Cairo University
Faculty of Computers and Information

An Efficient Advanced SQL–to-MapReduce Translator to Improve Big Data Analysis on the Cloud Computing Environment

By

Fawzya Ramadan Sayed Hassan

Under the Supervision of

Prof.\ Ibrahim Farag	Dr.\ Mohamed H. Khafagy
Computer Science Department	Computer Science Department
Faculty of Computers and Information	Faculty of Computers and Information
Cairo University	Fayoum University

A Thesis Submitted to the
Faculty of Computers & Information
Cairo University
In Partial Fulfillment of the Requirements for the Degree of

Master of Science
In
Computer Science

Egypt
April 201

Abstract

MapReduce has become an effective framework for processing and analysis huge data size in large systems. SQL Query is necessary to build an efficient and flexible SQL translator to MapReduce framework. The need of optimized SQL translator to deal with advanced queries which can increase data processing with Big Data growth. Hive supports queries which called HiveQL. HiveQL offers the same features as SQL, which still difficult to deal with complex SQL queries. Consequently, manual translation of HiveQL often leads poor performance.

Also, Flink has become an effective framework to Big Data analysis in large cluster systems. On the other hand, FLink doesn't support any Query language. So, designing and implementing SQL to FLink Translator is needed to execute SQL Query over FLink. This work of this thesis adopts these limitations of SQL translators and proposes two contributions which considered as SQL–to-MapReduce translators to improve Big Data analysis.

The first contribution is called QRMapper (Query Rewriting Mapper). It is developed to solve the problem of translating a complex SQL queries into HiveQL by utilizing and optimizing Query rewriting. This translator improves the performance of HiveQL without any change in Hive framework and provides the possibility of executing SubQuery and Advanced SQL Query. Our system performance has been evaluated using TPC-H Benchmark.

The second contribution is named SQL TO Flink Translator. A new system has been developed to define and add SQL Query language to Flink. This translator improves the performance of SQL without any change in Flink framework and provides the possibility of execute SQL Query on Flink by generating Flink algorithm that executes SQL Queries. Also, SQL TO Flink Translator has the capability of execute SQL when other systems support a low-performance. Our system performance has been evaluated using TPC-H Benchmark.

Generally, according to these two contributions, a new layer has been developed to execute advanced SQL Query over MapReduce translator. So it is considered a main contribution in the Big Data field.

Thesis is organized as follows:

Chapter 1- Introduction; provides an overview of the thesis, motivation, problem statement, objective and organization are described.

Chapter 2 – Background and Related Work; gives an overview of the related issues to the thesis's problem domain.

Chapter 3 – Proposed Sql–To-Mapreduce Translators; illustrates that there are two contribution (QRMapper), SQL to Flink Translator tool, and describes the Framework, Architecture and Methodology to each one.

Chapter 4 – Experimental Evaluation; describes the experimental results of both QRMapper tool and SQL to Flink Translator tool.

Chapter 5 – Conclusions and Future Work; concludes the thesis and discusses future work.