Tableau de bord budgétaire et des risques

Projet : Amazon Review Analysis

Auteur : Ismael SYLLA

Table des matières

1. Introduction et méthodologie

Le projet a été mené avec la méthodologie Agile Scrum, avec 8 sprints (1 sprint c'est 2 semaines).

1.1. Gestion du backlog

Le product owner est responsable du backlog produit dans Azure DevOps.
Le backlog est composé de User Stories fonctionnelles, de tâches techniques, de tâches de qualité et sécurité et de tâches de documentation et formation.
Chaque User Story possède une priorité (haute, moyenne, faible), une estimation en point d'efforts, des critères d'acceptation et un lien avec une des phases du projet (P1 à P5).
Le backlog est mis à jour à chaque Sprint Review, après chaque retour des Business Analysts ou en cas de nouveau risque ou besoin métier remonté.
Pour rappel voici la répartition des tâches entre les différents acteurs du projet :

- PO : priorisation et validation fonctionnelle
- Scrum Master : organisation, rituel, fluidité
- Data Engineers : ETL et data pipeline
- Data Scientist : algorithme zero-shot et algorithme de pondération
- DevOps : déploiement et automatisation
- Formateur : formation des utilisateurs
- Support : stabilisation après mise en prod

1.2. Backlog

Total points d'efforts : 92 points.

ID	User Story / Tâche	Priorité	Estimation (pts)	Responsable
US01	Construire le flux qui récupère les données Postgresql pour les déposer dans le Data lake.	Haute	8	Data Engineer
US02	Mettre en place le bucket S3 et les règles de sécurité	Haute	5	DevOps
US03	Nettoyer et normaliser les données clients	Haute	8	Data Engineer
US04	Construire le pipeline ETL automatisé (Airflow)	Haute	21	Data Engineer + DevOps
US05	Mettre en place le modèle ML zero-shot pour la classification	Haute	8	Data Scientist
US06	Implémenter le scoring et améliorer les models	Moyenne	8	Data Scientist
US07	Créer le modèle de données Snowflake	Haute	5	Data Architect
US08	Développer de dashboards Streamlit sur Snowflake	Moyenne	8	Data Engineer
US09	Modifier le backend de l'application Streamlit pour intégrer la récupération des commentaires les plus pertinents en fonction des buyer_id	Moyenne	8	Data Engineer
US10	Mettre en place les règles RGPD et sécurité	Haute	5	Securité/ Compliance
US11	Rédiger documentation technique & supports de formation	Moyenne	8	Data Engineer + PO

1.3. Planification des sprints

La répartition des US dans les sprints s'est faite ainsi :

Sprint	Objectif principal	US intégrées	Points
Sprint 1	Mise en place accès & sécurité data	- US02 : Bucket S3 et sécurité - US01 : Contruction flux Postgresql – S3 - US10 : RGPD & sécurité	18
Sprint 2	Préparation et qualité des données	- US03 : Nettoyage & normalisation	8
Sprint 3	Architecture data & modélisation	- US07 : Modèle de données Snowflake	5
Sprint 4	Pipeline & automatisation	- US04 : Pipeline ETL Airflow	21
Sprint 5	Maching Leaning (Zero-shot)	- US05 : Modèle ML zero-shot	8
Sprint 6	Maching Leaning + amélioration	- US06 : Scoring qualité	8
Sprint 7	Visualisation & applicatif	- US08 : Dashboards Streamlit	8
Sprint 8	Application & livrables projet	- US09 : App Streamlit - US11 : Documentation	16

1.4. Burnodown chart

Chaque acteur du projet enregistre le temps qu'il passe sur le projet dans un outil interne à l'entreprise, où il précise le projet, la phase sur laquelle il a travaillé et le temps qu'il a passé dessus par jour.

Un suivi est ensuite fait pour pouvoir comparer par rapport au temps estimé combien l'acteur a passé en temps réel.

