



Amazon – Analyse des avis produits

Projet Data Engineering & Analyse avancée



Ismaël SYLLA

Soutenance – Certification
Décembre 2025



PLAN DE LA PRÉSENTATION

1. Introduction & contexte du projet
2. Analyse stratégique et identification du besoin
3. Architecture data globale
4. Data Engineering & pipelines
5. Analyse avancée des avis & NLP
6. Gouvernance, déploiement & pilotage
7. Conclusion & perspective



Introduction & contexte du projet

Objectifs de la partie

- I. Présenter le contexte global du projet
- II. Exposer la problématique métier
- III. Définir les objectifs et la vision d'ensemble

Compétences RNCP mobilisées

C1 – Analyse de l'écosystème et du contexte
C8 – Cadrage et formalisation du besoin
C26 – Pilotage et vision globale du projet



- ❑ Amazon : acteur majeur du e-commerce et du cloud
- ❑ Les avis clients : levier clé de confiance et de performance
- ❑ Problématique : volume massif + données non structurées
- ❑ Besoin d'une exploitation data-driven des avis



Presentation d'Amazon et écosystème



- ❑ 3^e marque mondiale (Interbrand 2024)
- ❑ E-commerce, AWS, logistique, services numériques
- ❑ Culture data & IA au cœur des décisions
- ❑ Environnement international et réglementé



Pourquoi les avis clients sont stratégiques

☐ Impact sur :

- Confiance des acheteurs
- Classement des produits
- Recommandations algorithmiques

☐ Source directe de feedback métier

☐ Forte valeur mais faible structuration



- ☐ Trop d'avis → difficulté à identifier les plus pertinents
- ☐ Risque de faux avis ou avis peu utiles
- ☐ Manque de priorisation intelligente
- ☐ Besoin d'automatisation et de transparence



- ☐ Construire une solution data-driven
- ☐ Automatiser l'analyse et la priorisation des avis
- ☐ Intégrer NLP & scoring de pertinence
- ☐ Respecter les contraintes réglementaires



- ☐ Analyse des avis textuels
- ☐ Enrichissement par métadonnées (rating, image, achat)
- ☐ Pipeline data & NLP
- ☐ Restitution analytique



- ❑ Projet orienté **Data Engineering**
- ❑ Forte dimension **analyse et valeur métier**
- ❑ Prise en compte :
 - des contraintes techniques
 - des contraintes réglementaires
 - des contraintes organisationnelles



- ☐ Comprendre l'écosystème Amazon
- ☐ Identifier forces, faiblesses, opportunités
- ☐ Définir les besoins métier réels
- ☐ Prioriser les cas d'usage



Analyse stratégique et identification du besoin

Objectifs de la partie

- I. Analyser le contexte et les enjeux du projet
- II. Identifier les besoins métier prioritaires
- III. Définir le cas d'usage à plus forte valeur

Compétences RNCP mobilisées

C1 – C6
C8



FORCES	FAIBLESSES
<ul style="list-style-type: none">• Marque mondiale forte et reconnue• Réseau logistique très performant• Leadership e-commerce & cloud (AWS)• Capacités avancées en IA, NLP et Big Data• Large base de clients fidèles	<ul style="list-style-type: none">• Dépendance aux vendeurs tiers• Manque de transparence perçu des algorithmes• Complexité organisationnelle• Risques juridiques et réglementaires• Défis de sécurité à grande échelle
OPPORTUNITÉS	MENACES
<ul style="list-style-type: none">• Croissance des marchés émergents• Exploitation stratégique de l'IA et des LLM• Valorisation des retours clients• Innovation logistique et durable• Fidélisation et hausse du panier moyen	<ul style="list-style-type: none">• Concurrence intense (Alibaba, Walmart, Shein...) • Pression réglementaire (RGPD, CCPA, AI Act)• Cyberattaques et fuites de données• Perte de confiance dans les avis produits• Marché en évolution rapide



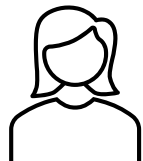
Approche méthodologique (Approche Design Thinking)

- ☐ Comprendre → Observer → Définir
- ☐ Idéation → Prototype → Test
- ☐ Approche centrée utilisateur
- ☐ Itération continue



IDENTIFICATION DES PARTIES PRENANTES

- ☐ Clients finaux
- ☐ Équipes produit
- ☐ Équipes data & analytics
- ☐ Équipes support et qualité



- Emma
- 21 ans
- Etudiante en psychologie

Besoin identifié :

- Bénéficier d'un service de livraison rapide.

