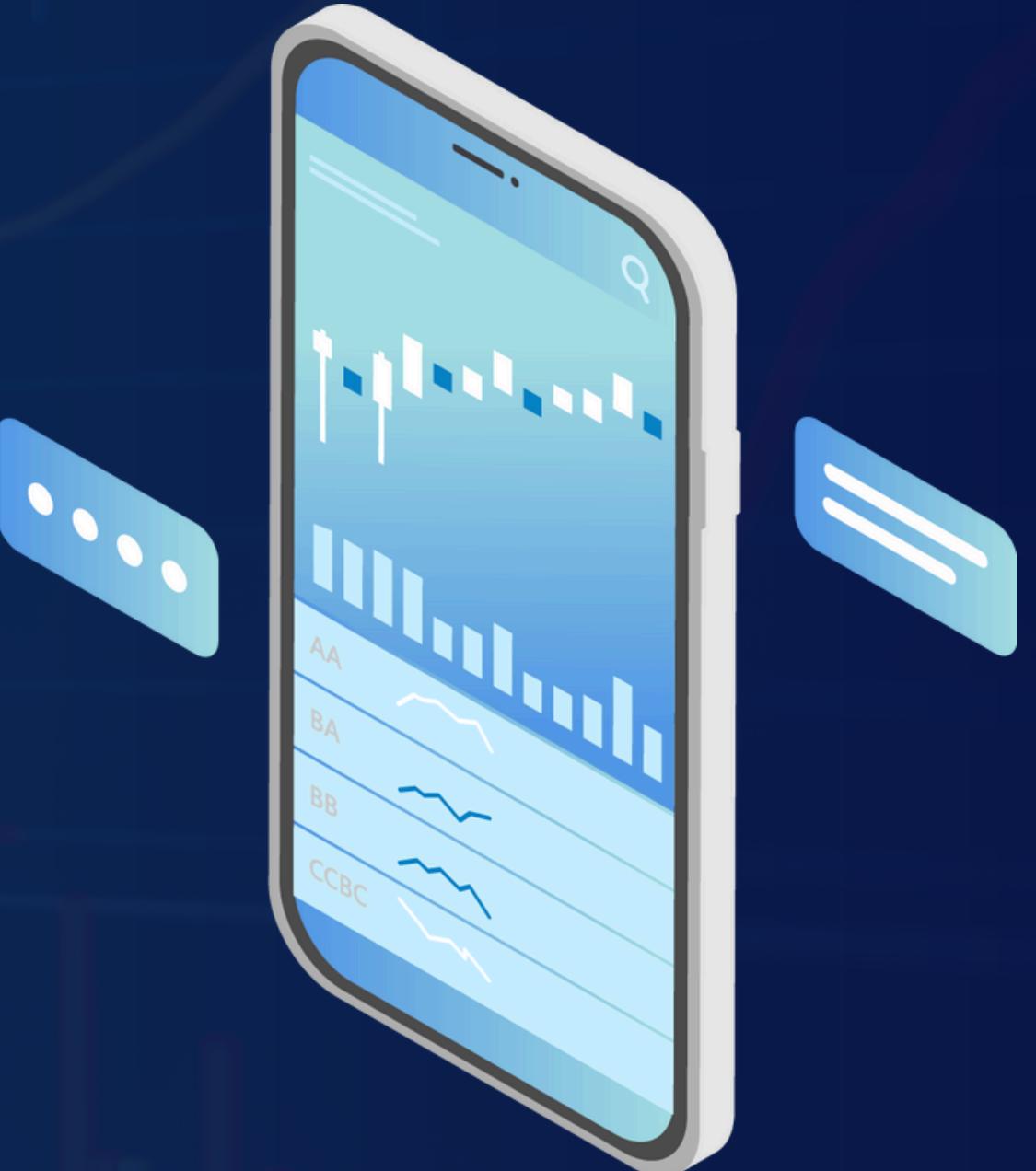


Projet Big Data Analysis

Prédiction du prix du Bitcoin à partir du sentiment Reddit

Réalisé par :

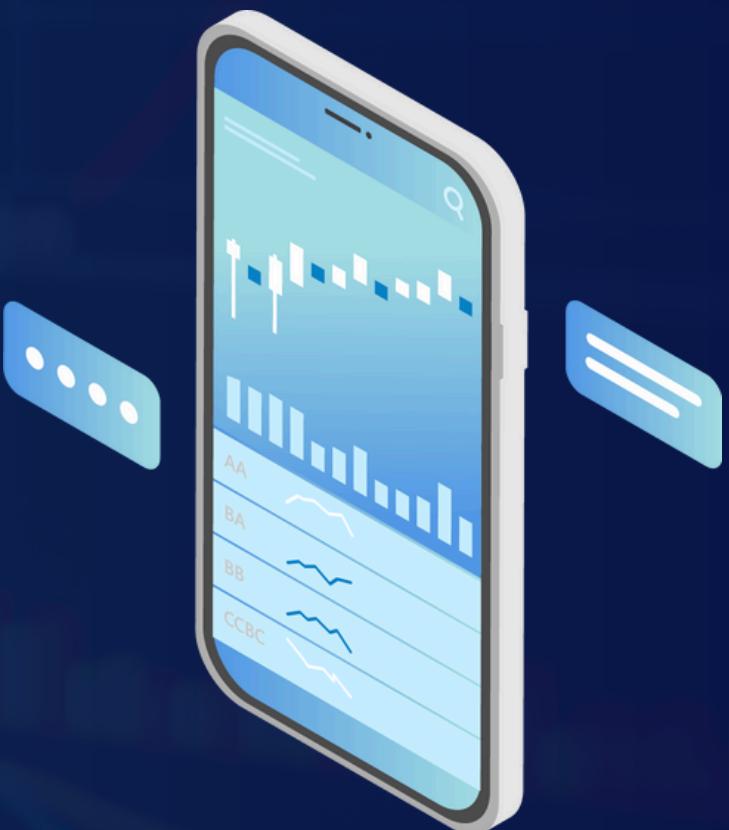
Ayman Fenkouch
Yahya MOUDRIK
Hiba ASGHAR
Youssef Rizki