Le burndown chart permet de visualiser le temps restant (JH), le travail réalisé et la dérive potentielle. Ceci permet aux décideurs de voir où nous en sommes dans le projet, est-ce qu'il faut prévoir plus de charge ?

Le tableau ci-dessous couvre uniquement le périmètre des activités réalisées par l'équipe produit en sprint. Pour rappel la charge totale du projet en incluant tous les acteurs s'èlève à 250 JH (incluant le sponsor, le métier…).

Voici les règles de calculs (suite Fibonacci) :
5 points d'efforts sur une US représentent ½ sprint.
8 points d'efforts sur une US représentent 1 sprint.
13 points d'efforts c'est 1 sprint + ½.
21 points d'efforts c'est 2 sprints et ½.

Sprint	JH prévus	JH réels	Ecart	Ecart %	Commentaire
Sprint 1	22,5	24	1,5	6,67%	Mise en place + accès complexes
Sprint 2	10	8	-2	-20,00%	Données plus propres que prévues
Sprint 3	6,25	6	-0,25	-4,00%	
Sprint 4	26,25	30	3,75	14,29%	Plus complexe que prévu
Sprint 5	10	10	0	0,00%	Conforme
Sprint 6	10	12	2	20,00%	Amélioration du ML a pris plus de temps que prevu
Sprint 7	10	9	-1	-10,00%	
Sprint 8	20	20	0	0,00%	Conforme
TOTAL	115	119	4	**3,48%**	**Dépassement de 3,48%**

2. Coûts ressources humaines

2.1. Coût total par rôle & global

Les chiffres JTM sont des hypothèses basées sur le marché actuel en France et peuvent varier selon le statut (salarié ou consultant ou freelance).

Cette estimation prend en compte des JTM variables selon le rôle et donne un budget ressources humaines globale d'environ 114 350, 00 euros.

Groupes	Rôle	JH (≈)	Hypothèse TJM (€)	Coût estimé	Coût estimé par groupe
Management	Sponsor (interventions ponctuelles)	5 JH	450 € (profil senior direction)	2 250,00 €	9 150,00 €
	Business Analysts	23 JH	300 €	6 900,00 €	
Delivery	Product Owner (PO)	40 JH	350 €	14 000,00 €	99 200,00 €
	Scrum Master	35 JH	400 €	14 000,00 €	
	Data Architect	15 JH	700 €	10 500,00 €	
	Tech Lead Data	25 JH	600 €	15 000,00 €	
	Data Engineers (x2)	50 JH total	500 €	25 000,00 €	
	Data Scientist	15 JH	550 €	8 250,00 €	
	DevOps Engineer	15 JH	550 €	8 250,00 €	
	Security / Compliance Officer	7 JH	600 €	4 200,00 €	
Support	Formateur	10 JH	300 €	3 000,00 €	6 000,00 €
	Support Client (post-prod)	10 JH	300 €	3 000,00 €	
	TOTAL estimé (RH)	~ 250 JH	—	114 350,00 €	114 350,00 €

3. Coûts Infrastructure & Outils Cloud

3.1. Description de l'infra utilisée

Pour rappel, voici l'architecture et composants du projet :

- Stockage fichiers, données brutes dans Amazon S3.
- Stockage des données rejetées et des logs dans MongoDB (self-hosted).
- Stockage des tables après nettoyage & traitement → Snowflake.
- Traitements ETL, ML et scoring avec Docker + Python (notebooks).
- Application de visualisation / dashboard → Streamlit
- Orchestration avec Airflow
- Eventuels conteneurs, déploiement, sauvegardes, logs, monitoring, etc.

3.2. Hypothèses de volume de données + croissance

Le volume initial de données stockées est de 1 To (avis brutes et transformés). Nous estimons la croissance de volume de 10% par mois. L'ETL se lance 1 fois par jour et les dashboards sont mis à jour quotidiennement.

3.3. Estimation du stockage, des ressources consommées et coûts

Stockage S3 : pour 1 To c'est 25 à 30 euros par mois, avec une croissance de 10% par mois, après 6 mois d'utilisation cela représente 45 à 50 euros par mois. Donc un total de 550 à 600 euros / an à peu prés pour la première année.

