Pipeline de Données d'Assurance

De MinIO à Power BI : Architecture Complète de Data Engineering

Découvrez comment transformer des données brutes d'assurance en insights stratégiques grâce à une architecture moderne de data engineering. Ce projet démontre l'implémentation complète d'un pipeline ETL avec MinIO, Dremio et Power BI pour optimiser la rentabilité et réduire les risques.

Architecture Technique du Projet

MINIO

dremio

Power BI

Couche Stockage

MinIO - Stockage objet S3-compatible pour l'ingestion des données brutes (clients, polices, sinistres, paiements, interactions)

Pourquoi ? Scalabilité, compatibilité S3, performance élevée pour le stockage distribué

Couche Transformation

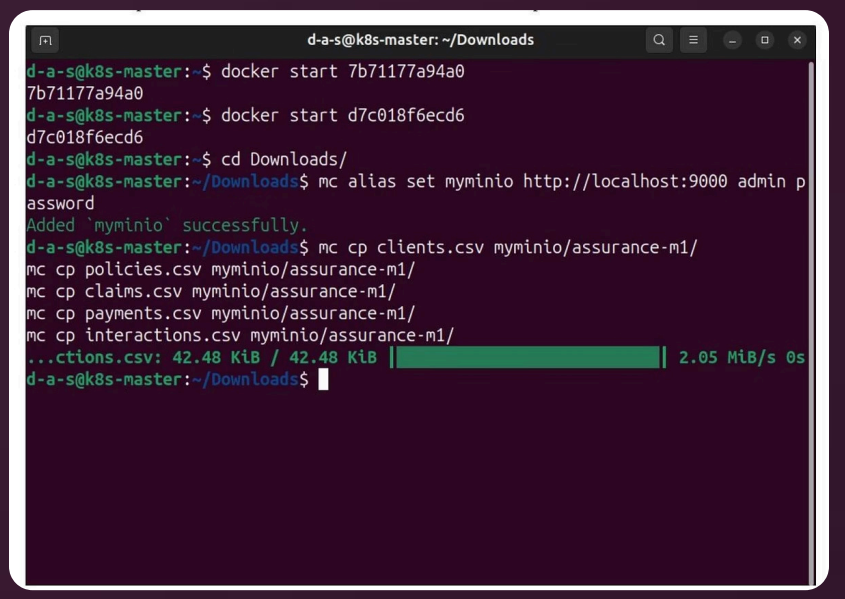
Dremio - Moteur SQL pour le nettoyage et la transformation des données en architecture médaillée (Bronze/Silver/Gold)

Pourquoi ? Requêtes SQL rapides, virtualisation des données, optimisation automatique

Couche Visualisation

Power BI - Plateforme d'analyse et de visualisation pour les KPIs métier et dashboards interactifs

Pourquoi ? Intégration native, visualisations riches, partage facilité



Étape 1 : Ingestion des Données dans MinIO

01

Configuration de l'accès MinIO

Établissement de la connexion avec les credentials admin et configuration de l'alias pour les opérations futures

02

Création du bucket assurance-ml

Initialisation du conteneur de stockage dédié au projet d'assurance avec les permissions appropriées