Plan de la présentation



1. Contexte, problématique et objectifs
2. Cas d'étude et sources de données
3. Architecture Big Data et pipeline de traitement
4. Analyse et modélisation
5. Visualisations et résultats
6. Conclusion et perspectives



Contexte et motivation

- Les marchés financiers sont sensibles à l'opinion publique
- Les réseaux sociaux génèrent des données massives et en temps réel
- Les cryptomonnaies sont fortement influencées par le sentiment des investisseurs
- Les technologies Big Data sont nécessaires pour traiter ce volume et cette vitesse de données



Problématique

Le sentiment exprimé sur Reddit à propos du Bitcoin a-t-il un impact sur l'évolution de son prix ?





OBJECTIFS

- Analyser le lien entre sentiment Reddit et prix du Bitcoin
- Mettre en place une architecture Big Data scalable
- Tester une approche prédictive à court terme

Choix du cas d'étude

Pourquoi Bitcoin ?

- Forte volatilité
- Actif très réactif aux news et au sentiment
- Données publiques et disponibles en continu



Pourquoi Reddit ?

- Plateforme majeure de discussion crypto
- Communautés très actives
- Indicateurs d'engagement (posts, commentaires, votes)



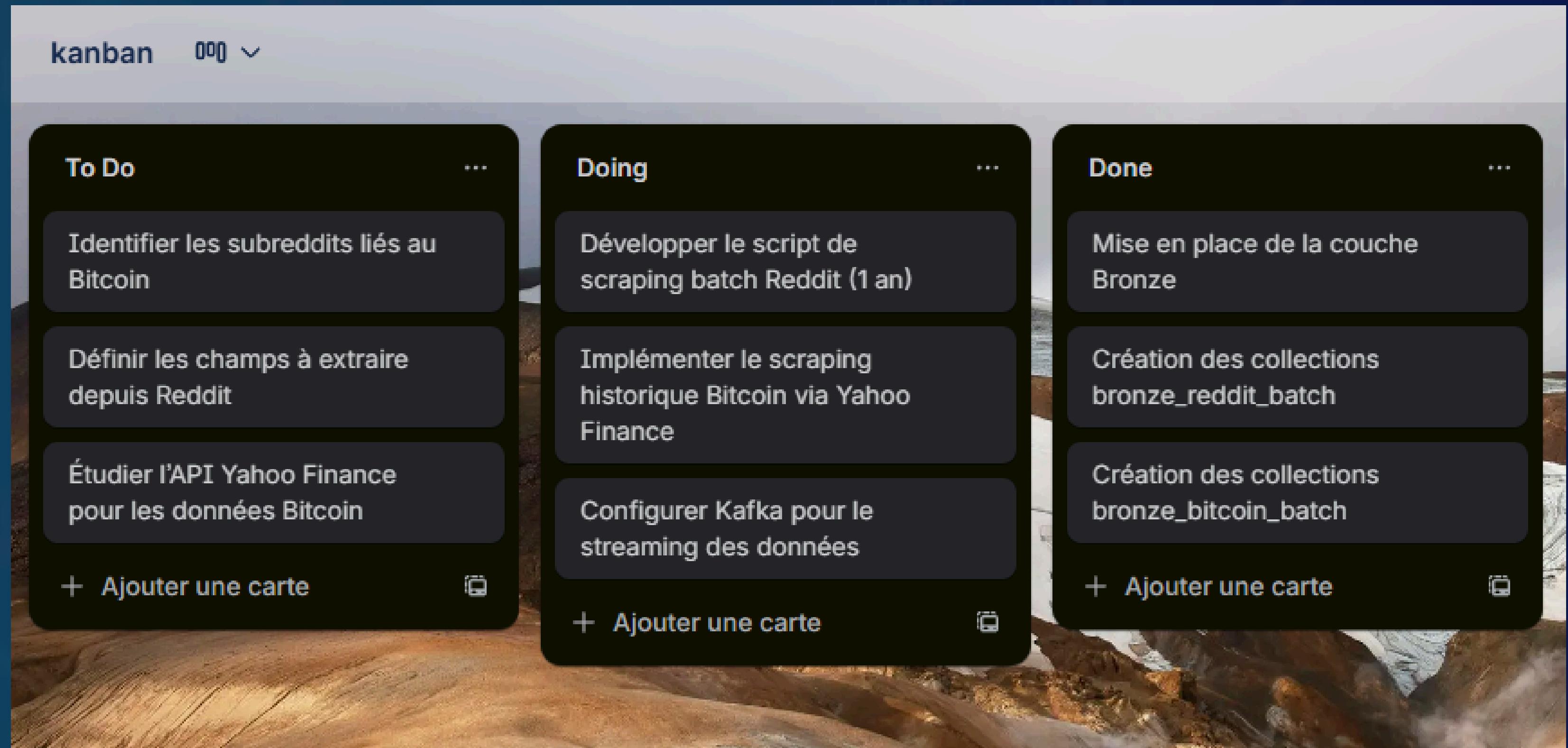
Organisation du travail

- Ingestion & Bronze : Yahya MOUDRIK
- Traitement Silver & sentiment analysis : Ayman FENKOUCH
- Modélisation ML : Youssef RIZKI
- Dashboard & visualisation : Hiba ASGHAR



Organisation du travail

kanban 000 ~



The image shows a digital kanban board with three columns: To Do, Doing, and Done. The board is set against a background of a wooden desk with various items like a keyboard, mouse, and papers.

