УДК 004.6

Методы и технологии для повышения производительности распределенных систем потоковой обработки данных

Иванов С.Е. (ИТМО)**, Конанов К.А.** (ИТМО)
Научный руководитель – кандидат физико-математических наук, доцент Иванов С.Е. (ИТМО)

Введение. Данные - ценный ресурс современного общества, они генерируются беспрецедентными и постоянно растущими темпами. Необходимость хранить, анализировать и оперативно предоставлять данные множеству пользователей ставит сложные задачи перед современными программными платформами. Каждая система, использующая большие объемы данных, имеет свои особенности в выборе модели данных, предположений об использовании, синхронизации, стратегии обработки, развертывания, гарантий согласованности, отказоустойчивости и упорядочивания. Многие компании предлагают свои решения, которые будут оптимальны под их задачи. Тем не менее, проблемы, с которыми сталкиваются системы, использующие большие объемы данных, и предлагаемые ими решения часто пересекаются [1]. В данном исследовании будут введены основные понятия современных систем для обработки потоковых данных, рассмотрены существующие решения. Также будут показаны существующие способы оптимизации отдельных компонент этих систем.

Основная часть. В рамках данной работы высоконагруженной системой будем называть систему, которая обрабатывает большое количество данных или выполняет множество операций одновременно.

Потоковая обработка данных - это способ обработки данных, при котором получение данных происходит в непрерывном режиме, по мере их появления. Данный подход обеспечивает непрерывное поступление данных, что дает возможности обработки данных и принятия решений на их основе с минимальной задержкой. Сегодня для потоковой обработки данных представлены несколько инструментов с открытым исходным кодом, что позволяет вносить в них доработки и настраивать их внутри компаний, подстраиваясь под конкретные требования и задачи. Наиболее популярные и часто встречающиеся в использовании инструменты это Spark Streaming и Apache Flink.

Существует ряд работ, где авторы предлагают различные оптимизации работы отдельных этапов работы инструментов, используемых в инфраструктуре системы обработки данных. Такими оптимизациями, например, являются

1) алгоритм StreamBed, который представляет из себя систему планирования мощностей для потоковой обработки [3]
2) алгоритм дискретизированных потоков (D-потоки), которые предназначены для проведения отказоустойчивых потоковых вычислений [4]

Такие методы и инструменты используются в компаниях с высокой нагрузкой и требованиям к консистентности, доступности [5].

Выводы. Проведен анализ методов и инструментов потоковой обработки данных.

Список использованных источников:

1. Margara, Alessandro & Cugola, Gianpaolo & Felicioni, Nicolò & Cilloni, Stefano. (2022). A Model and Survey of Distributed Data-Intensive Systems.
2. Aikins, M. V. (2023). Distributed storage systems and how they handle data consistency and reliability. Faculty of Natural and Applied Sciences Journal of Scientific Innovations, 5(1), 84–90. Retrieved from

https://www.fnasjournals.com/index.php/FNAS-JSI/article/view/206

3. Rosinosky, Guillaume, Donatien Schmitz, and Etienne Rivière. "StreamBed: capacity planning for stream processing." arXiv preprint arXiv:2309.03377 (2023).

4. Zaharia, Matei & Das, Tathagata & Li, Haoyuan & Hunter, Timothy & Shenker, Scott & Stoica, Ion. (2013). Discretized streams: Fault-tolerant streaming computation at scale. SOSP 2013 - Proceedings of the 24th ACM Symposium on Operating Systems Principles. 423-438. 10.1145/2517349.2522737.

5. Fu, Yupeng & Soman, Chinmay. (2021). Real-time Data Infrastructure at Uber.