A cela peuvent s'ajouter des coûts liés aux requetes (PUT, GET) si l'update est faite souvent, mais étant donné que les batchs ne tourne qu'une fois par jour, les coûts vont rester faibles.

MongoDB : Supposons qu'on utilise un cluster pour stocker les logs/rejets avec des volumes modestes avec quelques dizaines de GB, si le cluster tourne 24/7 c'est 0,08$/h, par mois ça serait entre 55 à 60 euros. Selon l'usage, nous allons partir sur 60 euros/mois à peu prés.

Snowflake : Snowflake facture au stockage et compute.
Pour 1 à 2 TB de stockage nous pouvons être aux alentours de 20 à 40 euros/mois.
Pour les requêtes on va être autours de 1500 euros par mois.
Le coût total de Snowflake serait estimé entre 1500 à 2000 euros par mois (soit entre 18 000 – 25 000 euros/an).
Compute, orchestration, serveurs et containers : dans ce projet, nous aurons besoin de plusieurs serveurs (Airflow, exécution ETL, hébergement de l'application Streamlit et API. On va estimer le coût à peu près à 100 euros/mois

Poste / composant	Coût estimé / mois	Coût estimé / an
Stockage S3	~ 50 €	~600 €
MongoDB (NoSQL logs)	~ 60 €	~ 720 €
Snowflake (stockage + compute)	~ 1500 €	~ 18 000€
Hébergement / compute / orchestration / front	~ 100 €	~ 1 200 €
TOTAL Infra estimée	~ 1 750 €/mois	~ 20 000 €/an

Nous arrivons à une estimation de **20 000** euros par an.

3.4. Suivi budgétaire

Coût ressources humaines et infrastructure estimés :

- Ressources humaines ~ 114 350 euros
- Ressources infra ~ 10 000 euros (/6mois)

Budget total estimé : 124 350 euros

Groupes	Rôle	Coût estimé	Coût réel	Ecart	Ecart %	Coût estimé par groupe	Coût réel par groupe
Management	Sponsor (interventions ponctuelles)	2 250,00 €	2 250,00 €	0,00 €	0,00%	9 150,00 €	9 150,00 €
	Business Analysts	6 900,00 €	6 900,00 €	0,00 €	0,00%		
Delivery	Product Owner (PO)	14 000,00 €	14 000,00 €	0,00 €	0,00%	99 200,00 €	102 350,00 €
	Scrum Master	14 000,00 €	14 000,00 €	0,00 €	0,00%		
	Data Architect	10 500,00 €	10 500,00 €	0,00 €	0,00%		
	Tech Lead Data	15 000,00 €	15 000,00 €	0,00 €	0,00%		
	Data Engineers (x2)	25 000,00 €	26 500,00 €	1 500,00 €	6,00%		
	Data Scientist	8 250,00 €	9 350,00 €	1 100,00 €	13,33%		
	DevOps Engineer	8 250,00 €	8 800,00 €	550,00 €	6,67%		
	Security / Compliance Officer	4 200,00 €	4 200,00 €	0,00 €			
Support	Formateur	3 000,00 €	3 000,00 €	0,00 €	0,00%	6 000,00 €	6 000,00 €
	Support Client (post-prod)	3 000,00 €	3 000,00 €	0,00 €	0,00%		
	TOTAL estimé (RH)	114 350,00	117 500,00	3 150,00 €	2,75%	114 350,00 €	117 500,00 €

Ressource	Budget prévu	Budget réel
Ressources humaines	114 350 €	117 500 €
Infra (6 mois projet)	10 000 €	11 500 €
TOTAL projet	124 350 €	129 000 €
Écart	**4 650 €**	**+ 3,7 %**

4. Suivi budgétaire & indicateurs (KPI)

Respect du planning : dans cet indicateur nous regardons combien de sprints ont été livrés à la date prévue au départ. Nous avons fixé un objectif de 90%, lorsque l'objectif est atteint cela veut dire que l'équipe data estime bien son travail et qu'il n'y a pas trop de retards ou bloages. Dans le cas inverse, il faut chercher à comprendre la problématique et proposer une solution, cela peut être lié à plusieurs factures : développements plus complexes que prévus, il n'y a pas de données propres, manques de ressources…

Ecart budget RH : indicateur pour comparer le budget des ressources humaines s'il est respecté, on compare l'estimation et le coût réel.

