

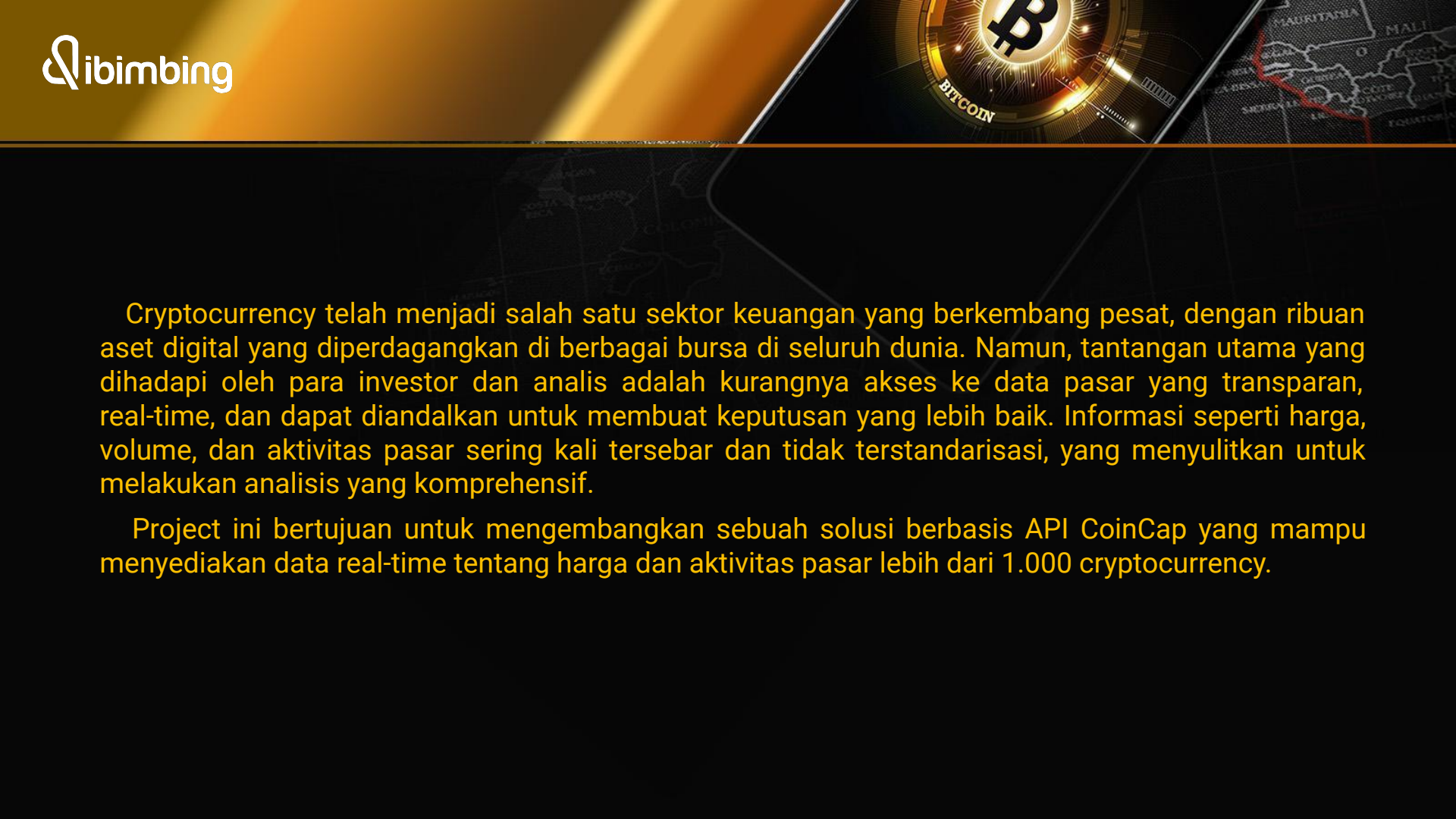
Hybrid Data Pipeline for Stream and Batch Processing of Cryptocurrency Market Data

Dhoifullah Luth Majied



Project Background



The background of the slide is a dark, textured surface. In the top right corner, there is a glowing Bitcoin symbol with the word 'BITCOIN' written below it. A faint world map is visible in the background, with a red line tracing a path across it. The text is in a bright yellow color.

