DOCUMENTATION D'ARCHITECTURE

Projet Amazon Reviews :

Analyse et classification des avis Clients

BONNAT Jonathan

Data Engineer - B2B-LP-DESFL

💬 INTRODUCTION GENERALE

CONTEXTE DU PROJET

Ce document s'inscrit dans la continuité du prototype développé précédemment, qui a démontré la faisabilité d'un modèle de classification automatique des avis produits à l'aide du NLP (zero-shot learning).

Cette nouvelle étape vise à transformer ce prototype local en une architecture complète et industrialisable, intégrée dans un environnement data conforme aux standards techniques et réglementaires d'Amazon.

OBJECTIF DU PROJET

L'objectif du projet est de concevoir une architecture de données robuste et automatisée permettant :

- d'extraire et de consolider des données issues du système transactionnel;
- de les transformer (nettoyage, enrichissement, classification);
- et de les charger dans des environnements de stockage adaptés à l'analyse (Data Lake) et à l'historisation.

L'ensemble du dispositif doit garantir la qualité, la traçabilité et la conformité du traitement vis-à-vis du RGPD, de la CNIL et du AI Act.

PÉRIMÈTRE FONCTIONNEL

Les données entrantes fournies et la solution développée ne portent que sur les avis produits (texte, note, images, métadonnées client et produit) en mode batch, sans traitement en temps réel.

ENJEUX DU PROJET

Les priorités choisies pour ce projet sont :

- la robustesse technique et scalabilité du pipeline;
- la fiabilité des données : contrôle des rejets et traçabilité complète;
- la conformité réglementaire : respect du RGPD et transparence des modèles IA;
- et l'interopérabilité avec les environnements cloud et analytiques d'Amazon.

🗂️ SOURCES ET DONNÉES DE RÉFÉRENCE

ORIGINE ET STRUCTURE

Le projet s'appuie sur une base de données relationnelle PostgreSQL issue de l'écosystème e-commerce d'Amazon. Cette dernière a été reconstituée à partir de fichiers bruts, puis normalisée en 3NF, garantissant la cohérence et la non-redondance des informations. Elle regroupe l'ensemble des entités nécessaires à la gestion des transactions, des produits, des clients et de leurs retours sous forme d'avis.

La volumétrie globale s'élève à plus d'1,4 million d'enregistrements répartis sur 27 tables. Les traitements réalisés dans le cadre de ce projet se concentrent sur les avis produits et leurs métadonnées, soit les tables les plus pertinentes pour notre cas d'usage "Analyse de la pertinence et de la thématique des avis clients".

TABLES PRINCIPALES UTILISÉES DANS NOTRE PIPELINE ETL

TABLE	DESCRIPTION	VOLUME	RÔLE
REVIEW	Contient le texte, la note et les métadonnées des avis clients	111 322	Source principale pour le traitement NLP et le scoring de pertinence
PRODUCT_REVIEWS	Table de liaison entre "REVIEW" et "PRODUCT".	111 322	Jointure qui permet d'associer chaque avis à un produit.
PRODUCT	Contient le nom, la description, le prix et la catégorie d'un produit	42 858	Permet d'enrichir l'analyse avec le contexte du produit
REVIEW_IMAGES	Contient la liste des images en pièces jointes des avis	119 382	Permet de pondérer le score de pertinence des avis (présence d'image dans l'avis)
ORDERS	Contient les dates et l'identifiant acheteur des commandes	222 649	Permet de pondérer le score de pertinence des avis (achat vérifié)
SUBSCRIPTION	Contient les informations d'abonnement Amazon Prime	50 091	Permet de pondérer le score de pertinence des avis (abonné Amazon Prime)

TYPOLOGIE DES DONNÉES MANIPULÉES

Le pipeline traite à la fois :

- Des données structurées : identifiants, dates, notes, prix, indicateurs binaires (abonnement, image jointe);
- Des données semi-structurées : catégories et descriptions produits;
- Des données non structurées : texte et titre des avis et images associées.

Cette diversité nécessite une architecture hybride, capable de gérer simultanément le stockage relationnel (PostgreSQL) et le stockage non structuré (NoSQL, Data Lake, S3 ou équivalent).

