Chapter 9

Analysis of Genomic Data in a Cloud Computing Environment

Philip Groth
Bayer Pharma AG, Germany

Gerhard Reuter
Bayer Business Services GmbH, Germany

Sebastian Thieme
Humboldt-University of Berlin, Germany

ABSTRACT

A new trend for data analysis in the life sciences is Cloud computing, enabling the analysis of large datasets in short time. This chapter introduces Big Data challenges in the genomic era and how Cloud computing can be one feasible approach for solving them. Technical and security issues are discussed and a case study where Clouds are successfully applied to resolve computational bottlenecks in the analysis of genomic data is presented. It is an intentional outcome of this chapter that Cloud computing is not essential for analyzing Big Data. Rather, it is argued that for the optimized utilization of IT, it is required to choose the best architecture for each use case, either by security requirements, financial goals, optimized runtime through parallelization, or the ability for easier collaboration and data sharing with business partners on shared resources.

INTRODUCTION

Big Data Challenges

In 2009, the total global amount of stored data was estimated to have reached 800 Exabyte (EB) (Association, 2010) and was increased by approximately 13 EB throughout the following year (Agrawal et al., 2012). It was recently estimated that the amount new data generated in 2013 alone has reached 900 EB, implying that the vast majority of data stored today have been generated in just the past two years (IBM). Of this, the global amount of healthcare data was estimated to have exceeded 150 EB in 2011 (IBM). There is a simple explanation for this strong increase: Data nowadays are generated anywhere and anytime in a mainly automated manner and storing them is relatively cheap (e.g. commercial data storage is offered for less than USD 0.01 per Gigabyte

DOI: 10.4018/978-1-4666-6611-5.ch009

and month) (AWS). The notion of 'Big Data' to describe this phenomenon that large amounts of data are generated within a specific domain or of a specific class has already been described in the mid-nineties of the last century when the term itself was first coined by John Mashey of Silicon Graphics (sgi) and since then been widely adopted (Lohr, 2013; Mashey, 1998).

Most types of Big Data have many characteristics in common, e.g. a typical life cycle. They are generated, copied and moved, processed and analyzed, versioned, archived and sometimes deleted. Each step brings up systematic issues, all of them involving IT. Handling of Big Data starts with the process of generation. To generate data and to keep them usable, it is important to document their existence. By whom, how, when and under what circumstances were they created? The lack of such attributes will reduce or even disable the usability of data. Already, this is not trivial; as such annotation should be stored in a searchable manner. A popular tool to handle Big Data, especially for data annotation, tracing and versioning is 'iRods' (www.irods.org), which is employed, for example by the CERN (for data from high energy physics experiments) and the Wellcome Trust Sanger Institute (WTSI) (for DNA sequencing data).

Assuming the data have been adequately annotated, they are oftentimes placed within the Internet for immediate global availability, creating the challenge to interested users of acquiring a local copy. Classical transfer methods based on FTP or HTTP were not designed to transfer large files (i.e. in the range of gigabytes or more). Tools like Aspera™ (using a proprietary protocol named FASP) or torrent-seeding based methods (cghub.ucsc.edu) remedy this issue to some extent but are not yet widely spread.

Finally, due to their size, Big Data are quite often stored on distributed architectures, bringing up another issue. Software meant to process Big Data must take into account partial failures of the underlying hardware (e.g. a single disk error) and communication latencies. Many popular software products are in the process of redesign to adapt to this change. To speed up the analysis of Big Data, they are processed in a parallel manner in such distributed environments. This can be done with tools like Hadoop™ (Apache, 2012), utilizing a proprietary file system to distribute data across the network and a so-called 'master-slave approach' to assign sub-tasks to interlinked compute nodes. Hadoop™ is a framework developed by Apache and used amongst others by Facebook™ and Yahoo™.

Genomic Data is Big Data

With the decoding of the human genome and the associated substantial progress in the development of laboratory and bioinformatics methods (Chen, Wang, & Shi, 2011; Kearney & Horsley, 2005; Wang, Gerstein, & Snyder, 2009) the data of known biological interrelationships has also increased dramatically. Such data comprise, for example, the complete information on an individual's genome, such as nucleotide variations, chromosomal aberrations or other structural changes within the genome, more generically known as mutations. The smallest mutation within a genome is the exchange of a single nucleotide within the DNA, the so-called 'building block of life' (see (Alberts et al., 2007; Strachan & Read, 2005) for more information). If such a mutation is shared by at least 1% of a defined population and not disease-causing per se it is called 'single nucleotide polymorphism' (SNP) (Barreiro, Laval, Quach, Patin, & Quintana-Murci, 2008; Risch, 2000). In 2005, it was estimated that there are approximately 10 million SNPs to account for variation in the human population (Botstein & Risch, 2003). But data from the 1,000 Genomes Project (Abecasis et al., 2010) have revealed that there are many more SNPs within the human genome. By 2011, more than 40 million SNPs had been identified (Eberle et al., 2011). Each SNP specifies a genotype, describing differences