



UTILISATION DE GOOGLE BIG QUERY POUR FAIRE DES RAPPORTS BI EN TEMPS RÉEL

Initiation au big data et au cloud computing

Présenté par :

Alioune Abdou Salam KANE
Khadidiatou COULIBALY
Francis Fromo HABA
Ameth FAYE
Awa DIAW

Sous la Supervision de :

Mme Mously DIAW,
Senior ML Engineer

ISE2 - 18 décembre 2025

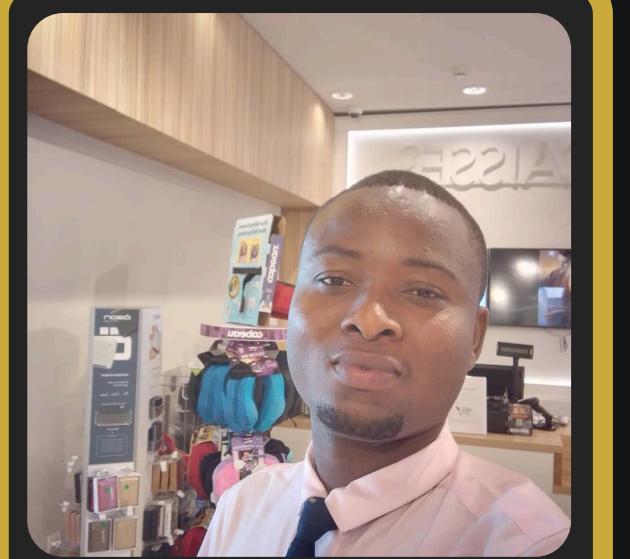
Présenté par



Alioune Abdou Salam
KANE



Khadidiatou
COULIBALY



Francis Fromo HABA



Ameth FAYE

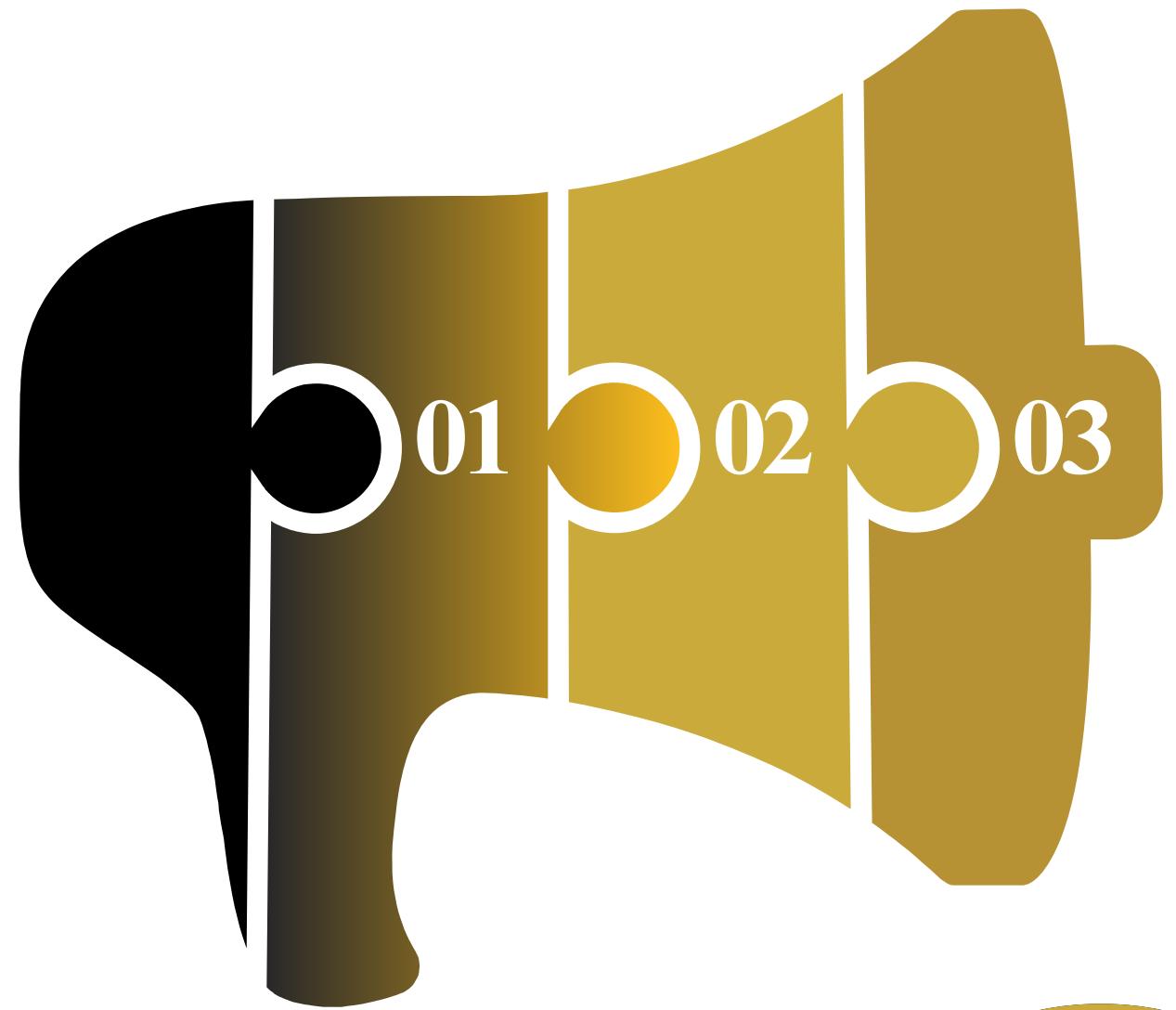


Awa DIAW

“In God we trust, others bring data”

PLAN

- 01 ANALYSE ET OBJECTIFS DU PROJET
- 02 MISE EN ŒUVRE ET MÉTHODOLOGIE
- 03 RÉALISATION ET BILAN



01 Analyse et Objectifs du Projet

Introduction
Contexte et problématique BI
Concepts clés
Données utilisées
KPI suivis (Indicateurs de performance)



Introduction



- Le marché de la cryptomonnaie est un environnement caractérisé par sa vitesse et sa volatilité extrêmes.
- Les actifs, comme le Bitcoin Cash (BCH), peuvent subir des variations de valeur majeures en quelques jours ou minutes.
- Pour les investisseurs, la clé du succès est la capacité à anticiper ces mouvements.
- ChainSight Solutions est une équipe d'analystes spécialisés dans l'exploitation du Big Data.
- Notre objectif : Proposer une solution de Business Intelligence (BI) en Temps Réel.

Contexte et problématique BI

Problématique de la Latence

