🏗️ Architecture moderne Big Data Analytics

Nouvelles architectures analytiques

🔷 Lambda Architecture

- **Combine batch (Spark SQL, ETL périodique) + streaming (Structured Streaming, Kafka)**
- **Avantage : précision + temps réel**
- **Spark 3+ simplifie la gestion du double flux avec le même code DataFrame**

🔷 Kappa Architecture

- **Tout passe par le flux temps réel (Kafka + Spark Streaming)**
- **Le batch est remplacé par la relecture du topic Kafka**
- **Avantage : plus simple, un seul pipeline ; support natif dans Spark 3+**

1️⃣ Ingestion & Intégration des données

Collecter les données depuis les sources (bases, API, fichiers, SaaS, Kafka, etc.)

Catégorie	Outils	Rôle
Ingestion batch & streaming	**Airbyte**, **Fivetran**, **Kafka**, **NiFi**, **Flume**	Connecter les sources et charger les données dans un data lake/warehouse
Message broker / streaming	**Apache Kafka**	Transport temps réel, file d'attente distribuée

2️⃣ Stockage & Data Lake

Lieu où sont stockées les données brutes et transformées.

Catégorie	Outils	Rôle
Object storage	**MinIO**, **Amazon S3**, **Azure Blob Storage**	Stocker de gros volumes de données non structurées (parquet, csv, json…)
Data Lakehouse	**Dremio**, **Databricks**, **Snowflake**, **Delta Lake**, **Iceberg**	Combiner lac + entrepôt (SQL + performance)

3️⃣ Orchestration & Workflow

Gérer les dépendances, planifier et exécuter les pipelines de données.

Catégorie	Outils	Rôle
Orchestration	**Apache Airflow**, **Dagster**, **Prefect**	Automatiser les pipelines (exécuter ETL, transformations, contrôles)

4️⃣ Transformation & Modélisation

Transformer les données brutes en données prêtes pour l'analyse.

Catégorie	Outils	Rôle
Data transformation (ELT)	**dbt (Data Build Tool)**	Gérer les transformations SQL dans l'entrepôt, versionner, documenter, tester
Compute engine	**Spark**, **Flink**, **Dremio**	Exécuter les transformations à grande échelle

5️⃣ Data Quality & Testing

Garantir la fiabilité et la cohérence des données.

Catégorie	Outils	Rôle
Contrôle qualité	**Great Expectations**, **dbt tests**, **Soda**, **Monte Carlo**	Vérifier la cohérence, la fraîcheur, la complétude des données

6️⃣ Gouvernance & Catalogage

Assurer la conformité, la traçabilité, et la documentation des données.

Catégorie	Outils	Rôle
Data catalog	**Collibra**, **Alation**, **DataHub**, **Amundsen**	Centraliser la documentation et la découverte des données
Gouvernance	**Apache Atlas**, **OpenMetadata**, **DataHub**	Gérer les politiques, lineage, conformité RGPD, rôles et sécurité

7️⃣ Data Consumption (Consommation)

Mise à disposition des données aux utilisateurs métiers, data scientists, BI, IA.

Catégorie	Outils	Rôle
BI / visualisation	**Power BI**, **Tableau**, **Superset**, **Metabase**	Créer des dashboards et indicateurs
Machine Learning	**Python (pandas, scikit-learn)**, **Spark ML**, **TensorFlow**, **Dremio**	Exploiter les données pour la prédiction et l'IA

8. Sécurité & Conformité (Security Layer)

Protéger les données, contrôler les accès et assurer la conformité réglementaire.

Catégorie	Outils / Technologies	Rôle
Authentification & IAM	**LDAP**, **Kerberos**, **OAuth2**, **Azure AD**, **AWS IAM**	Gérer l'identité et l'accès des utilisateurs
Chiffrement & Sécurité des données	**Ranger**, **Knox**, **Vault**, **SSL/TLS**, **KMS**	Chiffrement au repos et en transit, gestion des clés
Contrôle d'accès & Audits	**Apache Ranger**, **Atlas**, **DataHub**	Politiques RBAC, ABAC, logs d'accès, audit
Conformité & RGPD	**Collibra**, **DataHub**, **OneTrust**	Masquage des données sensibles, consentement, traçabilité
Monitoring & Sécurité réseau	**Prometheus**, **Grafana**, **ELK**, **SIEM**	Surveiller les anomalies, intrusion, performance

Étape	Catégorie	Exemples
1. Ingestion	Airbyte, Kafka, NiFi	Collecte des données
2. Stockage	MinIO, Dremio, S3	Data lake / lakehouse
3. Orchestration	Airflow	Planification et exécution
4. Transformation	dbt, Spark	Modélisation analytique
5. Qualité	Great Expectations, Soda	Vérification des données
6. Gouvernance	DataHub, Collibra	Traçabilité, catalogage
7. Consommation	Power BI, Tableau	Visualisation, IA

Étape	Catégorie	Exemples
8. **Sécurité**	Ranger, Knox, Atlas	Chiffrement, contrôle d’accès, conformité

🔐 **Flux visuel (mise à jour avec sécurité)**

Sources (API, fichiers CSV, logs)

↓

Kafka *(streaming temps réel)*

↓

Airbyte / Airflow *(ingestion + orchestration)*

↓

MinIO / S3 *(data lake sécurisé – chiffrement & IAM)*

↓

DBT + Dremio / Spark *(transformation & analyse)*

↓

Contrôles qualité & gouvernance *(Great Expectations + DataHub)*

↓

Sécurité & conformité *(Ranger, Knox, Atlas)*

↓

Power BI / Superset / ML *(consommation des données fiables et sécurisées)*