Projet de gestion d’infrastructure de données

## Données:

Source de données: [Brazilian E-Commerce Public Dataset by Olist](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)

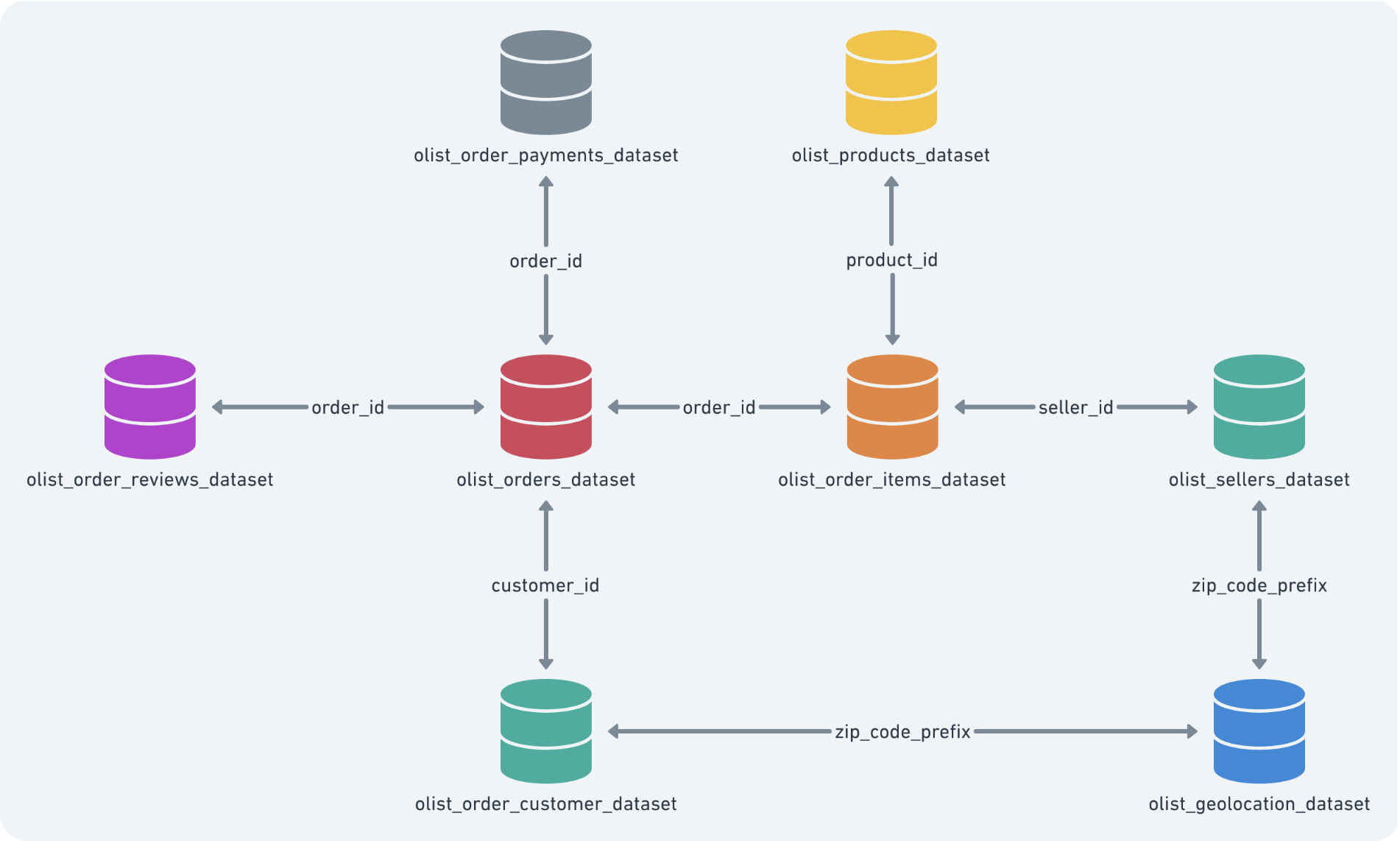
**Liste des données:**

* olist\_customers\_dataset.csv
* olist\_geolocation\_dataset.csv
* olist\_order\_items\_dataset.csv
* olist\_order\_payments\_dataset.csv
* olist\_order\_reviews\_dataset.csv
* olist\_orders\_dataset.csv
* olist\_products\_dataset.csv
* olist\_sellers\_dataset.csv
* product\_category\_name\_translation.csv