- Se fier aux données de prix publiques traditionnelles rend les décisions tardives et moins rentables.
- Le temps de latence est identifié comme l'ennemi de la performance pour l'investisseur.
- Notre positionnement repose sur l'Analyse "On-Chain".
- Cette méthode étudie les mouvements réels des transactions et des liquidités directement sur la blockchain (la source de vérité)



La Question Clé du Projet

Comment l'analyse en temps réel de l'activité transactionnelle sur la blockchain permet-elle d'anticiper les tendances de marché ?



L'objectif est de construire une solution analytique complète, s'appuyant sur l'architecture Cloud Serverless



Bénéfices Clients et Objectifs (Décisionnel)

- Maximisation du Profit : Savoir exactement quand acheter ou vendre du Bitcoin
- Anticipation des Tendances : Identifier les mouvements de "Whales" (gros porteurs) avant l'impact public sur les prix.
- Prise de Décision Éclairée : Disposer d'un rapport dynamique mis à jour en continu.
- Gestion du Risque : Être alerté immédiatement lorsque des seuils critiques (ex: sortie massive de liquidité) sont atteints

Business Intelligence & Big Data



Concepts clefs

Data warehouse & serverless

Analyse on chain & Whale

Pipeline & KPI

Business Intelligence (BI) : Processus d'analyse et de présentation des données brutes en informations exploitables.

