Documentation de Support Utilisateur

Projet : Analyse & Catégorisation des Avis Utilisateurs Amazon
Auteur : Ismaël Sylla
Organisation : Jedha x La Poste x Amazon
Date : Décembre 2025

Sommaire Premium

1. Introduction générale du support utilisateur
2. Guide de prise en main de l’environnement
3. Utilisation des outils et interfaces
4. Procédures opérationnelles quotidiennes
5. Procédures de support et de résolution de problèmes
6. Escalade, responsabilités et matrice RACI
7. Conclusion

1. Introduction générale du support utilisateur

Cette documentation a pour objectif d'accompagner l'ensemble des utilisateurs — analystes, data engineers, membres du support et équipes administratives — dans la bonne prise en main et l'utilisation du système **Amazon Review Analysis** en production. Elle permet de comprendre non seulement le fonctionnement technique du pipeline, mais aussi les responsabilités de chacun, les actions quotidiennes nécessaires, ainsi que la manière dont les incidents doivent être traités.

L'idée centrale de ce document est d'offrir un support clair, humain et orienté vers le terrain. Chaque section inclut des conseils pratiques, des exemples concrets et des procédures écrites de manière simple, compréhensible et directement applicables par des profils non techniques comme par des experts.

Le système reposant sur un ensemble d'outils connectés (PostgreSQL, Airflow, S3, Snowflake, MongoDB), il est essentiel que tous les utilisateurs sachent où agir, quoi surveiller et comment diagnostiquer un problème. Cette documentation est également conçue pour faciliter l'arrivée de nouveaux collaborateurs, accélérer la montée en compétence et réduire la dépendance aux équipes techniques.

2. Guide de prise en main de l'environnement

L'environnement de production s'appuie sur une architecture moderne DataOps intégrant un orchestrateur Airflow, une base PostgreSQL, un Data Lake S3, un Data Warehouse Snowflake et une base MongoDB destinée aux requêtes rapides. L'objectif de cette section est d'expliquer de manière simple comment s'y retrouver et comment interagir avec chaque composant.

Accès aux outils

Les accès sont fournis par l'administrateur système selon le rôle de l'utilisateur. Trois niveaux existent : **analyse**, **ingestion**, et **administration**. Chaque rôle récupère un accès limité à ce qui est réellement nécessaire, notamment pour respecter les bonnes pratiques de sécurité et le principe du moindre privilège.

Lorsqu'un nouvel utilisateur rejoint l'équipe, une procédure d'onboarding prévue dans les annexes permet de générer les comptes, attribuer les droits et valider qu'il dispose de toutes les ressources pour travailler immédiatement.

Environnement Airflow

Airflow est l'interface principale utilisée par les équipes opérationnelles. Grâce à sa vue graphique, un utilisateur peut rapidement comprendre l'état du pipeline, identifier l'étape en cours, consulter les logs ou relancer une tâche si nécessaire.

Un tutoriel d'utilisation est intégré dans cette documentation : comment lire un DAG, comment vérifier l'état d'une tâche, comment interpréter un échec et comment relancer un flux.

Accès aux données (S3 / Snowflake / MongoDB)

Chaque utilisateur dispose d'un guide pas à pas pour accéder à :

- **Snowflake**, afin de consulter les vues analytiques,
- **S3**, pour naviguer dans les zones RAW / PROCESSED,
- **MongoDB**, pour effectuer des recherches rapides ou alimenter des front-end.

Des captures d'écran et des exemples de requêtes sont fournis en annexe pour faciliter la prise en main.

3. Utilisation des outils et interfaces

Cette section détaille chaque interface utilisateur et explique comment exécuter des actions courantes. L'objectif est de rendre le système simple à utiliser, même pour des profils non techniques.

3.1 Airflow – L'interface principale des opérations

Airflow est utilisé au quotidien pour suivre l'exécution du pipeline. L'utilisateur peut y : - consulter le statut global du DAG,
- Naviguer entre les tâches,
- Visualiser les logs,
- Relancer une tâche en erreur,
- Vérifier l'historique complet du pipeline.

