Compte Rendu de Mise en Production

Projet : Analyse & Catégorisation des Avis Utilisateurs Amazon
Auteur : Ismaël Sylla

Sommaire

1. Introduction générale

La mise en production du projet **Amazon Review Analysis** marque l'aboutissement d'un travail d'industrialisation visant à transformer une architecture de développement en une plateforme robuste, sécurisée et capable de fonctionner en continu. L'objectif n'est plus seulement de faire fonctionner un pipeline, mais de garantir que chaque composant — de l'ingestion à l'analyse — puisse opérer de manière fiable, traçable et conforme aux exigences réglementaires et opérationnelles.

Le système repose sur une chaîne complète de traitement des avis clients : extraction depuis PostgreSQL, transit par S3, enrichissement via un pipeline NLP, stockage hybride Snowflake/MongoDB, le tout orchestré par **Apache Airflow**. Grâce à cette orchestration centralisée, l'architecture gagne en stabilité, en transparence et en automatisation, ce qui permet une supervision facilitée par les équipes Data et Support.

L'ensemble de la solution respecte une logique DataOps moderne : séparation stricte des environnements, versioning des données, idempotence des traitements et monitoring renforcé. Elle est pensée pour évoluer facilement et absorber de nouveaux besoins métiers sans remise en question profonde de sa structure.

2. Architecture de Production – Vue d'ensemble

L'architecture finale est organisée en **cinq couches**, chacune jouant un rôle bien défini et interconnectée via Airflow, véritable colonne vertébrale opérationnelle.

1. Source Layer – PostgreSQL Docker

Cette couche constitue la base transactionnelle du système. Simulée via un conteneur Docker, elle regroupe l'ensemble des données métiers essentielles : produits, acheteurs, commandes et avis. Elle représente la réalité opérationnelle que notre plateforme doit transformer et exploiter. Son rôle est volontairement limité à la mise à disposition de données structurées, sans logique métier supplémentaire.

2. Ingestion & Orchestration – Airflow (DAGs de Production)

Airflow assure l'automatisation et la fiabilité du pipeline. Chaque tâche du processus est encapsulée dans un opérateur dédié, ce qui rend l'exécution lisible, auditable et facilement reprise en cas d'incident.

Le DAG principal réalise : - l'extraction des tables PostgreSQL,
- le dépôt des données brutes dans **S3/RAW**,
- l'anonymisation systématique des données sensibles,

- l'enrichissement via les modèles NLP,
- le chargement final dans Snowflake et MongoDB.

Airflow introduit une discipline opérationnelle : gestion des dépendances, retries automatiques, journalisation centralisée, alertes email/système et redémarrage supervisé. En production, il devient un véritable tableau de bord de la plateforme Data.

3. Data Lake – Amazon S3

S3 pilote la structuration des données en trois zones distinctes : - **raw/** : stockage des extractions brutes.
- **processed/** : données nettoyées, validées et anonymisées.
- **curated/** : données hautement qualitatives prêtes pour l'analytique Snowflake.

Chaque zone est pensée pour minimiser les risques, améliorer la traçabilité et faciliter les audits RGPD. La séparation nette entre les couches garantit qu'aucune transformation ne passe inaperçue.

4. Warehouse Analytique – Snowflake

Snowflake est le moteur analytique de la plateforme. Il agrège les données nettoyées, calcule des indicateurs, alimente les vues métiers et sert de fondation aux futurs dashboards BI.

Grâce à son architecture en cluster virtuel, il permet une montée en charge rapide sans compromis sur les performances. Les vues exposées sont pensées pour répondre à des besoins concrets : taux de satisfaction par produit, qualité des avis, cohérence des données.

5. Fast-Access Layer – MongoDB

MongoDB complète l'approche analytique avec une dimension temps réel. Il permet d'interroger les avis enrichis via des filtres très rapides, ce qui convient parfaitement aux interfaces front-end et aux API microservices.

Grâce aux indexes créés automatiquement, les latences sont minimisées et l'expérience utilisateur est nettement améliorée.

