Towards an efficient evaluation of Graph Data Science Workflows

Advisors

Genoveva Vargas-Solar
CNRS, Database group, LIRIS
genoveva.vargas-solar@liris.cnrs.fr

Hamamache Kheddouci
U. Claude Bernard, Lyon 1, GOAL, LIRIS
hamamache.kheddouci@univ-lyon1.fr

1.1 Context

The project GALILEAN (Graph analytics workflows enactment on just in time data centres) gives context to this master project that will be performed in the database and the GOAL groups of the LIRIS lab.

Huge collections of heterogeneous data containing observations of phenomena have become the backbone of scientific, analytical, and predictive processes for solving problems in areas such as the connected enterprise, digital mesh, internet-connected objects and knowledge networks. Observations can be structured into networks that have their own interconnection rules determined by the variables (i.e., attributes) characterising each observation.

The notion of graph is a powerful mathematical concept for representing these networks as graphs, and the networks can be implemented by efficient data structures and exploited by applying different types of algorithms composed by workflows to solve data science problems.
When the graphs become large and even too large the algorithms used to process, explore, and analyse them become costly in terms of execution time, even if several cores are used, and in this case, given the characteristics of the algos, communication is also likely to be costly. So, workflows that exploit graphs become greedy consumers of computing resources.

1.2 Objectives and expected results

Graph processing and analysis workflows consist of tasks including

- deploying or retrieving graphs which are often distributed over an execution environment;
- applying complex algorithms in a distributed way;
- retrieving the results and making them available to other processes or to end-users.

In terms of infrastructure, execution is done on heterogeneous architectures that provide computing services with different capacities to execute them.
The execution of DS workflows on graphs consists of data processing tasks that can be scheduled on top of a target architecture.

The objective is to study the execution of DS workflows addressing graph analytics focusing on data processing, transmission and sharing across several resources.

- *Define an execution plan(s) exhibiting the data dependencies among tasks and the control flow to adopt for executing them considering the distribution of the data/execution workloads.*

Tasks

1. Study workflows using analytics graph algorithms used for answering community detection problems like page rank, Louvain [2,3].

2. Characterize DS workflows considering (i) the type of graph processing algorithms they address; (ii) the characteristics of the graphs (data) processed and results through these algorithms.
3. Generate execution plans from data science workflows specification including:
    a. tasks to be executed by the workflow (classic execution plan);
    b. the specification resources requirements associated with each task of the execution plan according to the algorithm it calls and the data injection function estimating the volume of data to process.
4. Experiment the execution of DS workflows execution plans on different target architectures configurations.

Expected results

1. Taxonomy of workflows using graph algorithms implemented in different platforms like Microsoft IA Gallery, Github and Kagggle.
2. State of the art of distributed query evaluation in relational databases and in graph management systems.
3. Algorithm for generating execution plans for data science workflows using graphs.
4. Experimental results of execution plans generated to be executed on different target architecture configurations proposed in [1].

References

[1] Ali Akoglu and Genoveva Vargas-Solar. Putting data science pipelines on the edge. *To appear in the proceedings of the 2021 International Workshop on Big data driven Edge Cloud Services (BECS 2021)*, May 18, 2021.

[2] Sarra Bouhenni, Saïd Yahiaoui, Nadia Nouali-Taboudjemat, Hamamache Kheddouci: A Survey on Distributed Graph Pattern Matching in Massive Graphs. ACM Comput. Surv. 54(2): 36:1-36:35 (2021)

[3] Assia Brighen, Hachem Slimani, Abdelmounaam Rezgui, Hamamache Kheddouci: A distributed large graph coloring algorithm on Giraph. Cloudtech 2020: 1-7