



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

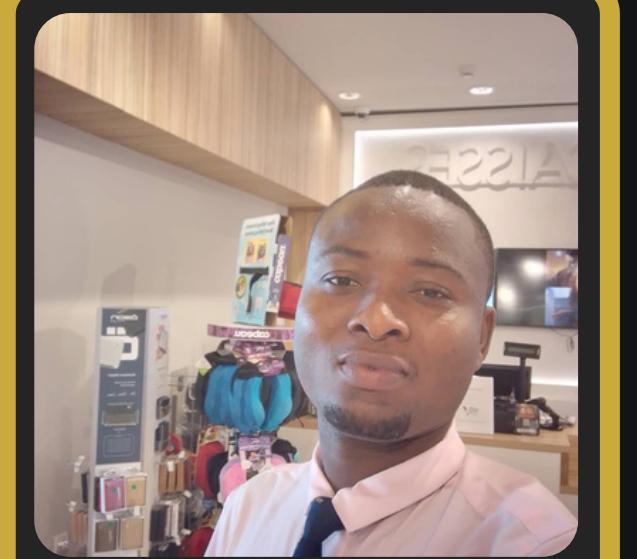
Présenté par



Alioune Abdou Salam
KANE



Khadidiatou
COULIBALY



Francis Fromo HABA



Ameth FAYE

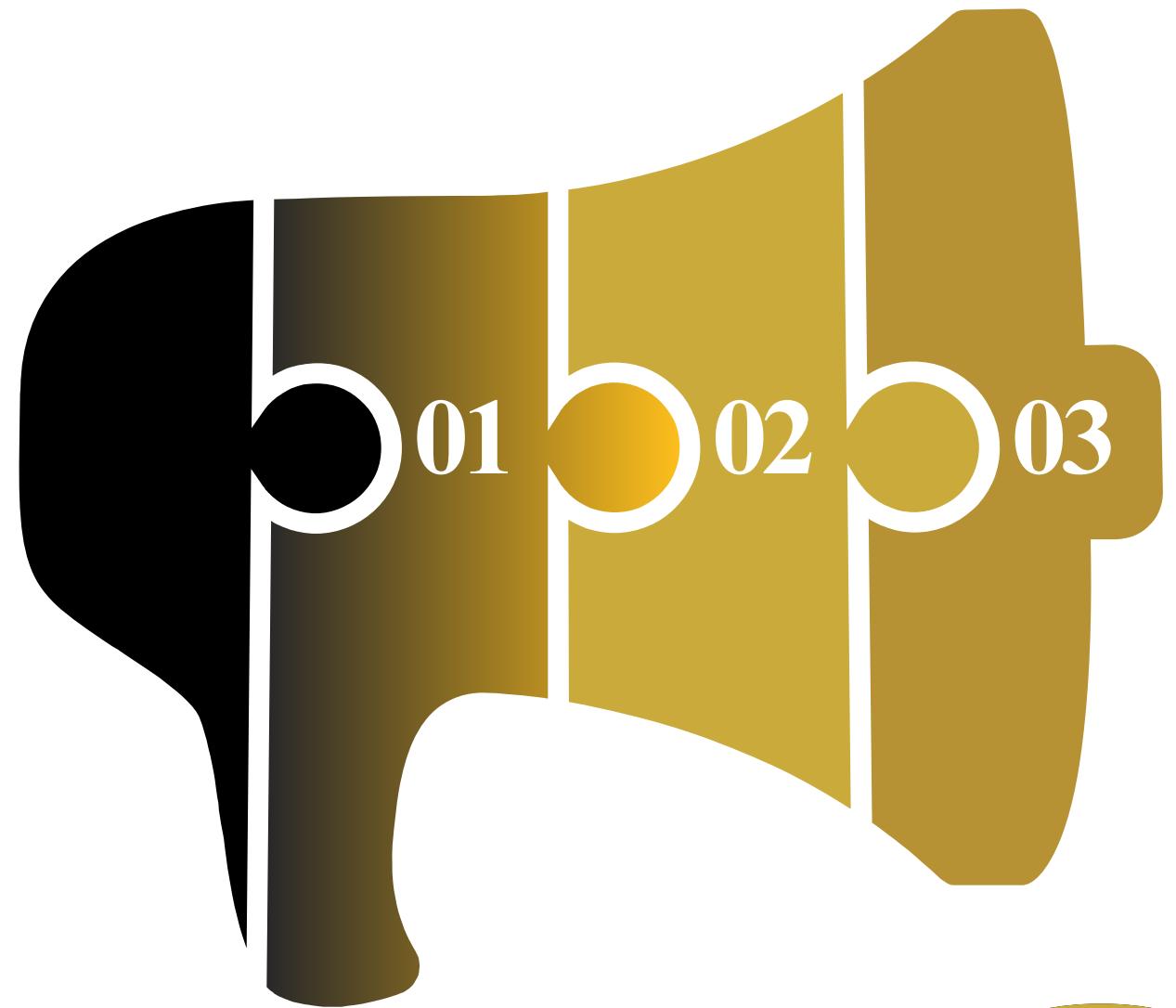


Awa DIAW

“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

Données utilisées



Source de Données Brutes

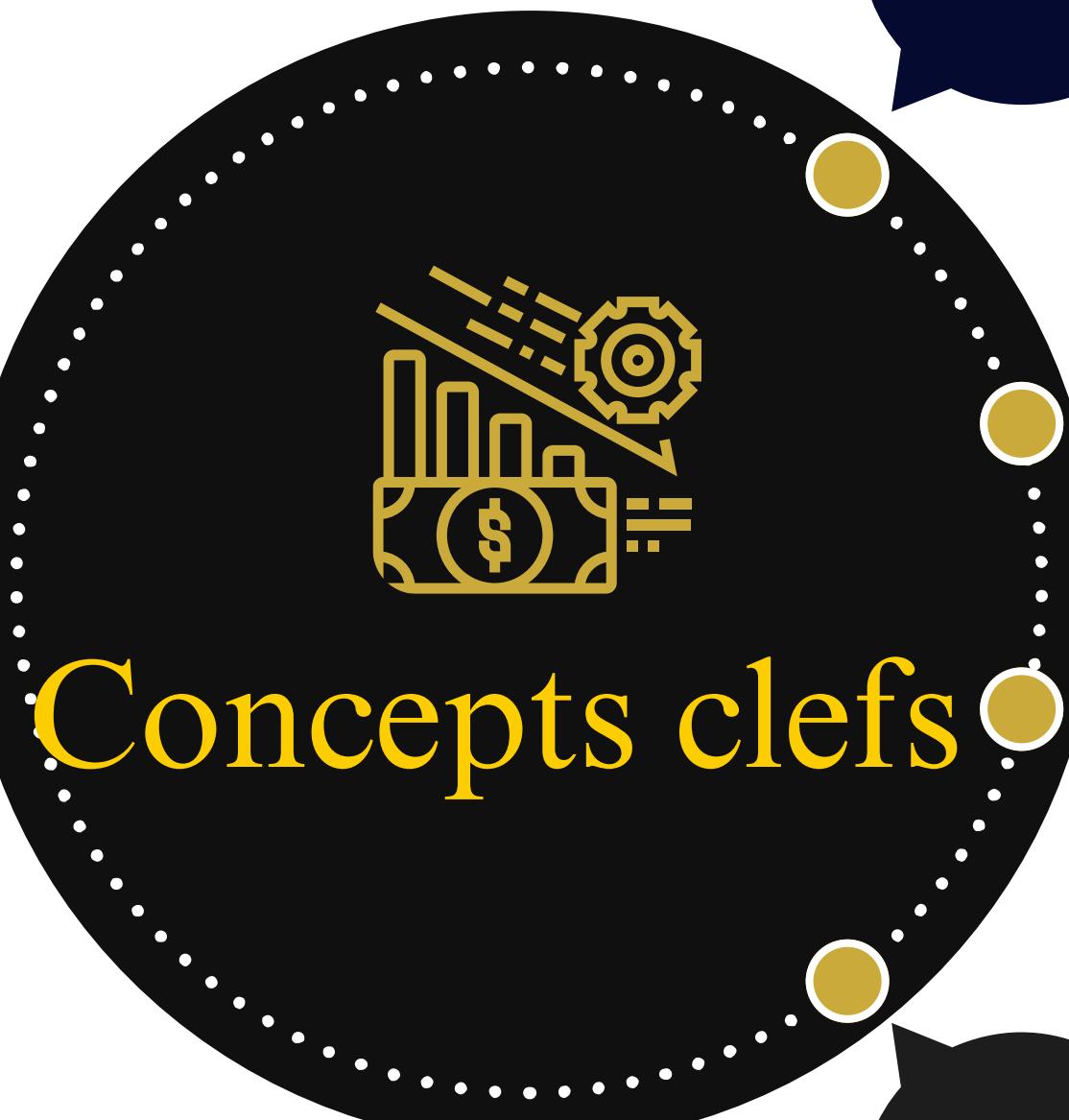
- Source officielle : Utilisation du Google Cloud Public Dataset `bigrquery-public-data.crypto_bitcoin_cashifiable`.
- Accessibilité : Dataset public hébergé sur BigQuery, permettant l'analyse complète de la blockchain Bitcoin Cash via des requêtes SQL.
- Architecture & volume : Environ 947 Go de données, organisées principalement autour des tables Blocks et Transactions, prétraitées au format Parquet.
- Couverture historique : Données couvrant toute l'histoire de Bitcoin Cash jusqu'en 2024, nettoyées et optimisées pour l'audit et la recherche économique.
- Mise à jour & exhaustivité : Synchronisation en temps réel avec le réseau BCH, contenant l'intégralité des données de la blockchain prétraitées.



