10ª Jornada Científica e Tecnológica
7º Simpósio da Pós-Graduação
INSTITUTOS FEDERAIS 10 ANOS 2008|2018
INSTITUTO FEDERAL Sul de Minas Gerais
ISSN 2319-0124

UTILIZAÇÃO DO BLAST EM NUVENS COMPUTACIONAIS

Marcelo Rodrigo de Castro [1]

RESUMO

Devido a redução dos custos e evolução dos mecanismos que efetuam o sequenciamento de DNA, obteve-se uma grande quantidade de dados referentes aos estudos em genômica e metagenômica. Esse fator de crescimento dos dados gerados pela genômica tendem a não acompanhar o poder computacional no que diz respeito processamento e análise de dados, conforme previsto pela lei de Moore. Para contornar a limitação quanto ao processamento e análise das informações geradas pelos sequenciadores e analisadores faz-se necessário o uso de sistemas distribuídos para computação em alta vazão (HPC), por exemplo Clusters, Grids e Computação em Nuvem. Com estes tipos de sistemas é possível escalabilidade e a utilização de diversas máquinas para obtenção de um desempenho melhor. Contudo, os algoritmos de bioinformática devem ser adaptados para utilizar o processamento paralelo e distribuído, como no paradigma MapReduce. Neste trabalho é investigado o Apache Hadoop e Spark e seu papel na Bioinformática.

Palavras-chave: Big Data; Apache Hadoop; Apache Spark; Sequenciamento genético, Bioinformática.

1. INTRODUÇÃO

Com a constante redução dos custos e evolução dos mecanismos que efetuam o sequenciamento de DNA (ácido desoxirribonucleico), obteve-se uma grande quantidade de dados gerados. O fator de crescimento dos dados gerados na bioinformática tendem a não acompanhar o poder computacional no que diz respeito custo da produção das informações, conforme previsto pela lei de Moore (CASTRO, 2017). Uma das alternativas utilizadas quanto ao processamento e análise das informações geradas pelos sequenciadores e analisadores é o uso de sistemas distribuídos para computação em alta vazão, por exemplo Clusters, Grids e Computação em Nuvem.

Contudo, os algoritmos de bioinformática devem ser reescritos ou adaptados para utilizar o processamento paralelo e distribuído. Como no paradigma MapReduce, onde é possível escalar centenas ou milhares de processadores para a execução das tarefas computacionais (PIREDDU, 2011).

O *framework* Spark que trabalha no processamento de grande quantidade de dados é uma alternativa para o paradigma MapReduce (Hadoop) e tem se mostrado importante e promissor na área de bioinformática para processamento muitas informações (ZAHARIA, 2011).

2. FUNDAMENTAÇÃO TEÓRICA

[1] - Marcelo Rodrigo de Castro - IFSULDEMINAS - marcelo.castro@ifsuldeminas.edu.br

2.1 BIOINFORMÁTICA

A bioinformática é uma ciência multidisciplinar (biologia, informática, etc.) que busca compreender as funcionalidades biológicas das células, mais especificamente, dos genes. O mapeamento de genomas gera diariamente um volume elevado de informações que são sistematicamente armazenadas em bancos de dados computacionais, que servem de fontes de estudo para biologia e medicina (NCBI, 2016).

Assim, a biologia alinhada à informática permite auxiliar os pesquisadores em bioinformática a criar, melhorar, desenvolver e manipular os dados obtidos com o sequenciamento genômico. Nesse contexto, uma das ferramentas mais utilizadas na bioinformática é o BLAST (NCBI, 2016). Porém, com o crescimento exponencial das informações cabe alternativas para melhorar seu desempenho na obtenção de resultados.

2.2 BIG DATA

Jagadish (2014) define o termo *big data* em relação à velocidade, quantidade e variedade de dados gerados o que ocasiona um gargalo no processamento de informações com as tecnologias atuais. Observa-se que os dados gerados pela bioinformática se enquadram nesta definição, pois crescem em ambas as características.

Para viabilizar o processamento de informações em tempo hábil *frameworks* como o Spark e o Hadoop são utilizados. Por exemplo, grandes corporações como Facebook, Yahoo!, Google e Twitter utilizam esses *frameworks* para as aplicações *big data* (CASTRO, 2017).

2.3 MAPREDUCE - HADOOP

O MapReduce é um modelo de programação funcional e paralela comumente utilizado para processamento de grandes quantidades de dados de forma distribuída, apresentado e popularizado pela Google em meados de 2004. Este modelo de programação já era utilizado em linguagens funcionais, como Lisp e Haskell. Mas, veio a ser mais utilizado com a implementação da Google. A manipulação e execução dos trabalhos feitos pelo MapReduce utiliza basicamente duas funções: *map* e *reduce*.

O MapReduce é um modelo de programação que permite o processamento de dados massivos em um algoritmo escalável, paralelo e distribuído, geralmente utilizando um *cluster* de computadores. Essa abordagem, mesmo não sendo fácil de ser utilizada, é muito útil para aplicações que envolvam dados massivos para processamento paralelo ou até mesmo para processamento de qualquer tipo de dado (PIREDDU, 2011).

