Flink + Cypher Thesis Proposal

This project is about developing a Cypher (see below) front end for Flink (see below).

Task

The main task for this thesis is to develop a Cypher query language implementation on top of the Flink (Scala) APIs for Data processing. This will involve researching which APIs are reasonable to support. Considering the data model mappings between the Flink data-set models and the Cypher property graph model and type system affects the realisation. It should be examined which Flink operations can be mapped directly to Cypher and which operations have to be added to either side or implemented in the integration. Same for data types. The implementation should consider using Flink in an idiomatic, efficient manner with as little overhead through object creation or memory usage as possible. As Flink is highly parallel it should be considered which implications this has for the implementation.

Apache Flink

Flink is a streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over streams. Flink exposes the following APIs:

1. DataSet API for static data embedded in Java, Scala & Python
2. DataStream API for unbounded streams embedded in Java & Scala
3. Table API with SQL-like expression language embedded in Java & Scala

Flink also bundles libraries for domain-specific use cases, including Gelly, a graph processing API and library.

Neo4j

Neo4j is a transactional graph database implemented in Java and accessible from software written in other languages using the Cypher query language.

Cypher is a declarative, SQL-inspired, language for describing patterns in graphs, allowing users to describe what to select, insert, update or delete from a graph database without expressing exactly how to do it.
Practical Things
- The Scala Cypher language spec, parser and reference parser implementation will be provided by openCypher
- Flink and Gelly have Scala bindings at all necessary levels of the stack, so Scala is likely to be the language used for this project

Student Skills
- Required
  - Scala/Java
  - Maven
  - Git
- Relevant Knowledge/education
  - Database Systems
  - Compilers
  - Distributed Systems
  - Query planning and optimization
- Beneficial
  - Experience with larger open source projects
  - Flink / Spark / Giraph / Hadoop / MapReduce / Pregel / Cypher (Neo4j)

Supervisor
- Supervision will be lead by a professor from the student’s university
- Members of the Neo Technology engineering team will co-supervise, partially via Skype and partially on-site if/when the student visits the Neo team in Malmö, Sweden
- Vasiliki Kalavri, Flink committer & creator of the Flink Gelly API, will co-supervise