

Budget et Risques

Projet Amazon Reviews :

Analyse et classification des avis Clients

BONNAT Jonathan

Data Engineer - B2B-LP-DESFL

INTRODUCTION GENERALE

Ce document présente le suivi budgétaire et le pilotage des risques liés au développement du système d'analyse et de catégorisation des avis clients d'Amazon. Dans le cadre de cette certification, je suis seul à réaliser l'ensemble des livrables. Toutefois, le budget est estimé comme s'il s'agissait d'un projet mené par une petite équipe, afin de refléter ce que coûterait réellement ce type de réalisation dans un contexte professionnel, et l'évolution que cela pourrait avoir.

Les charges exprimées en jours-homme proviennent du plan de projet et servent de base à l'estimation du coût de la main-d'œuvre. Les tarifs appliqués (TJM) sont des valeurs réalistes du marché français pour les rôles concernés (Data Engineer, Ops, Fullstack, CP, PO). L'objectif n'est pas de produire une simulation parfaite, mais d'obtenir une vision cohérente et crédible de l'effort nécessaire.

Le budget technique (cloud, stockage, exécution des pipelines) reste limité dans le cadre de ce projet, mais est extrapolé et intégré pour donner une image complète du coût de fonctionnement. L'ensemble du document vise à offrir un suivi simple : charges prévues, consommées, risques, actions et points d'attention.

ORGANISATION DE L'ÉQUIPE & CHARGES PRÉVUES

RÔLES PROJET

Dans le cadre d'une mise à l'échelle du projet, les rôles identifiés dans le Plan de Projet sont les suivants :

- Chef de projet / Scrum Master : Pilotage, coordination, suivi budget/risques, animation agile.
- Product Owner : Définition de la valeur métier, priorisation des besoins, validation des catégories d'avis.
- Data Engineer : Pipeline Airflow, extraction PostgreSQL, S3, scoring Zero-Shot, industrialisation.
- Développeur Fullstack : API FastAPI, intégration avec le front, tests d'exposition des données enrichies.
- Ops / SRE : Conteneurisation, orchestrations, monitoring, déploiement, supervision.

CHARGES PRÉVUES PAR RÔLE

(Issues du Plan de Projet sur les périodes S1 → S12)

RÔLE	CHARGE	DÉTAIL
Chef de Projet	10 JH	Pilotage, planning, coordination, reporting
Product Owner	6 JH	Cadrage fonctionnel, priorisation, validation métier
Data Engineer	20 JH	Pipeline ETL, scoring, DWH S3, tests
Développeur Fullstack	8 JH	Fast API, intégration front
OPS	6 JH	Conteneurs, déploiement, monitoring
Total	50 JH	Charge projet estimée

TARIFS JOURNALIERS MOYENS (TJM) ET ESTIMATION DE MAIN D'OEUVRE

(Basés sur les prix du marché français en 2024 → profils expérimentés, source Malt + Hays)

RÔLE	TJM	JH	COÛT
Chef de Projet	600 €	10	6000 €
Product Owner	550 €	6	3300 €
Data Engineer	600 €	20	12 000 €
Développeur Fullstack	550 €	8	4 400 €
OPS	650 €	6	3 900 €
Total	50 JH	-	29 600 €

BUDGET DE L'INFRASTRUCTURE + OUTILS

Ce budget est défini en gardant les mêmes outils que pour la première version (Airflow, Postgre, Mongo, Datalake S3, Fast API), avec un volume de données plus important, et grandissant dans le temps. Pour pallier à l'augmentation des données, l'infrastructure intégrera un instance GPU dédiée.

COÛTS SERVICES CLOUD

Nous avons décidé d'utiliser l'environnement AWS avec les machines virtuelles EC2, pour rester dans les outils Amazon et avoir une architecture scalable. Si besoin, nous pourrons changer de taille de serveur au cours du projet, ou ajouter d'autres machines pour travailler en architecture distribuée

SERVICE	CONFIG	JUSTIFICATION	COÛT MENSUEL	COMMENTAIRE
EC2 GPU	g4dn.xlarge	plus petite instance GPU, bon rapport performance / prix	380 €	pour l'inférence zéro shot
EC2 Airflow	2 x t3.medium	les 2 vCPU par machine permettront à Airflow de tourner correctement	80 €	pour Scheduler + Worker
EC2 API	t3.micro	plus petite instance CPU, suffisant pour notre API	15 €	pour Fast API
RDS Postgre	db.t3.micro	configurable jusqu'à 100 Go de stockage	15 €	suffisant pour les avis
MongoDB Atlas	M0	utilisation du service gratuit	0 €	Archivage léger
Amazon S3	10 Go	Pour le stockage des fichiers parquet	5 €	pour le Data Lake

