# APACHE KAFKA VA APACHE SPARK STRUCTURED STREAMING ASOSIDA OQIMLI MA'LUMOTLARNI QAYTA ISHLASH

*Farg'ona davlat texnika universiteti*
*510-23 guruh talabasi*
***Rasuljon Turg'unov Ravshanjon o'g'li***

**Annotatsiya:** Ushbu maqolada real vaqt rejimida oqimli ma'lumotlarni qayta ishlash uchun keng qo'llaniladigan Apache Kafka va Apache Spark Structured Streaming texnologiyalari yoritilgan. Maqolada ushbu texnologiyalarning asosiy tushunchalari, ularning o'zaro integratsiyasi, ishlash tamoyillari hamda amaliy qo'llash bosqichlari ko'rib chiqiladi. Kafka va Structured Streaming'ning birgalikda qo'llanilishi katta hajmdagi ma'lumotlarni samarali, ishonchli va kengaytiriladigan tarzda qayta ishlash imkonini berishi tahlil qilinadi.

**Kalit so'zlar:** Apache Kafka, Apache Spark, Structured Streaming, oqimli ma'lumotlar, real vaqt, Big Data.

**Annotation:**This article discusses Apache Kafka and Apache Spark Structured Streaming technologies, which are widely used for real-time stream data processing. The paper examines the fundamental concepts of these technologies, their integration, operating principles, and practical implementation stages. The joint use of Kafka and Structured Streaming enables efficient, reliable, and scalable processing of large volumes of streaming data.

**Keywords:** Apache Kafka, Apache Spark, Structured Streaming, streaming data, real-time processing, Big Data.

**Аннотация:**В данной статье рассматриваются технологии Apache Kafka и Apache Spark Structured Streaming, широко применяемые для обработки потоковых данных в режиме реального времени. В работе анализируются основные понятия данных технологий, их интеграция, принципы функционирования и этапы практического применения. Совместное использование Kafka и Structured Streaming позволяет эффективно, надёжно и масштабируемо обрабатывать большие объёмы потоковых данных.

**Ключевые слова:** Apache Kafka, Apache Spark, Structured Streaming, потоковые данные, обработка в реальном времени, Big Data.

## Kirish

Zamonaviy axborot tizimlarida ma'lumotlar nafaqat katta hajmda, balki uzluksiz va yuqori tezlikda hosil bo'lmoqda. Bank tranzaksiyalari, ijtimoiy tarmoqlar, IoT qurilmalari va onlayn xizmatlar real vaqt rejimida ma'lumotlar oqimini yaratadi.

Ushbu ma’lumotlarni kechikishsiz tahlil qilish tashkilotlarga tezkor qaror qabul qilish imkonini beradi.

Shu sababli oqimli ma’lumotlarni qayta ishlash uchun Apache Kafka va Apache Spark Structured Streaming kabi texnologiyalar keng qo‘llanilmoqda. Ularning kombinatsiyasi zamonaviy ma’lumotlar arxitekturasining ajralmas qismi hisoblanadi.

**Apache Kafka va Structured Streaming tushunchasi**

Apache Kafka

Apache Kafka — bu taqsimlangan oqim platformasi bo‘lib, real vaqt rejimida ma’lumotlarni qabul qilish, saqlash va uzatish imkonini beradi. Kafka yuqori unumdorlik, xatolarga chidamlilik va kengaytirilish xususiyatlariga ega.

Apache Kafka asosan quyidagi vazifalar uchun ishlatiladi:

hodisalar (event) oqimini boshqarish;

loglarni yig‘ish;

real vaqt tahlili;

ma’lumotlar quvurlarini (data pipeline) yaratish.

Apache Spark Structured Streaming

Apache Spark Structured Streaming — bu Apache Spark’ning oqimli ma’lumotlarni qayta ishlash uchun mo‘ljallangan API’si bo‘lib, u paketli (batch) va oqimli ishlov berishni yagona modelda birlashtiradi. Structured Streaming’ning asosiy afzalligi — oqimli hisob-kitoblarni odatiy DataFrame va SQL amallari orqali ifodalash imkoniyatidir.

Bu yondashuv ishlab chiquvchilarga oqimlarni qayta ishlashda murakkab past darajadagi mexanizmlarga chuqur kirmasdan samarali ilovalar yaratish imkonini beradi.

**Nima uchun Apache Kafka va Structured Streaming’ni birgalikda ishlatish kerak**

Apache Kafka va Apache Spark Structured Streaming o‘rtasidagi integratsiya juda samarali hisoblanadi. Kafka ishonchli va bardoshli xabar brokeri vazifasini bajarsa, Spark esa kuchli hisoblash va tahlil mexanizmini ta’minlaydi.

