A Infrastructure for Disease Prevention and Precision Medicine

Marco Moscatelli[1], Andrea Manconi[1], Matteo Gnocchi[1], and Luciano Milanesi[1]

Institute for Biomedical Technologies, National Research Council, Segrate (Mi), Italy

Abstract. Precision medicine is an emerging and novel approach for both disease treatment and prevention. Precision medicine allows to classify individuals into subpopulations that differ in their susceptibility to a particular disease with the aim to tailor the medical treatment to the individual characteristics of each patient. To provide precision medicine to patients researchers needs to analyze huge amounts of heterogeneous data from both biomedical research and healthcare systems. The growing amount of these data gives rise to the need for new research methods and analysis techniques. In this paper we present a infrastructure that exploits new strategies aimed at storing, accessing, and analyzing efficiently these heterogeneous data.

Keywords: Big Data, Disease Prevention, Precision Medicine, HPC

1 Motivation

Nowadays, advances in technology has arisen in a huge amount of data in both biomedical research and healthcare systems. This growing amount of data gives rise to the need for new research methods and analysis techniques. Analysis of these data offers new opportunities to define novel diagnostic processes. Therefore, a greater integration between healthcare and biomedical data is essential to devise novel predictive models in the field of biomedical diagnosis. In this context, the digitalization of clinical exams and medical records is becoming essential to collect heterogeneous information. Analysis of these data by means of big data technologies will allow a more in depth understanding of the mechanisms leading to diseases, and contextually it will facilitate the development of novel diagnostics and personalized therapeutics. The recent application of big data technologies in the medical fields will offer new opportunities to integrate enormous amount of medical and clinical information from population studies. Therefore, it is essential to devise new strategies aimed at storing, accessing, and analyzing the data in a standardized way. Moreover, it is important to provide suitable methods to manage these heterogeneous data.

2 Methods

In this work, we present a new information technology infrastructure devised to efficiently manage and analyze huge amounts of heterogeneous data for disease prevention and precision medicine. It should be noted that the rigidity of relational databases does not lend to the nature of these data. In our opinion, better results can be obtained using non-relational (NoSQL) databases. Starting from these considerations, the infrastructure has been developed on a NoSQL database with the aim to combine scalability and flexibility performances. In particular MongoDB [1] has been used as it fits better to manage different types of data on large scale. In doing so, the infrastructure is able to provide an optimized management of huge amounts of heterogeneous data, while ensuring high speed of analysis.

With the aim to enable researchers to perform their analysis through dedicated computing resources, the infrastructure has been built on a hardware platform intended to enable big-data classes of applications which consists of:

- a massive storage platform of 1.6 PB;
- 2040 CPU cores;
- 16 NVIDIA K20 GPUs;
- 2 big memory nodes (i.e., 1 node equipped with 1TB and 1 node equipped with 512GB of memory).

It should be pointed out that the concept of precision medicine is about the customization of healthcare, with decisions and practices tailored to an individual patient based on their genome sequence, microbiome composition, lifestyle, and diet in addition to medical and clinical data. Therefore, in addition to the data about the patient collected by healthcare providers, researchers need to incorporate many different types of data. To this end, a web-based platform built upon the Galaxy technology [2] has also been implemented to enable researchers to retrieve and analyze biological data in their analyses. Galaxy is an open web-based scientific workflow system for data intensive biomedical research accessible to researchers that do not have programming experience. By default, Galaxy is designed to run jobs on local systems. However, it can also be configured to run jobs on a cluster. The front-end Galaxy application runs on a single server, but tools are run on cluster nodes instead. To this end, Galaxy supports different distributed resource managers with the aim to enable different clusters. For the specific case, in our opinion SLURM [3] represents the most suitable workload manager to manage and control jobs on the above hardware infrastructure. SLURM is a highly configurable workload and resource manager and it is currently used on six of the ten most powerful computers in the world.

3 Results

The presented infrastructure exploits big data technologies in order to overcome the limitations of relational databases when working with large and heterogeneous data. The infrastructure implements a set of interface procedures aimed at preparing the metadata for importing data in a NoSQL DB. Moreover, data can also be represented as a graph using Neo4j [4]. The Neo4J DB allows you to emphasize and enhance the connections between the data and facilitate the retrieve and navigation of data. Experimental tests on huge amount of data show that our infrastructure exhibits performances in terms of speed and scalability unachievable with relational databases. These performances are mainly related to ability of the infrastructure to index any type of field as well as to customize the queries. In particular, the high flexibility to customize the queries increases the search performance and specificity of the results.

Moreover, the robust hardware infrastructure together with the Galaxy web-based platform allow to easily integrate and analyze heterogeneous data from different biological sources.

Currently, the infrastructure is used in a project aimed at implementing techniques to infer the predisposition to some cancer diseases. The project is funded by the Fondazione Bracco.

4 Supplementary Information

This work has been supported by the Fondazione Bracco, the Italian Ministry of Education and Research Flagship (PB05) InterOmics project.

References

1. http://www.mongodb.org
2. Goecks, J., Nekrutenko, A., Taylor, J., The Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25;11(8):R86.
3. http://slurm.schedmd.com/
4. http://neo4j.com/