2.4 APACHE SPARK

Com a evolução das tecnologias e novos desafios encontrados, o Spark tem surgido com uma nova ferramenta para processamento em larga escala. O Apache Spark é rápido e é um *cluster* de sistema de computadores genéricos (ZAHARIA, 2011) que tem se mostrado mais rápido que o Hadoop. Há bibliotecas específicas para trabalhar com SQL (Spark SQL), aprendizado de máquina (MLlib), processamento de grafos (GraphX) e streaming de dados (Spark Streaming) (ZAHARIA, 2011).

O Spark tem a execução mais rápida que o Hadoop em alguns casos porque tem algumas características para a execução de suas aplicações: *cache* de memória, tolerância a falhas e RDD (*resiliente distributed dataset* - conjunto de dados distribuídos).

3. MATERIAL E MÉTODOS

Foi feito um estudo sobre várias ferramentas, das quais pode-se citar: Bioinformática e nuvens computacionais, CloudBLAST, Biodoop, HBlast, SparkSeq, Adam API, entre outros (CASTRO, 2017). Com base no modelo de escalabilidade usado pelo Hadoop ao utilizar o BLAST foi desenvolvido uma aplicação em Spark.

A Figura 1 define como é o processamento de uma determinada entrada na aplicação.

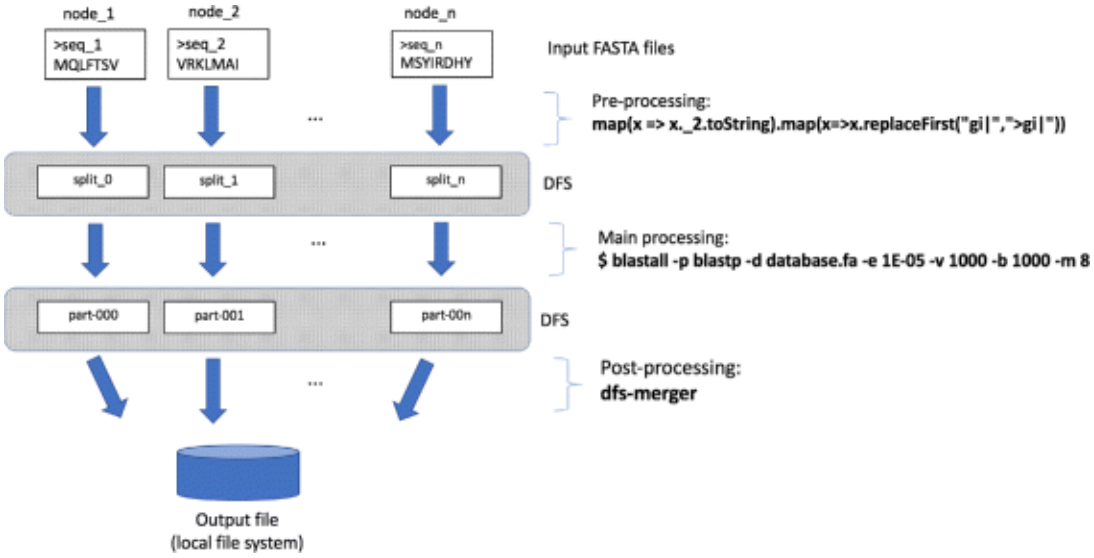

Figura 1: Cada nó processa parte de uma entrada, o resultado final é mesclado e exibido para o usuário. Fonte: própria do autor.

A aplicação foi publicada no GitHub para consulta de pesquisadores e o acesso está descrito nos resultados.

4. RESULTADOS E DISCUSSÕES

A aplicação e os resultados obtidos podem ser encontrados no seguinte endereço:

https://github.com/sparkblastproject/v2 bem como um artigo em revista Quali A1, no seguinte link: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1723-8.

A partir da ferramenta elaborada pode-se mensurar os tempos de processamento e aferir sua funcionalidade.

5. CONCLUSÕES

Com a crescente demanda no processamento e análise de dados, mais especificamente, na bioinformática. Surge a demanda em adequar/criar mecanismos que permitam a execução de aplicações em tempo hábil. O *framework* Apache Spark tem se mostrado promissor para avaliar e produzir resultados com base em análises genômicas e tem sido uma alternativa para quem usa o Hadoop.

AGRADECIMENTOS

Ao IFSULDEMINAS pelo PIQ, à UFSCar e FIOCRUZ pela ajuda na definição do problema e à Google e Microsoft pela disponibilização de ambiente em nuvem para execução do projeto.

REFERÊNCIAS

CASTRO, M. R. de, dos Santos Tostes, C., Dávila, A. M., Senger, H., & da Silva, F. A **Spark-BLAST: scalable BLAST processing using in-memory operations.** BMC bioinformatics, v. 18, n. 1, p. 318, 2017.

JAGADISH, H. V. et al. **Big data and its technical challenges**.Communications of the ACM, v. 57, n. 7, p. 86-94, 2014.

MASSIE, Matt et al. Adam: **Genomics formats and processing patterns for cloud scale computing**. University of California, Berkeley Technical Report, No. UCB/EECS-2013, v. 207, 2013.

NCBI. National Center for Biotechnology Information. **Nucleic acids researc. <** www.ncbi.nlm.-nih.gov> Acessado em 25 de agosto de 2016.

PIREDDU, Luca; LEO, Simone; ZANETTI, Gianluigi. **Mapreducing a genomic sequencing workflow**. In: Proceedings of the second international workshop on MapReduce and its applications. ACM, 2011. p. 67-74.

ZAHARIA, Matei et al. **Spark: cluster computing with working sets**. HotCloud, v. 10, p. 10-10, 2010.