RCMOS - Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

Big Data Tools Evaluation – Case Study: Spark Vs Flink

Big Data Tools Review – Case Study: Spark Vs Flink

David Francisco Cudijinguissa [1]
Private Polytechnic Institute of Kilamba,
davicudj@gmail.com

Abstract:

This article aimed to compare the performance of Big Data tools, Spark and Flink, considering five attributes that make these systems highly complex and demanding in terms of processing. As a result, the tools used to work with this data tend to be significantly more robust than conventional ones, and are often also more expensive. Open-source systems offer access to the source code, facilitating the understanding of systems and algorithms by collaborators and allowing them to adapt them to the needs of their projects. The methodology adopted in this study is based on descriptive research to relate the variables, with an exploratory and explanatory qualitative approach. A case study is presented, comparing the Spark and Flink platforms and considering factors such as scalability, data storage, complexity, and implementation options.

Keywords: Data analysis; Big data; Open source; Spark; Flink.

Abstract

This article aimed to compare the performance of Big Data tools, Spark and Flink, considering five attributes that make these systems highly complex and demanding in terms of processing. As a result, the tools used to work with this data tend to be significantly more robust than conventional ones, often being more expensive as well. Open-source systems provide access to the source code, making it easier for collaborators to understand the systems and algorithms, and allowing them to adapt them according to their project needs. The methodology adopted in this study is based on descriptive research to report the variables, with an exploratory and explanatory approach of a qualitative nature. A case study is presented, comparing the Spark and Flink platforms, considering factors such as scalability, data storage, complexity, and implementation options.
Keywords: Data analysis; Big data; Open source; Spark; Flink.

1 INTRODUCTION

The wider and deeper the diffusion of advanced information technology in factories and offices, the greater the need for an educated worker (Castells, 1999, p. 306). With the popularization of computing, internet of things and sensors, a large amount of data has expanded the need for analyze large volumes of data in batches or in real time. This skill has become a foundation fundamental to the competition that enhances and sustains new waves of productivity growth, innovation and consumer overactivity. Companies across all sectors will have to deal with the implications of large data, and not just with some data-oriented managers or some specific sector.

Big data techniques bring together methods from various areas that have developed over the years, such as statistics, artificial intelligence and machine learning (Torsten, 2018). These techniques allow the transformation of information into potentially useful knowledge.

Due to the complexity of processing this data, it is necessary to have an infrastructure and

[1] David Francisco Cudijinguissa, MSc. in Computer Engineering specializing in Computer Network and Communication Systems Management.

RCMOS - Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

more robust techniques than those used by traditional systems. In addition to performance, factors such as scalability and resilience need to be considered, as it is difficult to guarantee these aspects with conventional resources. The acquisition of these technologies can represent a major obstacle in the execution activities, as they often require high investments. The current financial crisis puts pressure on companies to save their resources as much as possible. With extremely controlled budgets, they find themselves obliged to continue delivering results and maintaining a competitive position in the market.

2 LITERATURE REVIEW

2.1 Big Data

The term “Big Data” was introduced by the head of SGI2 , scientist John R. Mashey (1998), we can describe as a massive collection of structured and unstructured data, availability and processing large volumes of streaming data in real time, in 2001 analyst Doug Laney defined this term as 3V: volume, velocity, and variety. Volume is linked to the large amount of data, velocity refers to data processing and variety refers to increasing complexity of the analyses. However, the literature shows that this concept is more related to 5Vs and more are added two characteristics. Veracity, which is directly linked to how true a piece of information is and value that is related to the value obtained from this data, that is, useful information. The term Big Data has always was a relative concept, as its size depends on who is using the data.