Cas d'utilisation :

- Optimisation logistique et amélioration de la livraison dans les zones rurales.



- Rayan
- 25 ans
- Business Analyste chez Amazon

Besoin identifié :

- Pouvoir automatiser l'analyse et la classification des avis clients sur les produits.

Cas d'utilisation :

- Classer les avis et identifier leur pertinence.



- Mathis
- 42 ans
- entrepreneur

Besoin identifié :

- Améliorer la visibilité et la performance de ses produits.
- Comprendre les critères influençant le classement des produits.

Cas d'utilisation :

- Analyse des avis clients et tendances de satisfaction.
- Transparence sur les algorithmes de classement produits.



- Rayan
- 25 ans
- Business Analyste chez Amazon

Besoin retenu :

- ✓ Classer les avis et identifier leur pertinence.

Valeurs ajoutées :

Analyse rapide et plus fiable.

Elimination d'un travail manuel et gain de temps.

Levier et première brique pour d'autres besoins identifiés.



Priorisation des cas d'usage

❑ Priorité : classification & pertinence des avis

- Fort impact métier
- Faisabilité technique
- Base pour d'autres cas d'usage



- ❑ Stockage & Data Lake : Amazon S3
- ❑ Data Warehouse : Snowflake
- ❑ Analyse & ML : NLP, Zero-Shot, VADER
- ❑ Plateformes avancées : Streamlit,
SageMaker



- ☐ RGPD : protection et anonymisation des données
- ☐ CCPA : droits des utilisateurs
- ☐ AI Act : transparence des algorithmes
- ☐ RSE : IA responsable et durable



Veille réglementaire statique



Texte	Sujet	Points clés	Date	Mots-clés utilisés
RGPD (Regulation (EU) 2016/679)	Protection des données personnelles	<ul style="list-style-type: none">- Cadre clé de conformité- Impact direct sur le NLP	Texte initial 27 avr. 2016 (applicable depuis 2018)	RGPD, data privacy, anonymisation, consentement, EU compliance
CCPA (California Consumer Privacy Act)	Réglementation américaine sur la protection des données personnelles	<ul style="list-style-type: none">- Cadre californien de protection des données- Impact opérationnel	Adopté en 2018 – renforcé en 2023 (CPRA)	CCPA, CPRA, data privacy US, consentement, droit à la suppression
AI Act (Règlement UE)	Régulation IA (transparence, obligations)	<ul style="list-style-type: none">- Transparence et contrôle- Impact technique	2024–2025 (mesures d'application 2025).	IA Acte, réglementation IA, transparence, conformité UE
RSE Numérique (Responsabilité Sociétale & Environnementale)	Réglementation / bonnes pratiques	<ul style="list-style-type: none">- IA responsable et sobre- Dimension éthique et sociale	2024- 2025	RSE numérique, CSRD, Green AI, IA responsable, durabilité



- ☐ Documentation centralisée
- ☐ Partage collaboratif (outils internes)
- ☐ Mise à jour continue





Conclusion de la phase stratégique

- ☐ Besoin clairement défini
- ☐ Cas d'usage priorisé
- ☐ Nécessité d'une architecture scalable
- ☐ Passage à la conception technique



Architecture data globale

Objectifs de la partie

- I. Présenter l'architecture data cible du projet
- II. Justifier les choix technologiques retenus
- III. Garantir la sécurité, la traçabilité et la conformité des données