Dans une logique de support, chaque page Airflow affiche un bouton vers les logs détaillés permettant de comprendre immédiatement la source d'un éventuel problème. Les logs sont rédigés en langage clair pour faciliter la compréhension.

3.2 Snowflake – Consultation des données et requêtes

Snowflake est la plate-forme où les données finales sont consultées. L'utilisateur peut accéder aux dashboards, aux vues métier et aux tables agrégées. Les requêtes SQL les plus fréquemment utilisées sont fournies en annexe.

Les vues principales exposées sont :

- *vw_reviews_by_product* : synthèse des avis par produit,
- *vw_reviews_by_category* : analyse par catégorie,
- *vw_data_quality* : qualité des données et reporting métier.

3.3 MongoDB – Consultation temps réel

MongoDB permet un accès rapide aux avis transformés, idéal pour des filtres dynamiques et une intégration front-end. Les utilisateurs peuvent y consulter les avis enrichis, les avis rejetés avec les motifs, et les métadonnées du pipeline.

4. Procédures opérationnelles quotidiennes

Les procédures opérationnelles quotidiennes jouent un rôle essentiel dans la stabilité de toute architecture en production. Elles permettent d'assurer que les flux fonctionnent comme prévu, que les données restent conformes, et que l'équipe détecte immédiatement toute anomalie avant qu'elle n'impacte les utilisateurs finaux ou les équipes métiers.
Cette routine s'apparente à une véritable « visite quotidienne » du système : rapide, incisive mais indispensable pour maintenir un niveau d'excellence opérationnelle.

4.1 Revue quotidienne du pipeline Airflow

Chaque matin, l'équipe Data commence par consulter l'interface Airflow. Il s'agit d'un réflexe incontournable : Airflow étant le point névralgique de l'orchestration, c'est souvent ici que l'on détecte les premiers signes d'un dysfonctionnement.

Cette revue inclut :

- l'état du DAG principal (succès, échecs, retries, tâches en attente),
- le temps d'exécution des tâches afin d'identifier des lenteurs inhabituelles,
- l'observation des dépendances, pour vérifier que les blocs critiques s'exécutent dans l'ordre prévu,

- la consultation des logs Airflow, permettant de repérer une erreur discrète qui n'a pas bloqué l'exécution globale mais pourrait devenir un problème futur.

Cette étape est rapide, mais elle donne une vision très fiable de la santé du pipeline.

4.2 Vérification des volumes et de la cohérence des données

Une fois Airflow validé, l'équipe passe à l'analyse de la volumétrie quotidienne : combien de produits, d'avis, de commandes ont été extraits ?
Ces chiffres doivent rester cohérents avec l'activité habituelle. Une baisse anormale peut indiquer :

- Un problème côté base PostgreSQL,
- Une table vide ou mal alimentée,
- Un problème réseau lors de l'extraction.

À l'inverse, un volume inhabituellement élevé peut cacher :

- Une duplication,
- Une réinjection erronée,
- Une logique métier qui aurait changé côté système source.

Cette étape joue un rôle crucial : elle permet de prévenir les dérives avant qu'elles ne remontent dans Snowflake ou MongoDB.

4.3 Surveillance de Snowflake et des performances analytiques

Snowflake étant la brique analytique principale, il est important de vérifier quotidiennement :

- La consommation du warehouse,
- Les durées d'exécution des vues analytiques,
- L'activation éventuelle de l'auto-scale,
- La cohérence des données chargées la veille.

Une dégradation des performances peut signaler :

- Un fichier S3 mal structuré,

- Une vue trop coûteuse,
- Une augmentation d'activité métier,
- Ou encore un paramétrage du warehouse à reconsidérer.

Cette surveillance garantit une expérience optimale pour les Data Analysts, responsables BI et équipes métiers.