## **2. Architecture de l'infrastructure**

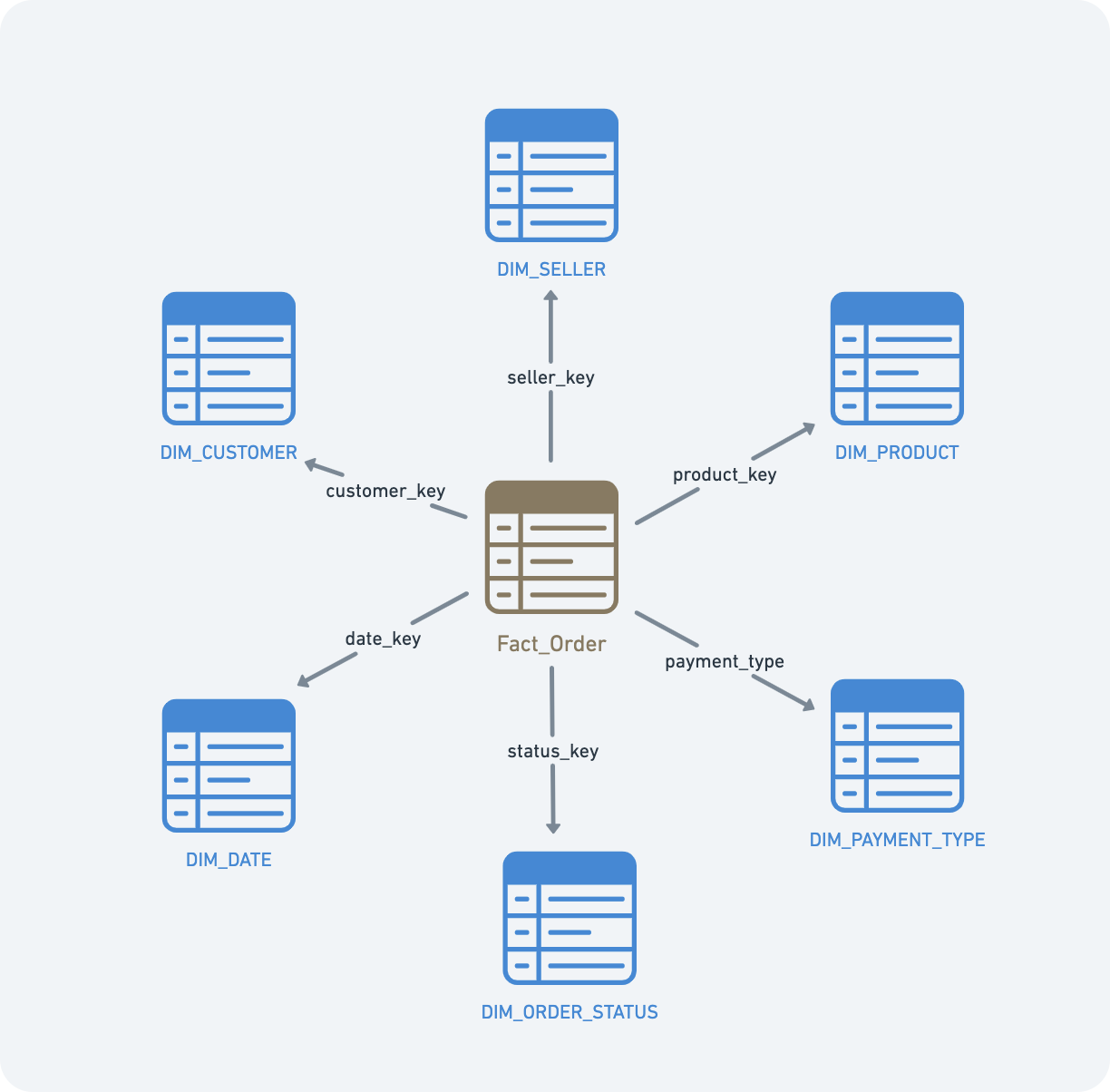
### **2.1 Base de données relationnelle (Cloud SQL MySQL)**

