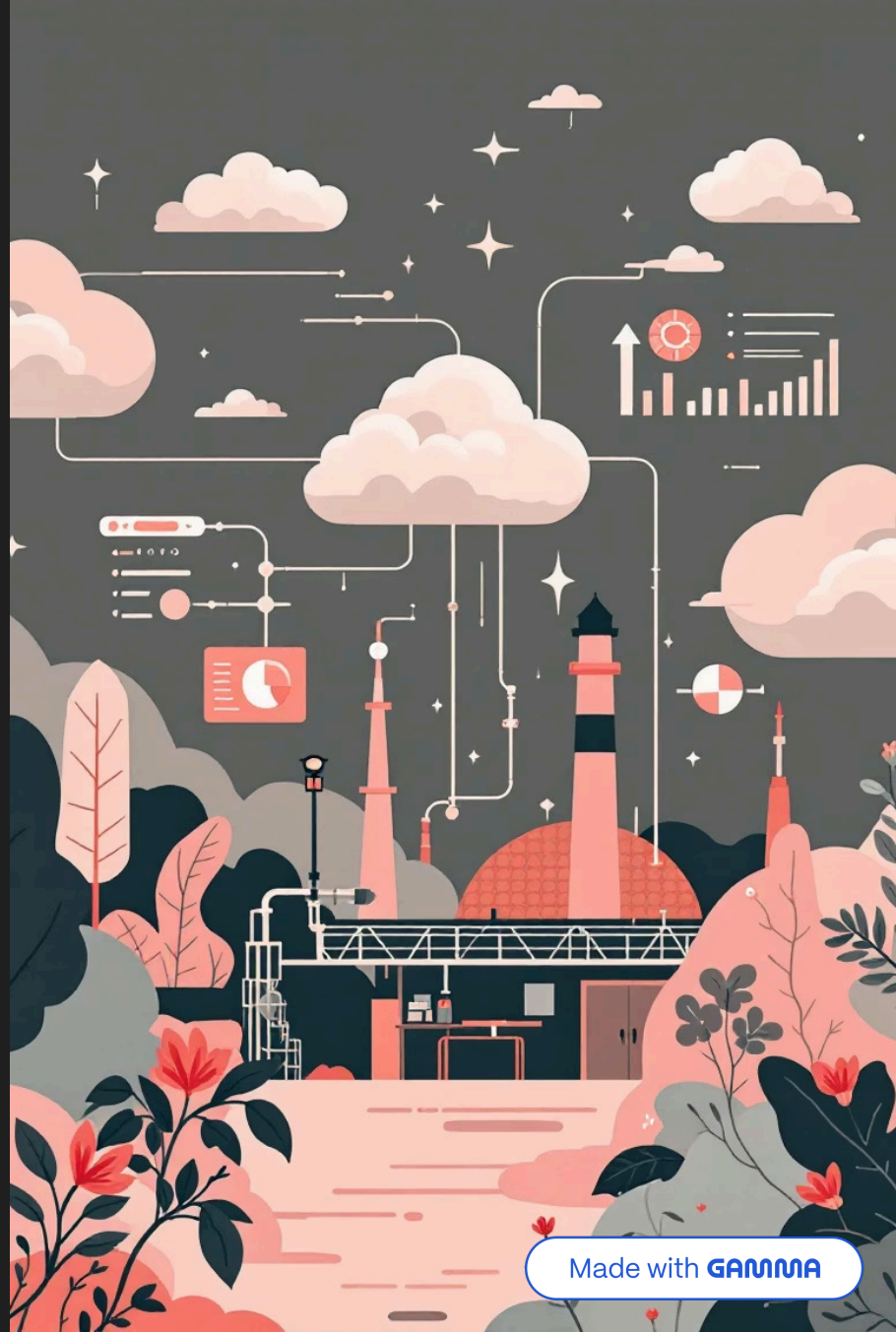
Pipeline ETL Cloud pour l'analyse de la qualité de l'air

Une solution de data engineering pour un futur plus sain.



