УДК 004.042

В. В. Стецюк
Т. В. Грищук

СУЧАСНІ МЕТОДИ ОБРОБКИ ПОТОКОВИХ ДАНИХ

Вінницький національний технічний університет

Анотація
У даній роботі розглянуто сучасні методи та технології для обробки потокових даних, які гарантують ефективність, неперервність, надійність та високу швидкодію.
Ключові слова: потокові дані, обробка даних, apache spark streaming, apache sazma. apache storm.

Abstract
In this paper we consider modern methods and technologies for processing streaming data, which guarantee efficiency, continuity, reliability and high performance.
Keywords: потокові дані, обробка даних, apache spark streaming, apache sazma. apache storm.

Вступ

Потокові дані стали частиною нашого повсякденного життя, оскільки вони збираються в режимі реального часу з e-commerce та соціальних мереж, онлайн-ігор, GPS і давачів. Зазвичай такі дані не вимагають довготривалого зберігання, а важливим є якомога швидший аналіз від моменту їх отримання. У даній роботі проаналізовано методи організації та обробки потокових даних.

Результати дослідження

Потокові дані – це такі дані, які генеруються безперервно тисячами джерел та зазвичай надсилають дані одночасно, і невеликими розмірами. Потокові дані включають велику кількість різноманітних даних, таких як логування дій створених клієнтами за допомогою мобільних або веб-додатків, купівля/продаж в e-commerce, активність гравців в іграх, інформацію з соціальних мереж, фінансові торговельні майданчики або також дані з давачів підключених пристроїв [1].

Прикладами областей, де застосування обробки потокових даних є доцільним:

– давачі транспортних засобів, промислового обладнання та сільськогосподарської техніки, що передають дані до системи, яка здійснює моніторинг та попередньо виявляє потенційні загрози та автоматично створює замовлення для заміни певних запчастин, що запобігає витримці обладнання;

– відстеження змін на фондовому ринку в режимі реального часу, обчислення ризику та автоматичний перерозподіл коштів на основі руху цін на акції;

– компанія, що займається розробкою онлайн-ігор збирає потокові дані про взаємодію між гравцями та грою і передає дані на свою ігрову платформу, яка аналізує дані в режимі реального часу та пропонує стимули для залучення своїх гравців на основі зібраних даних.

На сьогодні, існує широкий спектр технологій та засобів обробки потокових даних. Провідні компанії надають можливість здійснювати обробки таких даних за допомогою розроблених хмарних рішень, таких як: Amazon Kinesis, Apache Spark Streaming, Apache Storm, Apache Samza. На рисунку 1 зображено один зі способів організації обробки потокових даних з використанням Spark Streaming [2].

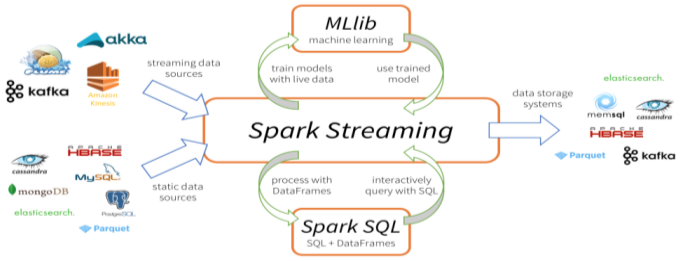

Рисунок 1 – Організація обробки потокових даних за допомогою Spark Streaming

При роботі з Apache Storm створюється граф обчислень реального часу, так звану топологію (topology), і передаємо його в кластер, де головний вузол розподіляє код між робочими вузлами для

виконання. Основними елементами топології є spout і bolt. Spouts генерують потоки даних в формі незмінних пар ключ-значення, які називаються кортежами (tuple), а bolts виконують перетворення цих потоків (підрахунок, фільтрація, тощо). Bolts, у свою чергу, можуть передавати дані іншим bolt для виконання послідовних стадій обробки [3].

Spark Streaming це розширення базового Spark API, що дозволяє організувати високопродуктивну обробку потокових даних. Дані можуть надходити із багатьох джерел, таких як Kafka, Flume, Kinesis або TCP сокетів, і можуть бути оброблені за допомогою складних алгоритмів, виражених функціями високого рівня, такими як Map, Reduce, Join та Window. По завершенню процесу обробки, дані можна вивести у різні файлові системи, бази даних, або на приборні панелі. На рисунку 2 зображено архітектуру Spark Streaming.