4.4 Contrôle de la santé de MongoDB

MongoDB est utilisé pour exposer les données enrichies à un usage applicatif temps réel. À ce titre, sa stabilité est essentielle.

Chaque jour, l'équipe vérifie :

- La fraîcheur des données dans les collections reviews et rejected_reviews,
- La bonne santé des indexes (absence de fragmentation excessive, taille cohérente),
- L'absence d'erreurs dans les logs Mongo,
- La consommation du disque et le taux de croissance de la base.

Ces contrôles garantissent que la plateforme répond instantanément aux requêtes, en particulier lorsqu'un front-end client ou une API s'y connecte.

4.5 Vérification centralisée des logs et alertes

Les logs représentent la mémoire du système. Même lorsqu'un pipeline semble avoir tourné correctement, les logs permettent d'identifier des comportements anormaux :

- Problèmes intermittents de connexion PostgreSQL,
- Latences dans les appels S3,
- Erreurs non bloquantes de parsing,
- Anomalies de validation dans MongoDB.

Chaque jour, l'équipe Data consulte :

- Les logs Airflow,
- Les logs d'exécution Snowflake,
- Les logs MongoDB (healthcheck),

- Les notifications du système de monitoring.

L'objectif n'est pas seulement de réagir, mais d'anticiper.
Un petit warning aujourd'hui peut devenir un incident majeur demain si rien n'est fait.

4.6 Gestion proactive des incidents mineurs

Les incidents mineurs sont fréquents en production : fichier légèrement corrompu, absence d'une ligne dans un CSV, micro-surtension réseau, etc.

Lorsqu'un incident de ce type est détecté, l'équipe procède immédiatement à :

- La relance d'une tâche Airflow,
- La correction manuelle d'un fichier RAW,
- La régénération d'un batch S3,
- La purge contrôlée d'un dossier,
- Ou la manipulation d'un index MongoDB.

L'objectif est d'éviter l'escalade en résolvant ces problèmes rapidement, souvent avant même que les utilisateurs métiers ne s'en rendent compte.

Synthèse

Les opérations quotidiennes ne sont pas seulement un ensemble de tâches techniques. Elles traduisent une philosophie : **prévenir plutôt que guérir**, et garantir un pipeline fluide, stable et résilient.

Elles permettent également :

- Un haut niveau de confiance dans les données,
- Une réduction drastique des pannes,
- Une traçabilité renforcée,
- Et une sérénité opérationnelle pour les équipes Data et Support.

5. Procédures de support et résolution d'incidents

5. Procédures de support et résolution d'incidents (Version Premium – Détaillée & Humanisée)

La mise en production d'une architecture data comme celle du projet Amazon Review Analysis impose de mettre en place un dispositif de support robuste, structuré et opérationnel. L'objectif n'est pas seulement de réagir lorsqu'un problème survient, mais d'assurer une continuité de service irréprochable et une gestion maîtrisée des imprévus.

Les procédures décrites ci-dessous s'inspirent des bonnes pratiques Amazon (Operational Excellence), du cadre DataOps, ainsi que des normes internes La Poste (MCO, traçabilité et conformité RGPD). Elles encadrent chaque étape de la prise en charge d'un incident : de la détection initiale à la résolution finale, en passant par la communication avec les parties prenantes.

5.1 Détection et identification des incidents

La détection est une étape clé. Elle peut provenir de plusieurs sources :

🔍 A. Détection automatique (monitoring & alerting)

Le système déclenche automatiquement une alerte en cas de :

- Tâche Airflow en échec ou en retard,
- Indisponibilité de PostgreSQL, MongoDB ou Snowflake,
- Dépassement du seuil d'erreurs NLP ou anonymisation,
- Anomalies de volume dans S3 (fichiers trop petits, manquants, corrompus),
- Chute de performance inhabituelle,
- Dépassement du quota Snowflake ou surcharge du warehouse,
- Indisponibilité du scheduler Airflow,
- Echec de connexion réseau ou timeout.