Total mensuel : 495 €

PROJECTIONS BUDGÉTAIRES SUR 12 MOIS (RUN)

Hypothèses réalistes :

- Volume d'avis multiplié par 4 sur 1 an
- Taille des fichiers parquet multipliée par 2
- Airflow peut monter sur t3.large au bout de 6 mois
- MongoDB peut passer sur M2
- API peut passer en t3.small selon trafic

SERVICE	MONTANT ANNUEL
EC2 GPU	4560 €
EC2 Airflow	960 €
EC2 API	180 €
RDS Postgre	240 €
MongoDB Atlas	54 €
Amazon S3	60 €

Les charges annuelles seraient donc situées entre 6000 et 6100 €

A noter : Si jamais le besoin se fait sentir, nous pourrions rajouter un DWH type Snowflake, mais une petite config coûterait déjà plus de 14 000€ à l'année.

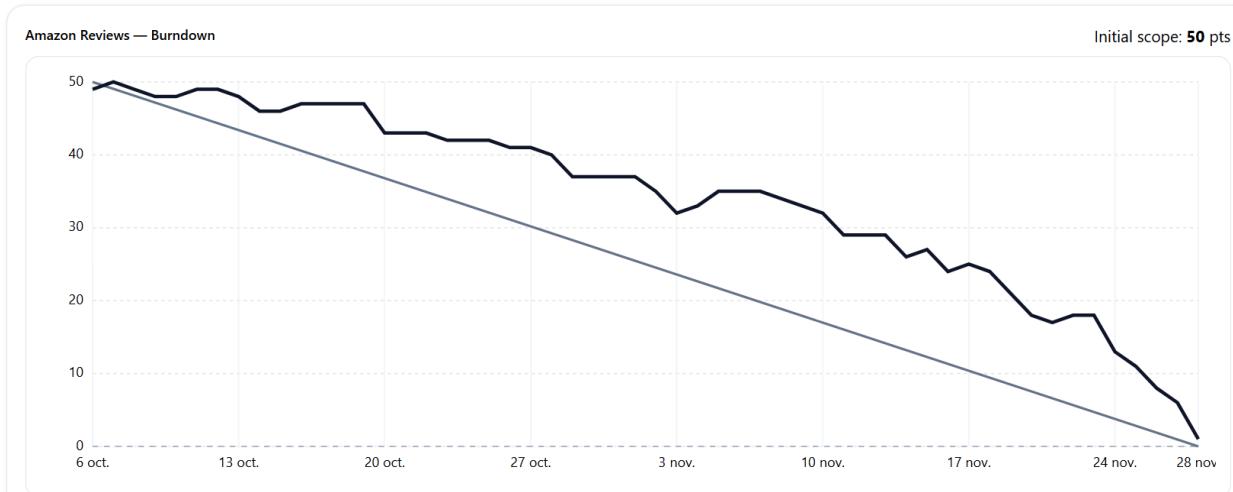
SUIVI BUDGÉTAIRE

SYNTHÈSE DU PRÉVU ET DU CONSOMMÉ

RÔLE	PRÉVU (JH)	CONSOMMÉ (JH)	ECART	JUSTIFICATION
Chef de Projet	10	11	+1 (+10%)	Coordination entre les équipes + rôle cumulé de Scrum Master
Product Owner	6	5	-1 (-16%)	Le périmètre fonctionnel est resté stable
Data Engineer	20	23	+3 (+15%)	gestion des dépendances, optimisation du scoring, benchmark de modèles
Dev FullStack	8	7	-1 (-12%)	API simple, pas de difficultés rencontrées
Ops	6	7	+1 (+17%)	Configuration de la machine GPU et mise en place de Airflow avec un worker dédié

L'écart total est de +3 jours, soit +6% par rapport à la charge indiquée. Le surcoût est uniquement lié à la main-d'œuvre, car l'infrastructure n'a pas subi de dépassement.