Cryptocurrency telah menjadi salah satu sektor keuangan yang berkembang pesat, dengan ribuan aset digital yang diperdagangkan di berbagai bursa di seluruh dunia. Namun, tantangan utama yang dihadapi oleh para investor dan analis adalah kurangnya akses ke data pasar yang transparan, real-time, dan dapat diandalkan untuk membuat keputusan yang lebih baik. Informasi seperti harga, volume, dan aktivitas pasar sering kali tersebar dan tidak terstandarisasi, yang menyulitkan untuk melakukan analisis yang komprehensif.

Project ini bertujuan untuk mengembangkan sebuah solusi berbasis API CoinCap yang mampu menyediakan data real-time tentang harga dan aktivitas pasar lebih dari 1.000 cryptocurrency.

Problem Statement



Saat ini, investor dan analis cryptocurrency menghadapi masalah utama berupa kurangnya alat yang efektif untuk mengakses data pasar cryptocurrency yang real-time, transparan, dan terorganisir. Data yang tersedia di berbagai bursa sering kali tidak konsisten, sehingga menyulitkan dalam melakukan analisis harga, tren pasar, dan pengambilan keputusan berbasis data. Selain itu, pengolahan data yang tidak efisien dapat menimbulkan keterlambatan informasi yang mengakibatkan kerugian bagi para penggunanya.

Project ini berfokus untuk memecahkan masalah berikut:

- **Kurangnya Transparansi Data:** Data pasar dari berbagai bursa sering kali sulit diverifikasi dan tidak terstandarisasi.
- **Keterbatasan Akses Data Real-Time:** Banyak pengguna yang tidak memiliki akses mudah ke data yang diperbarui secara real-time.
- **Pengolahan Data Manual yang Memakan Waktu:** Pengolahan data secara manual memperlambat analisis dan pengambilan keputusan.

Kesuksesan project akan diukur menggunakan metrik berikut:

- **Kelengkapan Data:** Berapa banyak pasangan aset yang berhasil diakses melalui API.
- **Kecepatan Respons:** Waktu rata-rata yang dibutuhkan API untuk merespons permintaan data.
- **Kemudahan Penggunaan:** Feedback dari pengguna terkait seberapa mudah mengakses dan memanfaatkan data.
- **Visualisasi:** metrik dapat divisualisasikan dengan tools seperti Looker studio, Tableau, Metabase atau Power BI

Data Platform Understanding



Teknologi



On-premise:

- Programming: Python and SQL
- Container: Docker
- Stream processing: Kafka dan Spark
- Batch processing: Spark
- Workflow orchestration: Airflow
- DBMS: PostgreSQL
- Monitoring: Grafana




Cloud:

- Cloud computing: Google Cloud Platform (GCP)
- Data warehouse: BigQuery
- Transform data: DBT
- Infrastructure as code (IaC): Terraform
- Visualization: Looker Studio

Data Understanding



The background of the slide is a dark, textured surface. In the upper right corner, there is a glowing Bitcoin coin with the word 'BITCOIN' inscribed on it. Below the coin, a faint world map is visible, with a red line tracing a path across it. The overall color scheme is dark with golden-yellow highlights from the Bitcoin coin and the text.

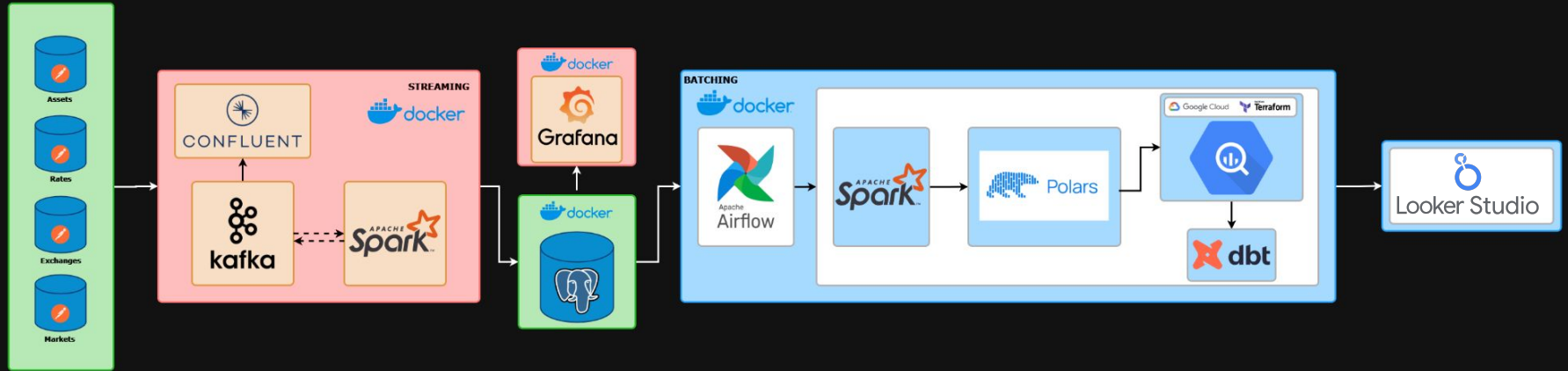
Pada Platform CoinCap API 2.0 menyediakan data terkait harga, kapitalisasi pasar, volume, dan informasi lain tentang cryptocurrency secara real-time. Sebelum melakukan proses extract, penting untuk memahami struktur dan cakupan data yang disediakan oleh API tersebut. CoinCap API memiliki berbagai endpoint untuk mengakses data tertentu:

- **Assets:** Memberikan data tentang cryptocurrency yang tersedia, termasuk harga saat ini, perubahan harga, kapitalisasi pasar, dan volume.
- **Rates:** Menyediakan nilai tukar cryptocurrency ke mata uang fiat atau cryptocurrency lain.
- **Exchanges:** Informasi tentang platform perdagangan cryptocurrency.
- **Markets:** Detail pasar tempat cryptocurrency diperdagangkan.

Transformation & Consideration



Data Architecture



Keterangan:

- Data Sources/Staging Area
- Stream Processing
- Batch Processing

Topic pada Confluent

CONFLUENT

Cluster overview

Brokers

Topics

Connect

ksqlDB

Consumers

Replicators

Cluster settings

Health+ New

markets

Configuration

Messages

Schema

Production

Consumption

Total messages

--

--

10,000

Filter by timestamp, offset, key or value

All partitions

Latest

Max 50 results

40 messages shown

Auto-refresh on

Timestamp	Offset	Partition	Key	Value
1734326229577	9989	0	"	{"exchangeId": "bibox", "rank": "89", "baseSymbol": "MBOX", "baseId": "mobox", "quoteSymbol": "tether", "priceQuote": "0.044483000000000000", "priceUsd": "0.0444528398652432", "volumeUsd24Hr": "770514.1074014368354458", "percentExchangeVolume": "0.3815146140759764", "tradesCount24Hr": null, "updated": 1734315269976}
1734326229576	9988	0	"	{"exchangeId": "bibox", "rank": "88", "baseSymbol": "ONG", "baseId": "ontology-g", "quoteSymbol": "tether", "priceQuote": "0.044483000000000000", "priceUsd": "0.0444528398652432", "volumeUsd24Hr": "770514.1074014368354458", "percentExchangeVolume": "0.3815146140759764", "tradesCount24Hr": null, "updated": 1734315269976}
1734326229575	9987	0	"	{"exchangeId": "bibox", "rank": "87", "baseSymbol": "OMG", "baseId": "omg", "quoteSymbol": "tether", "priceQuote": "0.044483000000000000", "priceUsd": "0.0444528398652432", "volumeUsd24Hr": "770514.1074014368354458", "percentExchangeVolume": "0.3815146140759764", "tradesCount24Hr": null, "updated": 1734315269976}
1734326229574	9985	0	"	{"exchangeId": "bibox", "rank": "85", "baseSymbol": "KSM", "baseId": "kusama", "quoteSymbol": "tether", "priceQuote": "0.044483000000000000", "priceUsd": "0.0444528398652432", "volumeUsd24Hr": "770514.1074014368354458", "percentExchangeVolume": "0.3815146140759764", "tradesCount24Hr": null, "updated": 1734315269976}
1734326229574	9986	0	"	{"exchangeId": "bibox", "rank": "86", "baseSymbol": "ETC", "baseId": "ethereum-c", "quoteSymbol": "tether", "priceQuote": "0.044483000000000000", "priceUsd": "0.0444528398652432", "volumeUsd24Hr": "770514.1074014368354458", "percentExchangeVolume": "0.3815146140759764", "tradesCount24Hr": null, "updated": 1734315269976}

Message details

Timestamp

Offset

Partition

12/16/2024, 12:17:09 PM

9959

0

(1734326229555)