Justification du Choix d'Architecture (BigQuery)

- Choix stratégique garantissant une réalisation rapide et performante de la BI
- Adoption d'un modèle ELT (Extract, Load, Transform) simplifié.
- L'élimination de la gestion complexe de l'ETL nous permet de nous concentrer sur la logique métier et l'analyse.
- BigQuery est un Data Warehouse Serverless géré par Google.
- Il offre des performances pour le scan et l'agrégation de données massives en quelques secondes.

Business Intelligence (BI) & Big Data



Concepts clefs

Temps Réel et Analyse
On-Chain

Whale Movement

Satoshis et Bitcoin Cash (BCH)

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



Temps Réel et Analyse On-Chain

- La solution est conçue pour la BI en temps réel, grâce à la connexion Direct Query entre Looker Studio et BigQuery.
- La base de données BCH est synchronisée en permanence (fréquence temps réel).
- Analyse On-Chain : Méthode étudiant les métriques extraites directement de la blockchain.
- Elle évalue l'activité fondamentale et la santé économique du réseau

Whale Movement (Mouvement des Baleines)

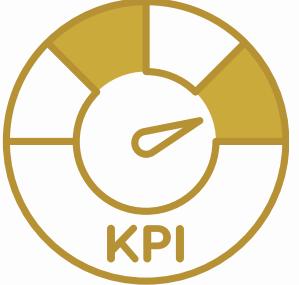
- Une Whale (Baleine) est un acteur du marché détenant une quantité significative de cryptomonnaie.
- Les Whales sont capables d'influencer les tendances du marché par leurs actions.
- Le Whale Ratio est un indicateur essentiel pour identifier ces mouvements avant qu'ils n'impactent publiquement les prix.
- Un seuil critique du Whale Ratio est fixé à $> 10\%$ du volume journalier.



