Evento	Salão UFRGS 2014: FEIRA DE INOVAÇÃO TECNOLÓGICA DA UFRGS – FINOVA
Ano	2014
Local	Porto Alegre
Título	Análise de DNA Humano com Computação Intensiva de Dados
Autores	JÚNIOR FIGUEIREDO BARROS Raffael Bottoli Schmmer
Orientador	CLAUDIO FERNANDO RESIN GEYER

Análise de DNA Humano com Computação Intensiva de Dados

O surgimento de volumes de dados na ordem de Petabytes cria a necessidade de desenvolver novas soluções que viabilizem o tratamento de dados através do uso de sistemas de computação intensiva em dados (BigData), tal como o MapReduce. O MapReduce é um framework de programação que permite abstrair do programador as etapas necessárias à paralelização de dados. Aplicações de bioinformática, responsáveis pelo alinhamento e analise de amostras de DNA sequenciadas são um tipo de aplicação intensiva em dados capazes de fazer uso do modelo MapReduce, e também demandar uso intensivo de processamento e armazenamento, dado o volume e o tamanho de entrada e das amostras.

O projeto de pesquisa em questão trabalha em conjunto ao hospital de clínicas de Porto Alegre (HCPA), de maneira que os dados utilizados pela aplicação MapReduce a ser desenvolvida, são obtidos diretamente de máquinas sequenciadoras de DNA humano de segunda geração (NGS), do tipo IonTorrent PGM. A aplicação MapReduce é executada sobre o framework Hadoop, onde é feito o pré-processamento, alinhamento e comparação das amostras utilizando como referência o genoma humano. O objetivo do pré-processamento e alinhamento é otimizar a detecção de possíveis mudanças nos genes sequenciados, ao passo que, em uma segunda etapa a comparação vem a detectar as mudanças (mutações) presentes nas amostras de pacientes sequenciados. Esta etapa ocorre com busca em bancos públicos de mutações, onde se tem catalogadas patologias genômicas ou relação de propoensão a desenvolvimento das mesmas.

O uso de computação de alto desempenho com ênfase em grandes volumes de informações têm correlação com máquinas do tipo Cluster e Cloud computing, tanto por serem distribuídas como por apresentar alta disponibilidade de processamento e armazenamento de informação.

Este trabalho tem também como objetivo realizar a implementação da aplicação de análise de DNA, sob uma perspectiva de melhor usufruir e utilizar as características apresentadas, tanto por máquinas do tipo Cluster quanto Cloud. As máquinas Cluster em sua grande maioria são privadas, possuindo comportamento previsível e controlado, em contrapartida Cloud em sua maior parte públicas, onde seu comportamento pode ser imprevisível pela volatilidade dos recursos e maneira como o compartilhamento dos mesmos é realizado.

Entender quais são os desafios impostos por cada arquitetura utilizada, e adaptar a aplicação de forma que sua arquitetura consiga executar de maneira ótima sobre ambos os tipos de máquinas utilizados pelo trabalho é também uma questão de interesse que é explorada por esse respectivo projeto de pesquisa.