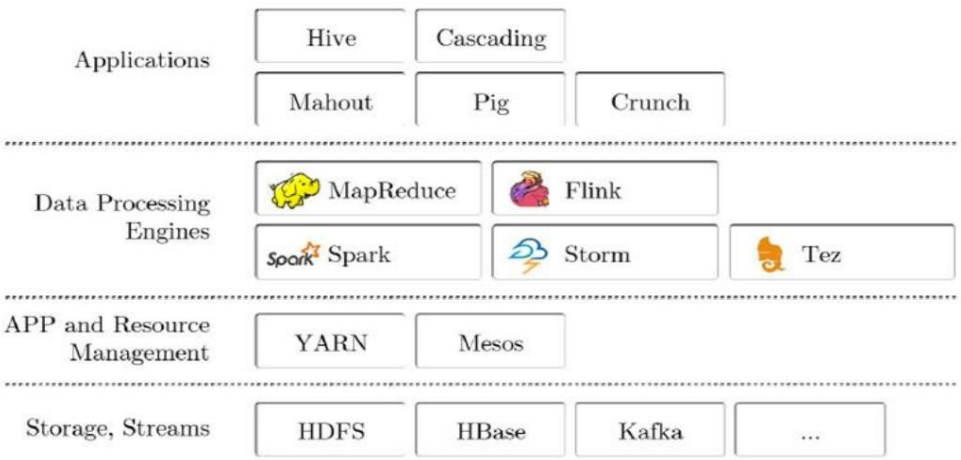

Figure 1 - Current Big Data structure

Source: Fayyad, Piatetsky-Shapiro, & Smyth, 1996.

Big Data is the set of large amounts of information generated every day – vast zettabytes of data that flow from our computers, mobile devices and machine sensors.

According to Miranda (2023, p.4), the traditional system uses the famous DBMS, or database management systems.

RCMOS - Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

of data, which store information in a structured way, in the format of tables, with rows and columns. They use machines with large storage and processing capacity. When there is a need for to expand the capacity of these machines, it is necessary to introduce new hardware components, so that have more memory and processing.

2.2 The characteristics of Big Data (five Vs)

Marques (2017) states that the volume, the large scales of the information processed can have orders of magnitude larger than traditional datasets, as the work requirements exceed the resources of a single computer, this becomes a challenge to pool, allocate and coordinate resources groups of computers.

Speed, data is flowing frequently into the Big data system from various sources and increasingly they are expected to be processed in real time to gain insights and update the current understanding of the system.

Variety, Big data problems are often unique due to the wide variety of sources processed and their relative quality.

Several individuals and organizations have suggested expanding the original three Vs, although these proposals tend to describe the challenges rather than the qualities of Big Data. Some common additions are:

Veracity: The variety of sources and the complexity of processing can lead to challenges in data quality assessment.

Value: The ultimate challenge of Big Data is delivering value. Sometimes, the systems and processes implemented are complex enough that utilizing the data and extracting real value can become difficult.

2.3 Classification of Analytical Processing

Regarding the Analytical Processing component, there are several architectures in the context of Big Date.

We classify the approaches according to their data storage support: (i) those based on disk data storage, which we divide into two groups according to the model of data, NoSQL databases (class A in Fig. 2) and relational (class B in Fig. 2); and those that support in-memory databases (class C in Fig. 2), in which data is kept in RAM reducing the I/O operations (Lee *et* al., 2012).

We detail each class as follows:

THE- NoSQL-based architecture: It is based on NoSQL databases as a model. storage, whether or not using DFS (Distributed File System). It is also based on models parallel processing, such as MapReduce, in which the processing workload is spread across many CPUs on common compute nodes. Data is partitioned across compute nodes in runtime and the underlined structure deals with machine-to-machine communication and machine failures. urals.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

The most famous embodiment of a MapReduce cluster is Hadoop. It was designed to run on many machines that do not share memory or disks (the shared-nothing model). This type of architecture is designed by Ain Fig.2.

B- Parallel relational database architecture: It is based, like classic databases, on tables relational files stored on disk. It implements features such as indexing, compression, views, materialized, sharing input or output, and caching results.