**Schéma de la base de données hébergée dans GCP Cloud SQL MySQL Instance pour supporter les opérations OLTP (Online Transactional Processing).**



### **2.2 Entrepôt de données (BigQuery)**

Stockage des données consolidées et transformées pour analyses avancées et reporting.



## **3. Services Cloud GCP utilisés**

* **Cloud SQL MySQL Instance** : Gestion des données transactionnelles pour le site e-commerce.
* **Dataflow** : Pipelines de traitement des données (batch & streaming) avec Apache Beam.
* **Cloud Composer** : Orchestration des workflows avec Apache Airflow.
* **BigQuery** : Entrepôt de données pour l'analyse et la business intelligence.
* **Looker Studio** : Outil de visualisation de données interactif.
* **Cloud Storage** : Stockage des fichiers CSV et des scripts de traitement.

## **4. Processus de migration vers GCP**

### **4.1 Configuration initiale**

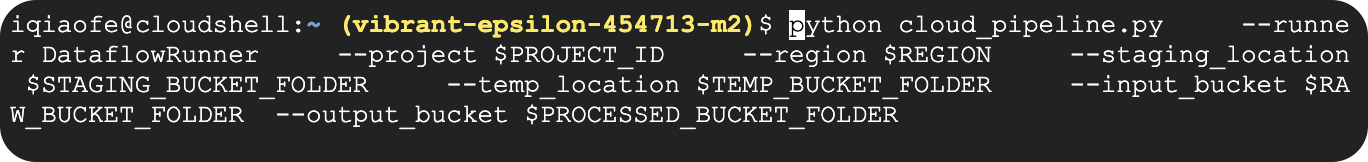
## Création d'un compte [GCP](https://cloud.google.com/gcp?utm_source=google&utm_medium=cpc&utm_campaign=emea-fr-all-fr-bkws-all-all-trial-e-gcp-1707574&utm_content=text-ad-none-any-DEV_c-CRE_529432261643-ADGP_Hybrid+%7C+BKWS+-+EXA+%7C+Txt+-+GCP+-+General+-+v2-KWID_43700060384861687-kwd-87853815-userloc_9056543&utm_term=KW_gcp-NET_g-PLAC_&&gclsrc=aw.ds&gad_source=1&gclid=EAIaIQobChMIw63Z-ou5jAMV9i0GAB3RNyxiEAAYASAAEgKknvD_BwE) et d'un projet.

