consulting Services at Moment Um Technologies 8 December 2015

HBase, Hadoop's NoSQL database: Concepts & Architecture, 21 September 2013

Henri Laude, "Data Scientist and R Language, Self-Study Guide to Big Data Mining," Eni: Paris, March 2016

Halima El Hamdaoui, E-commerce & E-marketing, Sup' Management, Fez, Morocco, 2022

Handl, Ant-based methods for tasks of clustering and topographic mapping: extensions, analysis and comparison with alternative techniques Master's Thesis, University of Erlangen-Nuremberg, Erlangen, Germany, 2003.

J. B. e. A. Richou, Big Data, Hadoop, MapReduce – Introduction for non-statisticians, University of Toulouse 2016.

Jean Privat Désiré BECHE., General information on big data, published on 9 February 2018, at 14:49:35.

Jain. A, K and Dubes, R.C "Algorithms for Clustering Data", Prentice Hall Advance Reference Series, 1988.

Jain A.K, Data Clustering 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8), pp: 651-666, 2009.

Mehdi Acheli and Selma Khouri. Tutorial to discover the world of Big Data: definition, applications and tools.

Medhioub H, "Description and grouping of resources for virtual networks" Master's thesis, University of Sfax, 2009.

Pang-Ning Tan, Steinbach M, and Vipin K, "Introduction to Data Mining", Addison Wesley, US edition, May 2005.

Rudi Bruchez, "NoSQL databases and Big Data", Eyrolles: Paris, 2015.

Webography

https://experiments.withgoogle.com/collection/a.i

https://www.talend.com/fr/resources/analytique-big-data/#

http://scikit-learn.org/0.10/modules/clustering.html.

https://www.vie-publique.fr/en-bref/284170-internet-des-objets-lue-presente-de...C:\Users\Congo Mobile\Desktop\hyperviseur-type1 (1).webp

https://openclassrooms.com/.../6313931-identifiez-un-hyperviseur-de-type-2

https://hadoop.apache.org/.../DockerContainers.html

https://hub.docker.com/r/apache/hadoop#!

Download the JDBC 3.0 driver for Microsoft SQL Server from http://download.microsoft.com/download/D/6/A/D6A241AC-433E-4CD2-A1CE-50177E8428F0/1033/sqljdbc_3.0.1301.101_enu.tar.gz.