Stream Data Processing

Field Session 2015
Dan Lynn
dan.lynn@codefutures.com

Background

AgilData is a framework for writing near-real-time data processing applications on a distributed cluster of commodity hardware. Users can specify streaming or batch computations using SQL, and can extend the platform by implementing user-defined functions. While SQL is great for describing a *logical* data manipulation, it doesn't specify the ideal *physical* plan for executing a given process. Relational databases use a query optimizer to translate the logical plan specified by SQL into the physical plan that is executable by the database engine. We would like to build a streaming query planner to apply this same idea to streaming data. This is especially interesting because streams aren't guaranteed to be finite.

Project Description

Team members will design and build a stream execution planner that can translate a logical query or stream computation into a physical execution plan and then run it using the AgilData framework. The team will interact heavily with the rest of the AgilData development team and will be exposed to concepts from database design to distributed computing. For bonus points, a visualization of both the logical plan and the execution plan using something like D3 would make a very nice demo. Team size of 3 is ideal.

About Us

- We are a startup trying to make it easier for developers to work with big data.
- We're a fast-moving company led by experienced entrepreneurs.
- We are a distributed team, spread out across multiple countries.
- We care deeply about software performance.
- We primarily use Java 8, and a mix of additional languages as needed.
- We're looking to form long-term relationships with students, potentially leading to career opportunities after school.

(continued)

About You

- You're interested in functional programming concepts.
- You're fascinated by data structures.
- You're interested in working with off-heap memory.
- You're curious about machine learning and real-time optimization algorithms.
- You've worked with Java before.
- You learn quickly; e.g. you have no qualms about digging around in the code of an open source project to figure out how it works.

Why you should choose this project

- You'll get to work on a fast-moving team.
- You'll be exposed to big data design principles.
- You'll learn concepts related to distributed computing.
- You'll get to work at a variety of levels in the application stack, from the disk to the network to the CSS.