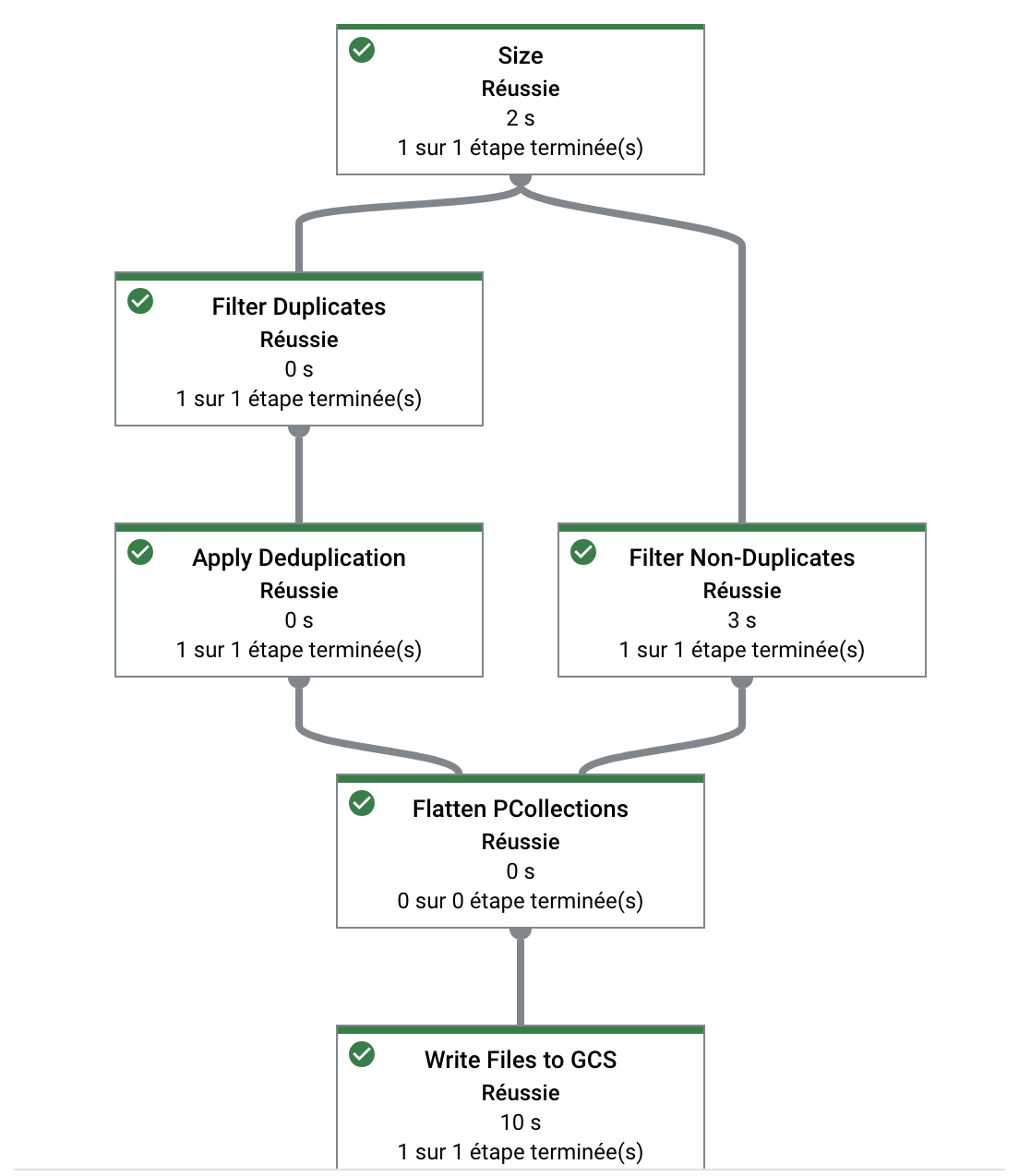
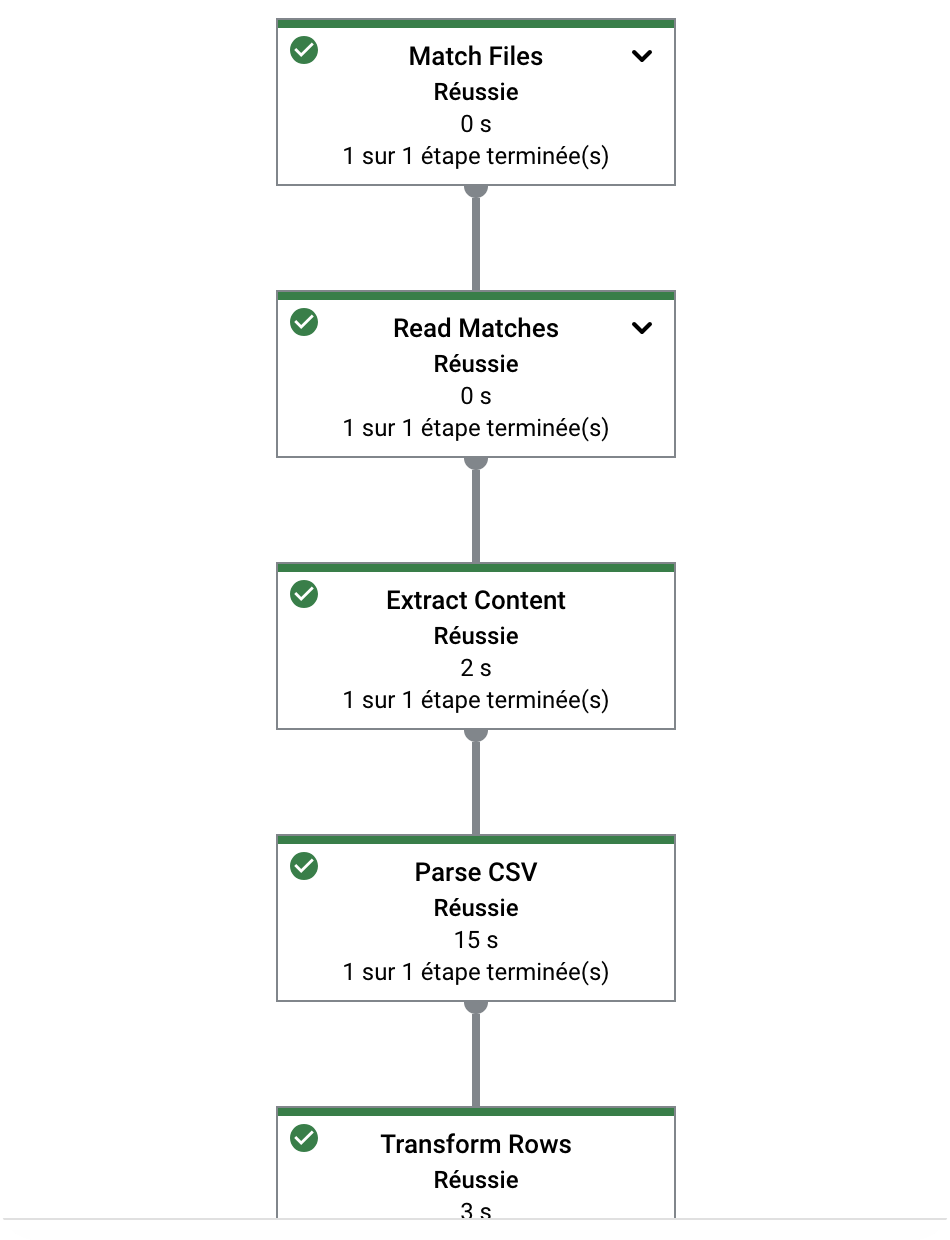
## Création d'un bucket Cloud Storage avec un répertoire raw pour les données brutes.