Le BI soutient la décision stratégique via des tableaux de bord

Business Intelligence (BI) & Big Data

Le Big Data est la capacité à gérer une quantité colossale d'informations, toujours plus rapides, variées, et souvent non structurées. Le Big Data est caractérisé par le Volume de données considérable à traiter, une grande Variété d'informations (venant de diverses sources, non-structurées, organisées, Open...), et un certain niveau de Vélocité à atteindre, autrement dit de fréquence de création, collecte et partage de ces données.

Le Cloud Computing permet à n'importe quelle entreprise, école ou start-up de disposer de la puissance de calcul d'un Google ou d'un Amazon... sans posséder un seul serveur





Data warehouse & serverless

- Data warehouse = Système d'entreposage de données optimisé pour les requêtes analytiques complexes sur de grands ensembles de données, tel que **Google BigQuery**.
- Serverless = la gestion de l'infrastructure serveur est entièrement déléguée au fournisseur cloud, permettant de se concentrer sur l'optimisation des requêtes et de la logique métier.

Analyse on chain & Whale

- Analyse on chain : Méthode d'analyse qui étudie les métriques extraites directement de la blockchain (transactions, volumes, adresses) pour évaluer l'activité du réseau.
- Une Whale (Baleine) est un acteur du marché détenant une quantité significative (+1000 BCH) de cryptomonnaie. Ils sont capables d'influencer les tendances du marché par leurs actions.



Pipeline & KPI

- Pipeline : Le flux de travail séquentiel par lequel les données sont acheminées, transformées et chargées depuis leur source (Blockchain) jusqu'à leur destination (Dashboard Looker Studio).
- KPI ou Key performance indicator = Métrique quantifiable utilisée pour évaluer l'efficacité de l'activité du marché (ex: vitesse de circulation).



Données utilisées



Source de Données Brutes

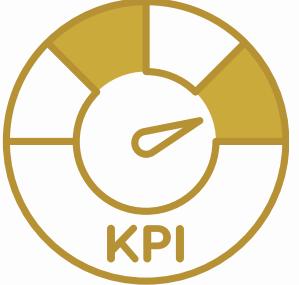
- Source officielle : Google cloud platform, dans l'espace google cloud market place (API, logiciels, fonctions, bases de données)
- Nom : crypto bitcoin cash après avoir appliqué 3 filtres : big data, données et gratuit
- volume : + 947 Go de données, +390,7 M de lignes
- Couverture historique : Données couvrant la période 2009-2024 (Dernière mise à jour en Mai 2024)



Services utilisés

- Architecture & Services utilisés : Google Cloud Platform – GCP qui est de type PaaS = Plateforme globale d'hébergement et de gestion des services cloud
- Traitement & Analyse des données : BigQuery qui est un Data Warehouse managé, Moteur d'analyse et de stockage des données, optimisé pour le traitement de grands volumes de données
- Simulation du temps réel : Cloud Functions de type FaaS (Function as a Service), fonction développée en Python qui permet simulation du temps réel sur la période 2009 – 2024
- Orchestration & Déclenchement : Pub/Sub qui est de type PaaS a un role de commandant, déclenche l'ordre d'exécution des fonctions
- Exécution des traitements : Cloud Run de type CaaS (Container as a Service) assure l'exécution des services conteneurisés, limite d'exécution : 9 minutes par traitement





KPI suivis



Nombre total de transactions

Mesure l'intensité d'usage du réseau



Adresses actives distinctes

Compte le nombre d'adresses uniques impliquées dans au moins une transaction (entrée ou sortie).



Volume total des transactions

Correspond à la somme des montants transférés on-chain sur la période. Il s'agit d'un proxy de l'intensité économique du réseau, indépendant du nombre de transactions.



Valeurs moyenne par transaction

Représente la taille moyenne des transactions effectuées sur la période. Cet indicateur est sensible aux transactions de grande taille et peut être fortement influencé par l'activité des grands détenteurs.



Valeurs médiane par transaction

Désigne la transaction « typique » de la période. Un écart important entre moyenne et médiane signale une forte hétérogénéité des montants échangés.