Compétences RNCP mobilisées

C9 – C11
C14
C15

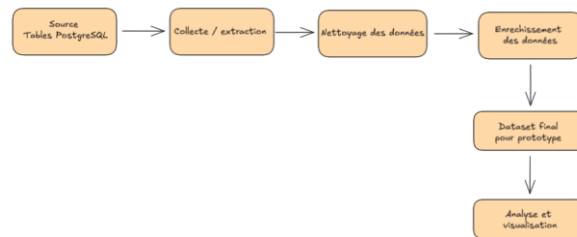


OBJECTIFS DE L'ARCHITECTURE

- ☐ Centraliser et structurer les données
- ☐ Garantir la qualité et la traçabilité
- ☐ Permettre l'analyse et la valorisation métier
- ☐ Assurer la scalabilité et la gouvernance

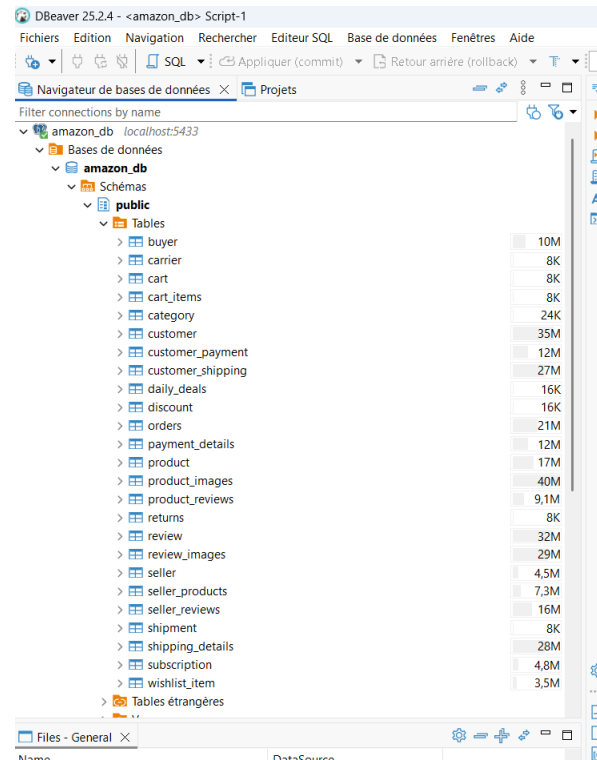


- ❑ Architecture orientée **Data Plateforme**
- ❑ Séparation claire des responsabilités
- ❑ Flux de données maîtrisé de bout en bout



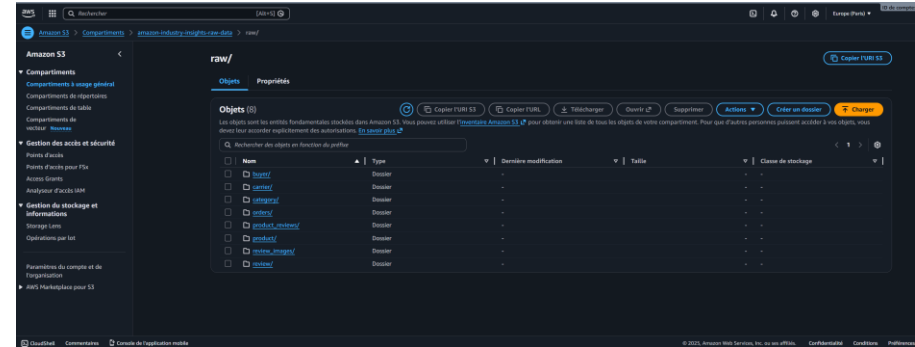


- ❑ Base transactionnelle PostgreSQL
- ❑ Données structurées (produits, acheteurs, avis)
- ❑ Données enrichies (images, métadonnées)



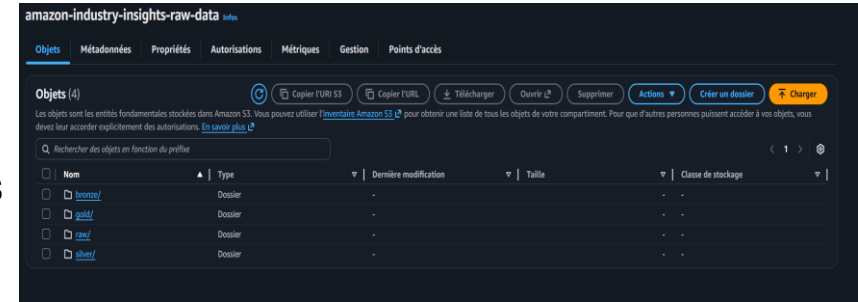


- ❑ Stockage des données brutes
- ❑ Historisation des données
- ❑ Découplage source / analytique





