Mérick Zounon
Jeremy Peterson Dutoya
Jessy Durca

AWS Cloud Architecture

L'objectif de ce projet est de concevoir et de mettre en place une architecture cloud capable de traiter et d'analyser des données de réservations aériennes, afin de mieux comprendre les comportements d'achat des passagers et d'optimiser les stratégies de vente. Nous partons d'une hypothèse où la compagnie aérienne dispose de ses propres données (initialement issues d'un jeu de données Open Data sur Kaggle), mais gérées comme si elles provenaient directement de ses systèmes de réservation (clients desktop, mobile, etc.).
Le jeu de données fournit une vision détaillée des comportements et des préférences des clients lors de leurs réservations de vols. Il inclut le nombre de passagers par réservation et l'origine géographique, permettant d'identifier les tendances (par exemple, voyages individuels ou en groupe). Les informations sur les canaux de ventes, le délai entre l'achat et le départ, ainsi que la durée du séjour offrent un aperçu précis des habitudes d'achat et des comportement de consommation.

Les caractéristiques des vols (itinéraire, horaire, durée, jour de départ) permettent d'analyser les préférences temporelles et d'évaluer l'influence de chaque paramètre sur la décision de réservation.
Les services additionnelles proposés (bagages supplémentaires, sièges préférentiels, repas spécifiques) soulignent l'intérêt des clients pour des options personnalisées et leur disposition à payer pour améliorer leur expérience. Enfin, le statut final de la réservation (booking_complete) sert d'indicateur clé pour distinguer les réservations aboutis de celles abandonnées, offrant une compréhension des facteurs de conversion.
L'objectif global est d'extraire des insight permettant d'améliorer l'expérience client et de booster les performances marketing.

Nous souhaitons concevoir et déployer une architecture data sur AWS, afin de collecter, stocker et analyser les informations de réservations aérienne dans le but d'optimiser les stratégies marketing, d'améliorer l'expérience client et d'accroître le taux de conversion.
Pour répondre à ces défis, nous avons défini des user stories spécifiques à plusieurs profils métiers (Data analyst, Marketing, CIO). Ces user stories décrivent les besoins concrets de chaque rôle. L'architecture que nous vous proposons doit servir autant pour l'analyse approfondie (Data analyst) qu'au pilotage des actions marketing ciblées (Marketing), qu'a la prise de décision stratégique (CIO).

Ce rapport présente notre démarche pour ingérer les données, les stoker de façon fiable, les préparer pour l'analyse et la visualisation, puis les exploiter selon les besoins de chacun de ces trois types d'utilisateurs.

Collecte des données

Hypothèse : Nous jouons le rôle d'une compagnie aérienne qui possède ses propres données de réservations (qui correspondent au jeu de données Kaggle que l'on a choisit). Les données peuvent être produites en continu ou en Batch à intervalles réguliers. Les clients (desktop, mobile) génèrent des données (réservations, infos de navigation, etc) qui sont stockées dans une base de données interne.

Ingestion des données

1. Amazon S3

Un bucket Amazon S3 stocke toutes les données brutes (cdv, exports, etc) avant tout traitement ou ingestion plus approfondie.

2. Kinesis Data Firehose

Nous avons mis en place un flux streaming avec Kinesis Data Firehose. Il ingère les données avant de les envoyer vers un nouveau bucket S3 après transformation des données, ce qui permet de stocker des fichiers exploitables (format propre) en continu.

Nettoyage et organisation des données

1. AWS Glue

Les données brutes peuvent contenir des erreurs ou nécessiter une mise au même format. AWS Glue réalise les transformations ETL (extract, transform, load).

2. Crawler

Un crawler va ensuite analyser les données dans Amazon S3 (notre fichier cdv par exemple) pour détecter automatiquement le schéma de la base de données. Le schéma est ensuite enregistré dans Glue Catalog ce qui permet de facilité l'extraction, la transformation et le loading. Cela permet aux autres services (Athena, quicksight) de comprendre la structure des données.

Nettoyage et organisation des données

Amazon Athena se connecte aux données dans S3, via les métadonnées du Glue Catalog. Ce qui permet aux Data Analyst d'écrire des requîtes SQL. Les équipes peuvent alors agréger, croiser et rapporter ces données afin de créer des vues utiles, par exemple les réservations par canal de vente.

Visualisation avec Amazon QuickSight

Amazon QuickSight s'appuie sur Athena et le Glue Catalog pour réaliser des tableaux de bord (dashboard) et des visualisations.

Le data analyst peut alors explorer et extraire des données précises, générer des fichiers excel, pdfs et autres rapports.