Key

Value

Headers

```

1 {
2   "exchangeId": "bibox",
3   "rank": "59",
4   "baseSymbol": "JST",
5   "baseId": "just",
6   "quoteSymbol": "USDT",
7   "quoteId": "tether",
8   "priceQuote": "0.044483000000000000",
9   "priceUsd": "0.0444528398652432",
10  "volumeUsd24Hr":
11    "770514.1074014368354458",
12    "percentExchangeVolume":
13      "0.3815146140759764",
14    "tradesCount24Hr": null,
15    "updated": 1734315269976
16  }

```

CSV

JSON

How to consume

Spark streaming dan Batch



Spark Master at spark://spark-master:7077

URL: spark://spark-master:7077

Alive Workers: 2

Cores in use: 2 Total, 0 Used

Memory in use: 4.0 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241213152541-172.19.0.7-39949	172.19.0.7:39949	ALIVE	1 (0 Used)	2.0 GiB (0.0 B Used)	
worker-20241213152541-172.19.0.8-33799	172.19.0.8:33799	ALIVE	1 (0 Used)	2.0 GiB (0.0 B Used)	

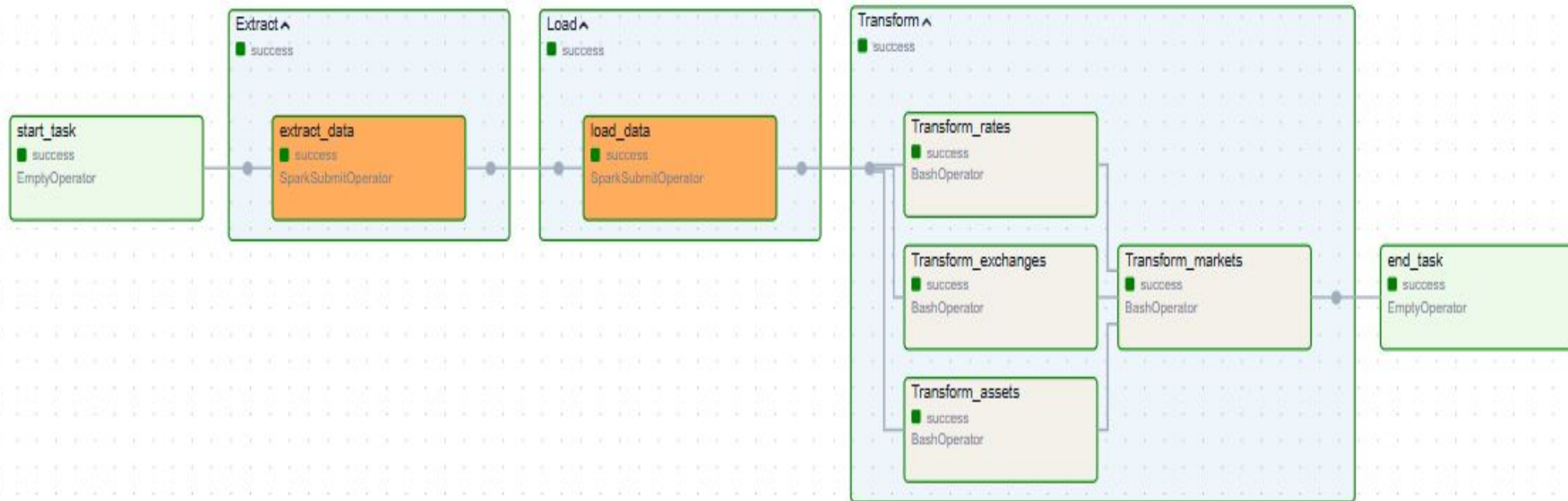
Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

