Hard or/and Soft tuning of HADOOP Ecosystem towards query energy efficiency on NoSQL Database

Subject proposed by Simon Pierre DEMBELE
Email: simon.pierre.dembele@ut.ee

1. Context:

With the emergence of the big data age, along with the data-centric and -intensive computing trend, the great amount of energy consumed by database systems has become a major concern in a society that pursues green information technology. Apache Hadoop framework supports the storing and processing of big data datasets using simple programming models.

It refers to the various components of the Apache Hadoop software library, as well as to the accessories and tools provided by the Apache Software Foundation. Hadoop is a Java-based framework that is extremely popular for handling and analyzing large sets of data.

The Hadoop ecosystem involves the use of different parts of the core Hadoop set such as MapReduce, a framework for handling vast amounts of data, and the Hadoop Distributed File System (HDFS). Hadoop, Apache has also delivered other kinds of accessories or complementary tools for developers. These include Apache Hive, a data analysis tool; Apache Spark, a general engine for processing big data; Apache Pig, a data flow language; HBase, a database tool; and also, Ambari, which can be considered as a Hadoop ecosystem manager, as it helps to administer the use of these various Apache resources together.

Today, in many compagnies Hadoop becoming the de facto standard for data collection and management. Research work on energy efficiency in the Hadoop ecosystem can be categorized into two approaches: software-based and hardware-based. The software approach covers Scheduling, data placement and resource management but does not make an in-depth analysis of the query processing system within Hadoop systems.

2. Objective

The objective of this master thesis work is to analyze the operation of the Hadoop ecosystem during the execution of analytical queries in order to identify configurations along several dimensions (database type, storage engine, partitioning form, etc.) that can better optimize energy consumption. To achieve these goals, we need to

- Perform a descriptive and diagnostic study of the Hadoop Ecosystem according to the energy dimension.
- Analyze and identify the most energy consuming components of the system that influence the execution of the query.
- Propose and evaluate the energy consumption of different system configuration scenarios during analytical queries execution.

3. Requis

The candidate must have skills in NoSQL databases, SQL language, Hadoop ecosystem.

Required technologies: Hadoop, Hive, HBase, Java.
Required profile: Master's student or equivalent diploma.
Contact: -
Application: -
Gratification: -
Internship period: -
Continuity in PhD: Yes, according to the results obtained

4. References:

a. CHEN, Yueguo, QIN, Xiongpai, BIAN, Haoqiong, et al. A study of SQL-on-Hadoop systems. In: Workshop on big data benchmarks, performance optimization, and emerging hardware. Springer, Cham, 2014. p. 154-166.
b. SHABESTARI, Fatemeh, RAHMANI, Amir Masoud, NAVIMIPOUR, Nima Jafari, et al. A taxonomy of software-based and hardware-based approaches for energy efficiency management in the Hadoop. Journal of Network and Computer Applications, 2019, vol. 126, p. 162-177.
c. FELLER, Eugen, RAMAKRISHNAN, Lavanya, et MORIN, Christine. Performance and energy efficiency of big data applications in cloud environments: A Hadoop case study. Journal of Parallel and Distributed Computing, 2015, vol. 79, p. 80-89.
d. Simon Pierre Dembele, Ladjel Bellatreche, Carlos Ordonez, Amine Roukh, Think big, start small: a good initiative to design green query optimizers, Cluster Computing Journal, Springer, 2019, https://doi.org/10.1007/s10586-019-03005-0
e. DEMBELE, Simon Pierre, BELLATRECHE, Ladjel, et ORDONEZ, Carlos. Towards Green Query Processing-Auditing Power Before Deploying. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020. p. 2492-2501.
f. Simon Pierre Dembele, Ladjel Bellatreche, Carlos Ordonez, al. ,Big Steps Towards Query Eco-Processing - Thinking Smart, journal ARIMA J., 34, 2020, doi 10.46298/arima.6767