Contexte et Objectifs du Projet

Module & Type

Data Engineering avec un pipeline ETL automatisé.

Source & Cloud

API Geodair, hébergé sur Google Cloud Platform.

Équipe Dév

Mehdi BEN CHEIKH, Priscilia GBOSSAME, Paul Thurin KENFACK, Younes BELBOUAB.

Objectifs Détaillés du Pipeline

1

Collecte Quotidienne

Acquérir des données sur la qualité de l'air chaque jour.

2

Traitement des Données

Nettoyage, structuration et enrichissement pour l'analyse.

3

Stockage Analytique

Optimiser le stockage pour des requêtes rapides et efficaces.

4

Dashboard Décisionnel

Alimenter un tableau de bord pour l'aide à la décision.

5

Pipeline Robuste

Assurer l'automatisation, la scalabilité et la robustesse du système.

Vision Globale de l'Architecture ETL

Extraction (E)

Récupération des données brutes.

Transformation (T)

Nettoyage et modélisation des données.

Visualisation BI

Looker Studio pour les rapports.

Chargement (L)

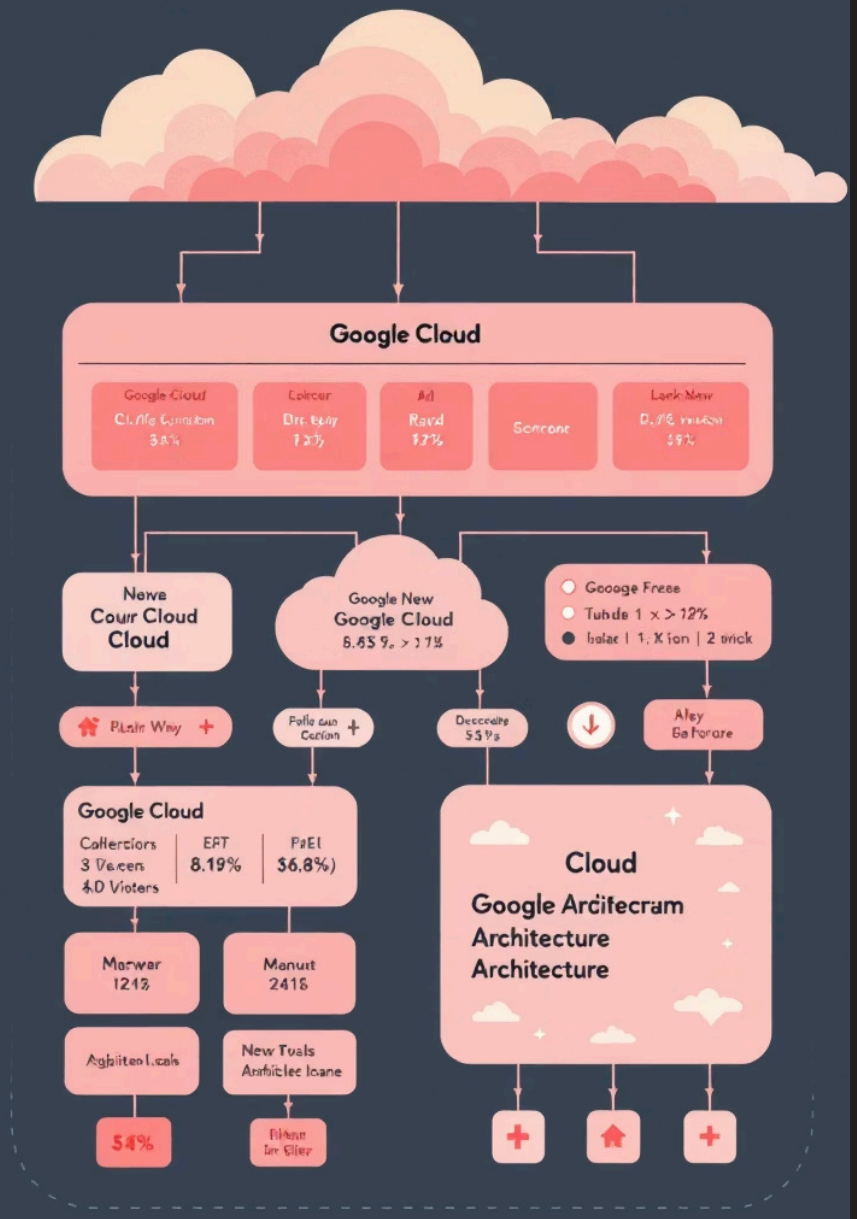
Injection dans l'entrepôt de données.