Research architectures include: shared nothing (multiple autonomous nodes, each having their own persistent storage devices and running separate copies of the DBMS), shared memory or anything shared (a global memory address space is shared and a DBMS is present) and shared disk (based on multiple processing nodes loosely coupled systems similar to shared nothing, but a global disk subsystem is accessible to the DBMS of any processing node).

In this architecture, the analytical solution is based on the conjunction of parallel databases (without sharing) with a parallel programming model. We can see this type of architecture designed by Bin Fig. 2.

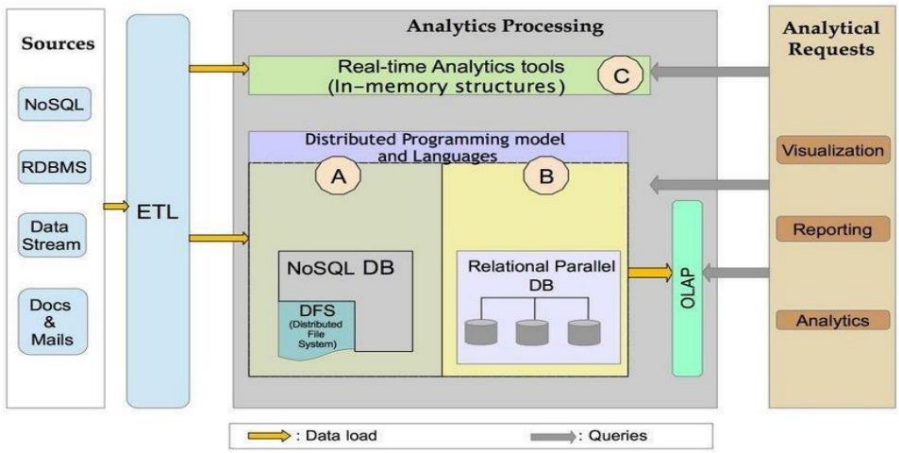

Figure 2 - Generic architecture for a big data approach

Source: Research Gate. (2021).

2.4 Types of processing

Big Data processing or big data analytics is a set of practices that analyze an immense volume of structured and unstructured data. The main objective of which is to obtain answers to face challenges and identify opportunities for your business.

“The number of devices added to the different file formats, and the need to extract of their value, showed the limitation of relational models, which served well for the treatment of structured data, but did not allow the processing of semi-structured or unstructured data.

Still in the NOSQL model, data originating from various devices, from mobile devices to servers, are replicated in clusters where they are processed through Analytics tools, and later

RCMOS - Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025
visualized through graphs, dashboards, among other analysis tools, just like in the ETL model, also used in relational models.

Structured information is information arranged in a rigid manner in an environment already designed for data analysis and quantitative analysis, such as Excel spreadsheets, legacy systems, or text files. If You create a cell in Excel that only accepts numbers, there's no point in adding text, understand? Data types structured:

- Spreadsheets like those in
- The bases themselves
- Files
- Files
- JSON files

Unstructured information. Unstructured information can also be used in your database. of data, however, these have a slightly more complex analysis, since they do not have a format for standardizing reading, such as: Internet pages; Videos, Audios, Telephone recordings, Documents from Word or Google Docs.

2.5 Batch Processing

Google's translation of the word "batch" is lote. Which reflects the definition well. batch processing, which involves processing a large amount of data at once. Microsoft (2025).

Batch Processing is a method of processing similar data in batches and executing them in sequence. It is very suitable for performing tasks that are repetitive and time-consuming or that require significant computational resources.

Batch processing is commonly used when faced with large amounts of data and in situations that do not require real-time analysis results. Some examples of use are situations financial aspects such as payroll or collections.

According to Wampler, (2018, p. 12), the three essential components of a batch architecture (Hadoop) are HDFS for storage, MapReduce and/or Spark for computing processes and YARN for control, while in the streaming architecture Kafka is used for storage, Spark, Flink, Akka Streams and Kafka Streams are used for computation, and control can be done by Kerberos, Mesos or YARN (with some limitations).