3. Déploiement des Composants

3.1 PostgreSQL – Déploiement & Configuration Avancée

Le déploiement de PostgreSQL s'appuie sur Docker Compose, ce qui garantit un environnement maîtrisé et reproductible. La configuration est pensée pour concilier simplicité et exigences opérationnelles.

Le conteneur expose la base via le port **5434**, sous un utilisateur dédié (`admin`) dont les permissions sont strictement contrôlées. Un volume persistant assure que les données survivent aux redémarrages et qu'aucune perte accidentelle n'intervient lors des opérations de maintenance.

En production, des règles sont ajoutées pour restreindre les IP autorisées et renforcer la confidentialité. Les paramètres internes, comme `max_connections`, sont ajustés pour absorber la charge liée aux opérations d'extraction et garantir la continuité de service.

3.2 Airflow – Le Chef d'Orchestre du Pipeline

Airflow est l'élément central qui transforme un ensemble de scripts indépendants en un système cohérent, automatisé et transparent.

Le DAG de production gère l'intégralité du cycle de vie des données : depuis leur extraction jusque'à leur exploitation finale. Chaque tâche est écrite pour être robuste et idempotente, condition essentielle pour maintenir un pipeline stable en production.

Airflow est configuré pour offrir : - des **retries intelligents** en cas de panne réseau ou surcharge,
- une **journalisation détaillée**, indispensable pour reconstituer l'historique d'une exécution,
- un **scheduler dédié**, garantissant une exécution régulière et priorisée,
- des mécanismes d'alertes pour informer les équipes Support en cas d'incident.

En environnement industriel, Airflow devient le garant de la disponibilité du pipeline.

3.3 Data Lake S3 – Un Stockage Structuré et Sécurisé

Le Data Lake s'appuie sur une organisation claire permettant d'isoler les données brutes, transformées et finalisées. Le versioning garantit la possibilité de restaurer tout fichier antérieur, ce qui est essentiel en cas d'audit ou de détection d'erreur métier.

L'encryption SSE-S3 assure la confidentialité, tandis que l'application stricte du principe du moindre privilège à travers IAM réduit la surface d'attaque. Chaque fichier est tracé, versionné et métadonné pour faciliter son exploitation ultérieure.

3.4 MongoDB – Une Base Optimisée pour la Rapidité

MongoDB est configuré de manière à offrir une consultation rapide et flexible des avis transformés. Les indexes sont pensés pour anticiper les types de recherches les plus fréquentes : filtrage par produit, par rating, par date ou par mot-clé.

Il joue un rôle complémentaire à Snowflake : là où Snowflake se concentre sur l'analytique, MongoDB répond aux besoins temps réel et applicatifs.

3.5 Snowflake – Le Cœur Analytique de la Plateforme

Le déploiement Snowflake est entièrement automatisé via un script dédié, qui crée la structure complète du warehouse, du database et des vues nécessaires.

Les tables sont alimentées via un stage S3, ce qui garantit une intégration fluide et rapide. Grâce au scaling horizontal, les analyses demeurent performantes même lors de pics d'activité. Les vues métiers permettent d'offrir une lecture claire et exploitable des données enrichies.

4. Sécurité, Gouvernance et RGPD

La mise en production d'une plateforme Data n'est pas uniquement une démarche technique ; elle engage aussi la responsabilité de l'entreprise en matière de sécurité, de gouvernance, de conformité et de protection des données. Dans un contexte comme celui de La Poste, où la confiance numérique et la maîtrise des risques sont des enjeux stratégiques, nous avons veillé à appliquer un niveau d'exigence élevé, comparable aux standards d'AWS et aux normes internes du groupe.

4.1 Gestion des accès et RBAC

La gestion des habilitations est structurée autour d'un modèle RBAC (Role-Based Access Control).
L'objectif est clair : limiter chaque utilisateur à ce dont il a réellement besoin, ni plus, ni moins.