03

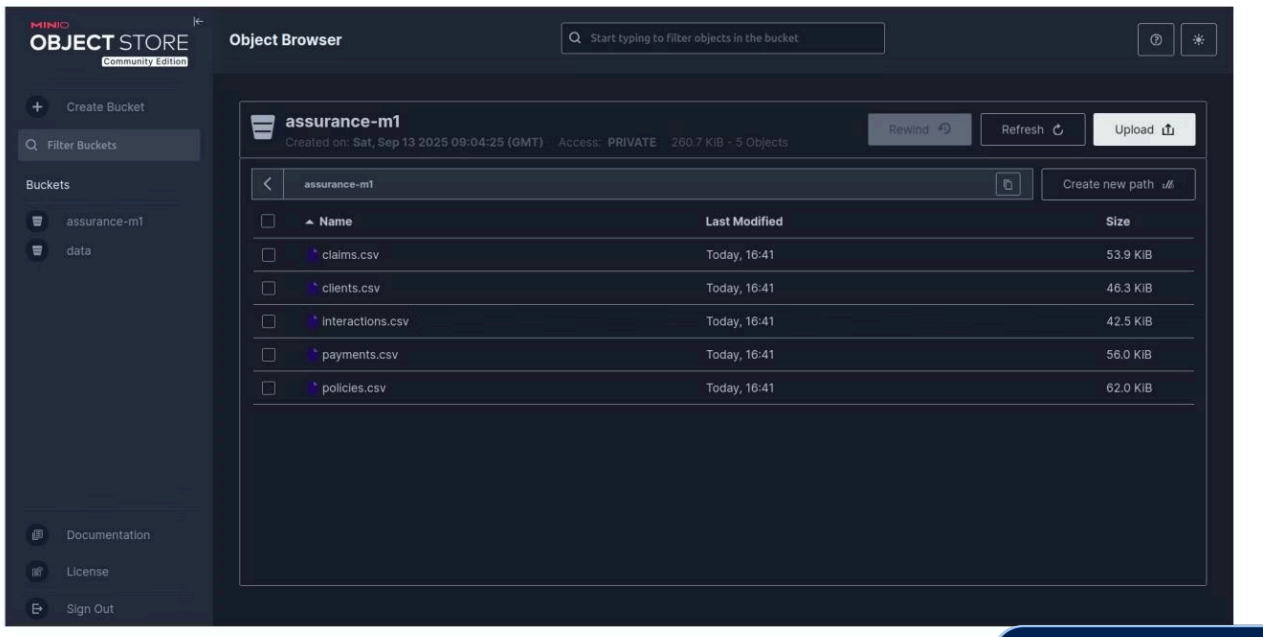
Upload des 5 fichiers CSV

Transfert des datasets : clients.csv, policies.csv, claims.csv, payments.csv, interactions.csv vers le bucket

04

Vérification de l'intégrité

Contrôle de la présence et de l'accessibilité de tous les fichiers dans le bucket via commande mc ls



Étape 2 : Architecture Médaillée dans Dremio

Couche Bronze

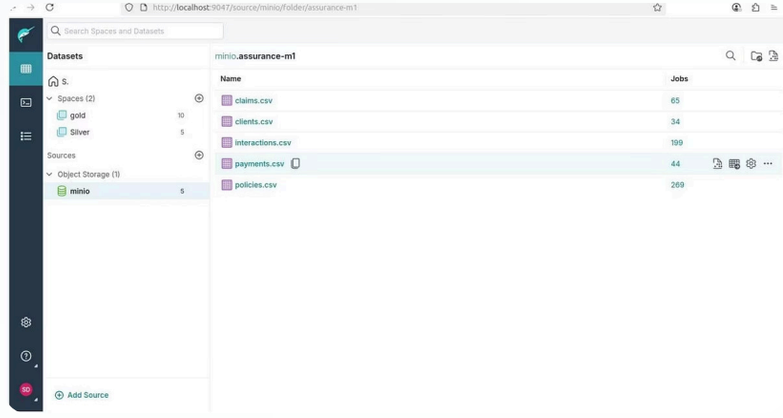
Données brutes ingérées depuis MinIO. Connexion S3 configurée avec endpoints personnalisés. Promotion des 5 datasets CSV en tables Dremio avec détection automatique des schémas.

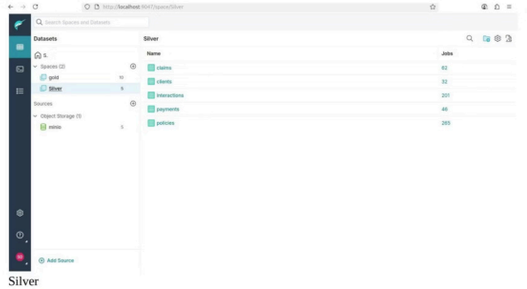
Couche Silver

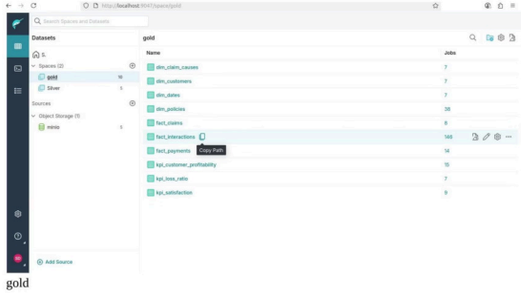
Nettoyage et standardisation : normalisation des formats de dates, conversion des devises en XOF (EUR×655.957, USD×600), harmonisation des catégories, suppression des doublons et valeurs nulles.