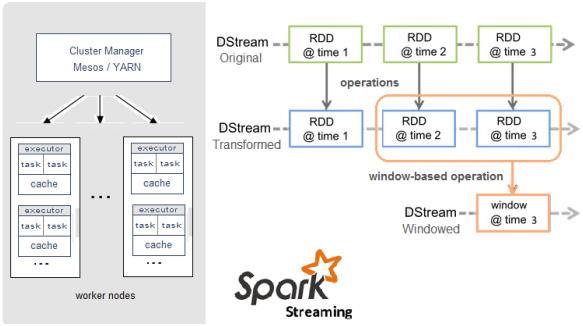

Рисунок 2 – Архітектура Spark Streaming

Концепція Apache Samza полягає в обробці повідомлень по мірі їх отримання. Потоковим примітивом Samza є не кортеж або DStream, а повідомлення (message). Потоки розбиваються на розділи (partition). Кожен розділ являє собою упорядковану послідовність доступних тільки для читання повідомлень. Кожне повідомлення має унікальний ідентифікатор. Система також підтримує пакетний режим (batching), який дозволяє послідовно приймати кілька повідомлень з одного розділу потоку. Модулі виконання і обміну повідомленнями Samza є можливими для підключення, тобто можуть бути замінені аналогами, але зазвичай використовуються YARN і Apache Kafka [3].

Всі три фреймворки чудово підходять для ефективної обробки потокових даних в реальному часі. Проте при виборі фреймворку слід керуватись наступними рекомендаціями.

Якщо потрібна високошвидкісна система обробки подій, що забезпечує інкрементні обчислення, Storm буде хорошим вибором. Якщо далі буде потрібно виконувати розподілені обчислення на вимогу, в той час як клієнт синхронно очікує результат, Storm надасть готову підсистему розподіленого віддаленого виклику процедур (distributed RPC).

Якщо необхідно збереження стану, в точності одноразова доставка повідомлень, і при цьому, не дуже турбує більш тривала затримка, тоді можемо обрати Spark Streaming. Цей фреймворк особливо підійде в тому випадку, якщо плануємо виконувати операції над графами, машинне навчання або доступ до SQL. Стек Apache Spark дозволяє використовувати спільно зі Streaming кілька інших бібліотек (Spark SQL, MLlib, GraphX) і реалізує зручну уніфіковану модель програмування. Зокрема, в поєднанні з потоковими алгоритмами, такими як потоковий метод k-середніх (streaming k-means), Spark може бути застосований для забезпечення прийняття рішень в реальному часі [3].

Висновки

У роботі розглянуто сучасні методи та технології для обробки потокових даних, які гарантують ефективність, неперервність, надійність та високу швидкодію. Наведено спектр галузей де застосування обробки потокових даних є доцільним. Проведено аналіз можливостей Spark Streaming, Apache Storm, Apache Sazma. Виявлено особливості їх застосування, основні переваги та недоліки.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. What is streaming data [Електронний ресурс] – Режим доступу до ресурсу: https://aws.amazon.com/streaming-data
2. Spark Streaming [Електронний ресурс] – Режим доступу до ресурсу: https://databricks.com/glossary/what-is-spark-streaming

3. Обработка потоковых данных: Storm, Spark, Sazma [Електронний ресурс] – Режим доступу до ресурсу: http://datareview.info/article/obrabotka-potokovyx-dannyx-storm-spark-i-samza
4. Mining of massive datasets — A. Rajamaran, J. Leskovec, J. Ullman, 2010. — 483 с.

Стецюк Вадим Валерійович— студент групи 3АКІТ-18м, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: stetsyuk.vadim@gmail.com

Науковий керівник: ***Грищук Тетяна Вікторівна*** — к.т.н., доцент, Вінницький національний технічний університет, м. Вінниця

Stetsiuk Vadym V. — Faculty of computer systems and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: stetsyuk.vadim@gmail.com

Supervisor: ***Gryshchuk Tetiana V.*** — Ph.D., Vinnytsia National Technical University, Vinnytsia.