RELATIONS ET INTÉGRITÉ RÉFÉRENTIELLE

Les principales relations logiques entre les tables sont :

- BUYER (1:N) REVIEW : un acheteur peut laisser plusieurs avis;
- REVIEW (N:N) PRODUCT VIA PRODUCT_REVIEWS : un produit peut recevoir plusieurs avis;
- REVIEW (1:N) REVIEW_IMAGES : un avis peut comporter zéro ou plusieurs images;
- BUYER (1:N) ORDERS : permet de vérifier l'existence d'un achat ;
- BUYER (1:1) SUBSCRIPTION : indique l'état d'abonnement du client.

Ces relations assurent la traçabilité complète du parcours client → produit → avis, essentielle pour la classification et l'évaluation de la pertinence.

DONNÉES PERSONNELLES ET ANONYMISATION

Certaines tables contiennent des données sensibles (identifiants client, adresse mail, nom, etc). Conformément au RGPD, toutes les données à caractère personnel sont pseudonymisées avant intégration dans le pipeline :

- remplacement des identifiants réels par des hash anonymes;
- suppression des champs non essentiels (noms, adresses mails, téléphones, etc);

Seules les données nécessaires à l'analyse (texte des avis, note, produit associé, statut

abonné) sont conservées pour le traitement.

SYNTHÈSE DES DONNÉES EXPLOITÉES

Le projet exploite un sous-ensemble représentatif de la base Amazon, couvrant environ 700 000 lignes de données issues des six tables principales citées au préalable.

Ce périmètre garantit un volume suffisant pour tester la robustesse du pipeline ETL et valider les performances de la classification NLP dans un contexte semi-industriel.

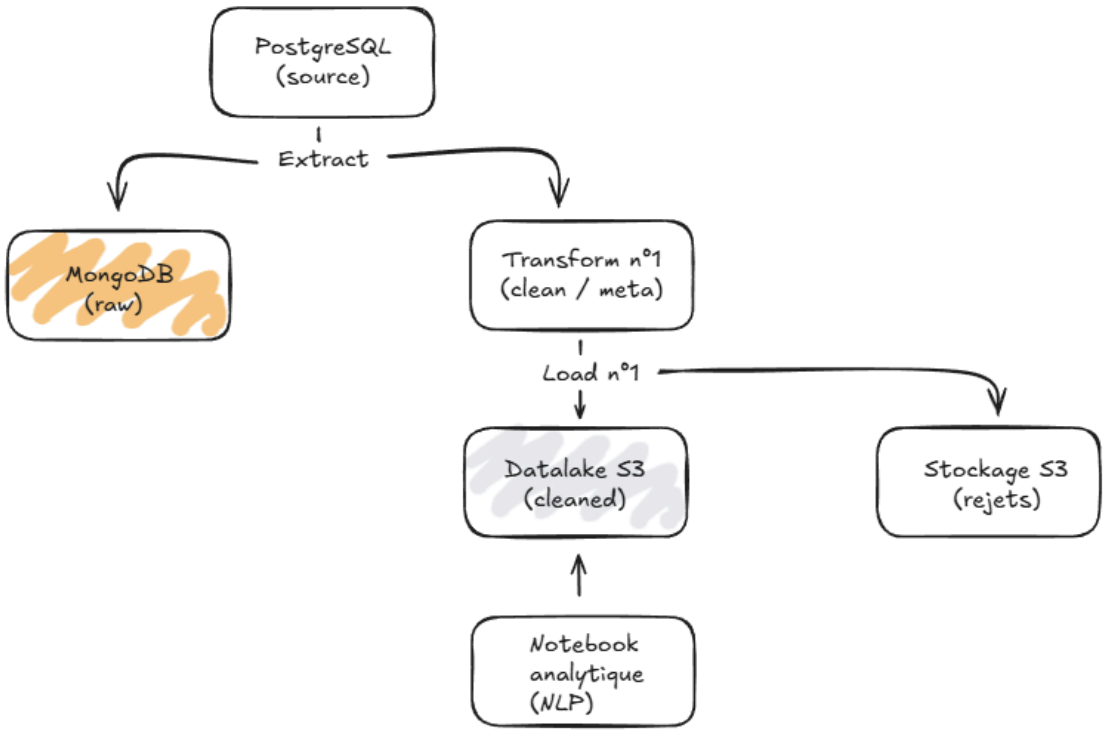
🏗️ ARCHITECTURE GLOBALE ET SCHÉMA DU SYSTÈME

L'architecture du système repose sur une chaîne de traitement complète, structurée autour des quatre grandes couches d'un environnement data classique :

- Source de données (Transactional Layer) → Base PostgreSQL normalisée issue du système e-commerce d'Amazon;
- Pipeline ETL (Processing Layer) → Processus d'extraction, transformation et chargement développé en Python;
- Stockage des données extraites sur MongoDB pour historisation et rejeu si besoin (sert de source de vérité à un instant T)
- Stockage et consolidation (Storage Layer) → Architecture hybride combinant entrepôt relationnel et fichiers bruts (parquet et / ou csv) sur Amazon S3;
- Exploitation analytique (Analytics Layer) → Accès aux données transformées pour analyse NLP, visualisation et reporting.