Couche Gold

Modèle en étoile optimisé : 4 dimensions (clients, polices, dates, causes) et 3 faits (paiements, sinistres, interactions) avec KPIs calculés et vues analytiques prêtes pour Power BI.




Silver


gold

Transformations SQL Clés de la Couche Silver

1

Clients

Normalisation des noms avec INITCAP, standardisation des formats de dates multiples (DD/MM/YYYY et YYYY-MM-DD), nettoyage des emails en minuscules, extraction des chiffres des numéros de téléphone, calcul de l'âge et segmentation en groupes d'âge.

2

Polices

Harmonisation des types de polices (Auto, Habitation, Santé, Vie), conversion universelle des primes en XOF avec taux de change fixes, standardisation des canaux de vente et statuts, gestion des doublons avec ROW_NUMBER() et critères de départage multiples.

3

Sinistres

Conversion des montants estimés et payés en XOF, catégorisation automatique des causes (Accident, Vol, Incendie, Dégât des eaux, Tempête), normalisation des statuts de sinistres, identification de la source de déclaration (Assuré/Tiers).

4

Paiements

Déduplication par payment_id avec conservation du plus récent, standardisation des méthodes de paiement (Carte, Virement, Chèque, Espèces), conversion des montants en XOF, nettoyage des références de transaction.

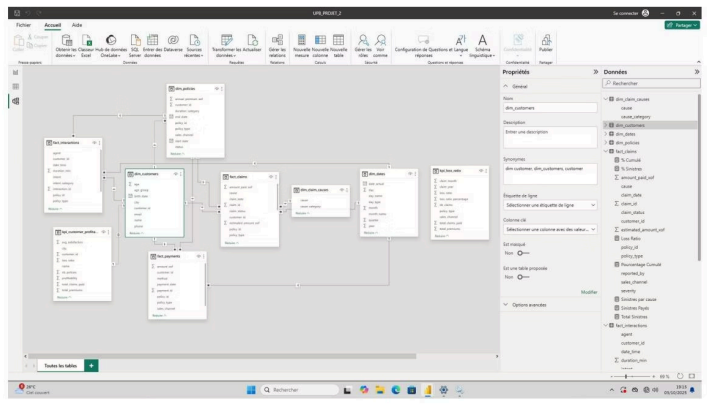
5

Interactions

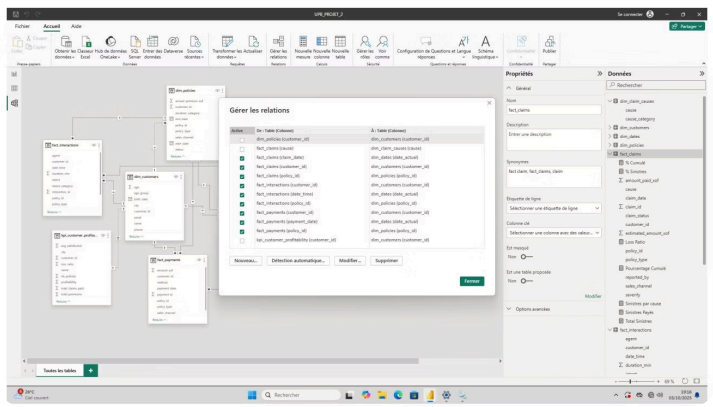
Parsing des timestamps multiformats, normalisation de la durée en minutes, validation et bornage de la satisfaction (1-5), catégorisation des intentions (Sinistre, Réclamation, Renouvellement, Information, Paiement).

