Departamento de Informática

Computação na Cloud (11481)

Mestrado em Engenharia Informática

Ano Letivo de 2017/2018, 2º Semestre

Projeto Laboratorial 1 – Desempenho de Plataformas Hadoop

Organização dos Grupos de Estudantes: Trabalho a realizar por um grupo de, no máximo, 2 estudantes.

Identificação dos Estudantes: Indique, na capa do relatório, o título do trabalho (acima indicado) e os nomes e os números dos estudantes que integram o grupo que realizou o trabalho.

Cotação do Trabalho: Este trabalho contribui com 3,0 valores (15%) para a classificação final do estudante expressa na escala de 0,0 a 20,0 valores.

Formato do Relatório: O relatório deve obedecer ao formato IEEE, conforme especificado na página de apoio a esta unidade curricular: http://www.di.ubi.pt/~mario/tcc.htm.

Aulas Práticas Para a Realização do Trabalho: O trabalho deve ser realizado nas aulas práticas entre 6 de março e 3 de abril (inclusive) de 2018.

Data e Forma de Entrega do Trabalho: Um estudante de cada grupo deve enviar o relatório, num ficheiro em formato pdf, até 10 de abril de 2018, por email para o endereço mario@di.ubi.pt e com subject: CC 2017/2018 – Trabalho 1.

Objetivos

O objetivo central deste trabalho consiste em instalar, analisar e comparar o desempenho da última release estável da plataforma Hadoop com o desempenho da versão do Hadoop tolerante a faltas, o Hadoop MR BFT.

Descrição

Em 2004, a Google apresentou o modelo de programação MapReduce e a respetiva implementação [1], o qual tem sido extensivamente usado pela Google nos seus datacenters para suporte a funções essenciais ao seu motor de busca, nomeadamente para processamento de índices. Contudo, a implementação deste sistema não está publicamente disponível.

Em 2009, foi proposta uma implementação da plataforma MapReduce num projeto de código aberto da Apache [2], designada por Hadoop, datando a release inicial de

Universidade da Beira Interior | Faculdade de Engenharia
Rua Marquês D'Ávila e Bolama
6201-001 Covilhã | PORTUGAL
Tel.: +351 275 242 081 | E-mail: secretariado@di.ubi.pt
www.ubi.pt

dezembro de 2011. Esta plataforma tem sido usada por muitas empresas de computação em nuvem, incluindo a Amazon, a IBM, a RackSpace e a Yahoo.

Neste trabalho pretende-se instalar, analisar e comparar o desempenho da última release estável (3.0.0 / December 13, 2017) da plataforma Hadoop [3] com uma versão do Hadoop tolerante a faltas, o Hadoop MR BFT [4], cujo código fonte se encontra disponível em [5]. O Hadoop deve ser instalado na configuração Single Node Cluster [6].

Despois de instaladas as plataformas Haddop, deve proceder-se à respetiva análise de desempenho recorrendo ao GridMix2 benchmark [7], apresentando e comparando, tanto para a última release estável do Hadoop como para o BFT MapReduce, o makespan de 6 aplicações do GridMix2 benchmark: WebdataScan, WebdataSort, Combiner, Javasort, Streaming, Monsterquery. No caso do BFT MapReduce pode ser usado apenas um dos modos especulativo (speculative) ou não especulativo (non-speculative) e deve ser considerado o caso $f = 1$.

Organização do Relatório

O relatório deve ser organizado de acordo com a informação apresentada na secção Template e Estrutura do Relatório do Projeto, disponibilizada na página de apoio a esta unidade curricular: http://www.di.ubi.pt/~mario/tcc.htm.

Referências

[1] Dean, J. & Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. In Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation, OSDI'04, 1–10.
[2] White, T. (2009). Hadoop: The Definitive Guide. O'Reilly, 1st edn.
[3] Apache Hadoop, http://hadoop.apache.org, último acesso: 27 de fevereiro de 2018.
[4] Costa, P., Pasin, M., Bessani, A.N. & Correia, M.P. (2013). On the Performance of Byzantine Fault-Tolerant MapReduc, IEEE Transactions on Dependable and Secure Computing, 10, 301–313, http://ieeexplore.ieee.org/document/6412676/, último acesso: 27 de fevereiro de 2018.
[5] Código Fonte do Hadoop MR BFT, https://bitbucket.org/pcosta_pt/hadoop-bft/, último acesso: 27 de fevereiro de 2018.
[6] Hadoop: Setting up a Single Node Cluster, https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html, último acesso: 27 de fevereiro de 2018.
[7] GridMix Benchmark for Hadoop Clusters, https://hadoop.apache.org/docs/r1.2.1/gridmix.html, último acesso: 27 de fevereiro de 2018.