BURNDOWN CHART SIMPLIFIE



Le burndown chart montre une progression globalement alignée sur la trajectoire, avec quelques événements correspondants aux différentes phases du projet :

- Début octobre : bonne avancée, clarification du périmètre, mise en place des environnements de travail
- Mi octobre : ralentissement car tâches complexes (configuration Airflow, utilisation du modèle de machine learning, dépendances non prévues, ...)
- Fin octobre : obstacles levés, premier pipeline simplifié qui fonctionne
- Début Novembre : nouveau ralentissement liés à l'intégration du scoring et aux tests ETL
- Mi novembre : tous les problèmes sont résolus, et l'équipe peut mettre les derniers efforts pour terminer toutes les petites tâches restantes jusqu'à la fin

ANALYSE D'ENSEMBLE

Le suivi montre que le projet a été bien contrôlé. Les dépassements sont concentrés sur les tâches les plus techniques (pipeline NLP + GPU + Airflow), ce qui est totalement cohérent avec le projet et ses exécutants. Les 3 jours de dépassement pour la main d'œuvre sont acceptables pour un projet à cette échelle. Aucun risque majeur n'a été déclenché et la roadmap a été respectée sans révision du périmètre.

REGISTRE DES RISQUES

TABLEAU DES RISQUES PRINCIPAUX

RISQUE	PROBA	IMPACT	NIVEAU	PLAN DE MITIGATION
Instabilité du modèle Zero-Shot	Faible	Elevé	Modéré	Mise en cache, exécution en batchs, optimisation du modèle
Défaillance du pipeline Airflow (erreurs DAG, dépendances)	Moyen	Elevé	Important	Activer les retry, surveiller les logs, tester unitairement les tasks
Dépendance aux services cloud (S3, MongoDB Atlas, EC2)	Faible	Elevé	Modéré	Sauvegardes automatiques des bases, snapshots S3

API indisponible ou lente sous charge	Faible	Moyen	Modéré	Activer l'auto-scaling, surveiller les montées en charge
Mauvaise interprétation métier des résultats	Moyen	Faible	Modéré	Documentation, exemples, feedback, itérations
Demandes RGPD (suppression de données utilisateurs)	Moyen	Elevé	Important	Purges dans Postgre et Mongo, et anonymisation / pseudonymisation
Rôles / besoins / retours flous	Moyen	Moyen	Modéré	Mise à jour du RACI, comités projets, validations PO / métier

Globalement, le projet contient trois types de risques :

- Les risques techniques (les plus présents) maîtrisés grâce à une architecture simple et réfléchie, une documentation fournie et une surveillance accrue au quotidien
- Les risques organisationnels (qui dépendent du nombre d'acteurs dans le projet) maîtrisés grâce au RACI, aux différentes réunions et à la stabilité du périmètre
- Les risques réglementaires (qui dépendent des institutions) qui sont maîtrisés par la surveillance de nouveaux besoins réglementaires et par l'application des mesures déjà en place (purge, anonymisation)

RISQUES RÉSIDUELS

Une fois les mesures de mitigation appliquées, il reste trois risques significatifs :

- Latence variable du modèle NLP si le volume d'avis croît fortement
- Airflow en tant que point de défaillance unique (SPOF) si mal configuré (scheduler ou worker isolés)
- Évolution des catégories métier (modification du NLP)

Ces risques ne remettent pas en cause la production mais doivent être suivis dans le tableau de risques hebdomadaire.

SYNTHÈSE DU REGISTRE DES RISQUES

Aucun risque critique n'empêche ou ne remet en question la mise en production et le bon fonctionnement du service. Les risques importants concernent principalement le pipeline (Airflow), la conformité RGPD et la performance du modèle Zero-Shot. Tous les risques disposent d'un plan de mitigation clair et d'un niveau résiduel acceptable.

CLAUSES DE CONFORMITÉ ET CONTRACTUELLES

ENGAGEMENTS DE DÉLAIS ET DE LIVRABLES

- Livraison du pipeline ETL/NLP : fin semaine 6
- Livraison de l'API d'exposition : fin semaine 10
- Mise en production complète : fin semaine 10
- Documentation utilisateur & support : semaine 11
- Documentation de maintenance : semaine 12

Clause définie : Tout retard supérieur à 10 % de la durée prévue doit être remonté au comité projet et validé par le PO. Une révision du périmètre ou un réalignement de planning peut être effectué si le retard menace la livraison globale du projet.