Vitesse de circulation – proxy

Mesure le ratio entre les flux sortants et les flux entrants agrégés. Il reflète la propension des fonds à circuler plutôt qu'à rester immobilisés. Il ne correspond pas à la velocity monétaire macroéconomique



Whale volume global

Quantifie le volume total échangé via des transactions de grande taille dépassant 100 BCH. Cet indicateur permet d'évaluer l'importance absolue des grands acteurs dans l'activité économique observée.



Whale ratio global

Mesure la part du volume total attribuable aux transactions de grande taille. Un ratio élevé traduit une activité concentrée, un ratio faible une participation plus diffuse.



Proportion de transactions coinbase

Représente la part des transactions liées à la création monétaire par le minage. Il ne reflète pas un usage économique classique du réseau et sert principalement de contrôle structurel et de qualité des données. Une valeur faible et relativement stable est généralement attendue.

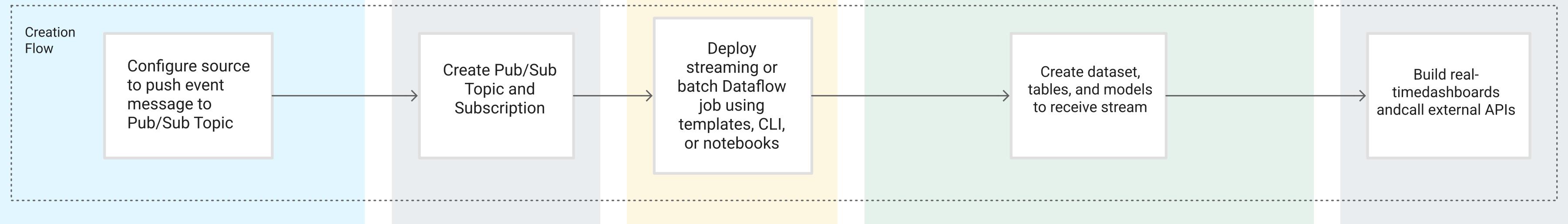
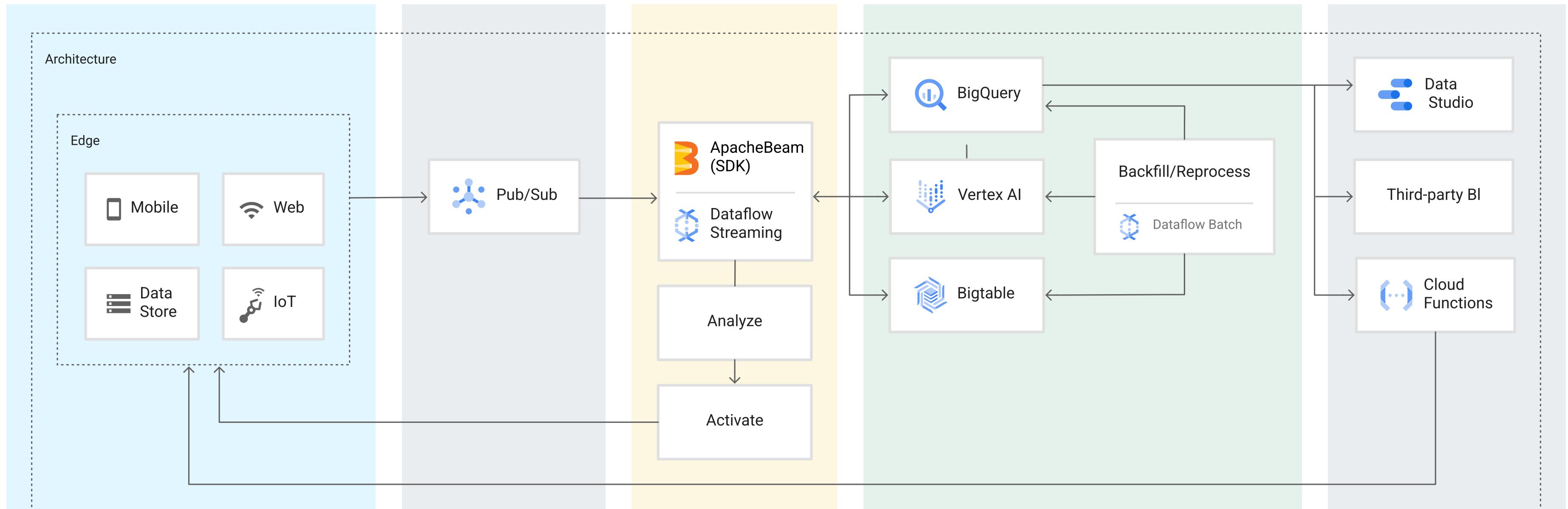
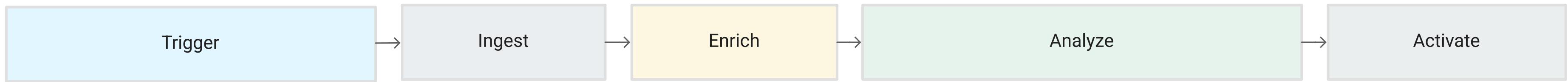
02 Mise en œuvre et Méthodologie