Modélisation Power BI : Schéma en Étoile

Tables de Dimensions

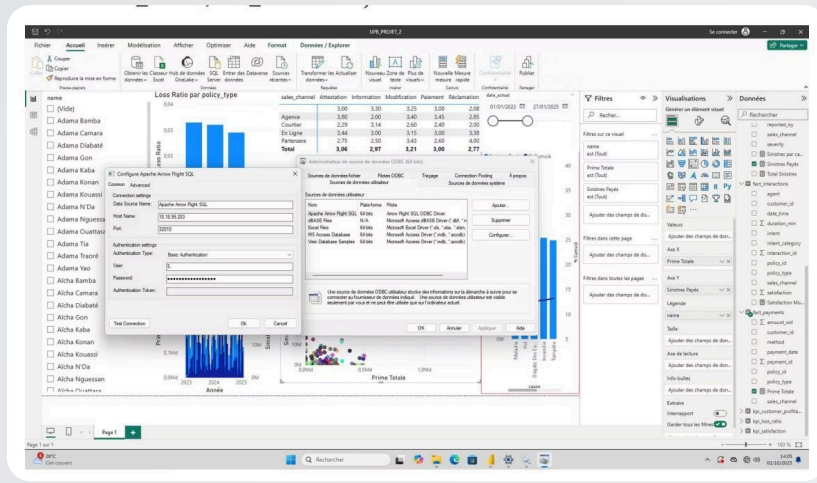


- **dim_customers** : Profils clients avec segmentation par âge (Jeune <25, Adulte 25-40, Mûr 41-60, Senior >60)
- **dim_policies** : Caractéristiques des contrats avec catégorisation par durée (Court/Moyen/Long terme)
- **dim_date** : Calendrier complet avec année, trimestre, mois, jour, type de jour (Weekend/Weekday)
- **dim_causes** : Taxonomie des causes de sinistres avec regroupement par catégorie



Tables de Faits

- **fact_payments** : Transactions de paiement avec montants en XOF
- **fact_claims** : Sinistres avec montants estimés/payés et niveau de sévérité
- **fact_interactions** : Historique des contacts clients avec scores de satisfaction



Mesures DAX Stratégiques

Loss Ratio

```
DIVIDE([Sinistres Payés], [Prime Totale], 0)
```

Indicateur clé de rentabilité mesurant le rapport entre sinistres payés et primes collectées. Un ratio >100% indique une perte.

Satisfaction Moyenne

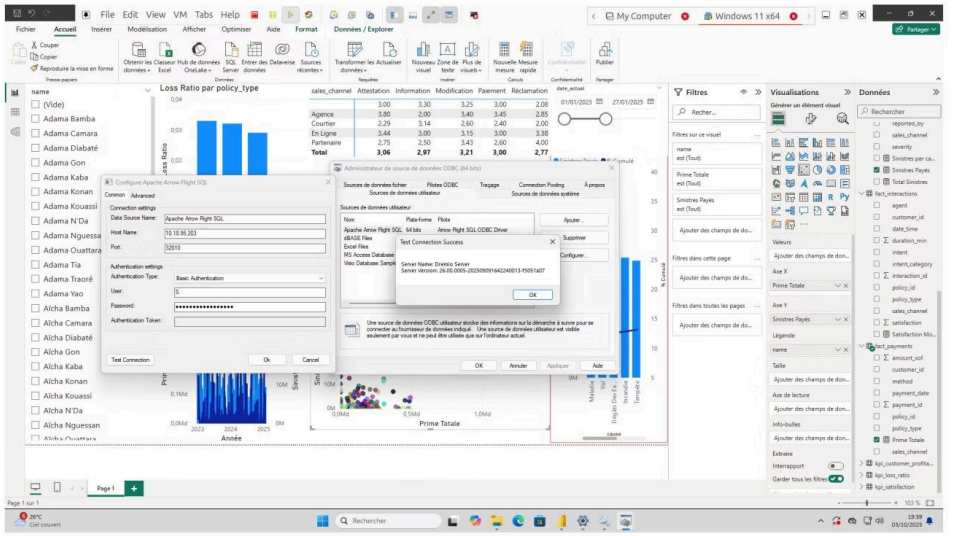
```
AVERAGE(fact_interactions[satisfaction])
```