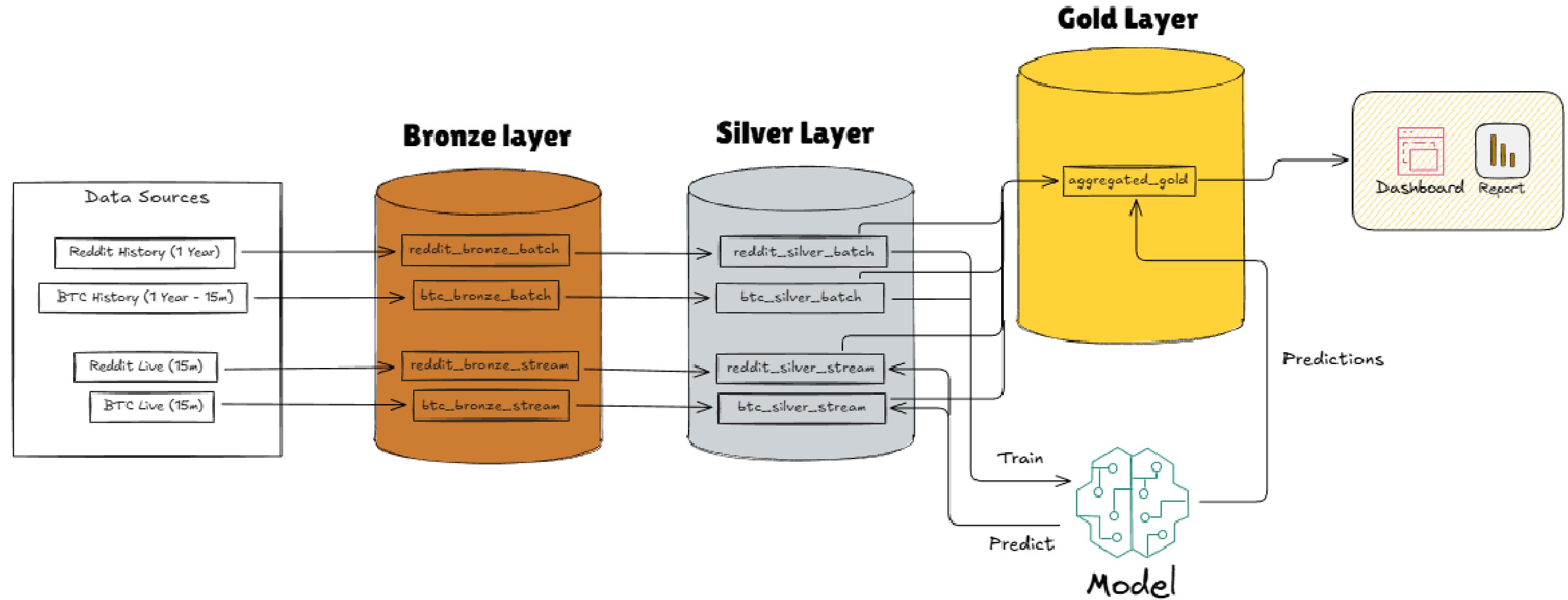
To Do	Doing	Done
Identifier les subreddits liés au Bitcoin	Développer le script de scraping batch Reddit (1 an)	Mise en place de la couche Bronze
Définir les champs à extraire depuis Reddit	Implémenter le scraping historique Bitcoin via Yahoo Finance	Création des collections bronze_reddit_batch
Étudier l'API Yahoo Finance pour les données Bitcoin	Configurer Kafka pour le streaming des données	Création des collections bronze_bitcoin_batch
+ Ajouter une carte	+ Ajouter une carte	+ Ajouter une carte

Architecture globale du projet

- Architecture Medallion :
 - Bronze : données brutes
 - Silver : données nettoyées et enrichies
 - Gold : données analytiques et prédition
- Séparation batch / streaming



Architecture globale du projet





Sources de données



reddit

- **Données Reddit**

- Posts liés au Bitcoin
- Plusieurs subreddits spécialisés
- Métadonnées (date, contenu, engagement)

