Comment on “SPARK Concept Review”

Rodrigues Artur, Correia Ricardo, Morelli Jean, Cruz Rúben, Rocha Ana

Abstract— The reviewed paper superficially approaches the topics of Parallel Computing and Cloud Computing, explaining its core concepts. It thoroughly explains the big data processing framework Apache Spark, elucidating on how it works, what it can achieve and what it can be used for.

1 Introduction

Spark Concept Review makes a generic approach on the Apache Spark engine, a leading framework to process Big Data, describing its features, applications and libraries, ending the article with a comparison against its main “competitor”, named Apache Hadoop. It reveals Spark superiority, outperforming the other in processing time by a huge margin. The article begins with an introduction to parallel and distributed systems and a brief description on Cloud and Grid computing. In the following chapters Apache framework is introduced, presenting in-depth its motivation, history, architecture along with its capital features and applications. The most eager chapter of the article comes in last, comparing Apache Spark vs Apache Hadoop-Map Reduce showing their prominent differences with Spark standing out as a ‘winner’ as it beats Hadoop in processing time, also offering a rich and integrated library adapted for Machine Learning, Data Streaming and Graphs making its use much more attractive.

The paper is able to describe, in depth, the tool Apache Spark, touching not only on the core concepts of the tool and explaining them using proper technical language, suitably introducing most of the cited mentions relating tools, libraries and frameworks, but also allowing the reader to have a better understanding of what makes it so special. However, it lacks depth on the performance section, where the matter was unfortunately lightly touched on, which is one of the most interesting and curious metrics to study.

2 Evaluation

The Abstract section of the paper does not provide any conclusions nor the potential implications of those conclusions. In some cases, the word ‘Spark’ is not capitalized when it should be, giving the reader a bad impression of the grammar of this paper. Following the Abstract, it is the Index Terms section, where it lacks important keywords that were mentioned on the past section, for example the following concepts: Apache Spark, Cloud, Grid and Apache Hadoop.

Regarding the introduction, the topic is poorly introduced with no background nor information about why the need for a data processing framework. The background of parallel computing and cloud computing is introduced but there is no mention about the Apache Spark background is given.

No previous research about the topic is mentioned and the problem that the Apache Spark framework is trying to solve is missing. The goal of the paper is well explained but, once again, there is no mention of the conclusion of the end result.

The next section is reserved to introduce the concepts of parallel and distributed systems. It begins introducing both concepts and its motivation in clear and concise and clear way. After the introduction, it goes through each of the main concepts related to parallel and distributed system.

Cloud Computing is the next section of the paper, it starts by explaining what is the cloud computing and the motivation behind it. It would be interesting to dive deep into the services provided by the cloud computing, such as IaaS, PaaS and SaaS, also the current state of the cloud computing is not mentioned. It is lacking a connection between these topics and the main topic of the paper, Apache Spark. There are no mentions of how cloud computing, or Grid, and Apache Spark are related.

The article does a good job in explaining what motivated the creation of big data processing frameworks, elaborating on the necessity of it with real examples to back it up. However, it does lack in highlighting why Spark in specific was created, especially when there were already many good frameworks in use. It is not, therefore, clear what specifically motivated the creation of Spark and what was the need behind it. We recommend that this matter is explored, to allow the readers to understand what need or goal was behind Spark creation, rather than just focus on big-data frameworks as a broad topic.

The Spark architecture section is well detailed and explained, allowing the user to understand how Spark, with a very high abstraction level, works. As for the RDD’s, the key-defining mechanism of Spark, it’s properly explained,

- *R. Artur is with the Faculty of Sciences abd Technology, University of Algarve, Portugal. E-mail: a64592@ualg.pt*
- *C. Ricardo is with the Faculty of Sciences abd Technology, University of Algarve, Portugal. E-mail: a64007@ualg.pt*
- *M. Jean is with the Faculty of Sciences abd Technology, University of Algarve, Portugal. E-mail: a64014@ualg.pt*
- *C. Rúben is with the Faculty of Sciences abd Technology, University of Algarve, Portugal. E-mail: a6591@ualg.pt*
- *R. Ana is with the Faculty of Sciences abd Technology, University of Algarve, Portugal. E-mail: a63971@ualg.pt*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

detailing the available operations, data-sharing and fault-tolerance mechanisms in a simple and understandable way, backed with quality and rich references.