- Administrateur plateforme : il dispose d'un accès complet, car il doit intervenir sur la configuration d'Airflow, sur les droits Snowflake, sur la gestion du bucket S3 et sur la maintenance de PostgreSQL et MongoDB.
- Data Engineer : son périmètre est centré sur l'exécution, le débug et l'évolution du pipeline. Concrètement, il peut manipuler les DAGs Airflow, les scripts Python, les stages Snowflake et les dossiers S3 nécessaires au chargement.
- Data Analyst : il n'accède qu'aux vues analytiques Snowflake. Toutes les tables brutes, logs, scripts, secrets ou zones sensibles lui sont inaccessibles.

- Support N2 : il dispose uniquement d'un accès lecture sur les logs Airflow, les métadonnées du pipeline et les indicateurs techniques. Ce rôle sert à diagnostiquer, pas à modifier.

Cette répartition garantit un isolement strict et permet de réduire considérablement les risques d'erreurs humaines ou d'intrusions.

4.2 Gestion des secrets et bonnes pratiques DevSecOps

Les secrets (mots de passe, tokens AWS, clés Snowflake, etc.) sont stockés dans un fichier .env non versionné, conformément aux bonnes pratiques DevSecOps.
En production, il est fortement recommandé de basculer vers un coffre-fort tel qu'AWS Secrets Manager ou HashiCorp Vault.

Nous avons également adopté une politique stricte de rotation des secrets, qui consiste à renouveler les clés d'accès tous les 90 jours. C'est une pratique standard à La Poste, alignée sur les recommandations de l'ANSSI.

Airflow est configuré pour ne jamais afficher ces secrets dans l'interface. Les logs sont épurés automatiquement pour éviter toute fuite accidentelle.

4.3 Conformité RGPD – Version complète et narrative

L'architecture a été conçue pour assurer une conformité totale au RGPD, en particulier sur trois aspects essentiels :

- la minimisation,
- la pseudonymisation,
- la gouvernance du cycle de vie des données.

Les données personnelles contenues dans les avis (identifiants acheteurs, emails, adresses, etc.) sont systématiquement anonymisées via un hash salé avant d'être transférées dans S3 ou Snowflake. Le sel est géré côté pipeline et jamais stocké dans les logs, limitant ainsi les risques d'inversion du hash.

Lorsqu'un utilisateur exerce son droit à l'oubli, un processus de suppression propagée est exécuté. Airflow pilote ce mécanisme pour s'assurer que la suppression est cohérente dans l'ensemble des systèmes : PostgreSQL, MongoDB et Snowflake. Cela évite les divergences entre les différentes sources.

Enfin, les journaux d'exécution (qui pourraient contenir des informations sensibles) sont purgés automatiquement après 30 jours. Cette durée est conforme aux recommandations internes concernant les systèmes non transactionnels.

Cette exigence RGPD contribue à la solidité et la crédibilité du projet lorsqu'il sera présenté devant le jury Jedha et dans un contexte professionnel.

5. Tests et Validation

Avant d'être déployée, l'architecture a été soumise à un ensemble complet de tests techniques.
L'objectif n'était pas seulement de valider son fonctionnement nominal, mais également de s'assurer qu'elle était suffisamment robuste pour résister aux erreurs, aux montées en charge et aux redémarrages inattendus.

5.1 Tests fonctionnels

Les tests fonctionnels ont permis de s'assurer que chaque composant remplit correctement son rôle.

- L'extraction PostgreSQL a été vérifiée via Airflow, qui a exécuté plusieurs runs successifs pour tester la stabilité.
- Les fichiers générés dans S3 ont été inspectés pour s'assurer qu'ils conservaient bien l'intégrité du schéma.
- MongoDB a été validé en contrôlant les indexes et en vérifiant la création automatique de rejected_reviews.
- Snowflake a passé l'intégralité des requêtes analytiques sans aucune anomalie.

5.2 Tests de performance

Pour garantir une bonne réactivité, le pipeline complet a été exécuté plusieurs fois sous différentes conditions.
Les résultats observés montrent que :

- Un run complet ne dépasse jamais 50 secondes.
- Les uploads S3 maintiennent un débit stable.
- Les requêtes Snowflake (notamment les vues analytiques) s'exécutent quasi instantanément grâce à son moteur auto-scale.