- ❑ **Bronze** : données brutes
- ❑ **Silver** : données nettoyées et enrichies
- ❑ **Gold** : données prêtes pour l'analyse





GESTION DES DONNÉES NON CONFORMES

(Gestion des rejets et logs)



- ❑ Données rejetées stockées séparément
- ❑ Journalisation des erreurs
- ❑ Traçabilité des traitements

Mongo Express Database: airflow_logs -> Collection: transform_load_logs -> Document 693eef787f4504d8e41b3520

Editing Document: 693eef787f4504d8e41b3520

```
1 {  
2   _id: ObjectId('693eef787f4504d8e41b3520'),  
3   message: 'loading tables from S3...',  
4   level: 'info',  
5   logger: 'utils.review_processor',  
6   timestamp: ISODate('2025-12-14T17:10:16.071Z'),  
7   dag_id: null,  
8   task_id: null,  
9   run_id: null  
10 }
```

← Back

Save

Delete

Mongo Express Database: airflow_logs -> Collection: extract_to_s3_logs -> Document 693eef6b2d306782f9466bcb

Editing Document: 693eef6b2d306782f9466bcb

```
1 {  
2   _id: ObjectId('693eef6b2d306782f9466bcb'),  
3   message: 'Starting anonymize data for table category...',  
4   level: 'info',  
5   logger: 'unusual_prefix_4c201c21eed689f79219dc2715e92d8f62d84ac5_extract_to_s3',  
6   timestamp: ISODate('2025-12-14T17:10:03.131Z'),  
7   dag_id: null,  
8   task_id: null,  
9   run_id: null  
10 }
```

← Back

Save

Delete



- ❑ Enchaînement automatisé des traitements
- ❑ Planification des tâches
- ❑ Gestion des dépendances

Mongo Express Database: airflow_logs

Viewing Database: airflow_logs

Collections

Collection Name [+ Create collection](#)

View	Export	[JSON]	Import	extract_to_s3_logs	Del
View	Export	[JSON]	Import	transform_load_logs	Del
View	Export	[JSON]	Import	transform_logs	Del

Database Stats

Collections (incl. system.namespaces)	3
Data Size	16.1 KB
Storage Size	61.4 KB
Avg Obj Size #	210 Bytes
Objects #	77
Indexes #	3



SÉCURITÉ & TRAÇABILITÉ (gouvernance)



- ☐ Contrôle des accès
- ☐ Traçabilité des transformations
- ☐ Conformité réglementaire intégrée



Conclusion de la partie architecture

- ☐ Architecture cohérente et scalable
- ☐ Séparation claire des rôles
- ☐ Base solide pour les pipelines



Data Engineering & pipelines

Objectifs de la partie

- I. Mettre en œuvre les pipelines de données automatisés
- II. Garantir la qualité, la fiabilité et la traçabilité des données
- III. Industrialiser les traitements et les flux data

Compétences RNCP mobilisées

C12 – C13

C18 – C19



RÔLE DU DATA ENGINEERING

- ☐ Transformer des données brutes en données exploitables
- ☐ Automatiser les traitements
- ☐ Garantir la fiabilité et la qualité des données
- ☐ Préparer l'analyse et la restitution métier



- ❑ Pipeline **ETL orchestré par Airflow**
- ❑ Architecture **modulaire et découplée**
- ❑ Séparation ingestion / transformation / restitution
- ❑ Traçabilité et qualité intégrées dès la conception



- ❑ DAG principal d'orchestration
- ❑ DAG d'extraction PostgreSQL → S3
- ❑ DAG de transformation & chargement
- ❑ Communication inter-DAG via XCom



- ❑ Base PostgreSQL transactionnelle
- ❑ Tables cœur métier :
 - avis,
 - produits,
 - acheteurs,
 - catégories,
- ❑ Données structurées et semi-structurées



INGESTION : POSTGRESQL → S3 (DATA LAKE)

