EÖTVÖS LORÁND UNIVERSITY

FACULTY OF INFORMATICS

MASTER THESIS TOPIC DECLARATION FORM

Name: Khuong Duy Vu — Neptun code: QFNFQT

Training: **Full-time** — Techical major: SDE

Name of the supervisor: Zoltán Istenes

Workplace: ELTE, Faculty of Informatics

Position: Associate Professor

Highest Degree: Doctorate

Title of the thesis: Empirical Assessment of SQL-like Continuous Queries Implementations over Data Stream

Topic of the thesis:

In the last few years Big Data generates a lot of buzz along with the launch of several successful big data products. The big data ecosystem has now reached a tipping point where the basic infrastructural capabilities for supporting big data challenges and opportunities are easily available. Entering the next generation of big data —so-called big data 2.0 — two of its concentrated areas are velocity and applications. The cause for the former is that data is growing at an exponential rate, and the ability to analyze it faster is more important than ever. At the same time real-time or (streaming) processing providing answers in low latency is get more and more focus even in proportion to the whole big data scene. The latter is helping to overcome the technical challenges of existing frameworks by making them easy to use and understand for everyone to benefit from big data.

As a result the demand for streaming processing is increasing a lot these days. Processing big volumes of data is not enough in the cases that infinite streaming data is arriving at high speed and have to be processed fast to react to any incident immediately. Stream processing solutions are designed to handle high volume in real time with a scalable, highly available and fault tolerant architecture. This enables analysis of data in motion.

Many Big Data frameworks provide the rich imperative code APIs only to process data stream on the fly. The APIs command the system to perform certain operations in a certain order but given that most big data applications are fairly simple algorithm-wise they might be less then optimal to use for most of the popular queries. In order to achieve "Applications" aspect of Big Data 2.0, we need to implement a high-level declarative language layer such as an SQL dialect. It hides all implementation details of the underlying services. Therefore, not only does it make possible for the system to introduce performance improvements without requiring any change to queries, but also more concise and easy to work with on end-user perspectives.

Streaming declarative queries are similar to database queries in how they analyze data; they differ by operating continuously on data as they arrive and by updating results incrementally in real-time. The queries are optimized automatically for high performance over distributed hardware but keep user free of complicated underlying services. Thus it poses the challenge that standard SQL semantics defined for relational databases can not be adopted to them in a straight-forward way and there is no codified model for stream processing yet.

The thesis assesses the current state of streaming SQL semantics in the literature and proposes a suitable one for big data applications. Implementation is to build a Query Compiler on top of the Apache Flink - an emerging big data analytics platform facilitating both batch and streaming processing.

Encryption of the topic is necessary: Yes/No

I approve of the suggested topic of the Master's Thesis:
Budapest, …./…./2015

……………………………………………
Zoltán Istenes

I approve of the suggested topic of the Master's Thesis:
Budapest, …./…./2015

……………………………………………

I approve of the suggested topic of the Master's Thesis:
Budapest, …./…./2015

……………………………………………
Dr. András Benczúr

I ask for the acceptance of my thesis topic.
Budapest, 25/02/2015

……………………………………………
Khuong-Duy Vu

The topic of the thesis is approved by the Department of ………………………
Budapest, …./…./2015

……………………………………………
Dr. Zoltán Horváth