PIPELINE BIG DATA IOT

TRAITEMENT EN TEMPS RÉEL DE DONNÉES CAPTEURS

Presentée par : Ben Imran Ahlam

Encadré par: Mr.Badir Hassan

PLAN

- Contexte et problématique
- Objectifs du Projet
- Architecture du Système
- Technologies Utilisées
- Implémentation Technique
- Résultats & Démonstration
- Défis Rencontrés
- Conclusion & Perspectives

Contexte : L'Explosion de l'IoT

25B+
Objets IoT Connectés dans le monde

< 1s
Latence requise pour décisions critiques

79 ZB
Données générées par an (IDC)

Problématique

Comment concevoir un pipeline Big Data capable de traiter des millions d'événements IoT en temps réel tout en garantissant :

✓ Fiabililité (pas de perte de données)
✓ Scalabilité (millions d'événements/jour)
✓ Performance (latence < 2 secondes)
✓ Résilience (tolérance aux pannes)

Objectif

1. Ingestion de Données

- Simuler 100 capteurs IoT
- Kafka pour ingestion
- Débit de 1000 evt/s

2. Traitement Temps Réel

- Spark Streaming
- Agrégations par fenêtre
- Latence < 2s

3. Stockage Distribué

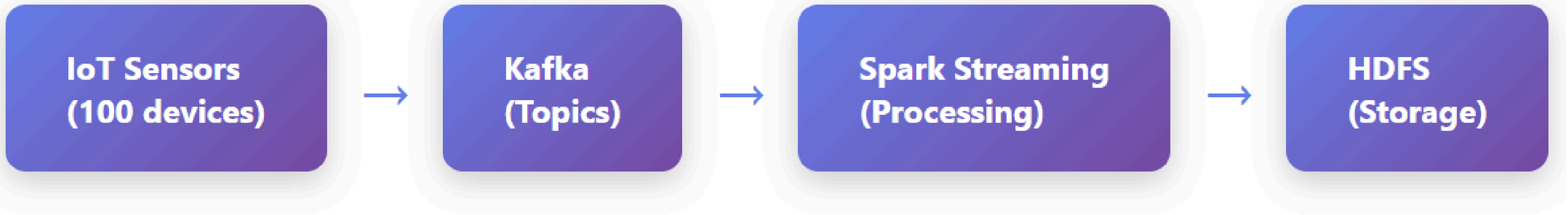
- Partitionnement par région
- Format Parquet
- HDFS

4. Résilience

- Récupération automatique
- Checkpointing Spark
- Réplication Kafka

ARCHITECTURE GLOBALE

100 capteurs



Pipeline complet : Ingestion → Traitement → Stockage

COMPOSANTS DU SYSTÈME


```
PS C:\Users\lenovo\iot-bigdata-pipeline> docker-compose up -d
[+] Running 6/6
 ✔ Container spark-master   Started
 ✔ Container namenode       Running
 ✔ Container zookeeper      Running
 ✔ Container kafka          Running
 ✔ Container datanode       Running
 ✔ Container spark-worker   Started
```

Rôle

- Coordination
- Message Broker
- Orchestration
- Exécution
- Métadonnées
- Stockage

IMPLÉMENTATION - IoT DATA PRODUCER

Simulation de 100 capteurs
- Répartis dans 4 régions (North, South, East, West)
- Génération de données réalistes

2. Débit
- 100 messages toutes les 100ms
- = 1000 événements/seconde

3. Configuration Kafka Producer
- Bootstrap servers: localhost:9092

IMPLÉMENTATION - SPARK STREAMING

Lecture du Stream Kafka
- Subscribe au topic: iot-sensors-raw
- Parsing JSON avec schéma défini

Agrégations par Fenêtre
- Fenêtre temporelle: 5 minutes
- Watermark: 10 minutes (gestion des retards)
- Groupement par région

```
PS C:\Users\lenovo\iot-bigdata-pipeline> sbt "runMain IoTDataProducer"
[info] welcome to sbt 1.11.7 (Oracle Corporation Java 21.0.9)
[info] loading settings for project iot-bigdata-pipeline-build from plugins.sbt...
[info] loading project definition from C:\Users\lenovo\iot-bigdata-pipeline\project
[info] loading settings for project iot-bigdata-pipeline from build.sbt...
[info] set current project to IoT-BigData-Pipeline (in build file:/C:/Users/lenovo/iot-bigdata-pipeline/)
[info] running IoTDataProducer
Starting IoT Data Producer - Topic: iot-sensors-raw
Sent 1000 messages
Sent 2000 messages
Sent 3000 messages
Sent 4000 messages
Sent 5000 messages
Sent 6000 messages
Sent 7000 messages
Sent 8000 messages
Sent 9000 messages
Sent 10000 messages
```