Ces alertes apparaissent dans :

- Airflow UI (suivi des DAGs),

- logs centralisés du pipeline,
- systèmes de notifications (email, Slack...),
- dashboards techniques.

🎧 B. Détection humaine (remontée utilisateur ou analyste)

Un utilisateur, un analyste ou un PO peut également signaler :

- un dashboard vide ou incohérent,
- Une latence anormale côté front-end (MongoDB),
- Des erreurs d'accès Snowflake,
- Une perte de données apparente,
- Une incohérence fonctionnelle.

Dans ce cas, l'incident est consigné et transmis à N1.

5.2 Qualification rapide

Dès qu'un incident est identifié, le support N1 réalise une qualification initiale afin de déterminer sa priorité et son périmètre.

N1 vérifie en priorité :

- le statut du DAG (failed / running / skipped),
- la disponibilité des services (Docker ps, connexions DB, health checks MongoDB),
- la présence de fichier corrompu ou manquant dans S3,
- la charge Snowflake (warehouse suspendu / en surcharge),
- l'exactitude des credentials (variable .env expirée),
- les logs d'anomalie les plus récents.

🔎 Objectif : déterminer si l'incident peut être corrigé par N1 ou s'il nécessite un passage immédiat au N2.

5.3 Actions de remédiation (Niveau 1 – interventions simples)

Si l'incident est mineur ou lié à des opérations courantes, N1 réalise :

- relance du DAG Airflow,
- redémarrage d'un service Docker (PostgreSQL, MongoDB, Airflow webserver),
- purge des fichiers temporaires bloquants,
- mise à jour des credentials expirés,
- vérification et réactivation du Snowflake Warehouse,
- relance d'un stage S3 en erreur.

Ces actions sont documentées et tracées dans le journal d'incidents.

5.4 Diagnostic approfondi (Niveau 2 – Data Engineering)

Lorsque l'incident dépasse les capacités du support N1, il est escaladé à l'ingénieur Data qui mène une analyse plus poussée.

Le N2 intervient sur :

- erreurs Python dans les transformations (anonymisation / NLP),
- corruption de données dans PostgreSQL,
- erreur de schéma dans Snowflake (incompatibilité de colonnes),
- problèmes d'index ou de volume dans MongoDB,
- incohérences métier dans les tables (doublons, valeurs nulles),
- absence de données dans une zone du Data Lake.

Le Data Engineer réalise :

- une analyse détaillée des logs Airflow (stack trace, timestamp, task_id),
- un replay de l'étape incriminée,

- une vérification de l'intégrité des données sur S3,
- un contrôle qualité de l'échantillon source,
- un test manuel de la transformation en local,
- des corrections ponctuelles si nécessaire.

📝 Toutes les analyses et corrections N2 sont consignées dans le rapport MCO.

5.5 Escalade critique (Niveau 3 – Cloud / Sécurité / Infrastructure)

Si le problème touche l'infrastructure, la sécurité ou le cloud, l'incident passe à N3 :

🚨 Situations typiques nécessitant N3 :

- S3 inaccessible ou permissions IAM rompues,
- base PostgreSQL corrompue ou volume Docker endommagé,
- cluster Snowflake défaillant ou authentification impossible,
- incapacité à accéder à MongoDB (panne majeure),
- rotation forcée des clés AWS ou incident de sécurité,
- suspicion d'exposition de données sensibles,
- problème réseau ou DNS.

N3 engage :

- diagnostics réseau,
- vérification IAM,
- restauration de snapshots,
- analyse de compromission (sécurité),
- rollback complet ou partiel si nécessaire.

Lorsque la situation est critique, une communication exceptionnelle avec le PO et le DPO peut être initiée.