Architecture technique (pipeline, outils, organisation)
Traitement des données (nettoyage, préparation)

Architecture technique (pipeline, outils, organisation)

L'architecture classique d'un projet Google Cloud Platform repose sur un pipeline de données managé qui transforme les flux bruts en informations exploitables. Ce processus s'articule autour de trois piliers : l'ingestion en temps réel ou en lot via Pub/Sub, le traitement continu ou par lot avec Dataflow, et enfin le stockage analytique dans BigQuery.

Les 3 slides suivants présentent visuellement dans un premier temps le pipeline utilisé pour le déploiement des données dans le cadre général puis dans le cadre de notre projet ensuite dans un second temps nous illustrerons l'étape centrale de notre pipeline sur laquelle repose le projet.





Notre pipeline

Traitement des données et Temps réel simulé



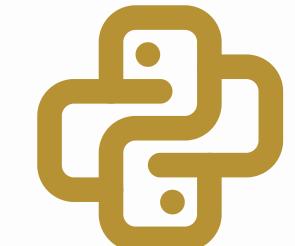
Initialisation

- Le script d'initialisation assure l'optimisation des performances Big Data.
- Création du Schéma : Le dataset de destination crypto_analytics est créé.
- Nettoyage/Enrichissement : Conversion BCH via SAFE_DIVIDE(t.output_value, 1e8).
- Préparation : Les structures imbriquées (inputs/outputs) sont désimbriquées (UNNEST) pour isoler les adresses uniques.



Optimisation du coût de requêtes SQL

- Partitionnement : Segmentation physique par date pour optimiser le data pruning et limiter les scans lors de la création de la table avec PARTITION BY DATE(horodatage_bloc).
- Clustering : Organisation par hash pour accélérer les point lookups et minimiser les entrées/sorties (I/O).
- Gouvernance : Filtres temporels et CTE pour garantir l'efficience du calcul et respecter les quotas de ressources.



Simulation

- Automatisation Python : Une Cloud Function lance le traitement SQL des observations directement sur la base agrégée.
- Translation de Table : Le système bascule les données pour ne conserver que l'historique glissant des 30 derniers jours.
- Latence Réduite : Ce flux assure une simulation proche du temps réel . On parle de NEART.



03 Réalisation et bilan

Dashboard interactif

Conclusion

Livrables

Documentation

Dashboard interactif**

Looker Studio est l'outil choisi pour le Dashboard dynamique.



- Un contrôle est effectué pour vérifier que les rôles IAM sont configurés.
- L'étape de connexion utilise le connecteur natif BigQuery pour une liaison la plus performante.
- Le processus consiste à sélectionner : le Projet Cloud, le Dataset (crypto_analytics) et les tables.