### **4.2 Traitement des données avec Dataflow**

1. Activation de l'API Dataflow.
2. Création d'un bucket pour les scripts Apache Beam.
3. Mise en place d'un pipeline de transformation :
   * Suppression des doublons.
   * Conversion des types de données (ex: date/heure).
4. Configuration des permissions IAM requises.
5. Exécution du pipeline Dataflow via Cloud Shell.

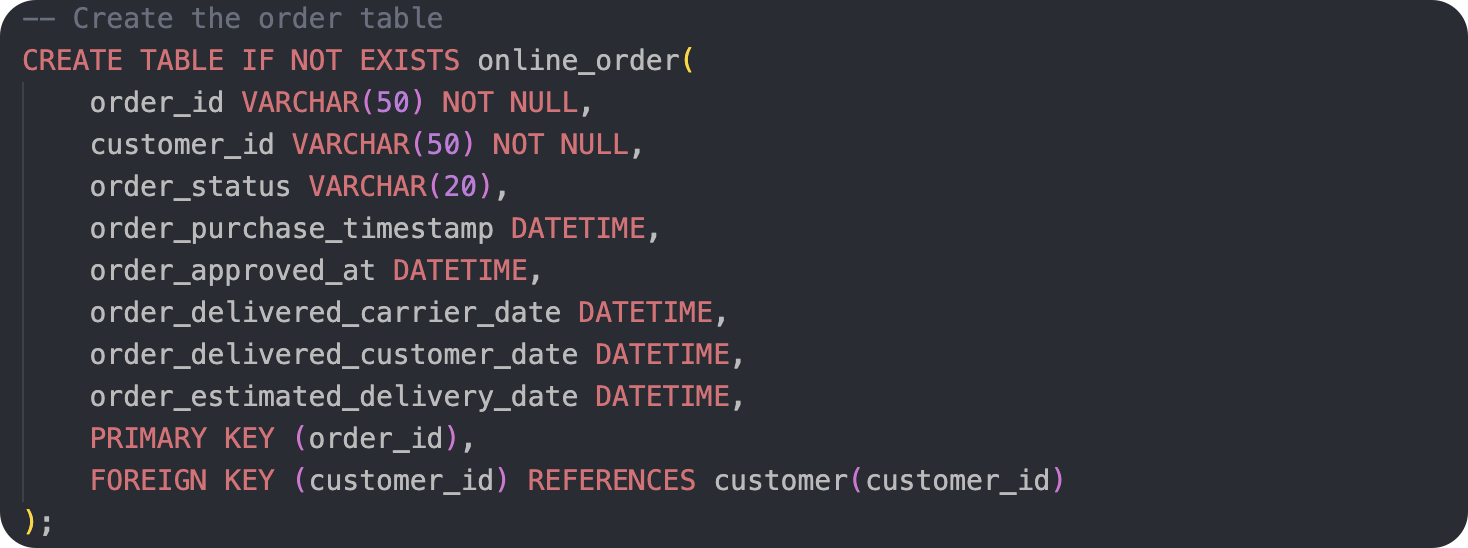