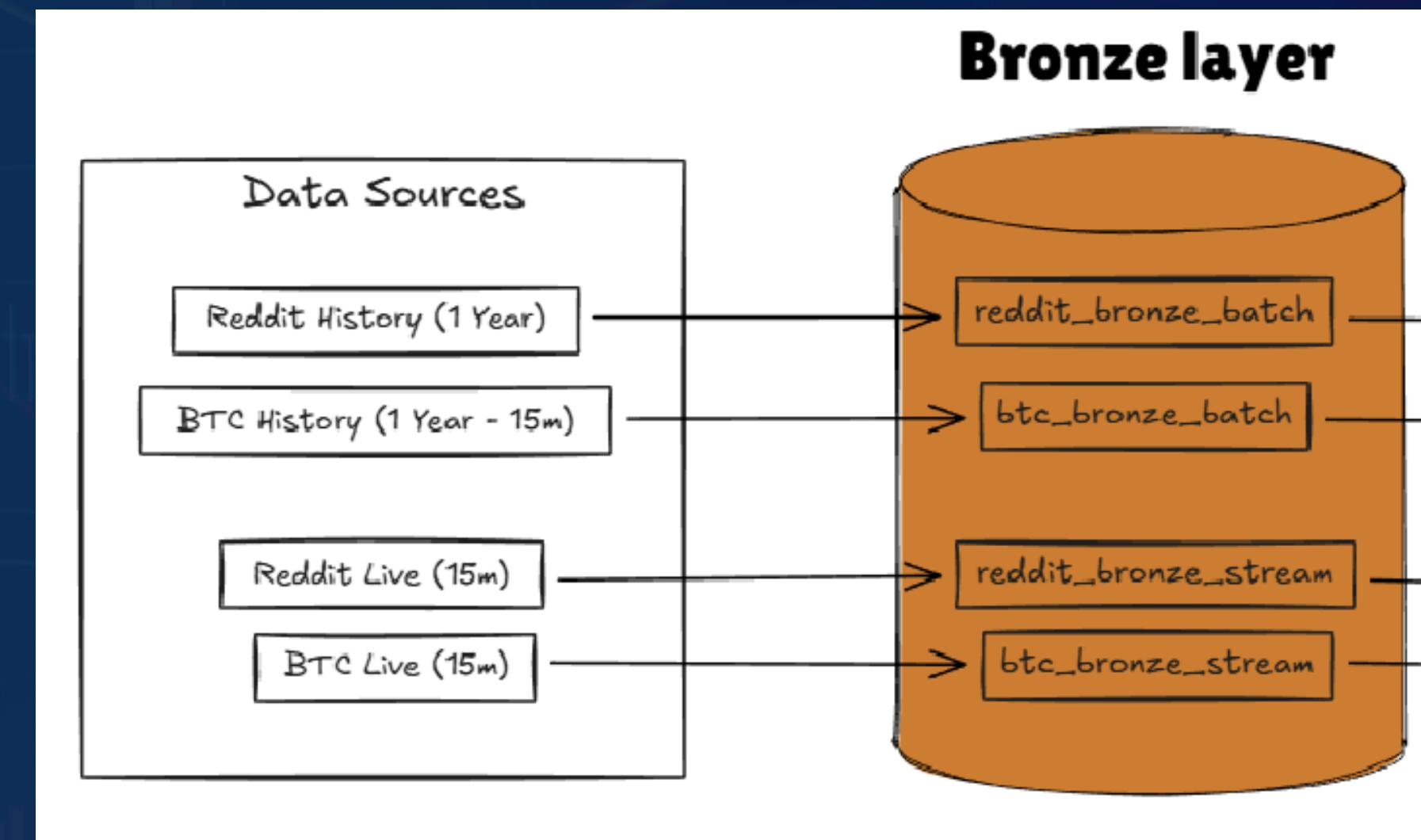


- **Données financières**

- Prix du Bitcoin
- Open, Close, High, Low
- Volume d'échange

Rôle de la couche Bronze

- Stockage des données brutes
- Aucune transformation
- Données conservées telles que collectées
- Séparation claire entre :
 - Batch
 - Streaming

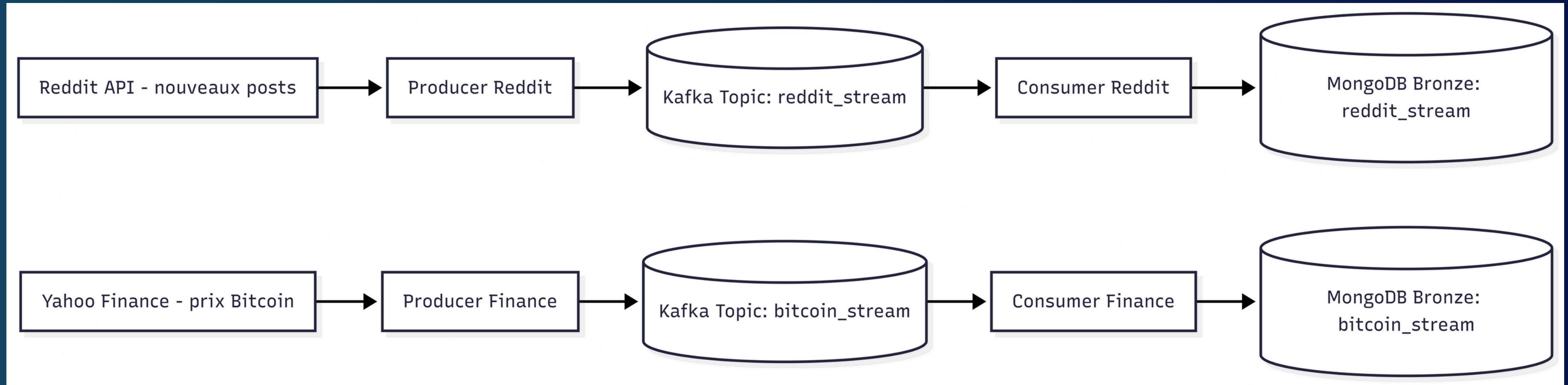


Ingestion Batch (Historique)

