Apache Spark Software Technology Evaluation Project

Jose Juan Pena Gomez and Enrique Vilchez Campillejo

November 27, 2019

Abstract

10-15 lines with the software technology and the highlights from the project that has been undertaken.

1 Introduction

The main purpose of the project is to use a unified analytics engine (Apache Spark) to compare how clustering works with machine learning and how the technology works itself. As we worked before in projects of Machine Learning, we think it would be interesting to work with distributed machine-learning framework.

In 2008, approximately,the basis for creating this technology appeared, Hive and HBase, which are two tools of the Hadoop ecosystem. Spark was founded at UC Berkeley in 2009, it was born practically from a Google paper and from there it evolved, going through the mapreduce processes.

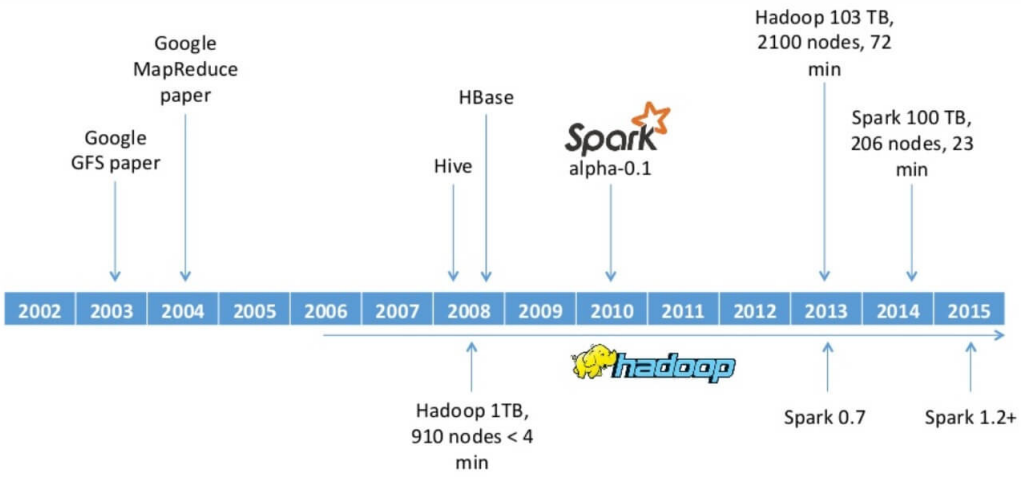

Figure 1: Apache Spark Development Timeline

Comparing this technology with others, for example with sklearn, the most well-known library for machine learning porpoises, we are satisfied with our results because the sklearn library is even not able to load big datasets. In addition we are satisfied with the clustering management, it is 60 % faster than without clustering management.

This rest of this report is organised as follows:

- Section 2 gives an overview of the key concepts and the architecture.
- Section 3 gives a view of the prototype and its implementation
- Section 4 gives an explanation of the test-bed environment used and the results of the experiments
- Section 5 gives a conclusion about the main concepts that we appreciated from this technology.

2 The Software Technology

About 4 pages that introduces in (sufficient) depth the key concepts and architecture of the technology. May use a running example to introduce the technology.

This part and other parts of the report probably needs to refer to figures. Figure 2 from [1] just illustrates how figure can be included in the report.

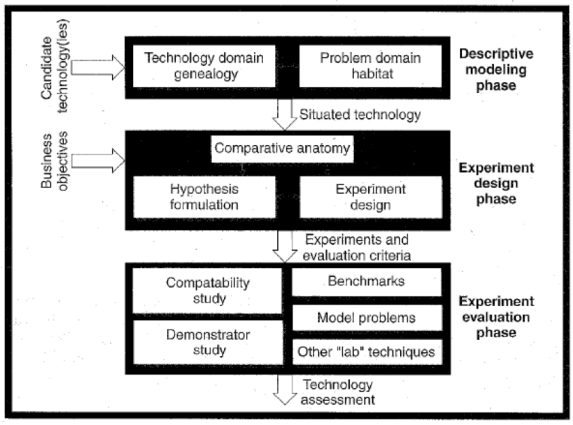

Figure 2: Software technology evaluation framework.

3 Demonstrator Prototype

About 5 pages that gives:

1. High-level view of the demonstrator and its purpose.
2. Details of how the demonstrator has been implemented.
3. May involve presentation of code snippets.

The example below shows how you may include code. There are similar styles for many other langages - in case you do not use Java in your project. You can wrap the listing into a figure in case you need to refer to it. How to create a figure was shown in Section 2.

Config	Property	States	Edges	Peak	E-Time	C-Time	T-Time
22-2	A	7,944	22,419	6.6 %	7 ms	42.9%	485.7%
22-2	A	7,944	22,419	6.6 %	7 ms	42.9%	471.4%
30-2	B	14,672	41,611	4.9 %	14 ms	42.9%	464.3%
30-2	C	14,672	41,611	4.9 %	15 ms	40.0%	420.0%
10-3	D	24,052	98,671	19.8 %	35 ms	31.4%	285.7%
10-3	E	24,052	98,671	19.8 %	35 ms	34.3%	308.6%

Table 1: Selected experimental results on the communication protocol example.

```
1  public class BoksVolum {
2
3     public static void main(String[] args) {
4
5        int b, h, d;
6        String btext, htext, dtext;
7
8        [ ... ]
9
10       int volum = b * h * d;
11
12       String respons =
13            "Volum [" + htext + "," + btext + "," + dtext + "] = " + volum;
14
15    }
16 }
```

4 Test-bed Environment and Experimental Results

About 3 pages that:

Describes the software used to establish the test-bed and for implementing the demonstrator prototype.

Explains what experiments have been done and the results.

For some reports you may have to include a table with experimental results are other kinds of tables that for instance compares technologies. Table 1 gives an example of how to create a table.

5 Conclusions

Concludes on the project, including the technology, its maturity, learning curve, and quality of the documentation.

The references used throughput the report should constitute a well chosen set of references, suitable for someone interesting in learning about the technology.

References

[1] A.W. Brown and K. C. Wallnau. A framework for evaluating software technology. *IEEE Software*, 13(5):39–49, September 1996.