Ularning birgalikda ishlatilishi quyidagi afzalliklarni beradi:

**Haqiqiy vaqtda ishlov berish** – ma’lumotlardan darhol xulosalar olish imkoniyati;

**Kengaytirilish (scalability)** – serverlar sonini oshirish orqali yuklamani boshqarish;

**Xatolarga bardoshlilik** – nosozliklar yuz berganda ham ma’lumotlar yo‘qolmasligi;

**Moslashuvchanlik** – turli manbalar va chiqish tizimlari bilan ishlash imkoniyati.

Natijada tashkilotlar real vaqt rejimida katta hajmdagi ma’lumotlarni ishonchli tarzda qayta ishlaydigan tizimlarni yaratishi mumkin.

**Atrof-muhitni sozlash**

Apache Kafka'da ma'lumotlarni Apache Spark Structured Streaming yordamida qayta ishlashdan oldin tizimni to‘g‘ri sozlash talab etiladi. Asosiy bosqichlar quyidagilardan iborat:

**Apache Kafka'ni o‘rnatish**

Kafka'ni lokal kompyuterda yoki serverda o‘rnatib, mavzular (topics) yaratish va xabarlarni nashr qilish imkonini beruvchi muhit tayyorlanadi.

**Apache Spark'ni o‘rnatish**

Kafka bilan ishlashni qo‘llab-quvvatlaydigan Spark versiyasi o‘rnatiladi va tegishli Kafka paketlari qo‘shiladi.

**Structured Streaming'ni Kafka'ga ulash**

Spark ilovasi Kafka brokeriga ulanadi, kerakli topic va ishlov berish mantiqi belgilanadi.

Ushbu sozlashlar oqimli ma'lumotlarni qayta ishlash uchun mustahkam texnik asos yaratadi.

**Ma'lumotlarni qayta ishlash bosqichlari**

Apache Kafka va Structured Streaming yordamida oqimli ma'lumotlarni qayta ishlash quyidagi bosqichlarda amalga oshiriladi:

1. **Kafka'dan ma'lumotlarni o‘qish**

   Spark'ning readStream metodi orqali Kafka topic'laridan ma'lumotlar olinadi.

2. **Ma'lumotlarni transformatsiya qilish**

   DataFrame va Spark SQL amallari yordamida filtrlash, guruhlash, agregatsiya va xaritalash ishlari bajariladi.

3. **Natijani saqlash (sink)**

   Qayta ishlangan ma'lumotlar ma'lumotlar bazasiga, boshqa Kafka topic'iga yoki fayl tizimiga yoziladi.

Transformatsiya jarayonida ma'lumotlar sifati va yaxlitligini ta'minlash muhim bo‘lib, xatolarni boshqarish mexanizmlari joriy etilishi lozim.

**Amaliy tajribadan o‘rganilgan darslar**

Apache Kafka va Spark Structured Streaming bilan ishlash jarayonida quyidagi muhim xulosalarga kelish mumkin:

**Kichikdan boshlash** – avval oddiy quvur yaratib, keyin murakkablashtirish;

**Monitoring va optimallashtirish** – tizim yuklamasi va ishlash samaradorligini doimiy nazorat qilish;

**Hujjatlardan foydalanish** – Kafka va Spark rasmiy hujjatlari muhim bilim manbai hisoblanadi.

Amaliy tajriba va nazariy bilimlar uyg‘unligi ushbu texnologiyalarni chuqur o‘zlashtirishga yordam beradi.

## Xulosa

Apache Kafka va Apache Spark Structured Streaming texnologiyalarining integratsiyasi real vaqt rejimidagi oqimli ma’lumotlarni samarali qayta ishlash uchun kuchli yechim hisoblanadi. Kafka ma’lumotlarni ishonchli uzatish va saqlashni ta’minlasa, Spark Structured Streaming ularni tezkor va moslashuvchan tahlil qilish imkonini beradi. Ushbu texnologiyalar zamonaviy Big Data tizimlarida muhim rol o‘ynaydi va kelajakda ham o‘z ahamiyatini saqlab qoladi.

## Foydalanilgan adabiyotlar

1. **Kreps J., Narkhede N., Rao J.** Kafka: A Distributed Messaging System for Log Processing // *Proceedings of the NetDB Conference*. – 2011.
2. **Apache Software Foundation.** Apache Kafka Documentation. – URL: https://kafka.apache.org/documentation
3. **Zaharia M., Xin R. S., Wendell P., et al.** Apache Spark: A Unified Engine for Big Data Processing // *Communications of the ACM*. – 2016. – Vol. 59(11). – P. 56–65.
4. **Apache Software Foundation.** Apache Spark Structured Streaming Programming Guide. – URL: https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html
5. **Karau H., Warren R.** *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*. – O’Reilly Media, 2017.