

Projet Big Data

Ingestion / Traitement / Visualisation



Par *S.Mahboubi / P.Zozor / F.Hammouch*
Sous la direction de Alex LIMA

Master 1 Data Engineer/Science
le 19/11/21

Sommaire

- Introduction
- Sources de données
- Architecture
- Traitement et Dataviz
- Conclusion et perspectives



Introduction

Présentation du sujet

Objectif est d'analyser l'impact des informations de l'univers cryptographique sur le marché des actifs numériques en temps réel ainsi que le ressenti des utilisateurs sur les informations pouvant orienter le marché.

- **Veille sur notre portefeuille d'actifs numériques** : Les news ont-elles un impact sur la valeur des cryptos suivies (Bitcoin, Ethereum, XRP).
- Développement d'un **algorithme d'achat et de vente d'actifs** en temps réel en croisant nos sources d'informations.
- Essayer de détecter si des bots automatisés de trading sont connectés à notre source d'information cryptopanic.
-

Qu'est-ce qu'une crypto monnaie ?

C'est un actif qui s'échange de pair-à-pair (P2P) sans tiers de confiance. Elles n'ont pas de support physique comme des pièces ou des billets, ne sont pas régulées par un organe central.

Tout comme le marché traditionnelle de la bourse, la valeur d'un actif numérique repose sur la confiance que les investisseurs ont de celui-ci, c'est donc une **guerre de l'information**.

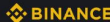


Sources de données



Sources de données

Binance


 Acheter des cryptos EUR Trader NFT New Dérivés Finance Earn

Connexion S'inscrire Téléchargements Français | EUR

Marchés


Rechercher le nom d'une monnaie

Meilleur gagnant (24h).

 GALA/BNB
0,00042305
€0


+116.67%

Meilleur perdant (24h).

 SNM/BTC
0,00000863
€0


-22.39%

Volume supérieur (24h).

 USDT/BIDR
14339
€1

-0.26%

Moments forts.

 MANA/USDT
3,2260
€3

-3.09%

★ Favoris Marchés Spot Marchés des Futures Nouveaux listings. Toutes les cryptos Aperçu du marché Les plus actifsPortefeuille Spot(3) Compte Futures(0) Marge(3)

Plateforme d'échange de crypto monnaies

Sources de données

CryptoPanic

- Outil qui absorbe tout type de données présentent sur la toile internet ayant pour sujet la crypto monnaie et les listes.
- Similaire à Bloomberg pour la bourse classique
- Ces informations sont essentiel pour les professionnels et les amateurs car il recense chaque news pouvant impacter le prix des actifs numériques.
- De plus les utilisateurs de la plateforme peuvent voter pour marquer leurs ressentiment sur une news : Bonne, mauvaise, troll
- Ainsi que le ressentiment sur la tendance de l'actif : importants, haussiers ou baissiers.

The screenshot shows the CryptoPanic website, a popular news aggregator for cryptocurrency. The interface is dark-themed and includes a navigation bar at the top with options like 'Top News', 'Show All', and a search bar. The main content area displays a list of news items, each with a headline, a timestamp, and a source link. On the right side, there is a 'Trending' section and a 'Recent Comments' section. The bottom of the page contains a footer with social media links, an 'Advertise' button, and a 'CryptoPanic.com' logo.

[CryptoPanic](https://cryptopanic.com)