Because this period brought new challenges to the heart of the discussion and communication played a role central to the development of the entire process. Organizations were forced to change processes and methods to work, practically from one day to the next, and remote work has become inevitable for the continuity of activities.

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

2.6 Streaming Processing

Streaming mechanisms with very low latencies such as Kafka allow similar functionalities to MapReduce performed in Hadoop environments, with the difference that they can process terabytes in microseconds instead of the high latencies of batch systems (Narkhede, Shapira, & Palino, 2017).

A Big Data architecture must configure both processing solutions, since neither all processes require an immediate response. Batch processing is highly efficient, highly scalable, low-cost, and processes data at rest, while streaming allows for responsiveness immediate as events continually unfold in organizations, improving the ability to responding to situations such as fraud detection, real-time price adjustments, clinical situation alerts among others (Narkhede, Shapira, & Palino, 2017).

2.7 Open source software

In this section we address the concept of open source software and its origins, and also mention some of the main advantages and disadvantages of using this type of model compared to others.

Open source software today constitutes a source, still little explored, of useful tools for the development of these new tasks. Currently, there is a set of open source software that allows the teaching and learning process to be developed not only in a more attractive way, but, above all, more effective.

Open source software, also known as free software, is an alternative to proprietary or commercial software. It is distributed under a set of licenses, including the GPL (General Public License) and BSD (Berkeley Software Distribution).

Pushkarev et al. (2010) evaluate seven open-source tools or those with free trial periods using criteria such as connectivity, management, interface and functionality. Gao et al. (2016) analyze eight commercial tools in relation to their functionalities.

3 - RESEARCH METHODOLOGY

This section describes the methodological aspects based on research regarding general classification, research approach, technical procedures, data collection technique.

“Method is the way to accomplish something and when you have the way, it becomes easier make trips knowing where you are and where you want to go and how to do it” (Pereira Et Al., 2018, p. 67).

The methodology used in preparing this study is based on descriptive research with the objective to relate the variables involved, in addition to an exploratory and explanatory approach. In terms of methodology, a qualitative approach was adopted. This method was chosen to facilitate understanding of Big Data tools, specifically on the Spark vs Flink platforms.

Various methods, theoretical procedures, techniques and types of research were used to

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

achieve the general objective, highlighting the following:

1. Theoretical Method:

   o Bibliographic Research: Used to analyze previously published bibliographic elements, such as books, magazines, publications and scientific articles. This method aimed to update knowledge on the issues to be resolved, ensuring identification of existing and non-existent information in the final validation of the proposal.

2. Empirical Methods:

   a) Observation: Applied in the verification process of open source Big Data tools. Furthermore, the interview technique was used in the requirements gathering phase to evaluate the level of knowledge about these tools.

   b) Experimentation: Used to validate the functionalities of some tools, such as Spark and Flink.

   c) Hypothetical-Deductive Method: Used in the formulation of hypotheses, starting from the premise that the evaluation of open source Big Data processing tools enables a better choice when faced with the specific need for Big Data processing in a given case.

This research has an inductive nature, as pointed out by Marconi and Lakatos (2007), given the observation systematic and classification of the selected phenomena. It is based on the circumstances and frequencies at which publications on this topic occurred and in measuring their different intensities.

4 - PRESENTATION OF RESULTS

4.1 Description

Tests with the Word Count application were carried out repeatedly until a level of 95% confidence interval, using a Student's t-distribution, with at least 90 runs. The maximum error allowed was 5%. Similarly, tests with the word count application were repeated until reaching 90%. confidence, also with a Student's t-distribution and at least 90 runs, allowing a maximum error of 10%.

The virtualization environment was configured in Hyper-V, containing two virtual machines, both running the Linux operating system - Ubuntu.