5.6 Documentation et communication

Une fois l'incident résolu, la documentation est obligatoire :

- cause racine (Root Cause Analysis / RCA),
- actions correctives appliquées,
- recommandations ou patch à prévoir,
- impacts éventuels sur les données,
- communication envoyée aux parties prenantes,
- mise à jour des procédures si nécessaire.

Cette étape est essentielle pour capitaliser sur l'incident et éviter sa répétition.

5.7 Amélioration continue (Post-mortem)

Chaque incident majeur donne lieu à un Post-Mortem, inspiré de la culture Amazon :

- sans blâme (« blameless post-mortem »),
- orienté amélioration continue,
- pédagogique pour les équipes,
- suivi par des actions concrètes (patch, documentation, monitoring).

L'objectif est de renforcer progressivement la plateforme.

6. Escalade, responsabilités et matrice RACI

La gestion des incidents et des responsabilités est un pilier essentiel du fonctionnement d'une architecture de production. Dans un environnement où plusieurs technologies interagissent — Airflow, PostgreSQL, Snowflake, S3, MongoDB — il est indispensable de

définir clairement **qui fait quoi**, qui intervient en premier, comment l'information circule, et à quel moment un incident doit être escaladé vers un niveau supérieur.

Cette démarche, héritée des pratiques Amazon ("ownership", "operational excellence") et des standards La Poste (traçabilité, conformité, continuité), permet d'éviter les zones grises, d'assurer une résolution rapide des problèmes et de garantir la stabilité du système dans la durée.

L'objectif n'est pas uniquement de répartir les tâches, mais de donner à chaque acteur une vision claire de son rôle dans le maintien en conditions opérationnelles (MCO) du pipeline.

1. Logique d'escalade (N1 → N2 → N3)

La gestion d'incident repose sur une chaîne d'escalade structurée en trois niveaux :

Niveau 1 – Support Opérationnel

Le support N1 est le premier point de contact lorsqu'une anomalie apparaît.
Il surveille les alertes Airflow, les logs d'exécution, les dashboards de monitoring et vérifie les disponibilités clés :

- statut des services (Airflow scheduler/webserver, PostgreSQL, MongoDB),
- charge Snowflake,
- état des buckets S3,
- saturation disque/container.

Son rôle est d'isoler rapidement la source visible du problème :
connexion perdue, espace disque, DAG en échec, erreur de credentials, etc.

N1 ne corrige pas le fond du problème mais applique un ensemble de procédures simples (redémarrage de service, purge de fichier temporaire, relance de DAG). Si l'incident dépasse son périmètre, il passe à N2.

Niveau 2 – Équipe Data Engineering

Le niveau 2 est généralement assuré par les ingénieurs Data responsables du pipeline.
Ils interviennent lorsqu'un incident nécessite une compréhension technique plus avancée :

- Analyse approfondie des logs Airflow
- Debug des scripts Python

- Vérification des transformations (anonymisation, NLP)
- Correction des schémas Snowflake / erreurs de staging
- Résolution de problèmes de performance sur les extractions PostgreSQL
- Analyse des données corrompues ou manquantes

N2 possède la vision complète du pipeline et peut apporter des correctifs ciblés.
Si la cause dépasse le cadre applicatif pour toucher l'infrastructure profonde (clusters Snowflake, permissions S3, réseau, Docker, sécurité), l'incident est escaladé à N3.

Niveau 3 – Équipe Infrastructure / Cloud / Sécurité

Le dernier niveau correspond aux experts infrastructure ou cloud.
Ils interviennent uniquement en cas de problème structurel ou critique :

- pannes réseau ou DNS
- défaillance du cluster Snowflake ou de l'authentification
- corruption profonde du volume Docker PostgreSQL
- politique IAM défaillante ou compromission de secrets
- incident de sécurité ou suspicion de fuite de données
- limitation S3 / quotas / droits bloquants

Leur rôle est de rétablir un environnement fonctionnel, de sécuriser l'ensemble des composants cloud (IAM, audit logs, rotation des clés) et de coordonner les actions de crise.