- ☐ Extraction depuis PostgreSQL
- ☐ Anonymisation des données sensibles
- ☐ Stockage en zone raw sur S3
- ☐ Gestion des erreurs et alerting



- ☐ Chargement des données depuis S3
- ☐ Jointures multi-tables
- ☐ Nettoyage et normalisation
- ☐ Séparation données propres / rejetées



- ☐ Tests de qualité automatisés
- ☐ Validation des règles métier
- ☐ Génération de rapports qualité
- ☐ Seuils d'acceptation contrôlés



- ☐ Chargement des données propres dans Snowflake
- ☐ Modèle orienté analytique
- ☐ Création de vues métiers
- ☐ Séparation stockage / calcul



- ☐ Logs centralisés dans MongoDB
- ☐ Métadonnées d'exécution
- ☐ Suivi des volumes traités
- ☐ Historisation des runs



- ☐ Détection des anomalies
- ☐ Alertes email automatiques
- ☐ Blocage contrôlé du pipeline
- ☐ Reprise sécurisée



- ☐ Conteneurisation Docker
- ☐ Variables d'environnement sécurisées
- ☐ Déploiement reproductible
- ☐ Séparation dev / prod



- ☐ Ajout de nouvelles sources
- ☐ Augmentation des volumes
- ☐ Nouveaux cas d'usage analytiques
- ☐ Passage à des traitements plus avancés



Conclusion Data Engineering



- ☐ Données fiables et prêtes à l'analyse
- ☐ Pipelines automatisés
- ☐ Base solide pour la valorisation métier



Analyse avancée des avis & NLP

Objectifs de la partie

- I. Analyser les avis clients à partir des données textuelles
- II. Mettre en place un scoring de pertinence des avis
- III. Exploiter le NLP pour produire de la valeur métier

Compétences RNCP mobilisées

C7
C16
C17



- ☐ Comprendre le contenu réel des avis clients
- ☐ Identifier les avis les plus exploitables
- ☐ Compléter la note par des signaux qualitatifs
- ☐ Aider la prise de décision métier



VUE D'ENSEMBLE DES DONNÉES ANALYSÉES

- ❑ 2 592 avis analysés
- ❑ Rating moyen : 4,46 / 5
- ❑ 31,6 % des avis contiennent des images
- ❑ Longueur texte très hétérogène



- ☐ Forte concentration en 5★ ($\approx 76\%$)
- ☐ Peu d'avis négatifs
- ☐ Déséquilibre naturel des classes



- ☐ Avis 2★–4★ → textes plus longs
- ☐ Avis 5★ → textes plus courts
- ☐ Avis négatifs souvent plus argumentés



PRÉSENCE D'IMAGES DANS LES AVIS

- ☐ Corrélation positive entre rating élevé et images
- ☐ 35,7 % des avis 5★ avec images
- ☐ Signal de preuve utilisateur



- ☐ Nettoyage du texte
- ☐ Normalisation
- ☐ Filtrage des avis trop courts
- ☐ Calcul de features textuelles



FEATURES UTILISÉES POUR L'ANALYSE

- ☐ Longueur du texte
- ☐ Présence d'image
- ☐ Rating
- ☐ Score de sentiment
- ☐ Indicateurs de pertinence



- Score positif / neutre / négatif
- Complément du rating
- Détection d'écarts note / sentiment



IDENTIFICATION DES AVIS PERTINENTS

- ☐ Combinaison longueur + image + sentiment
- ☐ Priorisation des avis utiles
- ☐ Réduction du bruit informationnel



- ☐ Classification thématique sans entraînementModèles
- ☐ NLP pré-entraînés (Transformers)
- ☐ Détection dynamique des sujets évoqués
- ☐ Forte flexibilité métier

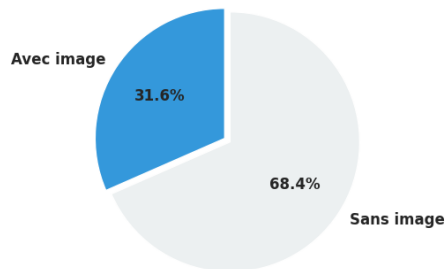


- ☐ Exploration interactive
- ☐ Filtres par score / rating
- ☐ Lecture directe des avis clés
- ☐ Approche orientée usage métier