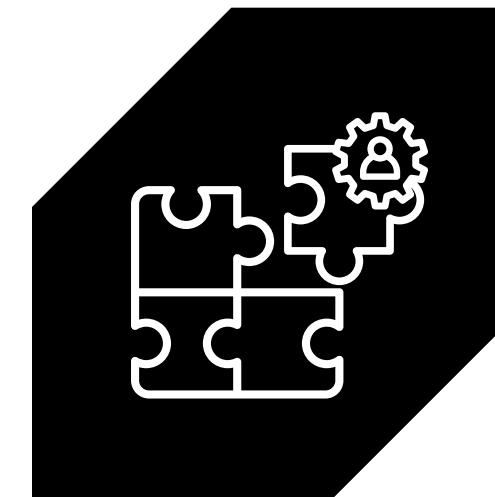
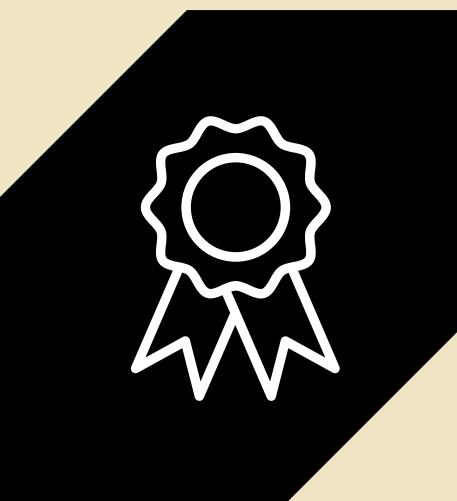


- Chaque visualisation exécute une requête optimisée sur BigQuery.
- La connexion établie en mode Direct Query garantit que les données affichées aux investisseurs sont constamment à jour.
- Le rapport conserve sa nature temps réel.

Conclusion

Limites

- Latence du traitement : L'utilisation des Scheduled Queries impose un rafraîchissement par micro-batches, empêchant ainsi l'obtention d'un flux de données en temps réel absolu à la seconde.
- Rigidité ergonomique : L'interface Looker Studio, étant une plateforme "No-Code", limite la personnalisation visuelle avancée en interdisant l'intégration de code CSS ou JavaScript personnalisé.
- Contraintes d'interactivité : La nature fermée de l'écosystème Google BI restreint les fonctionnalités dynamiques complexes, obligeant à privilégier la simplicité opérationnelle sur la précision esthétique.



Perspectives

- Transition vers le streaming : L'intégration de Google Cloud Dataflow permettrait de traiter les données en continu pour éliminer la latence des micro-batches et atteindre une réactivité immédiate voire même faire une analyse anticipационnelle avec du ML dans le cloud.
- Hybridation technologique : Le recours à des bibliothèques de visualisation tierces ou à des outils BI plus flexibles pourrait offrir la liberté de design nécessaire aux exigences professionnelles.
- Optimisation de l'architecture : La mise en place d'un pipeline hybride permettrait de concilier la puissance de calcul brute de BigQuery avec une interface utilisateur plus interactive et personnalisée.

Livrables

- 01 Support de présentation 
- 02 Une documentation technique 
- 03 Le lien vers les KPI expliqué en détails 
- 04 Le lien vers le Dashboard développé 
- 05 Un site web 

Documentation

Dans le cadre de la réalisation technique du projet, les ressources webographiques suivantes ont été exploitées :

- [Intégration Google BigQuery et Looker](#) : Un guide détaillé sur l'architecture de connexion entre l'entrepôt de données et l'outil de visualisation pour une analyse fluide.
- [Optimisation des rapports BigQuery dans Looker Studio](#) : Études de cas et meilleures pratiques pour structurer ses données BigQuery afin d'obtenir des tableaux de bord performants.
- [Premiers pas avec Google Cloud Platform](#) : Documentation officielle pour configurer l'environnement cloud et comprendre les fondamentaux de l'infrastructure GCP.
- [Introduction à Google BigQuery](#) : Présentation complète du moteur de base de données serverless, de son fonctionnement et de ses capacités d'analyse de données massives.
- [Guide de démarrage rapide Looker Studio](#) : Manuel d'utilisation pour créer ses premiers rapports et explorer les fonctionnalités de design de l'outil BI.
- [Connecter Looker Studio à Google BigQuery](#) : Procédure technique étape par étape pour établir la liaison entre les tables SQL et l'interface de visualisation.
- [Exécution et optimisation des requêtes SQL](#) : Documentation sur la syntaxe et les méthodes pour lancer des requêtes performantes au sein de la console BigQuery.

Utilisation de Google Big Query pour faire des rapports BI en temps réel

MERCI



Repository git hub



ISE2 - Décembre 2025