ENGAGEMENTS DE QUALITÉ

Le projet doit respecter les indicateurs suivants :

Qualité technique :

- Temps de réponse de l'API : < 300 ms pour une requête standard
- Disponibilité cible du service : 99 %
- Pipeline ETL : tolérance maximale de rejet < 3 %
- Robustesse : mécanismes de retry / monitoring obligatoires

Qualité fonctionnelle :

- Catégorisation NLP cohérente
- Mise à disposition du "most relevant review" pour chaque produit

- Données rafraîchies quotidiennement

Clause associée : Tout bug bloquant doit être traité sous 48h. Les anomalies mineures doivent être corrigées dans la prochaine itération du projet.

ENGAGEMENTS D'ÉVOLUTIVITÉ

Puisque le volume d'avis peut croître rapidement, le système doit permettre :

- la montée en charge sur le worker GPU
- l'augmentation de la fréquence d'exécution des pipelines
- l'ajout de nouveaux produits
- une éventuelle migration future vers Snowflake ou Redshift

Clause associée : Le système doit être dimensionné pour traiter au moins 4x le volume initial d'avis sans refonte majeure de l'architecture.

ENGAGEMENTS DE GOUVERNANCE

Pour éviter les blocages, les responsabilités décisionnelles doivent être définies :

- le PO arbitre les priorités fonctionnelles
- le PM valide les jalons et l'avancement
- le Data Engineer est responsable des pipelines
- le Dev Fullstack est propriétaire de l'API et du front
- Le Ops/SRE garantit la disponibilité et la sécurité

Clause associée : Toute décision impactant le délai ou le périmètre doit être documentée et validée en comité projet.

CONFORMITÉ RÉGLEMENTAIRE

RGPD :

- Données personnelles pseudonymisées
- Procédure de suppression (droit à l'oubli)

- Pas de logs contenant des informations sensibles
- Conservation limitée au nécessaire métier

Sécurité :

- Accès S3 en lecture seule pour l'API
- Pas d'accès public aux buckets
- Mots de passe chiffrés
- Architecture respectant le principe du moindre privilège (RBAC du Bloc 3)

Accessibilité :

- textes lisibles (AA)
- contrastes corrects
- navigation clavier possible dans l'interface
- images avec descriptions textuelles
- formation accessible (version PDF + version texte)

Clause associée : Toute demande de suppression ou d'accès aux données personnelles doit être traitée sous 30 jours, conformément à la loi.

RESTRICTIONS D'USAGE

- Le système ne doit pas être utilisé pour prendre des décisions automatisées ayant un impact légal ou significatif sur les utilisateurs.
- L'analyse NLP est considérée comme assistive (aide au tri), pas comme un moteur décisionnel.

Clause associée : Les résultats du modèle ne doivent pas être utilisés comme décision automatique sans validation humaine, mais seulement comme des indicateurs pour l'aide à la décision.

SYNTHÈSE CONTRACTUELLE

L'ensemble des clauses présentées établit un cadre clair garantissant :

- une livraison maîtrisée
- un niveau de qualité cohérent avec un système interne Amazon
- une gouvernance structurée

- et une conformité réglementaire solide.

Le périmètre reste réaliste et proportionné au projet : aucune clause n'est excessive, aucune exigence est non réalisable, dans le contexte de la certification comme dans le futur.

CONCLUSION

Le pilotage global du projet montre un déroulement maîtrisé, malgré quelques écarts naturels liés à la complexité du pipeline NLP et de son intégration dans un environnement de production.

La charge totale consommée atteint 53 JH pour 50 JH prévus, soit un écart de +6 %, parfaitement acceptable dans le cadre d'un projet data incluant Airflow, un worker GPU, et une API exposée en production.

Du point de vue budgétaire, l'infrastructure reste conforme aux attentes : les coûts cloud sont restés stables et ne présentent aucune dérive. Les dépassements sont exclusivement liés à la main-d'œuvre, principalement sur le volet Data Engineering, ce qui reflète la réalité terrain.

Le suivi opérationnel montre également que le projet a conservé un bon niveau de visibilité :

- les risques ont été identifiés tôt,
- les mesures de mitigation ont été appliquées,
- aucun risque critique n'a été déclenché,
- et la roadmap a pu être tenue sans réduction du périmètre.

Enfin, le niveau de qualité atteint correspond aux objectifs fixés : le pipeline est fonctionnel, l'API répond aux usages attendus, et la documentation de mise en production permet un transfert fluide à l'exploitation.

En résumé, le projet atteint ses objectifs dans de bonnes conditions, avec une légère dérive budgétaire maîtrisée et un pilotage cohérent. Les enseignements tirés permettront d'aborder les futures itérations avec une meilleure anticipation des charges liées au NLP et à l'orchestration.