FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

UNIVERSITEIT GENT

Invitation

You are cordially invited to the public defence to obtain the academic degree of

DOCTOR OF BUSINESS ECONOMICS

by Giselle van Dongen

Open Stream Processing Benchmark:
an Extensive Analysis of Distributed Stream Processing Frameworks

Supervisor:
Prof. dr. Dirk Van den Poel

Tuesday, 19 October 2021 at 17h
in 'Faculteitsraadzaal', 2nd floor, Tweekerkenstraat 2, 9000 Ghent
or virtually via the provided link
Please confirm your attendance no later than 10 October by email to
giselle.vandongen@ugent.be

EXAMINATION BOARD

Prof. dr. Patrick Van Kenhove
Dean of the Faculty of Economics and Business Administration
Ghent University

Prof. dr. Dirk Van den Poel
Supervisor
Ghent University

Prof. dr. Dries F. Benoit
Ghent University

Prof. dr. Matthias Bogaert
Ghent University

Prof. dr. Michel Ballings
University of Tennessee at Knoxville (USA)

Dr. Anita Prinzie
Omina Technologies

Abstract

The increasing demand for real-time insight in data, such as in the IoT domain, led to the development of several solutions that allow fast, accurate, and reliable processing of large quantities of data. Every solution has its own design with its advantages and disadvantages. For the users of these technologies, it has become increasingly unclear how to choose the most suitable technology for a use case. The goal of this dissertation is twofold. First of all, we want to gain insights into the differences in behavior and performance of these frameworks. After we have built up this knowledge, we want to formulate guidelines on how to select a suitable framework for a use case.

To meet the first goal, we present Open Stream Processing Benchmark (OSPBench), a benchmarking suite that allows testing several distributed stream processing frameworks on a list of key features. We use this system to increase our knowledge on the strengths and weaknesses of four state-of-the-art stream processing frameworks: Flink, Kafka Streams, Spark Streaming, and Structured Streaming. The codebase has been open-sourced and can be found here.

In Chapter 2, we measure the latency of several operations in Flink, Kafka Streams, Spark Streaming and Structured Streaming. Latency is the time the framework needs to generate an output event from one or more input events. In this chapter, we also dive deeper into the second category of workloads, which focuses on measuring performance under different throughput levels and data characteristics. Throughput is the number of messages that can be processed per second. We measure peak throughput in three scenarios: peak sustainable throughput under constant load, peak burst throughput at startup, and throughput under periodic bursts. In Chapter 3, we address another key feature of distributed systems: fault tolerance and recovery. Distributed systems have many different components that can all fail. We experiment with (1) master failure with and without a high-availability setup with Zookeeper, (2) worker failure with and without exactly-once semantics, and (3) application, job, stage, and task failures. Finally, we evaluate the scalability of these frameworks in Chapter 4. We measure scalability in two directions: horizontally and vertically. Horizontal scaling means increasing the number of workers, while vertical scaling indicates increasing the resources per worker. In this chapter, we determine the factors which influence the scalability of a processing job.

The knowledge base that was built up throughout Chapters 2-4 is then used in Chapter 5 to build up a comparison model that helps users to structure and guide the selection process. The model covers a broad spectrum of criteria related to general characteristics, time-related features, APIs, pipeline characteristics, latency and throughput performance, scalability, elasticity, parallelization, state management, fault tolerance, and deployment. With this model, we accomplish the second goal of this dissertation, i.e. educating users in the differences between these frameworks, and guiding users in the selection process by taking into account the most important factors. In summary, this dissertation gives the reader an immersion in the domain of stream processing in all its complexity. It provides insight into the behavior and performance of four prominent distributed stream processing frameworks and gives a foundation for a structured, use-case-tailored approach to framework selection.

Curriculum vitae

Giselle van Dongen (°1993, Turnhout) obtained the degree of Master of Science in Business Engineering (Option: Data Analytics) in 2016 at Ghent University. Afterwards, she started working as a PhD researcher at Ghent University, teaching and benchmarking real-time distributed processing systems such as Spark Streaming, Flink and Kafka Streams. Concurrently, she is Lead Data Scientist at Klarrio specialising in real-time data analysis, processing and visualisation. Giselle presented her academic work at various international conferences, such as IEEE Big Data Congress, Informs Annual Meeting and Spark Summit. The second chapter of her dissertation has been published in IEEE Transactions on Parallel and Distributed Systems. The third and the fourth chapter have both been published in IEEE Access. The fifth chapter is under revision at ACM Computing Surveys.