Cette organisation modulaire permet de séparer les responsabilités, de faciliter la maintenance, et de garantir la scalabilité de l'ensemble.

SCHÉMA LOGIQUE DE L'ARCHITECTURE



DESCRIPTION DES COUCHES TECHNIQUES

COUCHE	TECHNOLOGIES	RÔLES ET FONCTIONNALITÉS
Source	fichiers bruts mis sur PostgreSQL	Données transactionnelles brutes issues de la plateforme de e-commerce.
Pipeline ETL	Python, Pandas, SQLAlchemy, Transformers	Extraction, nettoyage, enrichissement et classification NLP des avis.
Stockage	PostgreSQL (Datalake), fichiers csv ou parquet sur S3 + MongoDB	Conservation et historisation des données traitées : données structurées, documents enrichis et archives.
Analyse	Jupyter Notebook, Streamlit	Visualisation et analyse des résultats (pertinence, catégories, volumétrie).

FLUX DE TRAITEMENT

1. Extraction : connexion à la base PostgreSQL via SQLAlchemy pour extraire les tables principales (review, product, buyer, etc).
2. Transformation : nettoyage des données, traitement de texte, constitution des rejets.
3. Chargement : export des résultats au format Parquet / csv et insertion dans une table analytique PostgreSQL (REVIEW_ANALYSIS). Les rejets sont consignés dans un fichier CSV (rejected_reviews.csv) et/ou sauvegardés sur S3
4. Exploration et visualisation via notebooks ou autres outils de visualisation (dont classification zero-shot et calcul du score de pertinence).

PRINCIPES D'ARCHITECTURE

- Séparation des environnements : distinction claire entre les zones sources, transformation, stockage et analyse.
- Évolutivité : possibilité d'intégrer ultérieurement d'autres informations (données d'images, vidéos, d'autres modèles NLP, avis d'autres produits, etc).
- Sécurité et conformité : chiffrement des flux, pseudonymisation des identifiants, application des règles RGPD et AI Act.
- Traçabilité : journalisation complète des étapes ETL, génération de données d'audit

⚙️ PROCÉDURES D'EXTRACTION, TRANSFORMATION ET CHARGEMENT (ETL)

Le pipeline ETL vise à automatiser le traitement complet des avis produits, de leur extraction depuis la base transactionnelle PostgreSQL jusqu'à leur chargement dans des environnements analytiques et de stockage.

Ce processus permet d'obtenir une donnée propre, enrichie et exploitable pour les analyses de pertinence et de satisfaction client.

ÉTAPES DU PIPELINE

ETAPE	OBJECTIF	PROCÉDURE	RÉSULTAT
Extraction	Collecter les données sources	Connexion à PostgreSQL Sélection des tables + jointures Export en Dataframe	Données brutes consolidées
Historisation	Stocker les données brutes	Connexion à Mongo + Création d'un document par table sélectionnée + Ecriture	Données brutes sauvegardées
Transformation	Nettoyer et enrichir	Nettoyage (doublons, valeurs nulles), ajout de métadonnées, détection des rejets	Données nettoyées et données rejetées
Chargement	Stocker les données traitées	Écriture des résultats dans S3 en parquet. Rejets consignés en csv sur S3 également	Données prêtes pour analyse

CONTRÔLES ET TRAÇABILITÉ

Chaque exécution du pipeline génère :

- un rapport de logs (date, volume traité, taux de rejet, durée d'exécution);
- un fichier de rejets listant les enregistrements non conformes (format CSV);
- et un rapport synthétique d'exécution à archiver (dans MongoDB ou dans S3).

Ces éléments assurent la traçabilité du traitement et la conformité aux exigences de qualité et d'auditabilité.