Entrepôt Analytique

BigQuery pour l'optimisation.

Stockage Cloud

Raw Zone et Transformed Zone.



Architecture Cloud sur Google Cloud Platform

Cloud Provider

Google Cloud Platform (GCP).

Services Clés

- Cloud Scheduler pour l'ordonnancement des tâches.
- Cloud Functions pour l'exécution sans serveur.
- Cloud Storage pour le stockage objet.
- BigQuery comme entrepôt de données analytique.
- Looker Studio pour la Business Intelligence.

Étape 1 : Extraction des Données (Extract)

1

Source de Données

API Geodair, fournissant des données environnementales.

2

Fréquence

Extraction quotidienne pour une actualisation régulière.

3

Format

Données brutes au format CSV.

4

Stockage Initial

Cloud Storage – Raw Zone pour l'historisation des données brutes.

Étape 2 : Transformation des Données (Transform)

Nettoyage & Normalisation

Uniformisation des formats et suppression des incohérences.

Typage des Colonnes

Assignation des types de données appropriés.

Gestion des Manquants

Stratégies pour les valeurs nulles ou manquantes.

Enrichissement Métier

Ajout de données contextuelles et calculées.

Modélisation en Étoile

Préparation pour le schéma en étoile.

Stockage Intermédiaire

Transformed Zone dans Cloud Storage.

Étape 3 : Chargement et Modélisation (Load)

Chargement Final

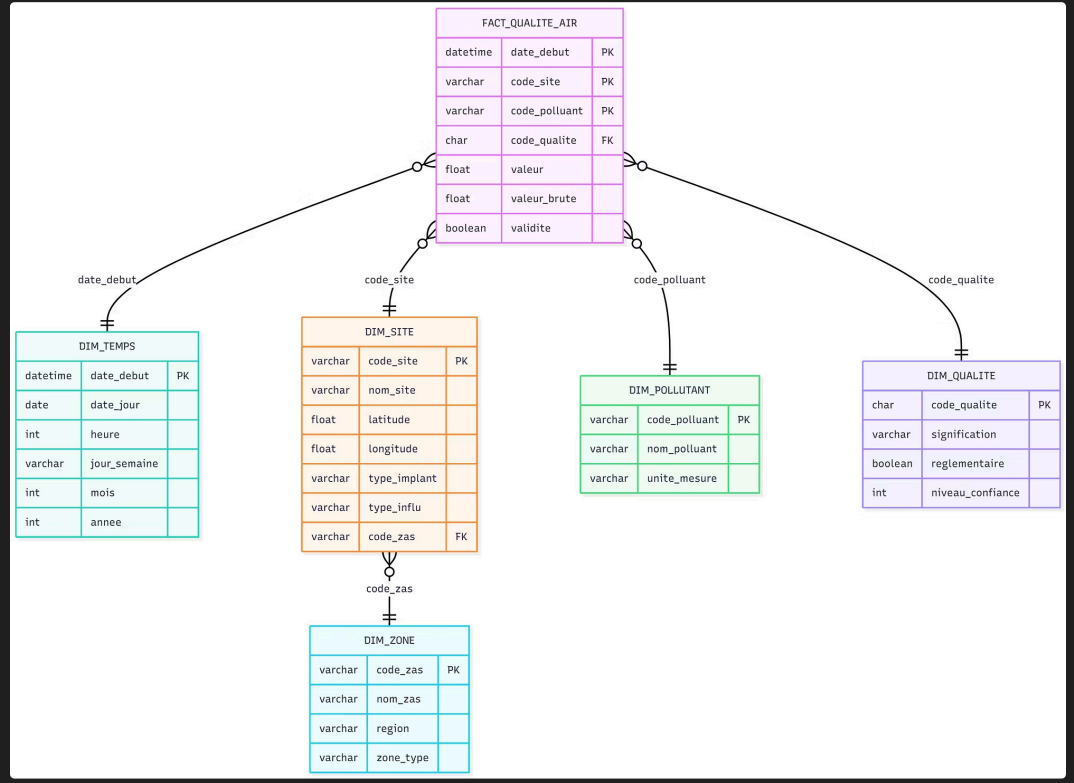
Les données transformées sont chargées dans BigQuery.