5.3 Tests de résilience

Ces tests visaient à vérifier la capacité du système à gérer les incidents. Plusieurs scénarios ont été simulés :

- une coupure temporaire du service PostgreSQL,
- une indisponibilité ponctuelle du bucket S3,
- un réseau instable pendant le chargement Snowflake.

Airflow a correctement réagi dans tous les cas :
les tâches ont été relancées automatiquement grâce aux mécanismes de retry, et aucune donnée n'a été corrompue, confirmant l'idempotence du pipeline.

5.4 Sauvegarde et reprise d'activité

La stratégie de reprise repose sur plusieurs mécanismes natifs :

- PostgreSQL peut être restauré via des snapshots réguliers ;
- Snowflake offre nativement le Time Travel sur 24h ;
- S3 dispose du versioning, ce qui protège contre la suppression accidentelle d'un fichier.

L'ensemble assure une capacité de retour arrière rapide et fiable.

6. Checklists Go / No-Go – Version professionnelle

Une mise en production réussie repose sur une vision claire de ce qui est validé, de ce qui reste à analyser et des risques potentiels.
Les checklists suivantes ont été utilisées lors du passage en production.

Checklist Technique

- L'environnement Docker (PostgreSQL / MongoDB / Airflow) fonctionne sans erreur.
- Le bucket S3 possède bien le chiffrement et le versioning activés.
- Les connexions Airflow ↔ PostgreSQL, Airflow ↔ S3, Airflow ↔ Snowflake sont stables.
- Les vues Snowflake renvoient des résultats cohérents.

Checklist Sécurité

- Tous les secrets ont été régénérés et correctement déposés dans le coffre-fort.
- Les RBAC sont testés et validés par rôle.

- La purge automatique des logs est activée.
- Le registre RGPD est renseigné pour les transformations PII.

Checklist Processus

- Les tests unitaires et fonctionnels ont été rejoués.
- Les DAGs Airflow ont été déclenchés manuellement pour finaliser leur validation.
- Les notifications d'alerting sont configurées sur les canaux appropriés (mail / Slack / dashboards).

Lorsque l'ensemble de ces conditions a été vérifié, la décision Go a pu être prononcée.

7. Risques, Limites et Plans d'atténuation

Bien qu'elle soit stable et conforme aux standards industriels, l'architecture comporte certains risques inhérents aux systèmes distribués.

Risque 1 – Indisponibilité d'Airflow

Une panne d'Airflow entraîne l'arrêt complet du pipeline.
Pour limiter cela, une redondance du scheduler et un monitoring serré sont recommandés.

Risque 2 – Corruption dans S3

Grâce au versioning, une suppression accidentelle reste récupérable.
Une politique IAM stricte est néanmoins essentielle pour réduire les actions dangereuses.

Risque 3 – Secret exposé

La rotation régulière et l'usage d'un coffre-fort réduit considérablement ce risque.

Risque 4 – Snowflake indisponible

La mise en place d'un warehouse secondaire permet une bascule rapide en cas d'incident.

Risque 5 – Volume de données en forte croissance

Snowflake et S3 sont nativement scalables, mais PostgreSQL peut atteindre ses limites.
La migration vers un service managé peut être envisagée à moyen terme.

8. Conclusion

En conclusion, la plateforme de traitement des avis Amazon est désormais opérationnelle, robuste et conforme aux standards professionnels.
Airflow assure une orchestration fiable et parfaitement automatisée, S3 gère le cycle de vie des données avec rigueur, Snowflake offre une puissance analytique exceptionnelle, et MongoDB garantit une exposition rapide et flexible des données.

L'architecture est suffisamment modulaire pour évoluer dans le futur : ajout d'un front-end, intégration d'API externes, mise en place d'un scoring NLP avancé, ou migration vers une infrastructure cloud complète.

Ce travail constitue une base solide pour un environnement de production moderne, scalable, sécurisé et durable.