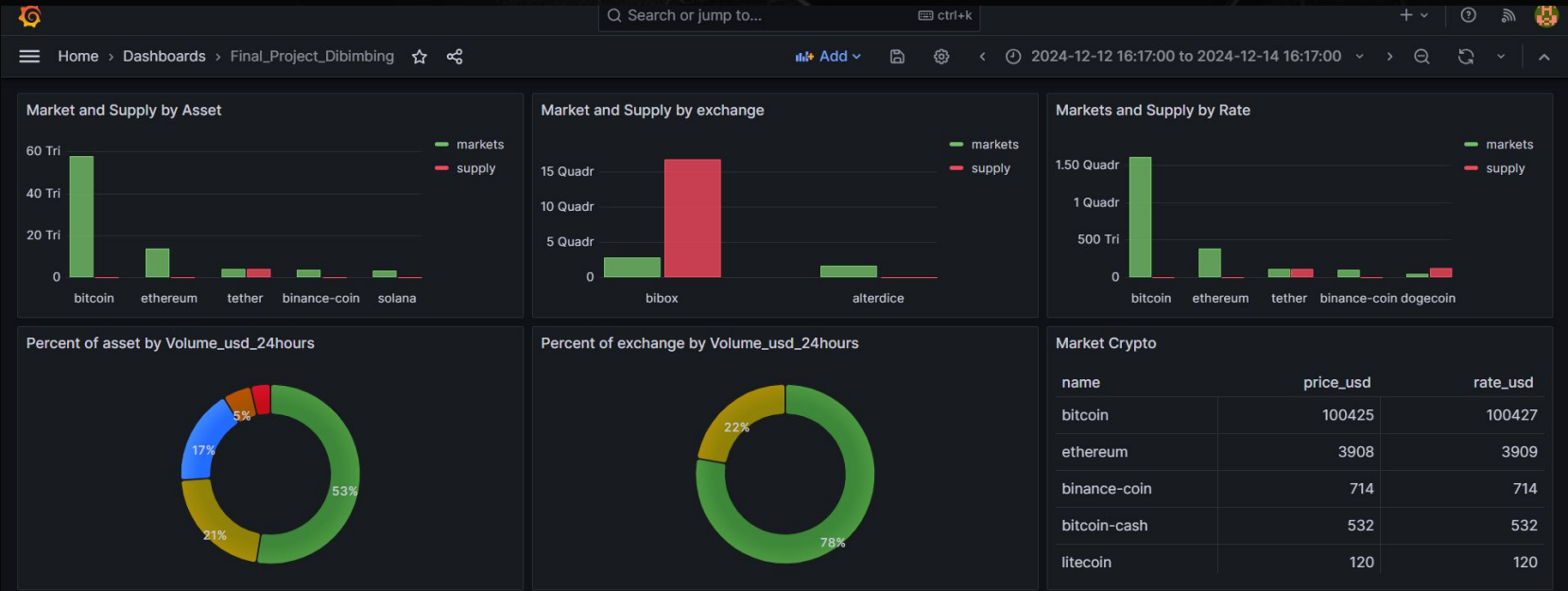
Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