Satoshis et Bitcoin Cash (BCH)

- Bitcoin Cash (BCH) : Cryptomonnaie conçue pour inclure plus d'octets par bloc que le Bitcoin, favorisant la vitesse des transactions.
- Les données brutes de valeur sont en satoshis (format natif).
- Pour l'analyse, une étape de nettoyage convertit la valeur des satoshis vers l'unité Bitcoin Cash (BCH).
- La table des transactions est fondamentale pour la traçabilité et l'horodatage (block_timestamp) des opérations.





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

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

Traitement des données (nettoyage, préparation)



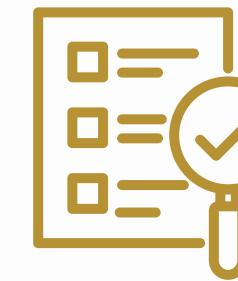
Initialisation

- 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 de l'Entrepôt

- L'optimisation réduit les coûts et garantit la rapidité des requêtes temps réel.
- Partitionnement : La table de base est partitionnée par jour (PARTITION BY DATE(horodatage_bloc)).
- Cela réduit le temps de scan lors de l'interrogation.
- Clustering : Le clustering est appliqué sur le hash de transaction.
- Ceci accélère les recherches ciblées au sein des partitions



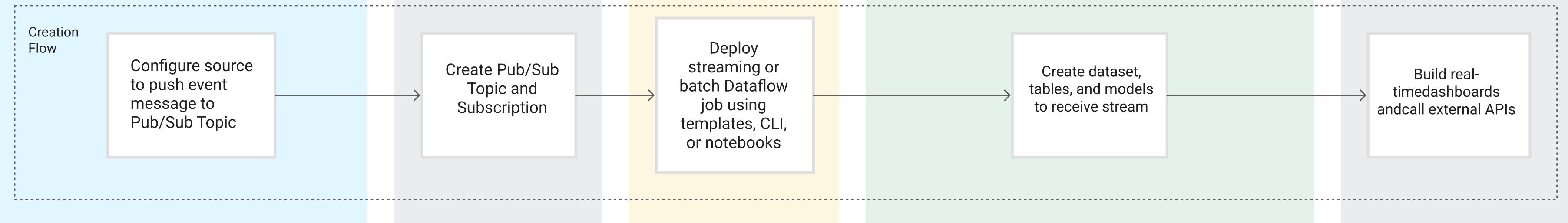
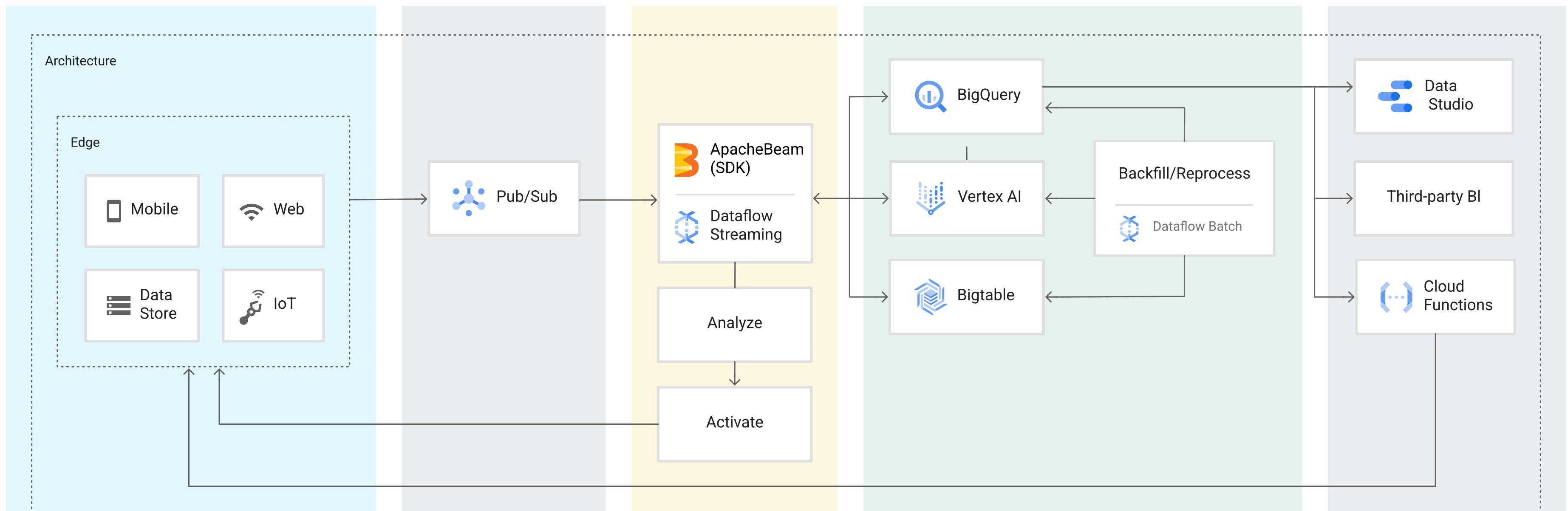
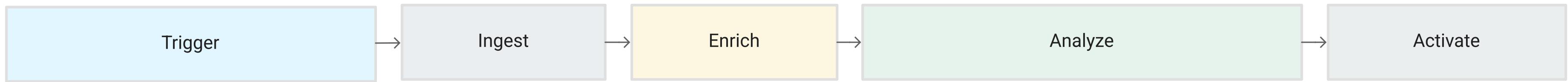
Filtrage et Sécurité