SYNTHÈSE

Le pipeline est exécuté en mode batch et produit trois flux :

- des données enrichies prêtes à l'analyse (Parquet);
- des données rejetées (CSV ou table de rejets);
- et des fichiers d'audit + logs d'exécution

Ce dispositif assure une préparation fiable et contrôlée des données, garantissant la qualité des indicateurs de pertinence avant leur exploitation analytique.

🧹 GESTION DES REJETS ET CONTRÔLE QUALITÉ

La gestion des rejets vise à garantir la qualité, la cohérence et la traçabilité des données traitées par le pipeline ETL. Elle repose sur l'identification, la consignation et l'archivage des enregistrements non conformes, afin de préserver l'intégrité des données chargées dans les environnements analytiques.

TYPES DE REJETS

Les rejets correspondent aux enregistrements qui ne respectent pas les règles de validation ou présentent des incohérences critiques. Ils sont classés selon trois grandes catégories :

- Technique : Problème de structure, d'encodage ou de conversion (type de données, structure des tables modifiées)
- Sémantique : Incohérence logique ou information manquante (avis sans description, sans identifiant produit ou sans acheteur)
- Modèle NLP (futur) : A terme, on pourrait imaginer le rejet des lignes avec classification impossible ou score de confiance < 0.25

Chaque rejet est tracé avec la cause identifiée et l'horodatage

MÉCANISMES DE DÉTECTIONS ET CONSIGNATION

Le pipeline inclut un module de validation automatique exécuté à la fin de la phase de transformation. Il applique les contrôles suivants :

- Vérification de la complétude des champs obligatoires (ID, texte, note);
- Validation des types de données et des formats (dates, nombres);
- Cohérence entre les identifiants (buyer_id, product_id).

Les éléments rejetés sont exclus de la phase de chargement, et sont ajoutés à un fichier

ou une table de rejets, comprenant les éléments suivants :

- l'identifiant de l'enregistrement;
- le motif du rejet;
- la date et l'heure;
- et le statut d'exécution du pipeline.

Ces données seront archivées pour des besoins de conservation et pour des évolutions futures (analyse de rejet, retraitement, etc).

SUIVI ET INDICATEURS QUALITÉ

À chaque exécution, le système calcule des indicateurs de qualité, qui sont intégrés au rapport d'exécution :

INDICATEUR	DESCRIPTION	OBJECTIF
Taux de rejets global	(Nombre de rejets / Nombre total d'enregistrements) x 100	< 2%
Taux de complétude	(Nombre de lignes sans valeurs manquantes / Nombre total d'enregistrements) x 100	> 98%
Taux de cohérence	(Nombre d'erreurs techniques + sémantiques / Nombre total d'enregistrements) x 100	> 99%

SYNTHÈSE

La gestion des rejets garantit la fiabilité et la transparence du processus ETL. Les enregistrements invalides sont détectés, alertés et archivés, assurant ainsi :

- la qualité des données produites;
- la traçabilité des erreurs;
- et la conformité aux exigences de gouvernance et d'audit.

💾 STOCKAGE ET ORGANISATION DES DONNÉES

Cette section décrit les choix de stockage et d'organisation des données issus du pipeline ETL. L'objectif est de garantir la pérennité, la performance et la traçabilité des données tout au long du cycle de vie : de leur production (avis client sur la plateforme de e-commerce) jusqu'à leur exploitation analytique.