- Récupération d'un an de données historiques
- Reddit :
 - Posts des 12 derniers mois
 - Plusieurs subreddits
 - Stockage dans bronze_reddit_batch
- Bitcoin :
 - Données historiques via Yahoo Finance
 - Stockage dans bronze_btc_batch



Ingestion Streaming (temps réel)



- Collecte en temps réel toutes les 5 minutes
- Nouveaux posts Reddit
- Prix du Bitcoin mis à jour
- Utilisation de Kafka pour le streaming

```
_id: ObjectId('695846a7aa0fa711490aaaae6')
datetime: "2025-01-01 00:00:00"
open: 93396.03
high: 94256.05
low: 93312.7
close: 94256.05
volume: 0
scraped_at: "2026-01-02T23:28:52.669222"
ticker: "BTC-USD"
interval: "60m"
```

Bitcoin Data

```
num_comments: 22
num_crossposts: 0
total_awards: 0
gilded: 0
created_utc: 1767225560
created_datetime: "2025-12-31T23:59:20"
edited: false
edited_datetime: null
over_18: false
spoiler: false
locked: false
archived: false
is_distinguished: false
distinguished_type: null
flair: "Gain"
flair_css: "profit"
author: "Entalope"
author_fullname: "t2_rmeb9"
author_flair: null
author_premium: false
is_video: false
is_gallery: true
media_type: "gallery"
author_link_karma: 2164
author_comment_karma: 7417
author_total_karma: 9581
author_created_utc: 1446309121
author_account_age_days: 3716
author_verified: true
author_has_verified_email: true
author_is_gold: false
author_is_mod: false
```

Reddit Data

Financial Engineering



Financial Engineering

$$SMA_n = \frac{P_1 + P_2 + \dots + P_n}{n}$$

$$RSI = 100 - \frac{100}{1 + RS}$$

Financial Engineering

$$TR = \max (\text{High} - \text{Low}, |\text{High} - \text{Prev Close}|, |\text{Low} - \text{Prev Close}|)$$

$$ATR_n = \frac{TR_1 + TR_2 + \dots + TR_n}{n}$$

$$MACD = EMA_{12} - EMA_{26}$$

Sentiment Engineering

The screenshot shows the homepage of the **r/Bitcoin** subreddit. At the top, there's a large orange button with a white Bitcoin symbol. Below it, the subreddit name "r/Bitcoin" is displayed in white text. There are two dropdown menus: "Best" and "New". A section titled "Community highlights" features a post by user **u/Reasonable-Team-1232** from 12 hours ago. The title of the post is **The Era of Bitcoin Abundance is Over**. The post content discusses the mining of 95% of the Bitcoin supply and includes a link to a chart at <https://en.macromicro.me/charts/29045/bitcoin-exchange-balance-total>. The post has received 461 upvotes and 119 comments. Below the post, there are buttons for "Share" and other interaction options.

Best ▾ New ▾

Community highlights

Bitcoin Newcomers
FAQ - Please read!
73 votes • 25 comments

Daily Discussion, January 22, 2026
3 votes • 4 comments

u/Reasonable-Team-1232 • 12 hr. ago

The Era of Bitcoin Abundance is Over

95% of Bitcoin supply has been mined. There will likely never be this much Bitcoin available to purchase ever again. <https://en.macromicro.me/charts/29045/bitcoin-exchange-balance-total> If you look at the entire history of the Bitcoin exchange balance you can literally see the exact date it peaked. Monday, July 26th, 2021. That day was the historical day the most Bitcoin was ever available to purchase. Since then, we have descended all the way back to 2018 level supply (nearly 8 years ago). From nearly 3.5 Million total available to purchase 1 year ago to 2.5 million today. All the while price has steadily risen from \$4000 to over \$120,000. It will likely continue gainin...

461 119 Share

Sentiment Engineering

u/Reasonable-Team-1232 • 12 hr. ago

Reasonable-Team-1232
u/Reasonable-Team-1232
Jan 15, 2026

293 Post karma 43 Comment karma

What is karma?

Follow Start Chat

likely never be this much Bitcoin available to purchase ever
[itcoin-exchange-balance-total](#) If you look at the entire history of
the exact date it peaked. Monday, July 26th, 2021. That day
available to purchase. Since then, we have descended all the way
from nearly 3.5 Million total available to purchase 1 year ago to
risen from \$4000 to over \$120,000. It will likely continue gainin...

Sentiment Engineering

Vader

$$\text{Compound Score} = \frac{\sum_{i=1}^n s_i}{\sqrt{\sum_{i=1}^n s_i^2 + \alpha}}$$

TextBlob

$$\text{Polarity} = \frac{\sum_{i=1}^n s_i}{n}$$

SocialIndex = log(post_score) + log(num_comments) + log(author_karma) + log(author_comment_karma) + upvote_ratio

MACHINE LEARNING

- Objectif:
 - La prediction le rendement du prix de Bitcoin dans l'horizon d'une heure en fonction l'historique et les données sociales dans Reddit.
- Input : Couche Silver
 - l'historique de prix de Bitcoin
 - des données de post de reddit.

MACHINE LEARNING : Feature Engineering

- engagement = score + num_comments + num_crossposts
- credibility = author_total_karma + author_account_age_days
- weight = credibility + 1
- on calcule :
 - VaderScore_weighted
 - Blobtext_weighted
 - le weighted socialindex
- on fait une agrégation sur les données de reddit par heure :
- on faire une jointure entre les deux datasets par la colonne datetime.