- Un filtre de sécurité est appliqué sur la date d'échantillonnage.
- Exemple : WHERE t.block_timestamp_month = '2019-01-01'.
- Ce filtre garantit la performance et le respect du quota gratuit de BigQuery lors du développement.
- Le script 2_kpi_journaliers.sql utilise ensuite les CTE pour calculer les indicateurs quotidiens

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 via Pub/Sub, le traitement et le nettoyage continu avec Dataflow, et enfin le stockage analytique dans BigQuery.

Les 2 slides suivants présentent visuellement le pipeline utilisé pour le déploiement des données dans le cadre général puis dans le cadre de notre projet





Notre pipeline



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.

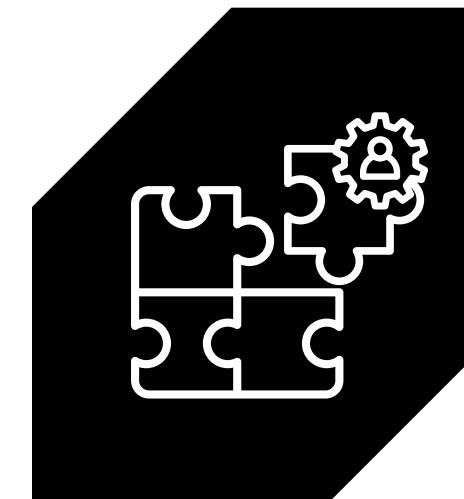
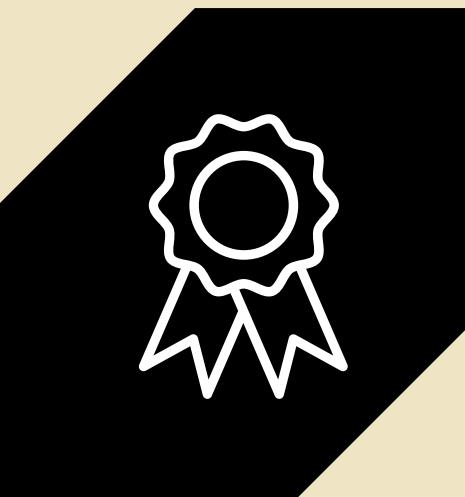


- Chaque visualisation exécute une requête optimisée sur BigQuery.
- La connexion établie en mode Direct Query garantit que les données affichées aux investisseurs sont constamment à jour.
- Le rapport conserve sa nature temps réel.

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.
- 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