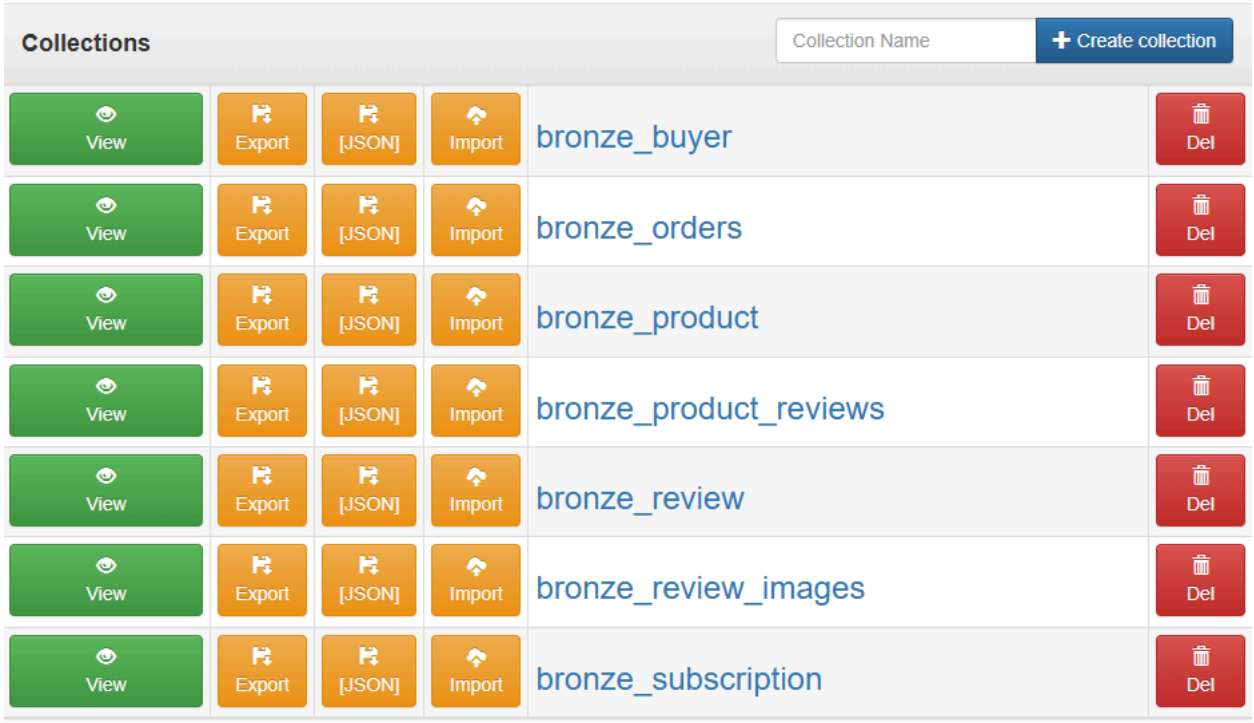
STRATEGIE GENERALE DE STOCKAGE

Le projet est constitué d'une architecture hybride et progressive, reposant sur trois environnements complémentaires :

ENVIRONNEMENT	RÔLE PRINCIPAL	TYPE DE DONNÉES	TECHNOLOGIE
Base NoSQL	Archiver et historiser les données brutes	Documents Json	MongoDB
Data Lake	Stocker les données enrichies et les rejets	Parquet, csv	Bucket S3

BASE NOSQL

Elle constitue la source de vérité de notre traitement. Cette base de données orientée documents nous permet de stocker les informations des différentes tables que nous avons extrait au début de l'ETL



Chaque élément contient les différentes lignes des tables, avec un champ supplémentaire “run_id” qui nous permet d’historiser plusieurs traitements et de rejouer un traitement en particulier si besoin.

DATA LAKE

Il est utilisé comme espace de stockage des données enrichies et des journaux d'exécution. Il peut être simulé en environnement local, mais doit respecter la logique d'un stockage objet pour faciliter l'installation sur Amazon S3. L'organisation des répertoires est la suivante :

📁 **/datalake/**

├── 📂 **raw/**

│ └── Données brutes (CSV / JSON, fichiers sources non transformés pour la création de la base de données dans notre projet)

├── 📂 **cleaned/**

│ └── Données enrichies (format Parquet, prêtes à l'analyse ou au reporting)

├── 📂 **rejected/**

│ └── Données rejetées (erreurs de format, de validation, ou d'ingestion)

EVOLUTIONS POTENTIELLES

À plus long terme, le dispositif de stockage pourra être étendu pour accueillir un DWH type Snowflake, qui nous permettra d'établir un cadre plus axé SQL aux données. Pour le moment, au vu de la complexité, de la volumétrie et du temps, nous n'en avons pas besoin.

🔒 SÉCURITÉ, CONFORMITÉ ET GOUVERNANCE

Cette partie vise à garantir que les opérations d'extraction, de traitement et de stockage respectent les exigences réglementaires (RGPD, AI Act, CNIL) et les standards internes de qualité et de sécurité des systèmes d'information.