MACHINE LEARNING:Feature Engineering:

- Calcul des retours horaires (variation du prix d'une heure à l'autre).
- Mesure de la volatilité (écart-type des retours).
- On crée des versions des variables sociales retardées avec un :
 - Décalage de 1 heure (impact immédiat).
 - Décalage de 2 heures (effet court terme).
 - Décalage de 6 heures (effet moyen terme).
 - Décalage de 24 heures (effet long terme).
- target_ret+1h : le retour horaire du Bitcoin à prédire, décalé d'une heure (valeur future).

MACHINE LEARNING : Construction du modèle LSTM:

- Format des données d'entrée.
 - Séquences de longueur fixe (ex. 48 heures).
 - Chaque séquence contient plusieurs features (prix BTC, indicateurs techniques, variables sociales).

MACHINE LEARNING : Construction du modèle LSTM:

- Architecture du réseau
 - Première couche LSTM :
 - 64 unités, return_sequences=True pour conserver la dimension temporelle.
 - Dropout (0.2) : régularisation pour éviter le surapprentissage.
 - Deuxième couche LSTM : 32 unités.
 - Dropout (0.2).
 - Couche Dense finale : 1 sortie (prédiction du retour horaire futur du BTC).

MACHINE LEARNING :Compilation du modèle

- Optimiseur :
 - Adam avec learning_rate=0.001.
- Fonction de perte :
 - MSE (Mean Squared Error).
- Métriques suivies :
 - MAE (Mean Absolute Error).

MACHINE LEARNING : Entraînement

- Données divisées chronologiquement :
 - train (0.70)/ validation(0.15) / test(0.15)
- Normalisation des features avec StandardScaler.
- Entraînement sur 30 époques, batch size = 32.
- Validation sur un sous-ensemble des données.

