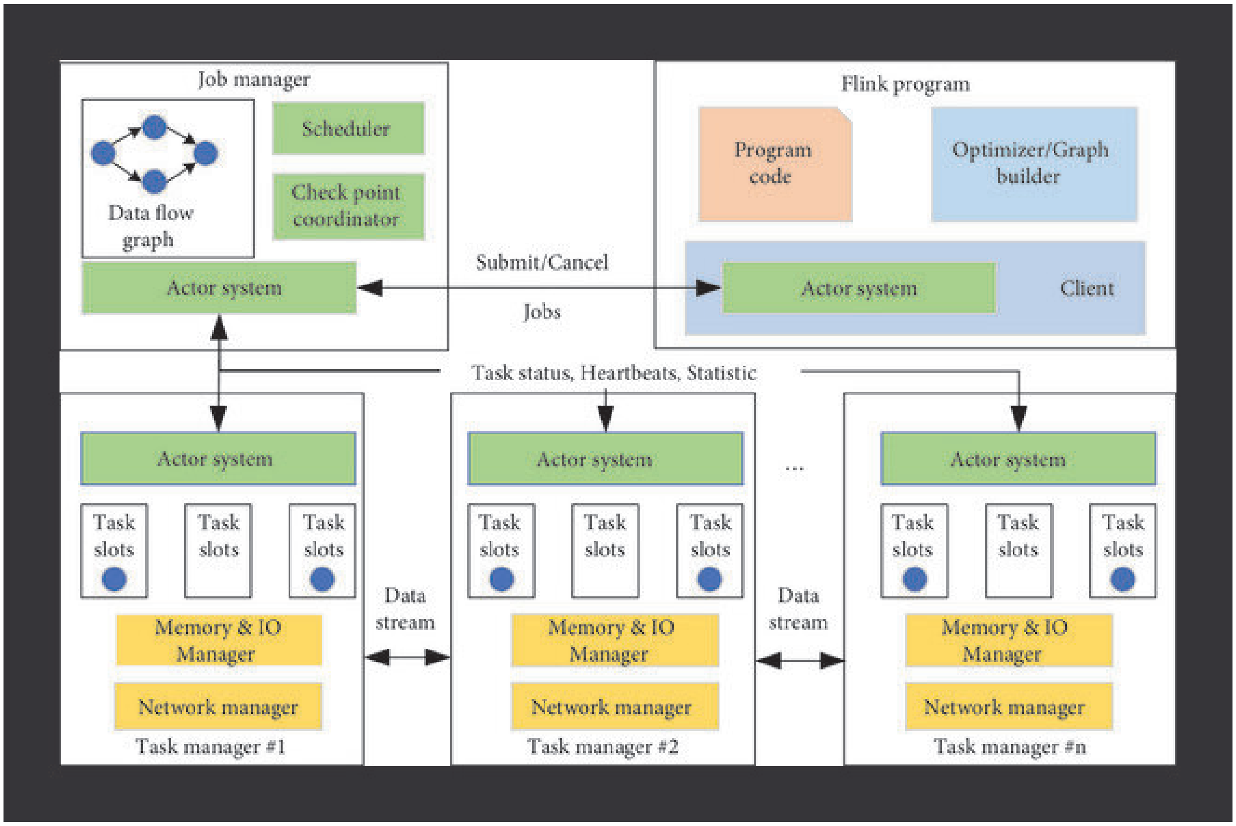
TechXplore

Automatically tuning the resource configurations for streaming data processing systems using machine learning

January 10 2023



Apache Flink architecture overview. Credit: *Intelligent Computing* (2022). DOI: 10.34133/2022/9820424

Data can be likened to a stream of water when a large amount of data is generated continuously. A variety of data including applications, networked devices, server log files, various online activities, and location-based data can form a continuous stream. We call such a form of data processing stream data.

In streaming data, various types of data sources can be collected, managed, stored, analyzed in real time and provided with information. For most scenarios where dynamic new data is continuously generated, it is beneficial to adopt streaming data processing, which is suitable for most industries and big data use cases.

Stream data processing systems are used to analyze stream data. There are already many stream data processing systems that are widely used by companies, such as Apache Flink, Apache Storm, Spark Streaming, and Apache Heron. These stream data processing applications are characterized by large deployments and long run times (months or even years) in applications, and each application runs with different data, so even small performance improvements can have significant financial benefits for companies.

To improve system performance, resource configuration parameters need to be tuned to specify the amount of resources such as CPU cores and memory used in tasks. But selecting key configuration parameters and finding their optimal values for stream data processing applications is very challenging, and manually tuning these parameters is extremely time-consuming.

For a single unknown application, a performance engineer, who has a deep understanding on the stream data processing system, may take several days or even weeks to find its optimal resource configuration.

In order to solve the above problem, researchers have started to apply machine learning methods to conduct research. A study was published in

Intelligent Computing. The authors used the Apache Flink program as an experimental stream data processing application.

The machine learning approach was used to automatically and efficiently tune the resource allocation parameters for the stream data processing application. It applies a Random Forest algorithm to build a highly accurate performance model for a stream data processing program that outputs the tail latency or throughput of the application, taking the speed of input data and key configuration parameters as input. In addition, the machine learning approach leverages the Bayesian optimization algorithm (BOA) to iteratively search the high-dimensional resource configuration space to achieve optimal performance.

This approach has been experimentally shown to significantly improve the 99th-percentile tail latency and throughput. The method proposed in this study is a parameter-tuning tool independent of the Flink system, and can be integrated into other stream processing systems, such as Spark Streaming and Apache Storm.

More information: Shixin Huang et al, Resource Configuration Tuning for Stream Data Processing Systems via Bayesian Optimization, *Intelligent Computing* (2022). DOI: 10.34133/2022/9820424

Provided by Intelligent Computing

Citation: Automatically tuning the resource configurations for streaming data processing systems using machine learning (2023, January 10) retrieved 9 April 2024 from https://techxplore.com/news/2023-01-automatically-tuning-resource-configurations-streaming.html