As for the various libraries that Spark contains, SparkSQL, SparkStreaming and Mllib are accurately explained, swiftly approaching how they work, what they contain and what they can be used to achieve. The only exception lies with the explanation of the library GraphX, where its details are extremely sparse and broad, not really exploring in any way the library. This library is compared to two other alternatives, which are not explained at all throughout the paper, giving the reader no information whatsoever about what this library is about. We suggest that the library GraphX should be explored and explained more through-fully, maybe using a high level approach to describe what sort of components it has, how they work and what they can be used for.

For the applications of Spark, the four big areas of application are thoughtfully explored, complementing the areas of interest with real applications of companies, allowing the user to have a better perspective of the utility of Spark in real case scenarios. However, some of the referenced applications are not described at all, making it hard to understand, for example, what exactly the domain specific interactive applications really are. Therefore, for this part of the section we recommend that the mentioned applications are elaborated a bit more, to allow the user to understand their functions and practical uses.

Regarding the section about Spark vs Hadoop-Map Reduce the authors make a vast analysis, approaching many important points. It starts with Performance, which is the least explored point, it lacks some support from graphics and images, the image shown is quite disappointing, being extremely small, making it hard to interpret and providing poor information. The authors could have had explored more applications of both frameworks, making the analysis of performance sturdier, it is shown a comparison on performance, but it is not known in which context the frameworks were used.

The next section is about pointing the main difference in the cost, it is well supported by bibliographic references, it is simply explicit about its content.

In the third section, Fault-tolerance the authors repeat the information stated in previous sections about the RDD mechanism to deal with faults, presenting no references.

In the next section, pointing out the differences between the Ease of use is concise and very well explained and in contrast the coming section about Security displays the most worrying difference between Spark and Hadoop-Map Reduce, the lack of security of Spark is approached, Kerberos authentication used by Hadoop-Map Reduce is here briefly introduced, however it feels like this topic would have needed a better clarification, since it is the main flaw of Spark compared to Hadoop, and more care with the grammar as some grammar error appear in this section.

Regarding the comparison on the Machine Learning topic, the authors cite that Spark offers more than 50 algorithms in its library and that Hadoop relies on an external library in order to use Machine Learning algorithm, however there is no information given about the library.

Scheduling and Resource Management, Data processing and Scalability are the last topics to be compared in the article and despite the authors had been clear on the details, there is not reference given to support them. On the last section, Scalability the authors should have been clearer. The authors imply that Hadoop has a better scalability than Spark due to being cheaper build clusters that are fit to work with.

3 Compliance with Author Guidelines

Regarding author compliance guidelines, the paper crosses several of the IEEE guidelines defined for regular papers for the IEEE Transactions on Parallel and Distributed Systems. Before enumerating the mentioned flaws, it is important to note that this paper does follow several of the stipulated guidelines, such as, for example: the size of the article, the type of letter and letter size, along with the line-spacing and paragraph spacing.

On the other hand, the usage of some of the references is not properly done. Whenever the author wants to cite references in text, they should appear on the line rather then outside of it [1], which is what happens in some parts of the paper, where the reference is cited after the sentence is over, for example ".[23]". Another violation of the guidelines present in the paper is the usage of apostrophes on acronyms, which should be not done at all [2]. The paper writes the acronymous "RDD's" multiple times, a clear violation of the rule mentioned above. One other aspect that appears and that is not in compliance with the guidelines is the appearance of the word "Spark" not always being capitalized, which should be the case given that it is the proper name of a technology [2](?).

