Ph.D. in Information Technology:

Thesis Defenses April 14th, 2021 online by TEAMS – at 10:00

Fai clic qui per partecipare alla riunione

GULINO ANDREA – XXXIII cycle

Advisor: Ceri Stefano

PhD Thesis Title : DISTRIBUTED PROCESSING AND OPTIMIZATION METHODS FOR QUERYING BIG GENOMIC DATASETS

Short Abstract:

Next-Generation Sequencing (NGS) is a revolutionary approach that allows reading the whole genome much faster and with significant cost reductions. Large and well-organized collections of sequence datasets and new technologies for integrating and querying them help answering fundamental biological questions and pave the way for personalized medicine. The data management technology developed within the Genomic Computing (GeCo) group at Politecnico di Milano, based on the "GenoMetric Query Language" (GMQL), is a high-throughput computational solution for integrating and processing big genomic datasets. This first contribution of this thesis describes significant optimizations and extensions made to the system, that relies on the Apache Spark framework. Specifically, it shows how the parallelism of GMQL's most complex operations, based on genome binning, was improved by adopting complex analytical models. Moreover, it shows how the system was extended to support the federated execution of GMQL queries, where multiple geographically distant sites share data and computational resources. Secondly, a broader category of Apache Spark application is considered and a performance prediction framework, based on machine learning and analytical models, is proposed. Such framework is then used to improve the execution planning of Federated GMQL queries. Eventually, it introduces MutViz, a complementary cloud-based tool for the visual analysis of somatic mutations. An ongoing work re-adapts Mutviz to support the visual analysis of SARS-CoV-2 variants.

HORLOVA OLHA – XXXII cycle

Advisor: Ceri Stefano

PhD Thesis Title: Exploiting the Array Data Model for Genomic Data Management

Short Abstract:

As DNA sequencing technologies (NGS) are evolving and becoming affordable for large adoption, genomic data became more utilized for diagnosing many diseases, such as cancer. As genomic applications are computationally intensive applications, the cost shifted from data acquisition to data processing. Therefore, there is a need for a compact and accurate representation of the genomic data; improving their memory management improves the overall performance, by reducing the memory size and the efficiency of data computations; in this thesis, I explore how to exploit the multi-dimensional (or array) data model for genomic data, structured in the form of regions.
The thesis is developed in the context of GenoMetric Query Language (GMQL), a high-level, declarative approach to data extraction and a cloud-based implementation for managing region-based genomic repositories, which have the typical big data dimensions. The GMQL engine, a system that is specifically

dedicated to region-based genomic calculus, uses the relational (row-based) model to represent data in memory, and uses Apache Spark as underlying data engine. Relational tables and multi-dimensional arrays have their pros and cons; the new data model presented in this thesis combines the pros of both the row-based and the array-based data models, to overall improve the engine performance.
We start by applying the array data model for the class of region-preserving operations, i.e. operations which do not create any new region but rather compose existing ones. As region-preserving chains of operations frequently occur in GMQL programs, we used data model transformations from row to array prior to executing the chains, and from array back to rows at the end of the chains. In this way, we discover the potential of using a multi-dimensional representation of genomic data on top of data flow engines.
Then, we show the full implementation on the array-based model of all GMQL operations, by also including the operations which are not region-preserving. This led us to get rid of redundant transformations between row and array-based data models, and finally demonstrates that a well-designed array-based implementation is very efficient, still using Apache Spark as execution engine. The thesis is completed by the introduction of a GMQL-like language for array databases, which uses the classic abstractions of array-based query languages.

Committee Members

Masseroli Marco	Politecnico di Milano - Deib
Missier Paolo	Newcastle University
Bozzon Alessandro	TU Delft