- First virtual machine: configured with Apache Spark for distributed processing of large data volumes.
- Second virtual machine: configured with Apache Flink, focused on stream processing real-time data.

This infrastructure allows the execution and comparison of workloads in different Big Data frameworks, evaluating the performance and efficiency of each one according to the test scenarios.

RCMOS - Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

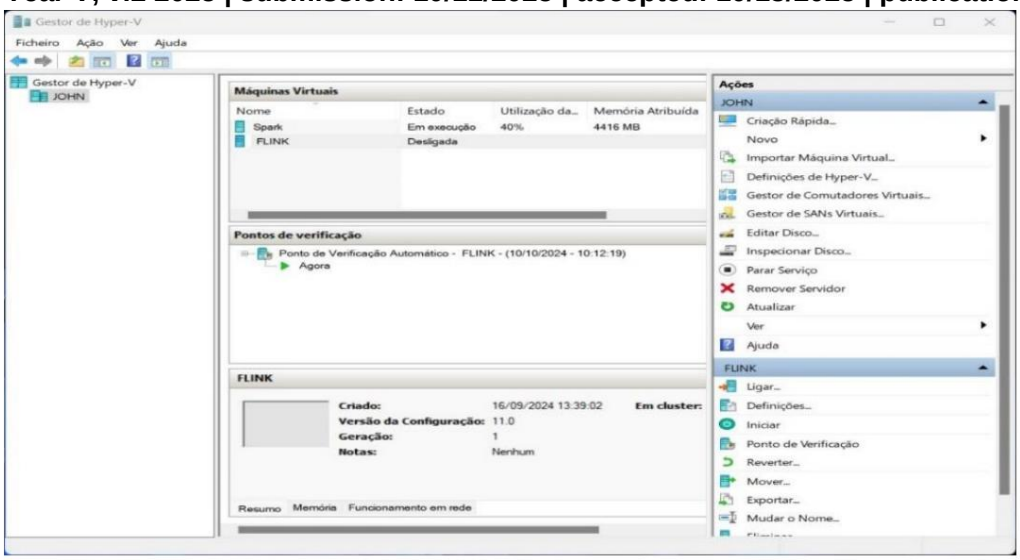

Figure 3 - Hyper-V environment with Spark vs Flink

Source: Author.

Figure 4 presents the Apache Spark graphical interface, displaying information about Jobs and Tasks, along with their respective duration times. This visualization allows you to monitor the execution of tasks, evaluate processing performance and identify possible bottlenecks in the distribution of the workload work.

