Revue architecture – Haute Vue (3h)

Public cible : comité architecture / métiers. Objectif : valider alignement et risques majeurs.

1. Flux de données

- Quelles sources/consommateurs prioritaires ? Volumes/jour et latences cibles.
- Chemins critiques : Ingress → NiFi → Kafka → Trino/Superset ; Airflow orchestrant.
- Mode de transport : lot vs streaming, contrats de schéma, gestion des évolutions (compatibilité).
- Points de contrôle : qualité (SODA), validation métier, observabilité des flux (lag, erreurs).

2. Stockage

- Données brutes/curées : MinIO (object), Longhorn (blocs/PVC), Postgres (métadonnées, métastore).
- Stratégies de réplication et RPO/RTO visés ; sauvegardes et tests de restore.
- Classes de stockage et coûts ; lifecycle et rétention (froids/chauds).

3. Traitement

- Pipeline batch (Airflow + Trino/Spark) vs temps réel (Kafka + NiFi).
- Dimensionnement : workers Airflow, executors Spark, workers Trino ; capacité pic et marges.
- Isolation des workloads (quotas/limites) et priorisation (jobs critiques vs ad hoc).

4. Exposition

- Points d'entrée : MetalLB VIP + Nginx Ingress ; FQDN, TLS terminés où ?
- Consommation : Superset (BI), Zeppelin (notebooks), API éventuelles (Trino/REST).
- Gestion des versions et compatibilité (clients JDBC/ODBC, API).

5. Sécurité

- AuthN : Keycloak OIDC (durée tokens, clients), Kerberos pour Kafka/Trino.
- AuthZ : OPA (policies, fallback), RBAC cluster/applicatif.
- Chiffrement : en transit (TLS interne/edge), au repos (Longhorn, MinIO, PG, Kafka).

- Secrets : stockage, rotation, audits ; durcissement images/OS, scans vulnérabilités.

6. Gouvernance

- Catalogue/metadata : OpenMetadata (sources couvertes, owners, lineage).
- Qualité : SODA (règles, alertes, remédiation).
- Traçabilité : logs/metrics/traces centralisés ? Rétention et accès.
- Conformité : données sensibles, masquage/anonymisation, droits d'usage.

7. Comparaison outils (à acter en séance)

- Ingestion : NiFi vs alternatives (Kafka Connect ?), critères : latence, ops, connecteurs.
- Traitement : Trino vs Spark pour interactif/ad hoc ; coûts et élasticité.
- Visualisation : Superset vs autres BI (features, SSO, cache).
- Stockage : MinIO/Longhorn vs options managées (si pertinentes) ; coûts, SLA, ops.

8. Risques majeurs à trancher

- Single points : coord Trino, Redis master, PG primary, Ingress VIP.
- Bascule testée ? (Redis/PG/Kafka/KDC/Ingress). Scénarios DR (perte VM, réseau, dépendance externe).
- Capacité : headroom vs pics prévus ; plan d'auto-scaling ou d'alerting.

9. Décisions attendues en fin de revue

- Paramètres cibles : RPO/RTO, latences, SLO par domaine (ingress, auth, Kafka, DB, jobs).
- Politique de schéma (compat, registry), politique de cache (Redis TTL, invalidation).
- Plan d'actions court terme (bascule/backup/tests) et jalons de mise en œuvre.