SÉCURITÉ DES ENVIRONNEMENTS

Les environnements de développement et de stockage appliquent des mesures de protection cohérentes avec les bonnes pratiques en architecture data, pour garantir la confidentialité, la disponibilité et l'intégrité des données tout au long du cycle ETL :

DOMAINE	MESURES MISES EN PLACE
Accès et authentification	Connexions sécurisées via identifiants individuels et variables d'environnement. Aucun mot de passe stocké en clair dans le code.
Périmètre d'accès	Principe du moindre privilège : accès en lecture seule à la base transactionnelle, droits d'écriture limités au Data Lake.
Chiffrement	Données sensibles (identifiants, logs) chiffrées via clés locales (AES / SHA256) ou fonctions PostgreSQL intégrées.
Sauvegarde et archivage	Sauvegarde automatique des fichiers de logs et exports (en local, mais prévision d'implémenter dans MongoDB)
Intégrité des données	Contrôles de cohérence et validation des clés primaires / étrangères à chaque exécution du pipeline.

SÉCURITÉ DES ENVIRONNEMENTS

Le traitement des données est conçu dans le respect des cadres légaux européens et internationaux. L'ensemble des traitements reste non nominatif et respecte le principe de proportionnalité : les données sont exploitées uniquement à des fins d'analyse interne et d'amélioration du service.

CADRE	ACTIONS AU SEIN DU PROJET
RGPD	Pseudonymisation des identifiants utilisateurs (buyer_id), conservation limitée des données, journalisation des traitements.
CNIL	Respect du principe de minimisation et d'exactitude des données : seules les informations nécessaires à l'analyse sont extraites.
AI Act	Suivi des performances et limites du modèle NLP (score de confiance, version du modèle, date d'exécution) afin d'assurer la transparence et la traçabilité des algorithmes utilisés.
RSE	Favorise une utilisation responsable des données et des modèles d'IA dans le respect de l'éthique numérique.

GOUVERNANCE DES DONNÉES

Une gouvernance simple mais claire est mise en place afin d'assurer la qualité et la maîtrise du cycle de vie des données. Elle assure la fiabilité opérationnelle du dispositif et son alignement avec les standards de gestion de la donnée en entreprise.

CATÉGORIE	DESCRIPTION
Propriété des données	Les données appartiennent à Amazon ; le traitement est effectué à des fins analytiques internes.
Rôles et responsabilités	Data Engineer → implémentation et maintenance du pipeline Data Analyst → Exploitation des données enrichies et validation Data Architect → Supervision des environnements, sécurité et conformité
Cycle de vie	Données brutes extraites → données enrichies (analysables) → archivage / suppression après usage.
Documentation et auditabilité	Conservation des logs, rapports d'exécution et fichiers de rejets pour audit interne.
Qualité de données	Indicateurs de suivi : taux de rejet, complétude, cohérence et performance NLP dans l'analyse.

SYNTHÈSE

La sécurité, la conformité et la gouvernance constituent les piliers du projet

Elles permettent de protéger les informations clients et produits, d'assurer la traçabilité complète du pipeline, mais aussi de garantir la conformité réglementaire face aux obligations européennes sur la donnée et l'IA.

L'ensemble du système est conçu selon une approche responsable, documentée et durable. Il est conforme aux bonnes pratiques d'architecture que l'on pourrait attendre d'un leader mondial comme Amazon.

🏁 CONCLUSION ET PERSPECTIVES

Ce document nous a permis de concevoir et formaliser l'architecture complète du projet Amazon Reviews (partie ETL), en assurant la cohérence entre les dimensions techniques, fonctionnelles et réglementaires.

À l'issue de cette étape, le pipeline de traitement est entièrement opérationnel : il extrait, nettoie et enrichit les avis avec différentes métadonnées, puis charge les résultats dans des environnements de stockage adaptés. Les procédures de contrôle qualité et de gestion des rejets garantissent la fiabilité des données produites, tandis que la structure de stockage (documents sur MongoDB + fichiers Parquet sur S3) assure la traçabilité et la performance analytique.

Pour la partie organisationnelle, l'ensemble du projet s'appuie sur des principes solides : séparation des environnements, sécurité, conformité RGPD, AI Act, et gouvernance claire des rôles. Ces éléments posent les bases d'une industrialisation progressive du pipeline et d'une future montée en charge vers des volumes massifs.

Les prochaines étapes du projet consisteront à :

- Analyser les données nettoyées avec un modèle de zero-shot et analyser les résultats

- renforcer la supervision et la planification du pipeline via un orchestrateur;
- et évaluer les performances globales du système à travers des tests d'acceptation formalisés.

Enfin, cette phase ETL démontre la capacité du projet à transformer la donnée brute en information exploitable, fiable et conforme. C'est une étape essentielle pour la mise en production d'un système analytique robuste au service de la valorisation des avis clients d'Amazon.