Comentário do artigo “Apache Spark”

Bruno Susana, Gonçalo Mascarenhas, Guilherme Correia, Henrique Cruz, Ricardo Rosa,

FCT UALG

Abstract— O presente documento visa expor a opinião crítica do grupo GC sobre o trabalho desenvolvido pelo grupo G3, que teve como objetivo de estudo o “Apache Spark”. O trabalho será avaliado em diferentes aspetos e será atribuída uma classificação entre 0 e 10 em cada aspecto. No final será realizado um pequeno comentário sobre o porquê de cada nota atribuída.

1 Introdução

O artigo em questão tem como tema o “Apache Spark”, este *Framework* é utilizado sobretudo em *Big Data*, termo este aplicado ao tratamento e análise de grandes volumes de informação e de dados que seriam demasiado demorosos e complexos de analisar pelos métodos mais tradicionais. O artigo em questão começa por abordar os conceitos de sistema paralelo e sistema distribuído na sua generalidade, apresentando vários conceitos relacionados, dos quais as diferenças entre programação concorrente e sequencial, região critica e *cloud and grid computing*.

De seguida é apresentado ao leitor o ponto fulcral do artigo que consiste no *framework* que está sendo análise, aqui é apresentada alguma informação relevante tal como a origem, a arquitetura, as bibliotecas implementadas e o seu funcionamento.

Após esta introdução é-nos apresentado alguns casos de uso em que este *framework* é utilizado e exemplos de grandes empresas que empregam esta tecnologia nos dias atuais, mostrando desta forma que esta tecnologia continua bem presente atualmente.

No final é feita uma comparação de *performance* em que se compara o *framework Spark* com o seu maior concorrente no mercado, *Hadoop-Map Reduce*, onde são avaliados alguns pontos fulcrais quando se escolhe qual a ferramenta mais adequada a cada caso.

Após uma primeira análise não realizada ainda com rigor, é possível constatar que o trabalho produzido pelo grupo apresenta bastante qualidade em termos de estrutura e em termos de escrita onde são apresentados vários conteúdos que foram lecionados em aula.

- *Bruno Susana, aluno de Licenciatura em Engenharia Informática, na Universidade do Algarve; email: a61024@ualg.pt*
- *Gonçalo Mascarenhas, aluno de Licenciatura em Engenharia Informática, na Universidade do Algarve; email: a64533@ualg.pt*
- *Guilherme Correia, aluno de Licenciatura em Engenharia Informática, na Universidade do Algarve; email: a61098@ualg.pt*
- *Henrique Cruz, aluno de Licenciatura em Engenharia Informática, na Universidade do Algarve; email: a61099@ualg.pt*
- *Ricardo Rosa, aluno de Licenciatura em Engenharia Informática, na Universidade do Algarve; email: a62461@ualg.pt*

2 Avaliação

Onde são abordados os sistemas paralelos e distribuídos os autores falam sobre os vários envolventes de forma clara e explicita, com o auxílio de imagens e tabelas para ajudar o leitor a compreender melhor os temas base do artigo.

Onde é abordado o tema *cloud computing*, os autores referem o impacto que este teve e é feita uma comparação com *grid computing* de forma a entender as semelhanças entre os conceitos. Desta forma, é possível ao leitor entender de forma supérflua a diferença entre os conceitos, não obstante o facto de serem feitas boas referências e introdução a temas que são falados mais à frente no artigo.

No ponto 4.2 no que se refere à história do Apache, é mencionado que ficou um Top-Level Apache Project em 2013, no entanto isso aconteceu em Fevereiro de 2014 [2].

De seguida é-nos introduzido o *framework* em estudo com uma esclarecedora e sucinta introdução onde são expostas as bases do *Spark* e a sua arquitetura. Aqui é mostrado como funcionam os *RDD´s* associados aos *framework*.

Na parte do *higher-level libraries*, os autores optaram por explicar as utilidades das bibliotecas no *Apache spark*. Começado pelo *Spark sql*, os autores explicam de forma geral em que consiste e ao longo do texto são indicadas as linguagens de programação utilizadas no desenvolvimento desta biblioteca, a figura 7 é bem escolhida demonstrando complementando graficamente a explicação dada.

Na parte sobre *ghrapx*, o tema poderia ser mais completo e mais explicativo, apesar do que já é apresentado ser o suficiente como forma de introdução do tema.

Relativamente ao *MLib*, o tema encontra-se bem explicado, os autores optaram por não colocar código, desta forma é mais simples para o leitor.

No que toca ao tema da comparação do Apache Spark

com o Apache Hadoop no geral foram feitas boas comparações, com bons fundamentos entre cada característica de cada framework de forma clara e compreensível explicando como funciona cada aspeto de cada framework.

Finalmente quando são referidas as aplicações, os autores escolheram temas de grande relevo no mercado e bastante apelativos à leitura, acompanhado de exemplos que captam o leitor.

3 Conformidade com as diretrizes do autor

Em termos de escrita, é utilizada um tipo de escrita agradável visualmente e de fácil leitura, o tamanho da mesma também não gera qualquer dificuldade visual. Não se detetou nenhum tipo de abreviação ou anotação que gerasse dúvida na ideia que se pretende transmitir.

Existe um pequeno reparo em que uma das figuras é repetida e a numeração das mesmas encontra-se errada, de tal forma que o leitor primeiro encontra a figura 3 e de seguida a figura 2 sendo que ambas as figuras constam também com a mesma descrição.

A figura 4 encontra-se num local inapropriado, esta inserida no tema *Spark*, mas deveria estar em *Cloud computing and Grid computing*.

A classificação das figuras não se encontra uniforme ao longo do documento, temos como exemplo a figura 1 e a figura 6 que não apresentam um mesmo tamanho na identificação da respetiva figura.

Na figura 8 é apresentada uma fase denominada *Spark engine* que não é explicada, a contrário das outras fases que tiveram uma explicação clara e sucinta.

A figura 9 encontra-se bastante reduzida, dificultando a sua leitura em formato físico.

4 Classificação

De seguida são atribuídas classificações de 0 a 10 em diferentes aspetos qualitativos:

Gráficos: 7

Referencias: 7

Organização (Ortografia e gramática): 8

Descrição do caso de estudo e enquadramento: 9

Clareza de organização: 9

Completude: 9

5 Dicas Úteis

Na parte do Graphx os autores poderiam ter colocado uma figura a mostrar como funciona a biblioteca de forma a transmitir melhor para o leitor. [1]

Deveriam ter indicado de onde vieram algumas referências sobre as aplicações.

Na parte da comparação entre o Spark e o Hadoop poderiam ter usufruído de mais figuras para ter uma comparação mais clara e mais consolidada em alguns dos aspetos e não apenas na performance, acrescentando assim uma maior legibilidade e consolidação da informação descrita.

6 Conclusão

Com este comentário, podemos concluir que o artigo analisado é um artigo bem estruturado, que apresenta os conteúdos de forma linear, começando com algo geral e ao longo do artigo torna-se mais concreto. Os autores também expõem o conteúdo de forma que o leitor a entenda sem ter qualquer tipo de dúvida.

Por outro lado, têm alguns erros que são necessários para compreensão do leitor e também possuem erros a nível de conteúdo.

Nota final: 9

Referências

[1] spark, a. (2020). *GraphX Programming Guide*. Obtido de spark apache: https://spark.apache.org/docs/latest/graphx-programming-guide.html

[2] *wikipedia* "*Apache Spark*". (23 de Abril de 2021). Obtido de https://en.wikipedia.org/wiki/Apache_Spark