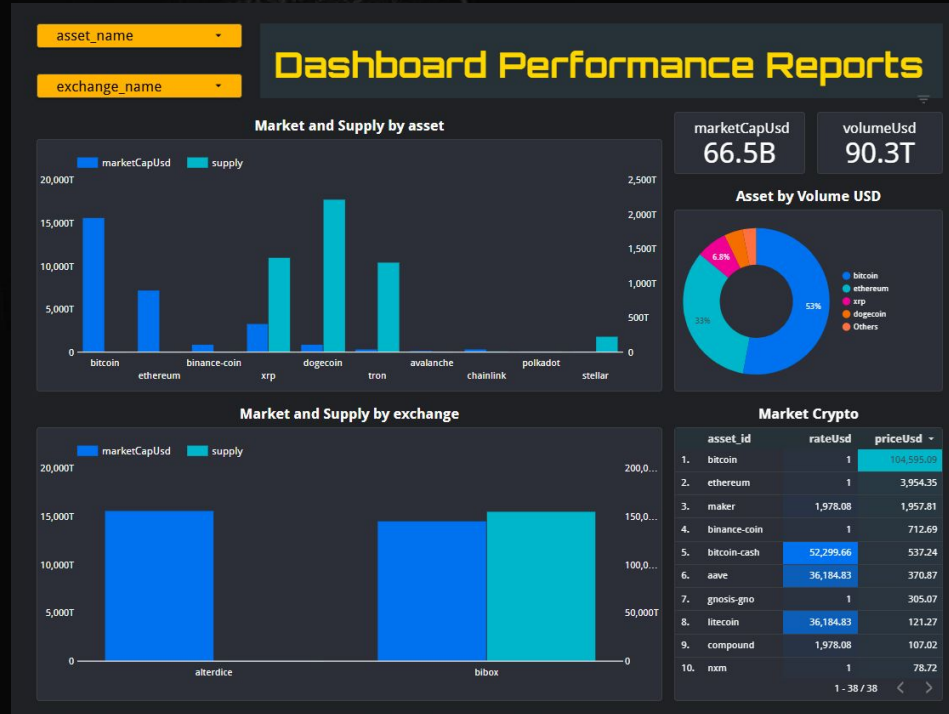
DAGs di Airflow



Dashboard Grafana

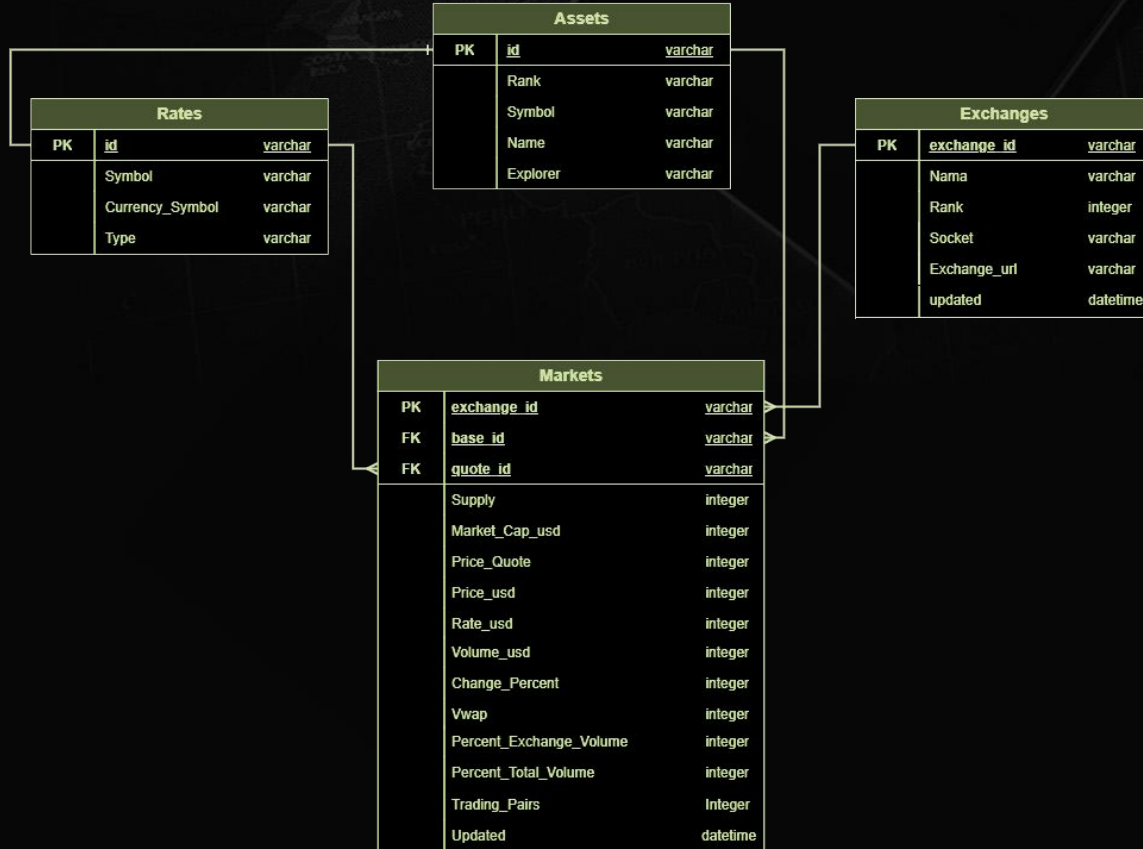


Dashboard Looker Studio



Data Modeling (Business)





Conclusion & Recommendation



Kesimpulan & Rekomendasi:

- Selama pengerjaan project ini, berhasil menjalankan container docker dengan success dan sedikit mengalami kendala saat proses pull.
- Data pipeline streaming dan batch berjalan sesuai rencana dan fungsi ya.
- Data yang digunakan mudah diakses dan dapat memberikan Insight untuk kebutuhan project
- Untuk mendapatkan data yang lebih baik dan akurat disarankan untuk melakukan metode streaming secara berkala.
- Penggunaan aplikasi untuk streaming perlu adanya penataan ulang Konfigurasi arsitektur ya karena seiring berjalan ya waktu akan terasa berat pada komponen komputer dan perlu ditinjau lebih lanjut untuk kedepan ya
- Data pipeline yang digunakan menggabungkan 2 metode stream dan batch, dimana harus dilakukan secara terpisah agar tidak memakan resources terlalu banyak.



**Terima
Kasih**
Contact



SCAN ME