Modèle en Étoile

Implémentation d'un Star Schema pour l'optimisation des requêtes BI.

- Tables de faits (FACT_QUALITE_AIR)
- Tables de dimensions (Temps, Site, Polluant, Qualité, Zone)

Ce modèle assure une performance analytique élevée pour Looker Studio.



Détails du Modèle de Données : Schéma en Étoile

FACT_QUALITE_AIR

Granularité : site, polluant, date.

Mesures : valeur, valeur_brute, validité.

Dimension Temps

Date, année, mois, jour, heure, etc.

Dimension Site

Nom du site, coordonnées géographiques, type de site.

Dimension Polluant

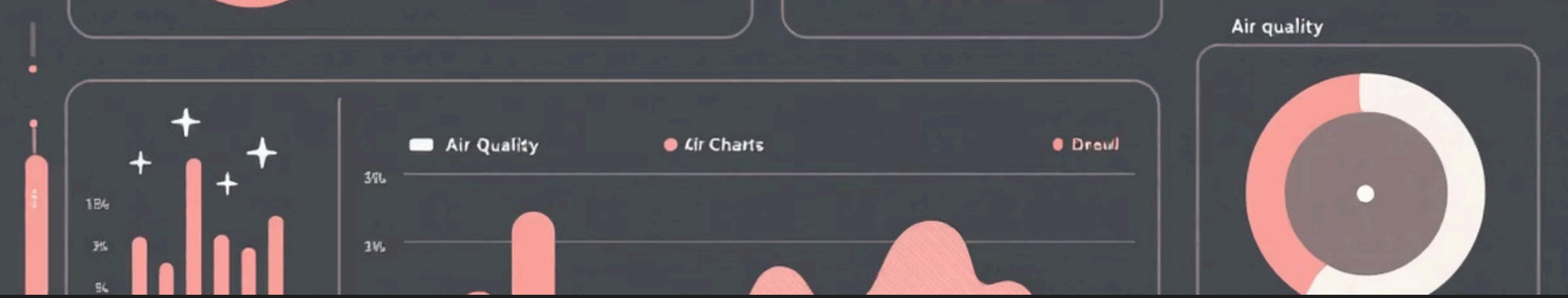
Nom du polluant, unité de mesure, seuils.

Dimension Qualité

Indices de qualité, catégories (bon, moyen, mauvais).

Dimension Zone

Région, département, ville, code postal.



Visualisation, Organisation, Difficultés & Perspectives

Visualisation avec Looker Studio

- Concentration par polluant.
- Évolution temporelle des données.
- Comparaisons géographiques pour l'analyse.
- Aide précieuse pour la décision environnementale.

Organisation & Difficultés

- Répartition des rôles et collaboration via Git.
- Gestion des données hétérogènes et des erreurs API.
- Défis de normalisation et déploiement Cloud.

Perspectives Futures

- Mise en place de monitoring et d'alertes.
- Intégration de nouvelles sources de données.
- Renforcement de la sécurité et gouvernance des données.