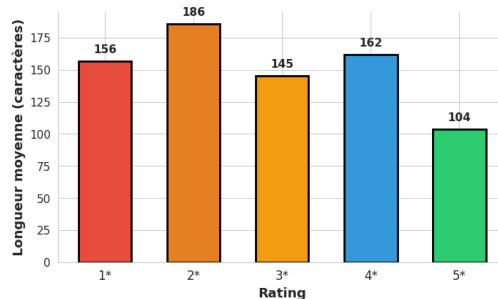


Analyse sur 10 produits (2592 avis)

Présence d'Images



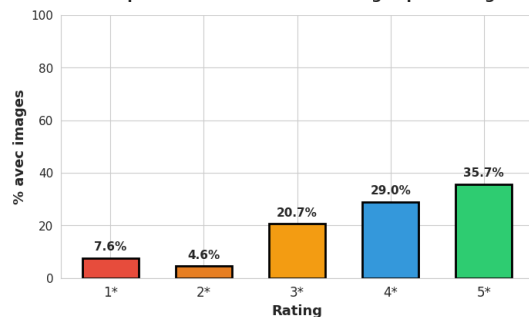
Longueur Moyenne du Texte par Rating



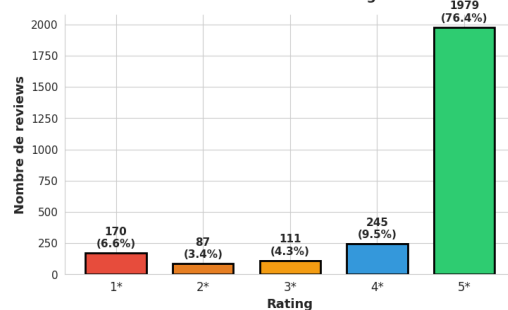
RÉSUMÉ STATISTIQUE

- Reviews totales : 2592
- Avec images : 818 (31.6%)
- Avec commandes : 2592 (100.0%)
- Longueur texte :
 - Moyenne : 117 caractères
 - Médiane : 44 caractères
 - Min : 4 | Max : 1815
- Rating moyen : 4.46/5

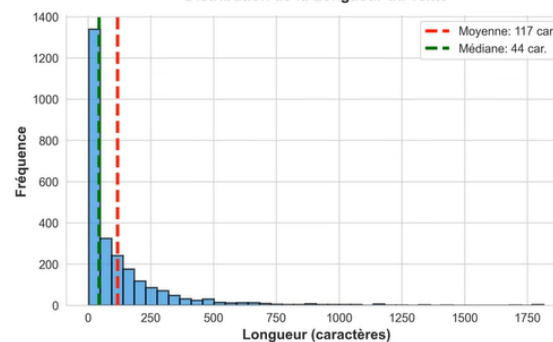
Proportion de Reviews avec Images par Rating



Distribution des Ratings



Distribution de la Longueur du Texte





Justificatif de pondération

30 %

Longueur de texte : plus l'avis est long, plus il y a de détail.

25 %

Sentiment (VADER) : analyse de sentiments pour détecter la satisfaction ou l'insatisfaction des clients.

20 %

Présence d'image : 31% des avis contiennent des images (indicateur d'engagement).