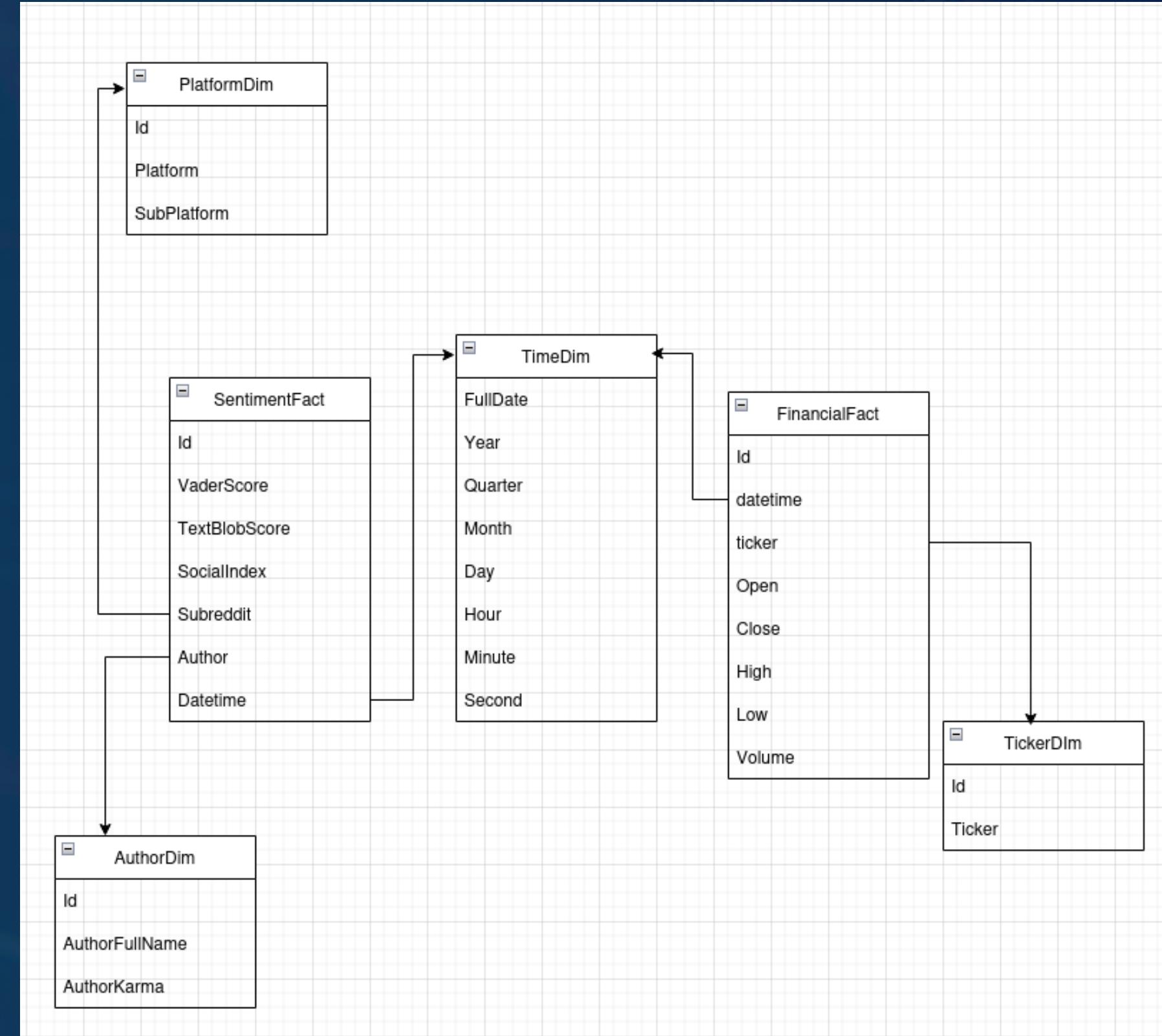
MACHINE LEARNING : Résultats

- RMSE: 0.008866
- MAE: 0.005953
- Directional Accuracy: 44.48%

Couche Gold



Dimensional Model



Dimensional Model

```
_id: "695846a7aa0fa711490aaaae6"
close : 94256.05
datetime : "2025-01-01 00:00:00"
high : 94256.05
low : 93312.7
open : 93396.03
ticker : "BTC-USD"
volume : 0
PrctChange : null
Direction : "No Change"
SMA20 : 94256.05
SMA50 : 94256.05
SMA200 : 94256.05
RSI : null
ATR : 943.3500000000058
MACD : 0
```

```
_id: "6958cd25bed92eeacdeffc8c"
author : "Entalope"
author_account_age_days : 3716
author_comment_karma : 7417
author_fullname : "t2_rmeb9"
author_has_verified_email : true
author_is_gold : false
author_is_mod : false
author_link_karma : 2164
author_premium : false
author_total_karma : 9581
author_verified : true
num_comments : 22
num_crossposts : 0
score : 241
 subreddit : "wallstreetbets"
title : "Year of the regard"
upvote_ratio : 0.89
VaderScore : -0.7221999764442444
TextblobScore : -0.05267857015132904
SocialIndex : 27.59373831757844
VaderFullScore : -19.92819716296379
TextblobFullScore : -1.453598679699972
datetime : 2025-12-31T22:59:20.000+00:00
author_created_ts : 2015-10-31T16:32:01.000+00:00
author_created_datetime : 2015-10-31T16:32:01.000+00:00
```

▲ Hide 1 field

Dashboards et Visualisations

Création de 3 dashboards interactifs dans Power BI pour fournir des insights actionnables sur le marché Bitcoin et le sentiment social.