2. Rôles et responsabilités détaillés

Afin d'assurer une répartition claire des responsabilités, voici la liste des acteurs impliqués :

Rôle	**Description**
Support Niveau 1 (N1)	Supervise Airflow, surveille les logs, identifie les symptômes, applique les relances simples
Data Engineer (N2)	Maintient et corrige les scripts, optimise les performances,

Rôle	Description
	analyse les erreurs métier
Cloud Engineer / Infra (N3)	Gère les environnements Docker, S3, Snowflake, sécurité IAM, incidents critiques
Product Owner	Priorise les anomalies, communique avec les parties prenantes, valide les impacts métier
Security Officer / DPO	Valide la conformité RGPD, supervise les demandes de suppression utilisateur
Business Analyst / Analyste Data	Remonte les anomalies de données ou incohérences analytics

Chaque acteur doit être mobilisé en fonction de l'incident et de son impact.

3. Matrice RACI

La matrice RACI clarifie précisément qui est **Responsable (R)**, **Accountable (A)**, **Consulté (C)** et **Informé (I)** pour chaque opération clé du système.

Je l'ai construite pour ton architecture complète :

- DAG Airflow (orchestration)
- PostgreSQL (source)
- S3 Data Lake (raw, processed, curated)
- MongoDB (NoSQL)
- Snowflake (warehouse)
- Monitoring / logs / RGPD

3.1 Tableau RACI – Gestion Opérationnelle

Activité / Composant	N1	Data Engineer	Cloud/Infra	Product Owner	Security/DPO	Analyste
Surveillance Airflow	R	C	I	I	I	I
Redémarrage d'un DAG	R	C	I	I	I	I
Correction de scripts Python	I	R/A	C	I	I	I
Problème d'extraction PostgreSQL	C	R/A	C	I	I	I
Maintenance Docker	I	C	R/A	I	I	I
Gestion credentials / .env	I	R	A	I	C	I
Gestion IAM (S3 / Snowflake)	I	C	R/A	I	C	I
Problème de performance Snowflake	I	R/A	C	I	I	C
Correction données corrompues	I	R/A	C	C	I	C
Demande RGPD (droit à l'oubli)	I	R	C	I	A	I
Mise à jour pipeline NLP	I	R/A	C	C	I	C
Monitoring / alerting	R	C	C	I	I	I
Documentation / MCO	C	R/A	C	C	C	I

3.2 Lecture claire de la matrice

- **Data Engineer** = rôle central du projet (pilotage technique global)
- **Cloud/Infra** = garant de la sécurité, du réseau, des accès et de la haute disponibilité
- **N1** = première réponse rapide aux incidents

- **Product Owner** = pilote priorisation + communication
- **DPO/Security** = responsable des impacts RGPD
- **Analystes** = utilisateurs finaux qui détectent les anomalies métier visibles

Cette structure est fidèle aux environnements professionnels modernes et prépare ton architecture à une évolution future (scalabilité, multi-services, nouvelles fonctionnalités).

Conclusion

La mise en place d'une chaîne d'escalade structurée et d'une matrice RACI complète n'est pas un simple exercice théorique. C'est une condition indispensable pour assurer la maturité opérationnelle du projet Amazon Review Analysis, surtout dans un contexte de production où interviennent plusieurs composants critiques et technologies complémentaires.

En clarifiant les rôles, en anticipant les incidents possibles et en donnant un cadre à l'action collective, on garantit une exploitation fluide, une meilleure collaboration entre équipes et une capacité à réagir rapidement en cas de problème. Ce modèle d'organisation, inspiré des meilleures pratiques Amazon et La Poste, renforce la fiabilité du pipeline, sécurise les données et améliore l'expérience des utilisateurs finaux.

Grâce à cette structure, le système est prêt à fonctionner durablement, à évoluer sereinement et à être opéré par différents profils sans perte de qualité ni de continuité.