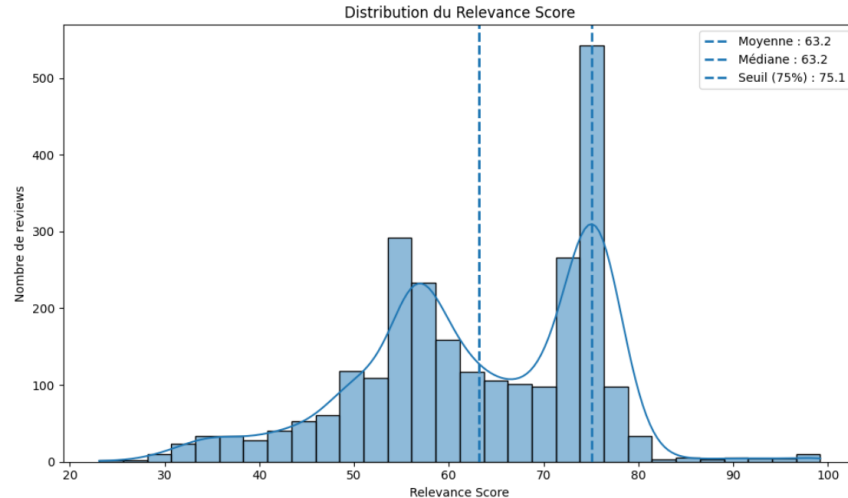
15 %

Note extrême : les avis très positifs ou très négatifs apportent souvent plus d'information.

10 %

Achat vérifié : tous les avis sont vérifiés

Sélection de seuil





Exemple 1 : avis pertinent

- Description : « Love it »
- Présence d'image
- Achat vérifié
- **Sentiment positif fort (VADER)**
- Score d texte (fonction centrée sur 300 caractères)

Exemple 2 : avis non pertinent

- Description : « This product is very good.... (~ 1000 caractères)»
- Présence d'une image
- Achat vérifié
- Sentiment moyen (VADER)
- **Score du texte (fonction centrée sur 300 caractères) qui fait baisser sa valeur**

Limites claires

- Score basé surtout sur la forme et l'émotion
- Survalorisation des avis 5★
- Avis courts parfois trop bien notés
- Avis longs parfois pénalisés
- Pas d'analyse sémantique
- Mesure l'engagement, pas la vraie pertinence



- ☐ Hiérarchisation claire des avis
- ☐ Identification des avis à forte valeur
- ☐ Réduction du bruit informationnel



Limites identifiées	Améliorations possibles
<input type="checkbox"/> Dépendance à la qualité du texte	<input type="checkbox"/> Modèles NLP plus avancés
<input type="checkbox"/> Pondérations perfectibles	<input type="checkbox"/> Apprentissage supervisé
<input type="checkbox"/> Volume de données limité	<input type="checkbox"/> Feedback utilisateur



- ☐ Tableaux analytiques
- ☐ Indicateurs clés par produit
- ☐ Aide à la décision



- ❑ Valorisation des résultats NLP
- ❑ Exploration interactive des avis clients
- ❑ Accès simple pour les équipes métier
- ❑ Approche produit data (MVP)



Amazon Product Reviews - Relevance Analysis

Browse products and view their most relevant reviews based on:

- ★ Rating score
- 📄 Text quality and length
- 🔍 Keyword relevance
- 🖼️ Presence of images

🔍 How do you want to search?

☒ Browse Products ☐ Search by Product ID

✅ 100 produits disponibles

🔍 Filter products by name:

Type to search...

📄 Showing 100 product(s)

🔍 Select a Product:

ISALE! SUPPLY Tallow Shave Soap (Juniper, 4.5 oz) (All Beauty)



- ☐ Connexion directe aux tables **Gold** (Snowflake)
- ☐ Lecture seule (pas d'impact sur les pipelines)
- ☐ Découplage traitement / visualisation
- ☐ Sécurité et traçabilité préservées

- ❑ Gain de temps pour les équipes
- ❑ Réduction de l'analyse manuelle
- ❑ Meilleure adoption de la solution data
- ❑ MVP évolutif (API, BI, authentification)

Most Relevant Reviews			
<i>Number of reviews to display:</i>			
<div> 1 review(s) per page(s) (page(s)) </div>			
<div> Review #1 (4/5) - Excellent product! Relevance: 63.84 </div>			
Description: <p>An exceptional product that performs as well as or better than advertised. The aroma is pleasing, but the moisturizing effect of the shave soap is very much appreciated. I bought two units and will undoubtedly repeat that order in the future. A very happy customer.</p> <p>Loungeston 3000 200 Characters</p>			
<div> Scores at Metrics: </div>			
 4/5	 95.069	 0.972	 No
 63.842	 0.985	 No	 Yes
<div> Information: </div>			
Buyer ID: 3d18c3e05044e4463a5f079194eb330ad38803a4ef38ac3ee4732da		Product / SKU / SUPPLY Follow Shave Soap (Loungeston 3000 200)	
Product ID: 911121PQCC		Category All Beauty	
		Review Category product quality or satisfaction	
		Status HELD	