Direct query



Power BI

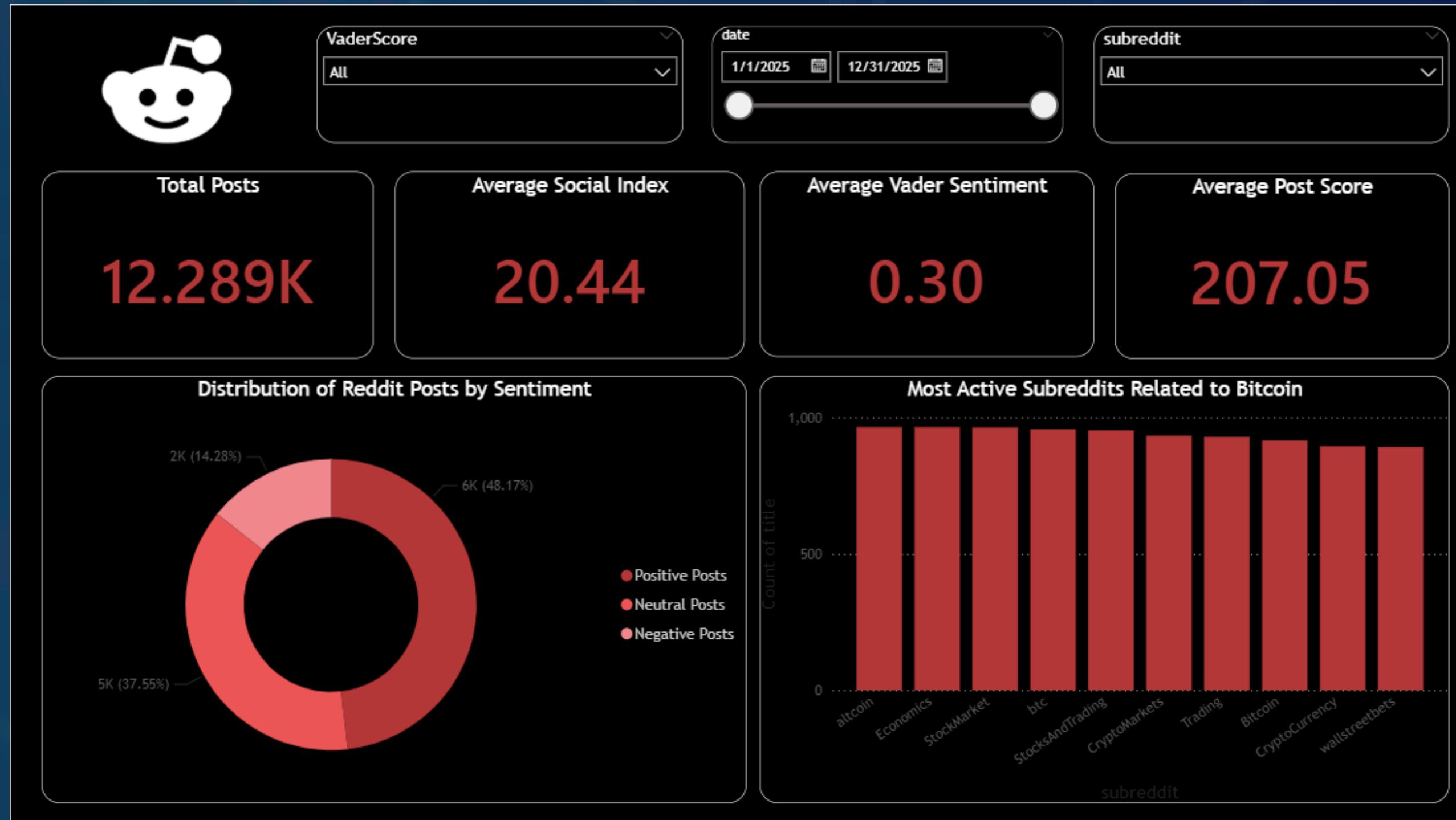
Dashboards et Visualisations

1- Bitcoin Market Overview



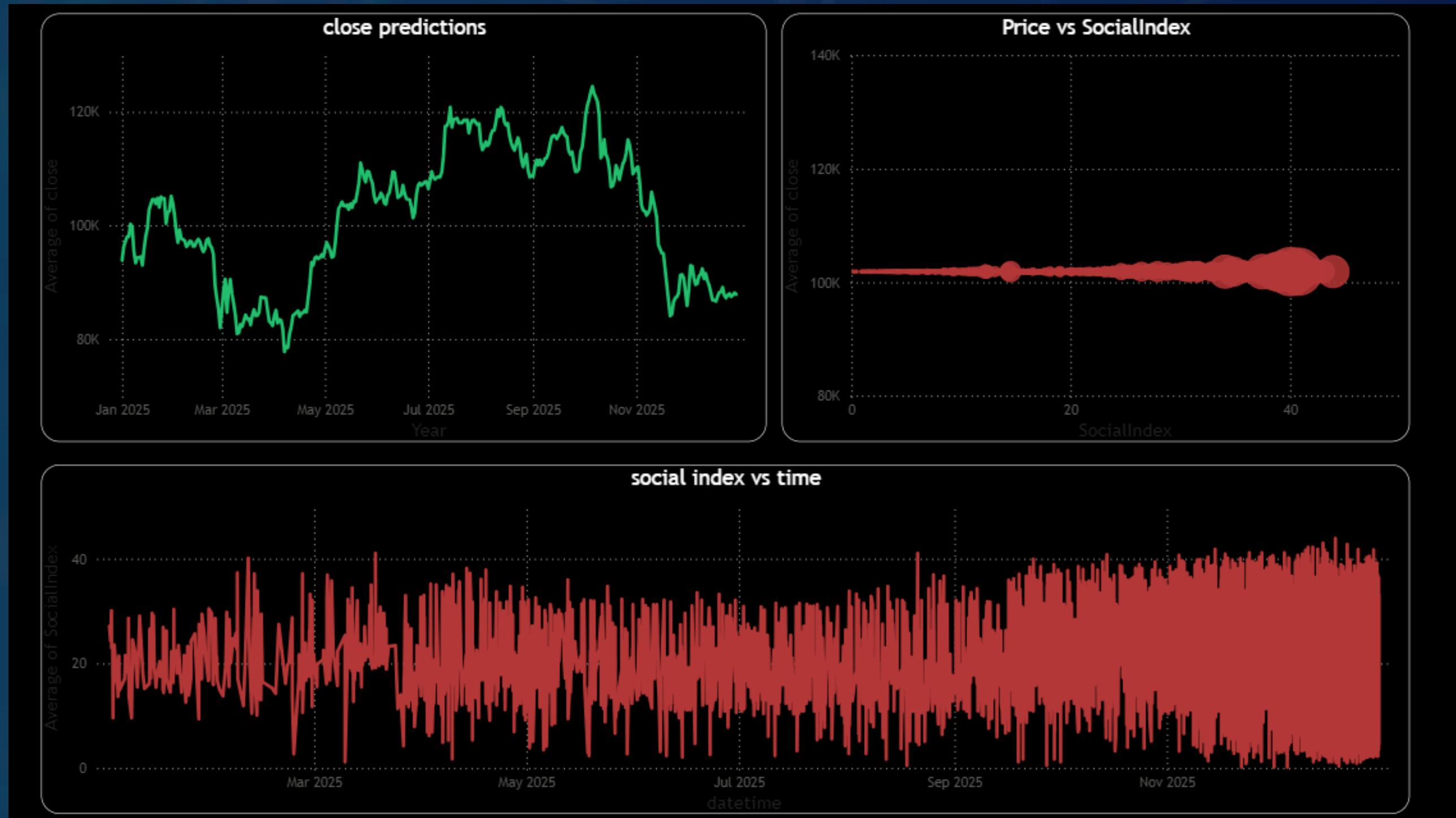
Dashboards et Visualisations

2- Social Sentiment & Influence



Dashboards et Visualisations

3- Social Impact on Bitcoin Price



Conclusion

- Architecture Big Data complète basée sur le modèle Medallion
- L'ingestion batch et streaming des données Reddit et Bitcoin a été mise en place avec succès
- L'analyse du sentiment social sur les données rassemblées montre une relation assez faible avec l'évolution du prix du Bitcoin
- L'analyse des sentiments seule n'a pas un effet visible sur les prédictions
- Reddit n'est pas le réseau social le plus efficient

Perspectives

- Intégration d'autres réseaux sociaux (X, Telegram, forums spécialisés)
- Extension à d'autres cryptomonnaies (Ethereum, Solana, etc.)
- Amélioration des modèles de prédiction avec des approches deep learning (LSTM, Transformers)
- Détection automatique d'événements majeurs via le sentiment