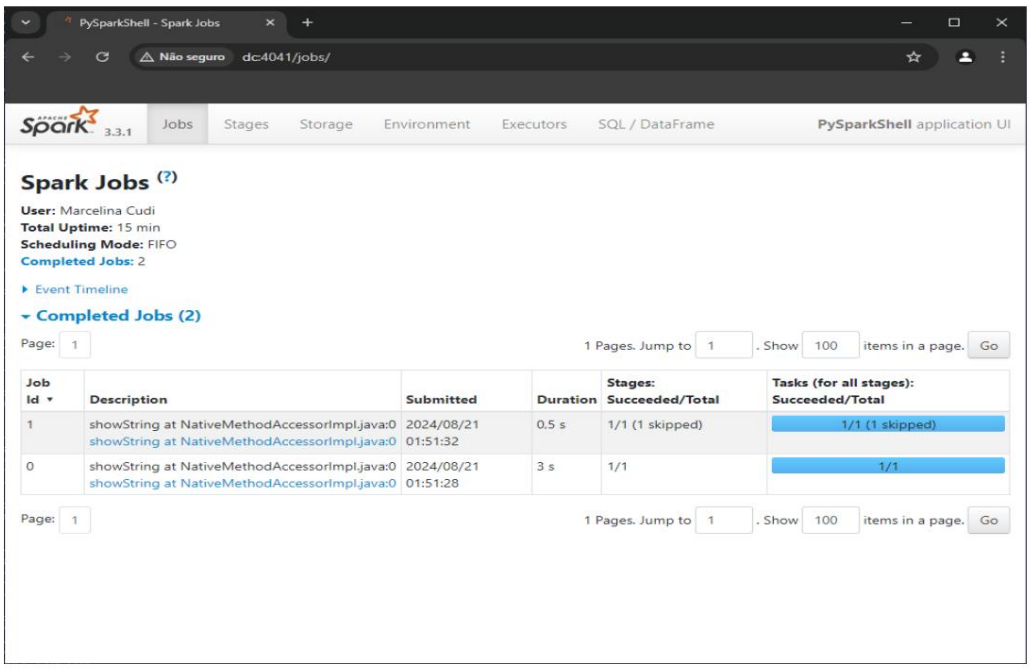

Figure 4 - Pysparkshel GUI Environment – Spark jobs

Source: Author.

Figure 5 shows the Apache Flink graphical interface, displaying the Job Manager metrics. These metrics provide detailed information about the performance and status of executions, including statistics of resource utilization, execution time and status of ongoing tasks.

RCMOS – Multidisciplinary Scientific Journal of Knowledge.
ISSN: 2675-9128. São Paulo-SP.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

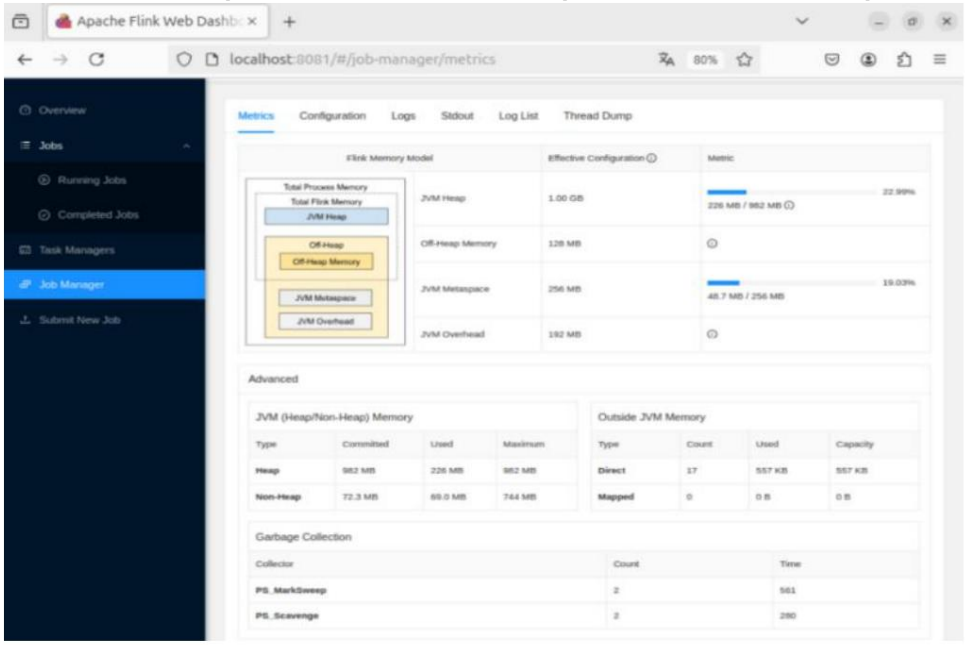

Figure 5 - Apache Flink Job Manager graphical interface.

Source: Author.

4.2 Description of results

$$= \frac{1}{}$$

The speedup of a parallel program is defined as:

Where:

- T1 is the execution time of the program on a single processor.
- Tn is the execution time of the program on nn processors.

The estimate for the speedup is given by:

$$= + \quad \frac{}{}$$

Being:

- Ts is the CPU time for the serial part of the program.
- Tp is the CPU time for the part executed in parallel.