Browse Products
 Search by Product ID

Enter Product ID

B0BM4GX6TT

Product ID: B0BM4GX6TT

★ Most Relevant Reviews

Number of reviews to display:

reviews | per (month) | trouble(s)

☐ Review #1 | ★★★★★ (5/5) - Easy to use! Highly recommended! | Relevance: 98.28

☐ Review #2 | ★★★★★ (5/5) - Works beyond belief for eyebores! | Relevance: 98.11

☐ Review #3 | ★★★★★ (5/5) - Brimstone(s) | Relevance: 96.33

☐ Review #4 | ★★★★★ (5/5) - Awesome product(s) | Relevance: 94.56



- ☐ Valeur métier concrète
- ☐ Aide à la décision
- ☐ Résultats exploitables



Gouvernance, déploiement & pilotage

Objectifs de la partie

- I. Assurer la gouvernance et la conformité réglementaire des données
- II. Déployer, superviser et maintenir la solution data
- III. Piloter le projet, les ressources et les risques

Compétences RNCP mobilisées

C20 – C32



- ☐ Gouvernance à 3 niveaux : stratégique, tactique, opérationnel
- ☐ Alignement métier / technique
- ☐ Décisions structurées et traçables
- ☐ Pilotage continu du projet



RÔLES, RESPONSABILITÉS & RACI

- ☐ Rôles clairement définis (PO, DE, DS, DevOps, Support...)
- ☐ Responsabilités formalisées
- ☐ Matrice RACI par phase
- ☐ Limitation des zones grises



- ☐ Méthodologie Agile / Scrum
- ☐ Sprints de 2 semaines
- ☐ Rituels : planning, daily, review, rétrospective
- ☐ Ajustement continu du backlog



- ☐ Déploiement progressif
- ☐ Séparation dev / prod
- ☐ Automatisation des traitements
- ☐ Sécurisation des accès



- ☐ Plan de formation structuré
- ☐ Documentation claire et versionnée
- ☐ Support à 3 niveaux
- ☐ Objectif : autonomie des utilisateurs



- ☐ Monitoring des pipelines
- ☐ Suivi des incidents
- ☐ Indicateurs de qualité
- ☐ Boucle d'amélioration continue



MAINTENANCE & ÉVOLUTIVITÉ

Maintien en conditions opérationnelles

- ☐ Maintenance corrective et évolutive
- ☐ Anticipation de la montée en charge
- ☐ Ajout de nouveaux cas d'usage



- ☐ Choix technologiques maîtrisés
- ☐ Priorité aux solutions open source / optimisées
- ☐ Arbitrage coûts / valeur



CONDUITE DU CHANGEMENT

Adoption par les utilisateurs

- ☐ Documentation claire
- ☐ Accompagnement des équipes
- ☐ Sensibilisation à l'usage des données



Conclusion & perspectives

Objectifs de la partie

- I. Présenter le bilan global du projet
- II. Mettre en évidence les apports et les limites
- III. Ouvrir sur les perspectives d'évolution

Compétences RNCP mobilisées

Synthèse globale du cycle de vie d'un projet
data

Capitalisation des compétences acquises
Vision long terme et amélioration continue



- ☐ Projet data de bout en bout
- ☐ Architecture robuste et scalable
- ☐ Valeur métier démontrée
- ☐ Gouvernance intégrée



- ☐ Enrichissement des modèles NLP
- ☐ Intégration de nouveaux jeux de données
- ☐ Déploiement à plus grande échelle



- ☐ Data Engineer orienté architecture
- ☐ Sensibilité gouvernance et conformité
- ☐ Capacité à piloter un projet data



amazon

CONCLUSION GÉNÉRALE

- ☐ Problématique métier adressée
- ☐ Solution technique industrialisable
- ☐ Vision long terme et responsable

Ismaël SYLLA
Soutenance – Certification
Décembre 2025





Remerciements



Merci aux membres du jury
Merci à l'organisme de formation / accompagnement
Merci pour le temps et l'échange



Merci !