Les équipes marketing, peuvent utiliser QuickSight pour consulter et suivre des indicateurs comme le taux de complétion (booking_complete), la répartition par sales_channel ou encore la demande pour les services additionnels. Les décideurs (CEO) peuvent accéder à des synthèses plus globales, des rapports annuels et des tableaux de bord stratégiques pour piloter l'activité.

Machine Learning avec Amazon SageMaker

Nous avons ajouté dans notre architecture un service Amazon Sage Maker pour créer des modèles prédictifs (exemple : la probabilité qu'un client finalise sa réservation, adoption d'un service supplémentaire, etc.)

Sage Maker peut tirer ses données d'entrainement depuis Amazon S3 ou via Athena, il peut ensuite renvoyer ses prédictions dans la pipeline ou vers QuickSight pour être incluses dans les tableaux de bord. Canvas est utiliser comme interface no-code/low-code pour les utilisateurs non expert. NoteBook est l'environnement Python pour la data science et enfin le service Train afin d'entrainer le model.

Rôles et accès

1.Data analyst

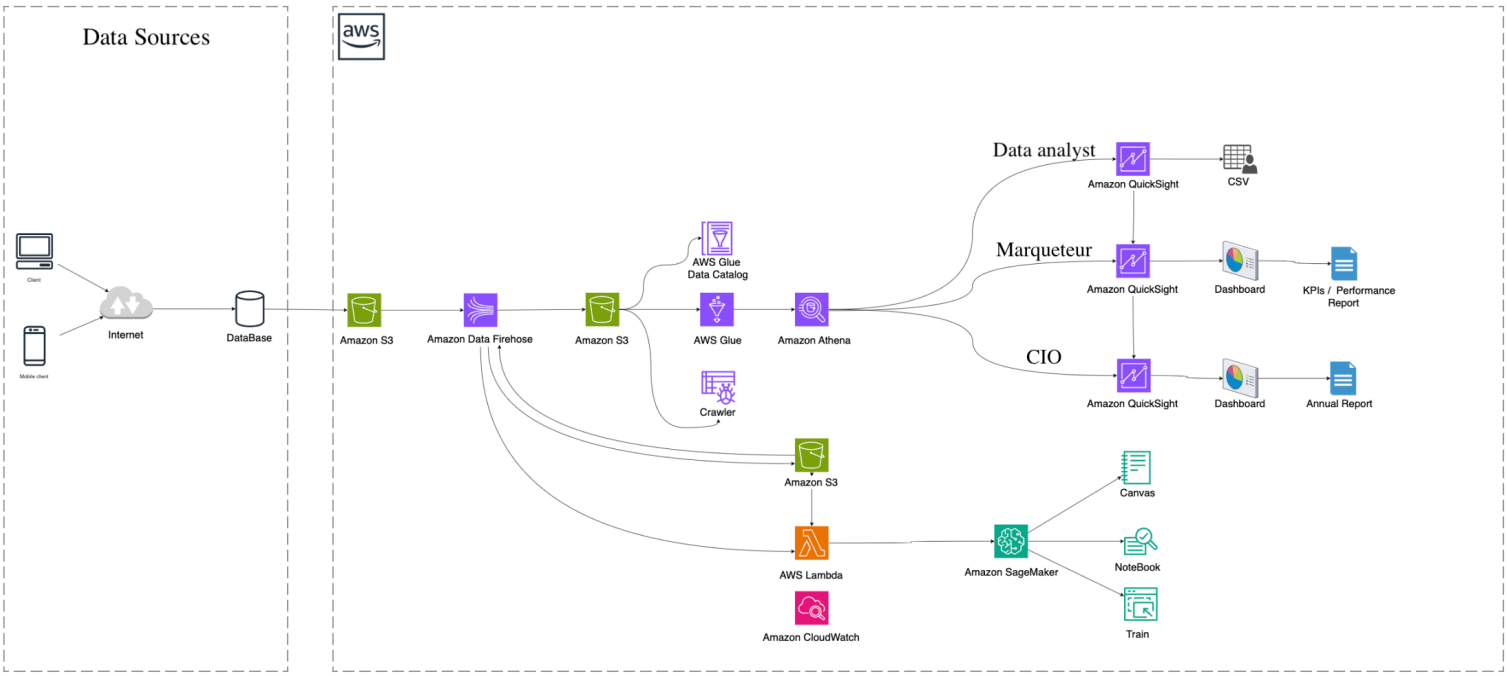
- Se connecte à Athena pour executer des requetes SQL
- Créer des rapports, bilans

2. Equipes Marketing

- Veulent connaitre les habitudes des voyageurs afin de leur proposer des solutions adaptes, des promotions, les préférences
- Accèdent à QuickSight pour visualiser les indicateurs clés, configurer des alertes ou des insight automatique

3. Direction / CIO

- Besoin d'une vue stratégique en temps quasi réel de son activité.
- Consulte QuickSight pour visualiser les tableaux de bord de synthèse fournit par les équipes marketing, reçoivent des rapports (dashboard, rapports, etc)



Cette architecture cloud AWS se veut scalable et flexible. Centrée sur Amazon S3, kinesis Firehose, Glue Athena, QuickSight et SageMaker, elle couvre toutes les étapes :

- Ingestion
- Stockage
- Organisation
- Analyse
- Machine learning

Elle permet de fournir aux différents métiers (data analyst, marqueteur, direction) des outils adaptés, qu'il s'agisse d'analyses ad hoc, de reporting marketing ou de décisions stratégiques basées sur la donnée.