Therefore, the speedup can be expressed as:

$$= \frac{1}{} + \frac{+}{+ \frac{}{}}$$

For ÿ ÿ, the equation approaches:

$$' = \frac{+}{}$$

4.3 Experimental results:

- Total time (real): 4.472 s (includes other processes in the system).

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

- User time: 7.243 s (time spent by the processor executing the program).
- System time (sys): 0.360s (time spent by the operating system kernel for the program).

These values indicate that the total execution time includes operating system overhead and other overhead. concurrent processes. For a more accurate speedup analysis, user time should be considered. as a reference for program execution.

5 - FINAL CONSIDERATIONS

5.1 Conclusions

Although they are still relatively new platforms, Flink has shown results impressive when compared to Apache Spark. For example, in an application that deals with large volumes of input and output, such as WordCount, the speedup achieved up to 1:86. In an application like IP Estimation, Flink's runtime remained almost constant even with the increase in the number of samples, contrary to what was observed with the Spark, where the maximum speed gain was 7:30.

However, the Flink installation faced great difficulties due to lack of memory, which prevented the collection of execution time. Based on the literature search, it can be inferred that Flink is optimized for streaming data processing, making intensive use of memory. This results in a significant reduction in execution time when dealing with large data flows.

5.2 Recommendations for Future Research

We suggest that future researchers continue this comparative study of tools of *open source Big Data* prepared for real-time projects, aiming to offer an alternative to commercial tools.

Furthermore, it would be enriching to follow the development of open source tools. Spark and Flink, comparing their performance results with commercial mining tools of data. This way, it will be possible to establish a direct comparison between the two realities and understand the costs involved in achieving these results.

REFERENCES

1. Castells, M.A. (1999), *Network Society.* v. 1.
2. Fayyad, UM; Piatetsky-Shapiro, Gregory; Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* USA: MIT Press.

Year V, v.2 2025 | submission: 10/11/2025 | accepted: 10/13/2025 | publication: 10/15/2025

3. Gao, J.Z., Xie, C., and Tao, C. (2016). *Big data validation and quality assurance issues, challenges, and needs.* In SOSE, pages 433-441. IEEE Computer Society.
4. John, RM (1998), Big Data, 1.3 X/year CAGR: historical trendline 1.6 X/year since 1990 2.0 X/year leap 1998/1999.
5. Lee, YW et al. (2012). *A methodology for information quality assessment. Information & Management:* Aimq: 40(2):133-
6. Marconi, M. de A., & Lakatos, E.M. (2017). Scientific Methodology: (7th ed. Updated by João Bosco Medeiros). Atlas.
7. Marques AV (2017), *importance of big data in the sector.*
8. Microsoft. (2025). Big Data Architectures. Microsoft Learn. Available at: https://learn.microsoft.com/en-us/azure/architecture/databases/guide/big-data-architectures Miranda, J.V. (2023). *Big data:* https://www.alura.com.br/artigos/big-data.
9. Narkhede, N., Shapira, G., & Palino, T. (2017). *Kafka: The Definitive Guide. Gravestein Highway North, Sebastopol, CA:* O'Reilly Media, Inc.
10. Pereira, AS et al. (2018). *Scientific research methodology:* Santa Maria: UAB / NTE / UFSM, Doi: http://repositorio.ufsm.br/handle/1/15824.
11. Pushkarev, V. et al. (2010). An overview of open source data quality tools. In IKE, pages 370-376. CSREA Press.
12. ResearchGate. (2021). *Generic-Architecture-for-a-Big-Data-Analytical.* https://www.researchgate.net/publication/318870641/figure/fig1/.
13. Torsten H. (2018). *Machine Learning & Statistical Learning.* Available at: https://www.pt.m.wikipedia.org/wiki/Big_data.
14. Wampler, D. (2018). *Fast Data Architectures for Streaming Applications* - Second Edition. Gravenstein Highway North, Sevastopol: O'Reilly Media, Inc.
15. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3 edn, Morgan Kaufmann Publishers.