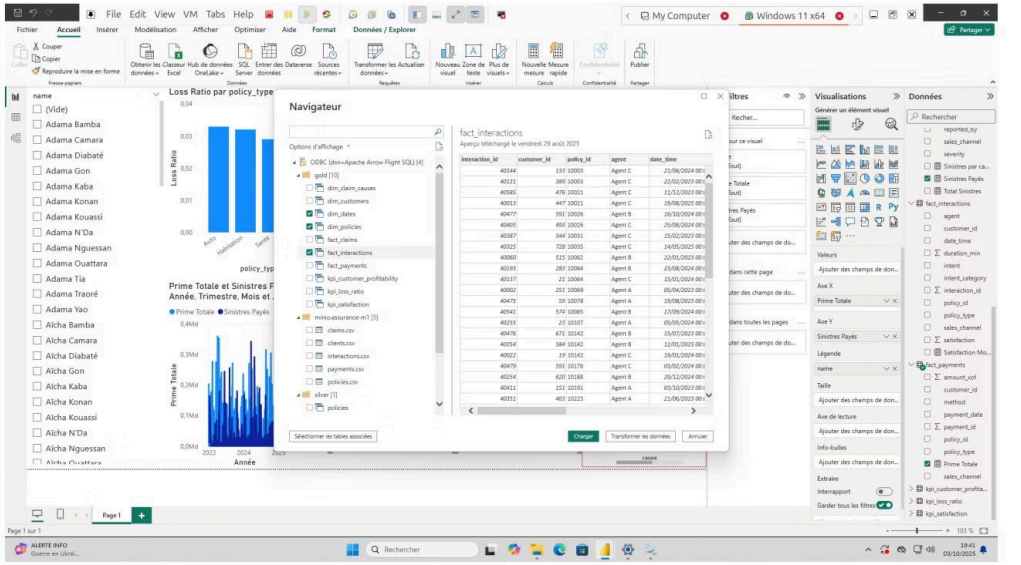
Score moyen de satisfaction client sur échelle 1-5, permettant d'identifier les canaux et produits nécessitant des améliorations.

Pourcentage Cumulé

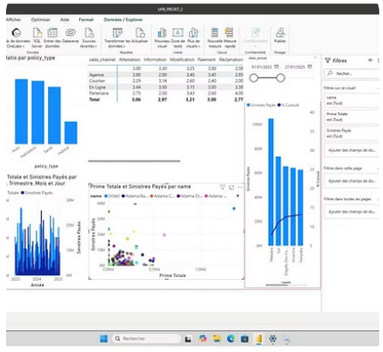
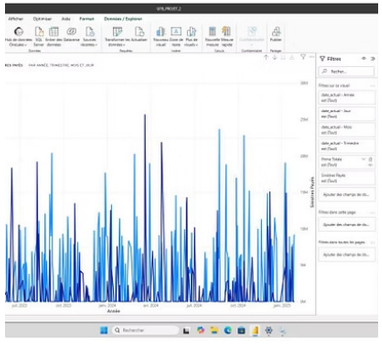
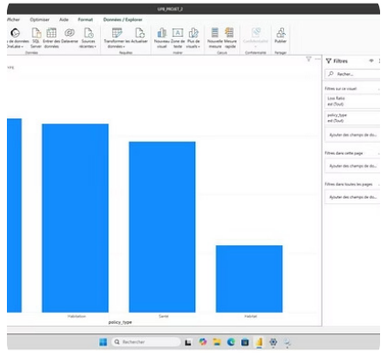
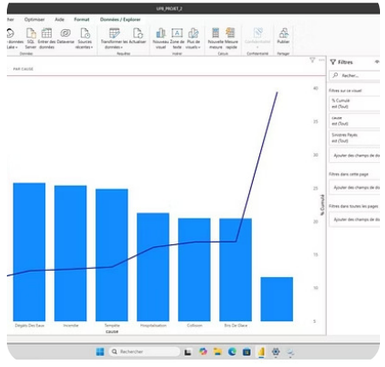
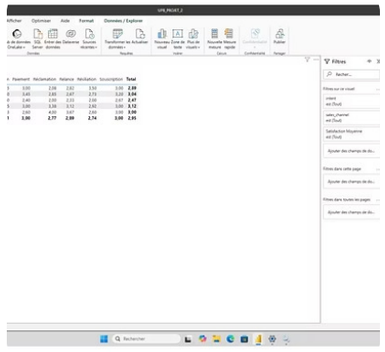
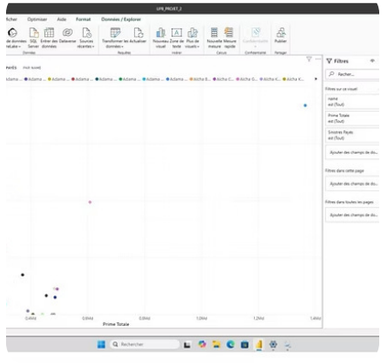
```
DIVIDE([Sinistres Payés], [Total Sinistres], 0)
```

Analyse de Pareto pour identifier les 20% de causes générant 80% des coûts de sinistres, guidant les actions préventives.