Coût infra / mois : cet indicateur nous permet de suivre la facturation mensuelle totale de l'infrastructure et de voir s'il n'y a pas de dépassements importants.

Taux de complétion des sprints : cet indicateur nous permet de suivre les user stories/tâches prévues dans un sprint, est-ce qu'elles ont toutes été réalisées ? On regarde le nombre de tâches réalisées / le nombre prévues.

Qualité de données : cet indicateur nous permet de suivre la qualité des données après nettoyage et de regarder s'il y a des valeurs manquantes, des doublons et si les règles sont respectées. C'est important car on a un risque que les algorithmes soient faussés et que les décisions métiers soient mauvaises.

Nombre d'incidents critiques : cet indicateur nous permet de suivre les incidents critiques, par exemple l'ETL qui tombe en panne, Streamlit qui n'est plus accessible…

Taux d'adoption métier : est-ce que les Business Analysts utilisent la solution ?
On fixe un seuil de 70%, s'il y a moins de 70% d'utilisation de la solution, on doit commencer à se poser des questions, est-ce que finalement la solution répond bien au besoin ? Est-ce que les utilisateurs ont été bien formés ?

Indicateur	Objectif	Résultat	Statut
Respect du planning (sprints livrés à temps)	> = 90 %	87,5 % (7/8 sprints)	Alerte
Écart budget RH prévu vs réel	< 5 %	2,75 %	Alerte
Coût infra / mois	< 2k €/mois	~ 1 900 euros	
Taux de complétion des sprints	> 95 %	93%	Alerte

Qualité des données (après nettoyage)	>= 98 %	97,50%	Alerte
Nombre d'incidents critiques (pipeline / production)	< 5	4	
Taux d'adoption métier (utilisateurs)	>=70 %	78 %	

5. Registre des risques

Le registre est mis à jour à chaque sprint et revue en comité de projet/COPIL avec un suivi des actions

ID	Risque	Probabilité	Impact	Niveau	Responsable	Plan d'action / mitigation
R1	Retard réception données ou volumétrie plus grande que prévue	Moyen	Fort	Élevé	Project Manager / Data Architect	Prévoir données tests + buffer marge volume
R2	Coût infra ou stockage plus élevé que prévu (data volume, usage intensif)	Moyen	Fort	Modéré	DevOps / Tech Lead	Suivi conso mensuelle + optimisation stockage / nettoyage / archival
R3	Performance / coût élevée de Snowflake (compute, requêtes complexes)	Fort	Fort	Élevé	Data Architect / Tech Lead	Optimisation requêtes, planification jobs, batch vs on-demand
R4	Modèle IA (zero-shot + scoring) peu performant → mauvaise qualité de résultats	Moyen	Fort	Élevé	Data Scientist	Tests, validation, itérations, feedback métier, retrain
R5	Problème sécurité / conformité (RGPD, accès données)	Faible	Très fort	Critique	Security / Compliance Officer	Audit, chiffrement, contrôles accès, documentation, conformité RGPD
R6	Sous-estimation des charges RH (tâches plus longues)	Moyen	Fort	Élevé	Project Manager	Buffer marge + veille sur la charge + ajustement en COPIL

R7	Faible adoption métier (dashboards non utilisés)	Moyen	Moyen	Modéré	Change Manager / Product Owner	Formation, communication, accompagnement, feedback utilisateur
R8	Dette technique ou maintenance infra complexe	Faible	Moyen	Modéré	DevOps / Tech Lead	Documentation, standardisation, automatisation, monitoring

6. Clauses contractuelles & contraintes

Voici les éléments qui sont couverts :

Délais de livraisons :

La livraison finale du projet est prévue à la fin du sprint 8, conformément au planning validé au lancement du projet.