### **4.2 Traitement des données avec Dataflow**

1. Activation de l'API Dataflow.
2. Création d'un bucket pour les scripts Apache Beam.
3. Mise en place d'un pipeline de transformation :
   * Suppression des doublons.
   * Conversion des types de données (ex: date/heure).
4. Configuration des permissions IAM requises.
5. Exécution du pipeline Dataflow via Cloud Shell.



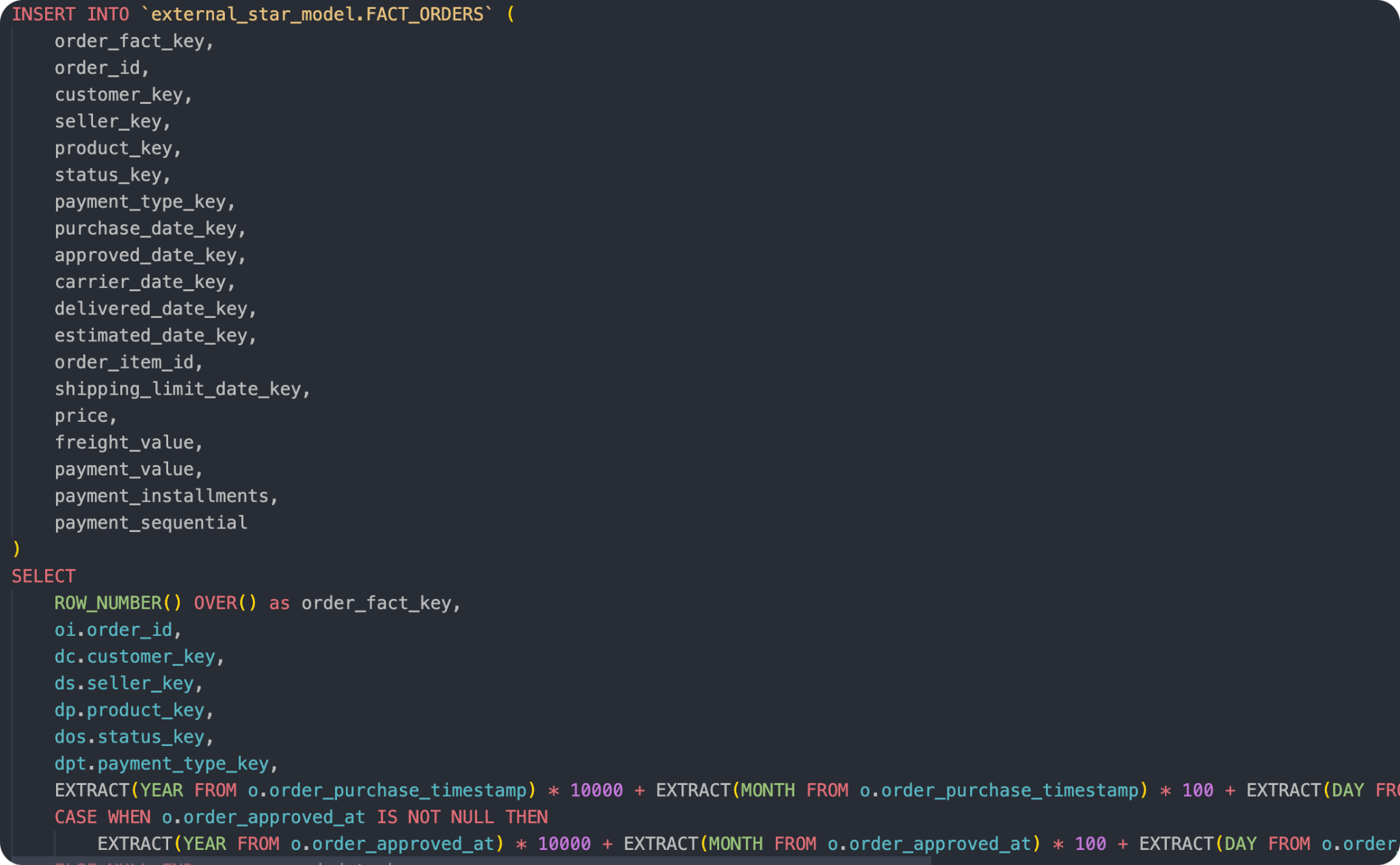
### **4.3 Importation des données dans Cloud SQL**