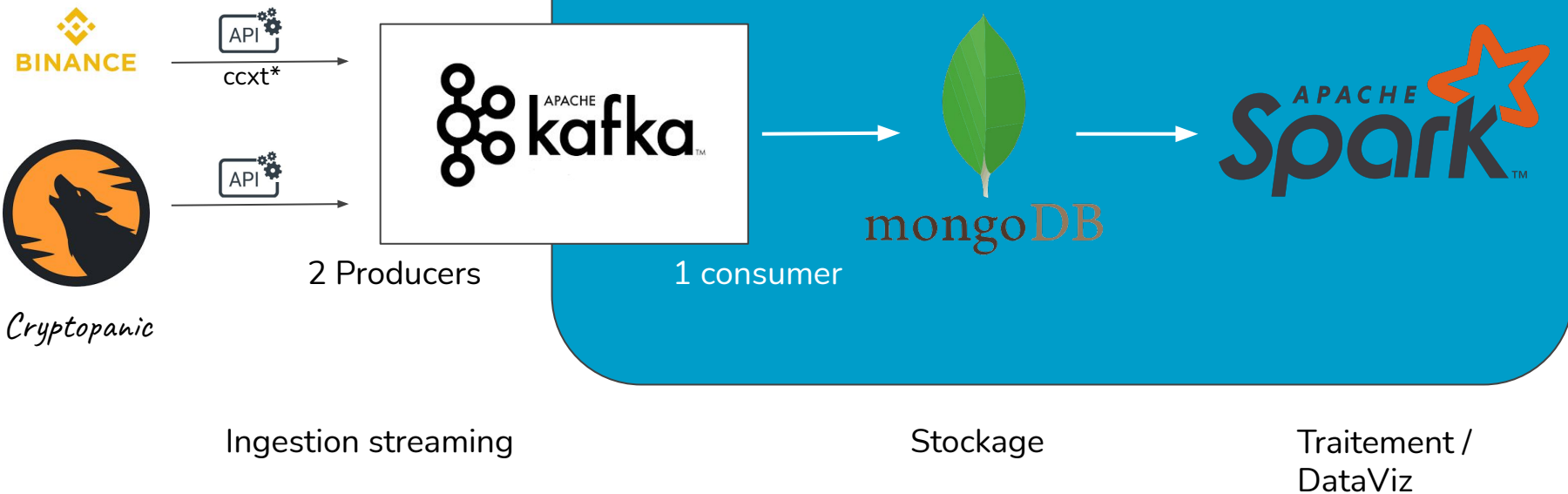
Architecture





Architecture

Vue générale



**Library Java/ Python / PHP pour le commerce des crypto-monnaies avec un support pour de nombreux marchés d'échange de bitcoin/ether/altcoin et des API de commerçants.*



Architecture

Composition du docker

- Kafka
- Mongo
- Spark



docker

RUNNING



docker_spark-worker_1 [bde2020/spark-...](#)

RUNNING



docker_kafka_1 [wurstmeister/k...](#)

RUNNING PORT: 9092



docker_mongo_1 [mongo](#)

RUNNING PORT: 27017



pyspark_notebook [jupyter/pyspark...](#)

RUNNING PORT: 8888



docker_mongo-express_1 [mongo-express](#)

RUNNING PORT: 8081



docker_spark-master_1 [bde2020/spark-...](#)

RUNNING PORT: 7077



docker_zookeeper_1 [wurstmeister/z...](#)

RUNNING



Architecture

Producer CryptoPanic

- Cryptos suivies : **BTC, XRP, ETH**
- Récupération des données “**Hot**” de nos 3 cryptos
- Fréquence de streaming : **1 heure**
- Topic du producer : **CryptoPanic**
- Données recueillies :

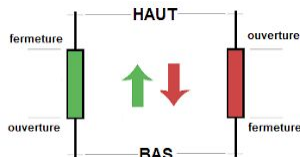
```
{'id': 13384320,  
'currencies': [{'code': 'BTC',  
  'title': 'Bitcoin',  
  'slug': 'bitcoin',  
  'url': 'https://cryptopanic.com/news/bitcoin/'},  
{'code': 'ETH',  
  'title': 'Ethereum',  
  'slug': 'ethereum',  
  'url': 'https://cryptopanic.com/news/ethereum/'}],  
'title': 'Binance Smart Chain sets record after 14.7 million daily  
transactions in one day',  
'published_at': '2021-11-18T20:00:43Z',  
'votes': {'negative': 1,  
  'positive': 6,  
  'important': 3,  
  'liked': 5,  
  'disliked': 1,  
  'lol': 0,  
  'toxic': 0,  
  'saved': 0,  
  'comments': 1},  
'url':  
'https://cryptopanic.com/news/13384320/Binance-Smart-Chain-  
-sets-record-after-147-million-daily-transactions-in-one-day'}
```



Architecture

Producer Binance

- Cryptos suivies : **BTC, XRP, ETH**
- Récupération des données **ohlcv** de nos cryptos, grâce à la méthode `fetch_ohlcv` de ccxt
- Fréquence de streaming : **1 min**
- Topic du producer : **Binance**



id	BTC	Timestamp	Open	Low	High	Close	Volume
	ETH	Timestamp	Open	Low	High	Close	Volume
	XRP	Timestamp	Open	Low	High	Close	Volume

```
{'id': 0, 'BTC': [[1637235720000, 59591.21, 59601.8, 59515.01, 59523.35, 19.73556]], 'ETH':  
[[1637235720000, 4219.75, 4219.99, 4215.15, 4215.61, 84.5654]], 'XRP': [[1637235720000, 1.093, 1.0931,  
1.0904, 1.0909, 269886.0]]}
```



Architecture

Consumer

- Réception des données des **2 producers** (Binance et cryptoPanic)
- Enregistrement des données dans MongoDB (4 DataBases):
 - BTC
 - ETH
 - XRP
 - Hot spot CryptoPanic *

* Stockage uniquement des post parus au cours de la dernière heure

- Exécution de deux script avec deux thread



Traitement et Dataviz



Traitement et Dataviz

- Traitement avec spark
- Représentation graphique interactive
- Filtre sur période/crypto et calcul statistique
- Affichage des post par Crypto ou date
- “Jauge” positif /négatif
- Calcul de la variation des cryptos quelques heures après la parution d’un Hot post
- ...