Estimation des couts

Récapitulatif des coûts potentiels (scénario simple)

Service	Coût mensuel estimé
Amazon S3	3 $ à 5 $
Kinesis Firehose (opt)	0 $ (si non utilisé) ou ~5 $
AWS Glue (Crawler/Jobs)	6 à 10 $ (petites charges)
AWS Lambda (opt)	0 $ à 5 $ (selon usage)
Amazon Athena	0,50 $ à quelques dollars
Amazon QuickSight	20 à 30 $ (pour 2-3 utilisateurs)
Amazon SageMaker (opt)	2 $ à 70 $ (selon usage ML)

Notre architecture cloud AWS a été conçue avec un souci d'optimisation des coûts tout en maintenant les performances nécessaires pour répondre aux besoins de l'entreprise. Voici une analyse détaillée des différents composants et leurs coûts associés.

1. Stockage et ingestion des données

Le stockage des données repose principalement sur Amazon S3, avec un coût estimé entre 3$ et 5$ par mois pour un volume de données modéré (environ 100 Go). Ce coût inclut à la fois le stockage (environ 2,30$ pour 100 Go) et les opérations de lecture/écriture qui représentent généralement 1$ à 2$ pour un usage standard.

Pour l'ingestion en temps réel, Kinesis Data Firehose représente un coût optionnel d'environ 5$ par mois pour 100 Go de données en streaming, comprenant l'ingestion (2,90$) et la transformation via Lambda (2$). Ce service n'engendre des coûts que lorsqu'il est effectivement utilisé.

2. Traitement et organisation des données

AWS Glue, notre solution ETL, engendre des coûts de 6$ à 10$ par mois pour des charges modérées, répartis entre :
- Le Crawler (environ 2,20$ pour une exécution quotidienne de 10 minutes)
- Les jobs ETL (environ 4,40$ pour 10 DPU-heures par mois)

AWS Lambda, utilisé pour les transformations en temps réel, bénéficie d'un niveau gratuit généreux (1 million d'invocations et 400 000 Go-secondes). Les coûts réels varient de 0$ à 5$ selon l'utilisation au-delà du niveau gratuit.

3. Analyse et visualisation

Amazon Athena, notre service d'analyse SQL, présente une tarification basée sur le volume de données analysé (5$ par To). Pour notre cas d'usage avec 100 Go de données mensuelles, le coût est d'environ 0,50$ à quelques dollars par mois. La conversion des données en format Parquet permet d'optimiser significativement ces coûts.
Amazon QuickSight, essentiel pour la visualisation, représente le coût le plus prévisible :
20$ à 30$ mensuels pour une petite équipe (2-3 utilisateurs)
Cette estimation inclut les licences pour un Data Analyst, un Marketeur (tous deux éditeurs) et un Directeur (lecteur)

4. Machine Learning avec SageMaker

Pour les fonctionnalités de machine learning, SageMaker représente un investissement variable selon l'usage :
- Configuration minimale : environ 2,30$ pour 10 heures d'entraînement mensuel
- Déploiement en production : jusqu'à 70$ par mois avec un endpoint d'inférence en continuRécapitulatif et optimisation
Le coût total mensuel se situe entre :
- 30$ à 50$ pour une configuration de base (S3 + Glue + Athena + QuickSight)
- 100$ ou plus en incluant le streaming temps réel (Kinesis) et le machine learning (SageMaker)

Pour optimiser ces coûts, nous recommandons de :
1. Convertir les données en format Parquet pour réduire les coûts d'Athena
2. Utiliser le mode batch plutôt que temps réel quand possible
3. Monitorer et arrêter les endpoints SageMaker inutilisés
4. Adapter la fréquence des crawlers Glue et des mises à jour QuickSight

5. Considérations supplémentaires

Le transfert de données entrant vers AWS est gratuit, tandis que le trafic sortant est facturé (0,09$/Go après 1 Go gratuit). Pour notre cas d'usage analytique, ces coûts restent généralement minimes.

Pour une évaluation plus précise adaptée à vos volumes exacts, nous recommandons d'utiliser l'AWS Pricing Calculator qui fournira une estimation personnalisée selon vos besoins spécifiques.