Tout retard devra être justifié et fera l'objet d'une replanification validée par les parties prenantes.

Résultat : la solution a été livrée avec un retard d'une semaine.

Qualité des données :

Le projet devra garantir un taux de qualité minimal de 98 % après traitement et nettoyage des données.

La qualité est mesurée à partir de règles de contrôle automatisées (détection de doublons et suppression, vérifications des formats, vérifications de valeurs manquantes).

Si la qualité de données est inférieure au seuil fixé, une correction devra être effectuée avant la mise en production.

Sécurité & conformité / RGPD :

Les données utilisées dans le cadre de ce projet incluent les identifiants des acheteurs, le projet devra donc respecter les consignes suivantes :

- Anonymisation des données sensibles.
- Respect des principes du RGPD (minimisation, durée de conservation, droit à l'effacement)
- Traçabilité des connexions et des accès (logs)
- Gestion des rôles et permissions (RBAC)
- Accès limité aux seules personnes habilitées

Une attention particulière sera portée à la sécurisation des données stockées et en transit.

Documentation :

Le projet devra être livré avec une documentation complète comprenant :

- Architecture globale de la solution
- Description du pipeline de données
- Modèle de données
- Description des outils et technologies utilisées
- Procédure de déploiement et de mise à jour
- Documentation utilisateur :
    - Guide d'utilisation
    - Explication des KPI
    - Support de formation

Disponibilité / SLA:
Après mise en production, la solution devra respecter les critères suivants : disponibilité 99%, temps maximum d'indisponibilité par mois est fixé à 1 heure. Le temps de chargement des tableaux de bord est inférieur à 5 secondes et le délai de mise à jour des données est quotidien. Tout écart ou retard devra être signalé et investigué.

Maintenance & support :
Une phase de maintenance post-production de 1 mois sera assurée et inclura :

- Un échange entre PO, Support client et métier tous les jours de 15 mins pour échanger sur les problèmes rencontrés et bugs remontés.
- Correction des anomalies et bugs (en priorités sur les autres sujets).
- Ajustements selon le retour du métier.

Après le mois d'accompagnement, le support métier sera assuré par les agents du Support (niveau 1) et des demandés seront remontés à l'équipe Data (niveau 2).

Budget & ressources allouées :
Accord sur ressources, JH, budget infra, marges.
Toute modification de périmètre devra être accompagnée d'un ajustement du budget.

Mesures correctives en cas de non-conformité :
En cas de non-respect des engagements (délais, qualité, sécurité, documentation ou performance), les mesures suivantes pourront être appliquées :

- Révision du planning
- Renforcement ou réallocation des ressources
- Mise en place d'un plan d'actions correctives
- Réduction ou révision du périmètre fonctionnel

Ces mesures visent à garantir la conformité finale de la solution aux exigences définies.

7. Revue finale & recommandations

Points de vigilance et risques :

- Le coût infra peut fortement augmenter si le volume de données explose, si le nombre d'utilisateurs monte, ou si les traitements et requêtes sont lourds.
- Les estimations de TJM / JH peuvent varier selon le profil (senior, junior), le statut (salarié, consultant, freelance) et les charges sociales.
- Le risque qualité des données et performance des algorithmes ML — prévoir des phases tests / ajustements.
- L'adoption métier : si la solution n'est pas adoptée, le ROI du projet peut diminuer.

Recommandations :

Il faut surveiller l'infra régulièrement (stockage, requêtes, logs, usage) pour éviter des surprises sur les coûts. Il faut prévoir des jalons de revue pour évaluer les coûts, la charge, risques et ajuster le scope si besoin.

Il est important de bien documenté l'architecture, les flux, les responsabilités de chacun, les accès et conformités réglementaires