The usage of contractions is also extremely common in this paper, with nearly every section using one at least a single time, for example the contraction of "it is" as "it's". This violates one guideline, in which contractions are not used in technical papers.

The paper also have several grammar errors, especially when it comes to lapses of clarity, verb conjugations and misspelling of words, which can be considered another violation of the author guidelines.

Some of the used words are written using British spelling, which is also considered another violation, as all the spellings should be American exclusive. The labeling of one of the lists present in this paper uses "+" to order its items, which is considered to be a violation of the author guidelines, as the order of labeling of all lists always has to follow 1),2),3), followed by a),b),c) and then i),ii) and iii) [2].

Regarding the consistency of the paper, it shows some irregularities. First is the indentation that can be noticed on the fifth section, *Spark vs Hadoop-Map Reduce*, some of those indentations is different from the rest of the paper. Another irregularity is the space difference between the beginning of a section and the end of the paragraph before that section. This difference can be seen when the fifth section is compared with the last section, *Conclusion*, where the space above the last section is almost doubled the space above the *Spark vs Hadoop-Map Reduce* section. Last irregularity is the space between paragraphs, in the section 5.3, the third paragraph has a space around it that is not present on the other paragraphs of this paper.

One other clear lying violation in this paper is how the numbers in the order of tens of thousands are being written like, because the guidelines refer that such numbers should always be separated using thin spaces. This is most definitely not this paper, as such numbers are written using either commas or are not separated at all, being therefore considered a clear violation of the guidelines. Also related to math, another existing violation is when the paper makes references to values in euros, in which they explicitly type “euros” rather than using the symbol €.

Finally, one other violation present in this paper is the lack of references in some of the mentioned data, since that no source is given for some of the cited information. This also comes across as a violation of the author guidelines.

4 Grading

In this section we will grade individually some important aspects of the article, considering a scale that goes from 1 to 10, where 1 means poor and 10 excellent, as such we present the following gradings:

1. Organization of article: 7
2. Theoretical or practical significance: 7
3. Clarity of presentation: 8
4. Adequacy of background. 6
5. Adequacy of literature review: 7
6. Appropriate approach: 8
7. Adequacy of analysis of issues: 7

5 Helpful Hints

Consistency is an important concept that should have been taken more seriously when this paper was developed. When there is more than one person that contributes to paper it is hard to maintain a certain level of consistency. Communication among the contributors is essential when writing a research paper. The grammar is another important factor that it was not well written.

When the authors describe the motivation on building the Spark framework, the author focused more on the motivation behind Big Data processing frameworks instead of describing the motivation specifically that lead to the creation of Apache Spark.

In the section where the authors compare Spark vs Hadoop-Map Reduce, it could have been more explored in detail the differences between both frameworks when the performance was being analyzed, it is shown a comparison without any context, refering that Spark is about 100 times faster than Hadoop-Map Reduce but it would be interesting to see how it would perform in many different situations/applications.

6 Conclusion

In conclusion, taking all the aspects refered in the previous chapter in consideration we give this paper a final grade of 7.

References

[1] NIST, “SI Unit rules and style conventions - Check List for Reviewing Manuscripts” available online at http://physics.nist.gov/cuu/Units/checklist.html since February 1998, last accessed June 2009.

[2] Journals.ieeeauthorcenter.ieee.org. 2021. [online] Available at: https://journals.ieeeauthorcenter.ieee.org/wp-content/uploads/sites/7/IEEE-Reference-Guide-Online-v.04-20-2021.pdf

[3] Journals.ieeeauthorcenter.ieee.org. 2021. [online] Available at: https://journals.ieeeauthorcenter.ieee.org/wp-content/uploads/sites/7/IEEE-Editorial-Style-Manual-for-Authors-Online-v.04-20-2021.pdf