Le NoSQL

1. Description détaillée des pipelines

A. Pipeline 1 — Ingestion Temps Réel (CDC OLTP → Kafka → NoSQL/Datalake)

Sources

- Transactions (OLTP)
- Refunds, disputes
- Customers, merchants

Mécanisme

- **CDC** (Change Data Capture) via Debezium ou native engine
- Écrit dans **Kafka** sous forme d'événements transaction_created, refund_created, etc.

Sorties

- NoSQL (fraud_events, features, mirrors)
- Data Lake (zone *raw* sous format parquet/JSON)

SLA

- Latence < 3 secondes

B. Pipeline 2 — Clickstream (App/Web → Kafka → NoSQL)

Sources

- Événements de navigation
- Conversions
- Abandons
- Contenu vu

Destination

- NoSQL (collection clickstream_events)
- Data Lake (raw → staged)

Techno

- Kafka Connect
- Schemas Avro / Protobuf
- Partitionnement par user_id ou date

SLA

- Débit élevé (100k events/sec)

C. Pipeline 3 — Logs (Microservices → Kafka → NoSQL)

Sources

- Logs applications
- Errors
- Events de monitoring (security, auth, audit)

Destination

- NoSQL : stockage temps réel
- Data Lake : archivage long terme

SLA

- Latence < 2 secondes

---

2. Pipelines Batch (Airflow)

A. Pipeline : Raw → Staged → Curated → DWH

Étapes Airflow :

1. **Ingest Raw**
    - récupère data du Data Lake (format brut)
2. **Clean / Validate**
    - normalisation (monnaies, statuts, champs)
3. **Deduplicate**
4. **Enrichissement**
    - jointures (geo, merchants, devices…)
5. **Load DWH (OLAP)**

    - tables fact & dimensions

Fréquence

- batch toutes les 15 minutes
- nightly pour rébuild complet

B. Pipeline ML — Feature Engineering (Airflow)

Inputs

- NoSQL features
- OLTP CDC
- Clickstream
- Historique DWH

Transformations

- Window functions (7d, 1h)
- Scores comportementaux
- Vectorisation
- Normalisation

Output

- Feature Store (NoSQL ou Hudi/Iceberg)

Fréquence

- 5 minutes (fraude)
- 1h (marketing)
- nightly (modèles long terme)

---

3. Pipelines ML Temps Réel

A. Training Pipeline

- Airflow / MLflow
- récupère features
- train modèles (fraude, churn, scoring utilisateur)

- validation + tests + drift detection

B. Inference Pipeline

- API REST ML ou Flink/Spark Streaming
- reçoit un événement Kafka → calcule un score → renvoie vers :
    - NoSQL (fraud_events)
    - OLTP (update real-time status)
    - Kafka (alerting)

SLA

- temps d'inférence < 100 ms

---

4. Pipelines OLAP → BI

DWH vers dashboards

- Looker, Metabase, PowerBI

Tables servies :

- fact_transactions
- dim_customer
- fact_fraud
- fact_clickstream

SLA

- fraîcheur toutes les 15 minutes / 1h selon contexte.

---

5. Stratégies d'observabilité des pipelines

Monitoring :

- Airflow : SLA alerts
- Kafka : consumer lag
- Prometheus : metrics ingestion
- Grafana : dashboards pipeline

Alerting :

- latence > threshold
- erreur parsing
- drift ML