Entrée [56]: `#Statistique sur la periode choisie
filtered_period.describe()`

Out [56]:

	open	high	low	close	volume
count	7.000000	7.000000	7.000000	7.000000	7.000000e+00
mean	1.102414	1.103229	1.100771	1.101686	6.402276e+05
std	0.001772	0.001779	0.002131	0.001678	4.040545e+05
min	1.100400	1.101300	1.098000	1.099900	3.724070e+05
25%	1.100850	1.101300	1.098000	1.099900	3.724070e+05

Filtrer sur une periode

Entrée [55]: `# selection du debut et de la fin de la periode
start= '2021-11-18 01:00'
end= '2021-11-18 01:20'
df2 = df.set_index(['dateheure'])
filtered_period = df2.loc[start : end]
del filtered_period['timestamp']
filtered_period`

Out [55]:

	open	high	low	close	volume
2021-11-18 01:00:00	1.1047	1.1061	1.1041	1.1045	614096.0
2021-11-18 01:03:00	1.1045	1.1046	1.1021	1.1027	398525.0
2021-11-18 01:06:00	1.1027	1.1037	1.1023	1.1028	372407.0
2021-11-18 01:09:00	1.1029	1.1035	1.1002	1.1003	521846.0
2021-11-18 01:12:00	1.1004	1.1013	1.0993	1.1009	660493.0
2021-11-18 01:15:00	1.1009	1.1020	1.0980	1.1007	1519906.0
2021-11-18 01:18:00	1.1008	1.1014	1.0994	1.0999	394320.0





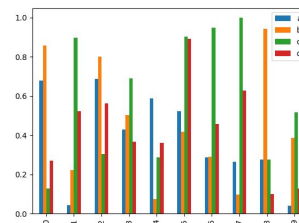
Traitement et Dataviz

Site dynamique pour suivis
d'une liste de crypto

Entrée [56]: `#Statistique sur la periode choisie`
`filtered_period.describe()`

Out[56]:

	open	high	low	close	volume
count	7.000000	7.000000	7.000000	7.000000	7.000000e+00
mean	1.102414	1.103229	1.100771	1.101686	6.402276e+05
std	0.001772	0.001779	0.002131	0.001678	4.040545e+05
min	1.100400	1.101300	1.098000	1.099900	3.724070e+05
25%	1.100850	1.101700	1.099350	1.100500	3.964225e+05
50%	1.102700	1.103500	1.100200	1.100900	5.218460e+05
75%	1.103700	1.104150	1.102200	1.102750	6.372945e+05
max	1.104700	1.106100	1.104100	1.104500	1.519906e+06





Conclusion & perspectives






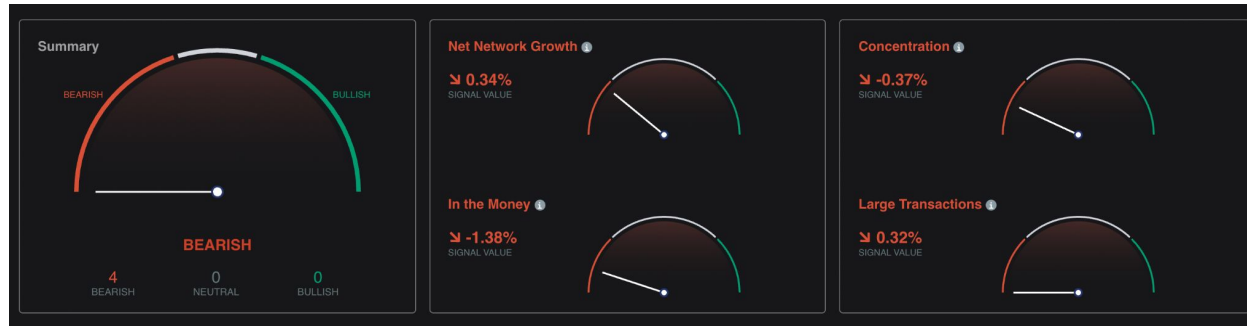
Conclusion

& Difficultés rencontrées

- Architecture terminée
- La complexité du marché fait en sorte que certaines informations sont informelles et biaisent nos analyses et prédictions (d'où l'idée d'ajout de modèles de Machine Learning dans nos traitements)
- beaucoup d'apprentissage...
 - Fonctionnement et connexion à une API
 - Streaming de données provenant de sources web.
 - Utilisation de MongoDB
 - Python (pymongo, pyspark)
 - Déploiement d'un environnement complet sur Docker et interaction entre ses différentes composantes
- Difficultés rencontrées et niveau de difficulté :
 - Envoi et stockage des données sur MongoDB (2/5)
 - Connexion MongoDB - Spark pour le traitement de données (5/5)
 - Bannissement de l'API de Binance (0/5)

[illegible]

- Connecter d'autres sources d'information sur les Cryptos
ex: Twitter, CoinmarketCap
- Ajouter des méthodes de Machine / Deep Learning à nos traitements
- Automatiser tout le process en Cloud
- Améliorer la visualisation en rendant dynamique notre rapport avec  Power BI





Merci pour votre attention