1. Création d'une instance MySQL Cloud SQL.
2. Configuration des paramètres (zone, type de stockage, sauvegarde, etc.).
3. Création de la base de données.
4. Connexion à la base avec gcloud sql connect $MYSQL\_INSTANCE --user=root.
5. Création d'un utilisateur avec les permissions nécessaires.
6. Définition des tables via des requêtes DDL SQL.
7. Importation des fichiers CSV via Cloud SQL Studio.



### **4.4 Migration des données vers BigQuery**

1. Création d'un dataset BigQuery.
2. Chargement des données via plusieurs méthodes :
   * bq load depuis Cloud Storage.
   * Datastream pour la réplication en temps réel.
   * Requêtes Federated Queries depuis Cloud SQL.
   * Dataflow pour ETL avancé.
3. Conception d'un modèle étoilé :
   * Tables de dimensions et table de faits.
   * Transformation des données via SQL.

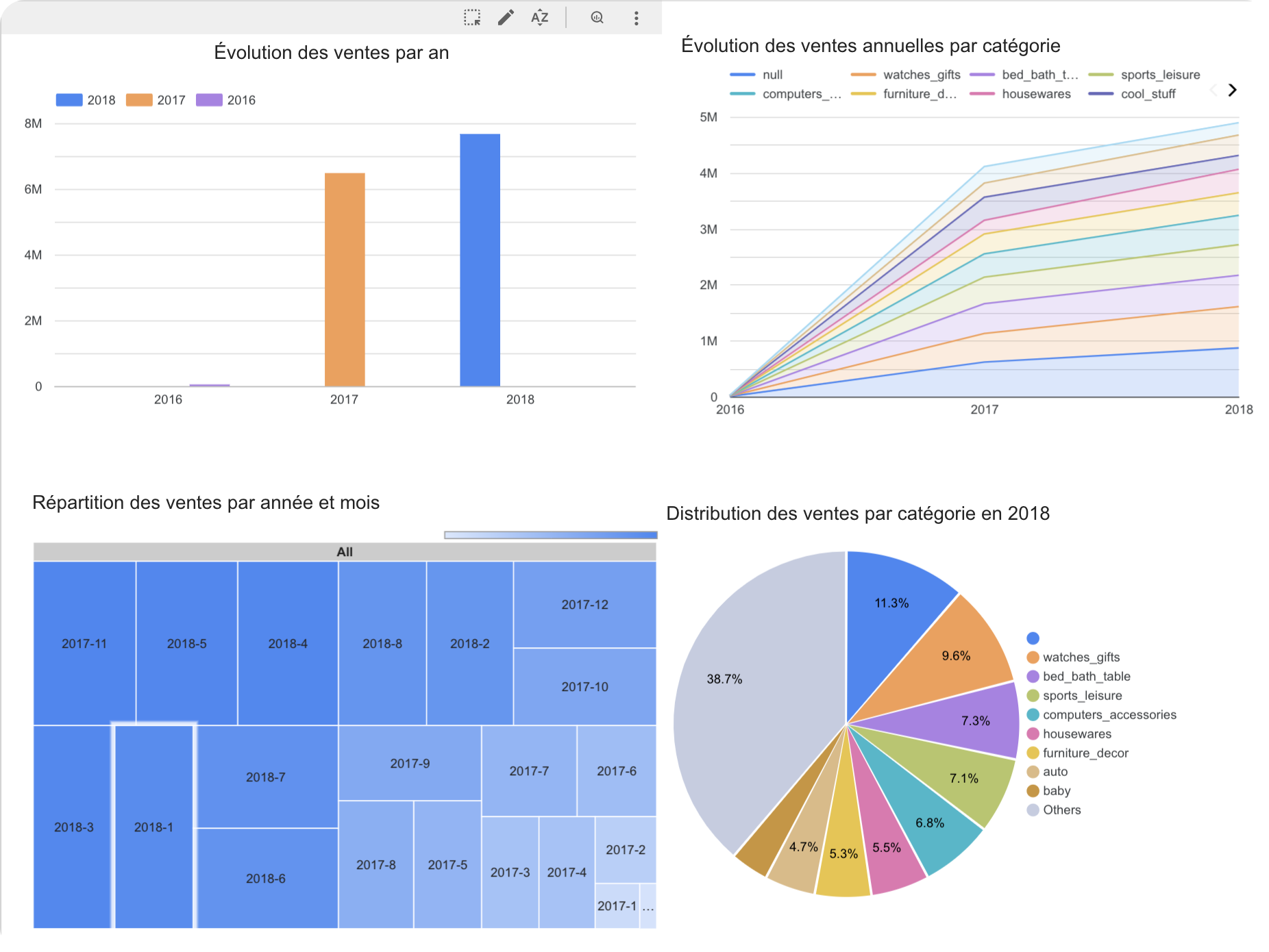


### **4.5 Orchestration avec Cloud Composer**

1. Création d'un environnement Cloud Composer.
2. Configuration des connexions et des variables d'environnement dans Airflow.
3. Écriture d'un DAG Apache Airflow avec DataflowCreatePythonJobOperator.
4. Envoi des fichiers Python dans le bucket Cloud Storage :
   * Pipeline Dataflow dans /scripts
   * DAG dans /dags
5. Exécution et monitoring du pipeline via l'interface Airflow.

### **4.6 Visualisation avec Looker Studio**

1. Connexion à Looker Studio.
2. Création d'un rapport et ajout de données.
3. Connexion à BigQuery et sélection du dataset du modèle étoilé.

Conception de dashboards interactifs.

### **4.7 Catalogage et gouvernance des données**

1. Utilisation de **Data Catalog** pour la classification et la recherche des métadonnées.
2. Application des règles de conformité RGPD.
3. Gestion des accès et de la sécurité des données.

## **5. Conclusion**

Ce projet met en place une infrastructure de données robuste et scalable sur GCP pour un site de e-commerce, en optimisant le stockage, le traitement et l'analyse des données. L'utilisation de services comme Dataflow, BigQuery et Cloud Composer permet d'automatiser et d'optimiser le pipeline de données, offrant ainsi des performances accrues et une gestion efficace des flux de données.