جـــامعة الفـــيوم
Fayoum University

Name of Candidate: Hassan Salem Hussein Saad

Degree: Master of Science

Title of Thesis: Cloud Computing Applications for Large Scale Data Analysis

Supervisors: 1- Prof. Dr. Amr Mohamed Gody

2- Associate Prof. Dr. Rania Ablsoud

Department: Electronics and Communication Engineering

Approval-: 30-10-2018

ABSTRACT

The recent revolution in sequencing techniques has resulted in exponential growth in DNA sequence data. As a result, most of the existing bioinformatics tools are obsolete because they do not differentiate with data. To deal with "data overload", we present here the analysis and studies of the pig-tool for biological sequences, such as DNA sequences and protein comparisons. The purpose of the sequences comparison is to discover the similarities between different biological sequences then creating clustering tree. For example, the human and mouse genomes are 85% identical, so they will encode very similar proteins. Next-Generation sequencing produces huge collections of strings to be analyzed. This massive dataset challenges traditional analytics tools and increasingly requires novel solutions adapting to big data platforms. MapReduce software framework presents a viable solution to large-scale sequence analysis in terms of efficiency and scalability. Hadoop as an open source implementation of MapReduce framework is designed to run applications on large scale clusters built on commodity hardware. Hadoop distributed file system (HDFS) and Hadoop MapReduce are two important components of Hadoop framework. HDFS provides scalable, fault-tolerant and distributed data storage, while MapReduce is the core concept of Hadoop framework and provides a scale-out data processing solution across hundreds or thousands of nodes in Hadoop cluster. The Fayoum University cloud computing environment cluster is built over a scientific Linux cluster for Big Data analysis (SLBD). SLBD runs open source software with large computational capacity and high performance cluster infrastructure. SLBD composed of one cluster contains identical, commodity-grade computers interconnected via a small LAN. Cloudera Manager is used to configure and manage an Apache Hadoop stack. Hadoop is a framework allows storing and processing big data across the cluster by using MapReduce algorithm. MapReduce algorithm divides the task into smaller tasks which to be assigned to the network nodes. SLBD clustering system allows fast and efficient processing of large amount of data resulting from different applications. SLBD also provides high performance, high throughput, high availability, expandability and cluster scalability.

The proposed system used Pig tool that provided by the SLBD system to write a pig's scripts for making a comparison between DNA sequences. Pig has major advantage: Pig's programmability greatly reduces development time for parallel bioinformatics applications. Sequence comparison is a fundamental task in computational molecular biology that aims to discover similarity relationships between molecular sequences. Many current comparison methods based on word frequencies (kmers) are studied. The proposed system use these methods to construct evolutionary trees of the mitochondrial genome of 11 vertebrates known as the real tree. Our analysis shows that the correlation distance for the dimers (K = 2) produces the best trees.

The summary not more than 500 words