

№14 Дәріс. Big Data және Stream Processing технологиялары (Kafka, Spark Streaming, Flink)

Дәріс мақсаты:

Big Data және ағындық деректерді өңдеу (stream processing) концепцияларын түсіндіру, Kafka, Spark Streaming және Apache Flink технологияларының архитектурасы, мүмкіндіктері және қолданылу салаларын талқылау.

Кілт сөздер:

Big Data, Stream Processing, Apache Kafka, Spark Streaming, Apache Flink, batch processing, real-time analytics, event processing, distributed computing.

Дәрістің жоспары:

1. Кіріспе: Big Data ұғымы
2. Big Data-ның сипаттамалары (3V/5V)
3. Stream Processing vs Batch Processing
4. Kafka архитектурасы және жұмыс принципі
5. Spark Streaming технологиясы
6. Apache Flink жүйесі
7. Қолдану салалары
8. Қорытынды
9. Бақылау сұрақтары
10. Пайдаланылған әдебиеттер

1. Кіріспе

Big Data — көлемі өте үлкен, құрылымы әртүрлі және жылдам өзгеретін деректер жиыны. Дәстүрлі жүйелер мұндай деректерді өңдей алмайды, сондықтан арнайы үлестірілген деректер платформалары қолданылады.

2. Big Data сипаттамалары

Классикалық 3V моделі:

V	Түсініктеме
Volume	Үлкен көлем (терабайт – петабайт)
Velocity	Жоғары жылдамдықта келетін деректер
Variety	Түрлі форматтар: мәтін, видео, сенсор деректері

Кеңейтілген 5V моделі:

V	Түсініктеме
Veracity	Деректердің сенімділігі
Value	Деректерден құн (аналитика, шешім қолдау)

3. Stream Processing vs Batch Processing

Салыстыру	Batch Processing	Stream Processing
Мақсаты	Үлкен деректерді пакеттік	Нақты уақыттағы деректерді

Салыстыру	Batch Processing	Stream Processing
	өңдеу	өңдеу
Уақыт талабына	Минут/сағат/күн	Миллисекунд/секунд
Мысал	Hadoop	Kafka + Flink/Spark Streaming

Мысал:

- YouTube статистикасы — stream processing
- Журнал логтарын талдау — batch

4. Apache Kafka

Kafka — үлкен ағындық деректерді буферлеу, сақтау және жеткізу үшін қолданылатын messaging платформасы.

Kafka архитектурасы.

Компонент	Сипаттамасы
Producer	Деректерді жібереді
Consumer	Деректерді оқиды
Broker	Kafka сервері
Topic	Хабарламалар логикалық арнасы

Ерекшеліктері:

- Жоғары өткізу қабілеті
- Үлкен көлемді stream деректерді өңдеу
- Fault tolerance және replication
- Real-time messaging

Қолдану мысалдары

- Банктік транзакциялар мониторингі
- IoT сенсор деректері
- Clickstream (веб қолданушы әрекеті)

5. Spark Streaming

Apache Spark Streaming — деректерді микроблоктар түрінде өңдейтін real-time жүйе.

Артықшылықтары:

- Micro-batch моделі (тұрақты, тез)
- MLlib арқылы машиналық оқыту мүмкіндігі
- Fault tolerance
- Ядросында Spark Engine → жоғары өнімділік

Мысал қолдану:

- Нақты уақыттағы лог-аналитика
- Желілік шабуылдарды анықтау (IDS)
- Тікелей dashboard аналитика

6. Apache Flink

Apache Flink — нақты real-time stream engine, event-time processing және low-latency қолдвйды.

Артықшылықтары

Артықшылығы	Сипаттамасы
True streaming	Нақты event time
Low latency	Миллисекунд деңгей
Stateful stream processing	Жай-күй сақтау
Exactly-once guarantee	Қателерден қорғау

Flink vs Spark

Салыстыру	Spark Streaming	Flink
Механизм	Micro-batch	True streaming
Кешігу	100ms–секунд	миллисекунд
Нақты уақыт	Жақсы	Керемет

7. Қолдану салалары

Сала	Мысал
Банкинг	Алаяқтықты анықтау (fraud detection)
IoT	Сенсор деректерін өңдеу
Көлік	Навигация, жол қозғалысын басқару
Интернет-маркетинг	Реал-тайм ұсыныстар жүйесі
Киберқауіпсіздік	Атака детекциясы
Телеком	Трафик талдау

8. Қорытынды

Big Data мен Stream Processing технологиялары қазіргі цифрлық әлемнің жүрегі. Kafka, Spark Streaming және Flink сияқты жүйелер нақты уақыт аналитикасын іске асыруға мүмкіндік береді.

Тиімді қолдану арқылы ұйымдар:

- жедел шешімдер қабылдайды
- операциялық тиімділікті арттырады
- тәуекелдер мен алаяқтықтың алдын алады

9. Бақылау сұрақтары

1. Big Data ұғымы қандай ерекшеліктерге ие?
2. Batch және Stream өңдеу айырмашылығы қандай?
3. Kafka архитектурасының негізгі элементтері?
4. Spark Streaming қалай жұмыс істейді?
5. Flink-тің Spark-тан артықшылығы неде?
6. Ағындық деректер қандай салаларда қолданылады?

10. Пайдаланылган әдебиеттер

1. Marz N., Warren J. *Big Data: Principles and best practices*.
2. Kreps J. *Kafka: The Definitive Guide*. O'Reilly.
3. Karau H. *Learning Spark*. O'Reilly.
4. Tzoumas K., *Streaming Systems with Apache Flink*.
5. Apache Documentation: Kafka, Spark, Flink official docs.
6. MIT Big Data & Real-Time Analytics Lecture Notes.