Dashboards Power BI : Insights Visuels

Évolution Temporelle

Courbes comparatives des primes collectées vs sinistres payés par mois, révélant les tendances saisonnières et pics d'activité

Loss Ratio Segmenté

Histogrammes détaillant la rentabilité par type de produit (Auto, Habitation, Santé, Vie) et canal de distribution

Pareto des Causes

Identification des causes de sinistres les plus coûteuses : Maladie, Objets volés, Dégâts des eaux représentent 60% des coûts

Matrice Satisfaction

Analyse croisée de la satisfaction client par canal de vente et type d'interaction, score moyen entre 2.08 et 4.00

Détection d'Outliers

Nuage de points identifiant les contrats à risque : primes basses mais sinistres élevés nécessitant une révision tarifaire

Insights Stratégiques et Recommandations

Performance par Type de Police

Constat : Auto, Habitation et Santé affichent un loss ratio similaire de 3%, tandis qu'Habitat ne présente que 1%.

Action : Analyser en profondeur la performance exceptionnelle d'Habitat (peu de sinistres ou primes élevées ?) et revoir les politiques de souscription des produits à 3% pour optimiser la rentabilité.

Optimisation des Canaux de Vente

Constat : Le canal "En Ligne" montre le ratio le plus bas (2.08) tandis que "Partenaire" atteint 4.00 sur les réclamations.

Action : Investir massivement dans les canaux digitaux rentables, auditer les partenaires pour comprendre leur forte sinistralité, et reformer les commissions des courtiers à risque élevé.

Segmentation Client Avancée

Constat : Certains clients ont des primes élevées avec sinistres élevés (risqués mais rentables), d'autres ont des sinistres quasi-nuls malgré des primes significatives.

Action : Implémenter un scoring de risque personnalisé basé sur l'historique pour identifier les profils rentables et ajuster les tarifs des segments à risque.

Gestion de la Saisonnalité

Constat : Pics visibles de sinistres fin 2024 et mi-2025, avec une tendance cyclique stable.

Action : Analyser les causes saisonnières (météo, événements), constituer des réserves préventives et lancer des campagnes marketing ciblées pendant les périodes à forte sinistralité.

Prévention par Cause

Constat : Maladie, Objets volés et Dégâts des eaux concentrent la majorité des coûts de sinistres selon l'analyse de Pareto.

Action : Prioriser les actions de prévention sur ces causes (campagnes santé, sécurité domestique), revoir les conditions générales (plafonds, franchises, exclusions) pour limiter l'exposition.

Synthèse : Impact Business et Prochaines Étapes

Résultats Clés du Projet

5

Datasets Intégrés

Clients, polices, sinistres, paiements, interactions unifiés en un modèle cohérent

3

Couches de Données

Architecture médaillée Bronze/Silver/Gold garantissant qualité et traçabilité

7

KPIs Stratégiques

Loss ratio, satisfaction, rentabilité client, sévérité des sinistres calculés en temps réel

Technologies Maîtrisées

- **MinIO** : Stockage objet distribué et scalable
- **Dremio** : Virtualisation et transformation SQL avancée
- **Power BI** : Modélisation DAX et visualisations interactives
- **Architecture Médaillée** : Best practices data engineering

Plan d'Action Stratégique

Domaine	Recommandation
Type de police	Revoir produits avec ratio > 3%
Canal de vente	Promouvoir le canal en ligne
Clientèle	Segmentation par rentabilité & risque
Saisonnalité	Analyser pics pour actions préventives
Causes sinistres	Réduire impact des causes fréquentes

> "Cette solution complète permet d'identifier les segments à risque et de proposer des actions correctives concrètes pour améliorer la rentabilité. L'architecture en trois couches assure la qualité des données et facilite l'analyse décisionnelle."