Abstract

With the growing importance of real time data processing, human and technical latencies became a major risk factor. But through optimization frameworks, such as Apache Flink, which is specifically designed to process unbounded data streams, these latencies can be reduced. This poster will explain the anatomy of Apache Flink and its functionality.

The Anatomy of Apache Flink

Since unbounded data has to be continuously processed, the resulting technical latency is high. By using so called savepoints, persistent state images, Apache Flink is able to process unbounded data streams. With its own set of state primitives, handling vast amounts of data is made possible by the framework.

Apache Flinks programming allows an integration with common cluster resource managers as well as the deployment of Flink as a standalone cluster. It automatically identifies needed resources and manages these for the application. Even in case of failure, Flink can replace failed containers.

Reducing human latency is achieved by Apache Flinks API being build upon two models. The cost model for automatic query optimization and the algebraic model with second- and third-order functions. This layers the API into three distinct use cases. By that, join logics and the logic for query processing and data analysis are strictly separated. The code becomes more readable and thus the human latency is reduced.

Reducing technical latency with Apache Flink is made possible by its handling of iterations, which are optimized in two ways. First, unlike other Big Data Systems, iterations are not only looped but implemented as data flows with a direct feedback loop in the initial looping function. The second way is the usage of delta iterations that cause fewer dependencies. This optimizes the performance of applications using Apache Flink and reduces the technical latency.

Use cases of Apache Flink

Apache Flink being an open source framework with the functionality to process unbounded data streams, it can be used for event-driven applications, data analytics and data pipeline applications. The framework supports programs written in Java, Python, Scala and SQL.

References:
https://flink.apache.org/
"Mosaics in Big Data Data Infrastructures: Stratosphere, Apache Flink, and Beyond" a lecture held by Prof. Dr. Volker Markl, TU Berlin, on 04.01.2022