Containers Docker (docker ps)

```
PS C:\Users\lenovo\iot-bigdata-pipeline> docker run -d --name spark-master -p 7077:7077 -p 8080:8080 apache/spark:3.5.0 tail -f /dev/null
>>
fdef00fcf98f7a25c45a6ce52a9f3b9267cf1fa0aa5122cd7bc045eca31e2f06
PS C:\Users\lenovo\iot-bigdata-pipeline> docker exec -d spark-master /opt/spark/sbin/start-master.sh
PS C:\Users\lenovo\iot-bigdata-pipeline> docker rm -f spark-worker
Error response from daemon: No such container: spark-worker
PS C:\Users\lenovo\iot-bigdata-pipeline> docker run -d --name spark-worker --link spark-master:spark-master apache/spark:3.5.0 tail -f /dev/null
c149679f6d49f3e30b3469258a380897c2e89fe079bd47caf60bb391fea23148
PS C:\Users\lenovo\iot-bigdata-pipeline> docker exec -d spark-worker /opt/spark/sbin/start-slave.sh spark://spark-master:7077
PS C:\Users\lenovo\iot-bigdata-pipeline> docker ps
>>
CONTAINER ID   IMAGE                                          COMMAND                  CREATED             STATUS                       PORTS
                     NAMES
c149679f6d49   apache/spark:3.5.0                             "/opt/entrypoint.sh …"   15 seconds ago      Up 13 seconds
                     spark-worker
fdef00fcf98f   apache/spark:3.5.0                             "/opt/entrypoint.sh …"   50 seconds ago      Up 48 seconds                0.0.0.0:7077->7077/tcp, 0.0
.0.0:8080->8080/tcp  spark-master
39579b061471   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   About an hour ago   Up About an hour (healthy)   9864/tcp
                     datanode
746eece60be3   confluentinc/cp-kafka:7.5.0                    "/etc/confluent/dock…"   About an hour ago   Up About an hour             0.0.0.0:9092->9092/tcp
                     kafka
21a0330325b6   confluentinc/cp-zookeeper:7.5.0                "/etc/confluent/dock…"   About an hour ago   Up About an hour             2888/tcp, 0.0.0.0:2181->218
1/tcp, 3888/tcp      zookeeper
9cadee1a271d   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   About an hour ago   Up About an hour (healthy)   0.0.0.0:9000->9000/tcp, 0.0
.0.0:9870->9870/tcp  namenode
```

Producer en Exécution

```
PS C:\Users\lenovo\iot-bigdata-pipeline> sbt "runMain IoTDataProducer"
[info] welcome to sbt 1.11.7 (Oracle Corporation Java 21.0.9)
[info] loading settings for project iot-bigdata-pipeline-build from plugins.sbt...
[info] loading project definition from C:\Users\lenovo\iot-bigdata-pipeline\project
[info] loading settings for project iot-bigdata-pipeline from build.sbt...
[info] set current project to IoT-BigData-Pipeline (in build file:/C:/Users/lenovo/iot-bigdata-pipeline/)
[info] running IoTDataProducer
Starting IoT Data Producer - Topic: iot-sensors-raw
Sent 1000 messages
Sent 2000 messages
Sent 3000 messages
Sent 4000 messages
Sent 5000 messages
Sent 6000 messages
Sent 7000 messages
Sent 8000 messages
Sent 9000 messages
Sent 10000 messages
```

Résultats Spark Streaming

```
-------------------------------------------
Batch: 1
-------------------------------------------
+------------------------------------------+------+------------------+------------------+------------+
|window                                    |region|avg_temperature   |max_temperature   |sensor_count|
+------------------------------------------+------+------------------+------------------+------------+
|{2025-12-24 08:30:00, 2025-12-24 08:35:00}|South |24.570041451948413|40.79081423662285|110         |
|{2025-12-24 08:30:00, 2025-12-24 08:35:00}|North |15.202634667431969|25.87142435353177|89          |
|{2025-12-24 08:30:00, 2025-12-24 08:35:00}|West  |17.465077444881125|34.73636808438689|103         |
|{2025-12-24 08:30:00, 2025-12-24 08:35:00}|East  |19.444966654653427|33.23103458601831|98          |
+------------------------------------------+------+------------------+------------------+------------+
```

Données HDFS Stockées


```
root@ce8e90c25b7f:/# hdfs dfs -ls /iot/raw-data
Found 5 items
drwxr-xr-x   - spark supergroup          0 2025-12-24 08:51 /iot/raw-data/_spark_metadata
drwxr-xr-x   - spark supergroup          0 2025-12-24 08:51 /iot/raw-data/region=East
drwxr-xr-x   - spark supergroup          0 2025-12-24 08:51 /iot/raw-data/region=North
drwxr-xr-x   - spark supergroup          0 2025-12-24 08:51 /iot/raw-data/region=South
drwxr-xr-x   - spark supergroup          0 2025-12-24 08:51 /iot/raw-data/region=West
```

Performances du Système

1000 evt/s

Débit d'ingestion stable

100 capteurs × 10 mesures/s

< 2s

Latence de traitement Spark

Agrégations + shuffle

86.4M

Événements/jour traités

~15 GB non compressé

65%

Réduction de taille

Parquet Snappy vs JSON

MERCI POUR VOTRE ATTENTION !