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

Dhoifullah Luth Majied

ibimbing

Project Background

ibimbing

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.

ibimbing

Problem Statement

ibimbing

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

ibimbing

Data Platform Understanding

ibimbing

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

ibimbing

Data Understanding

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:

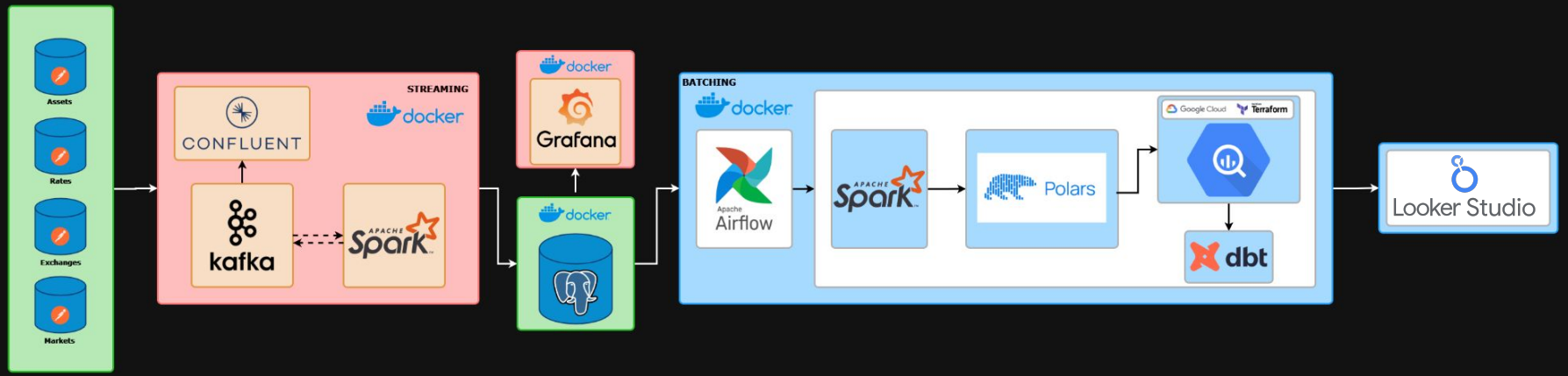
- **Assets:** Memberikan data tentang cryptocurrency yang tersedia, termasuk harga saat ini, perubahan harga, kapitalisasi pasar, dan volume.
- **Rates**: Menyediakan nilai tukar cryptocurrency ke mata uang fiat atau cryptocurrency lain.
- **Exchanges**: Informasi tentang platform perdagangan cryptocurrency.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

ibimbing

Transformation & Consideration

ibimbing

Data Architecture

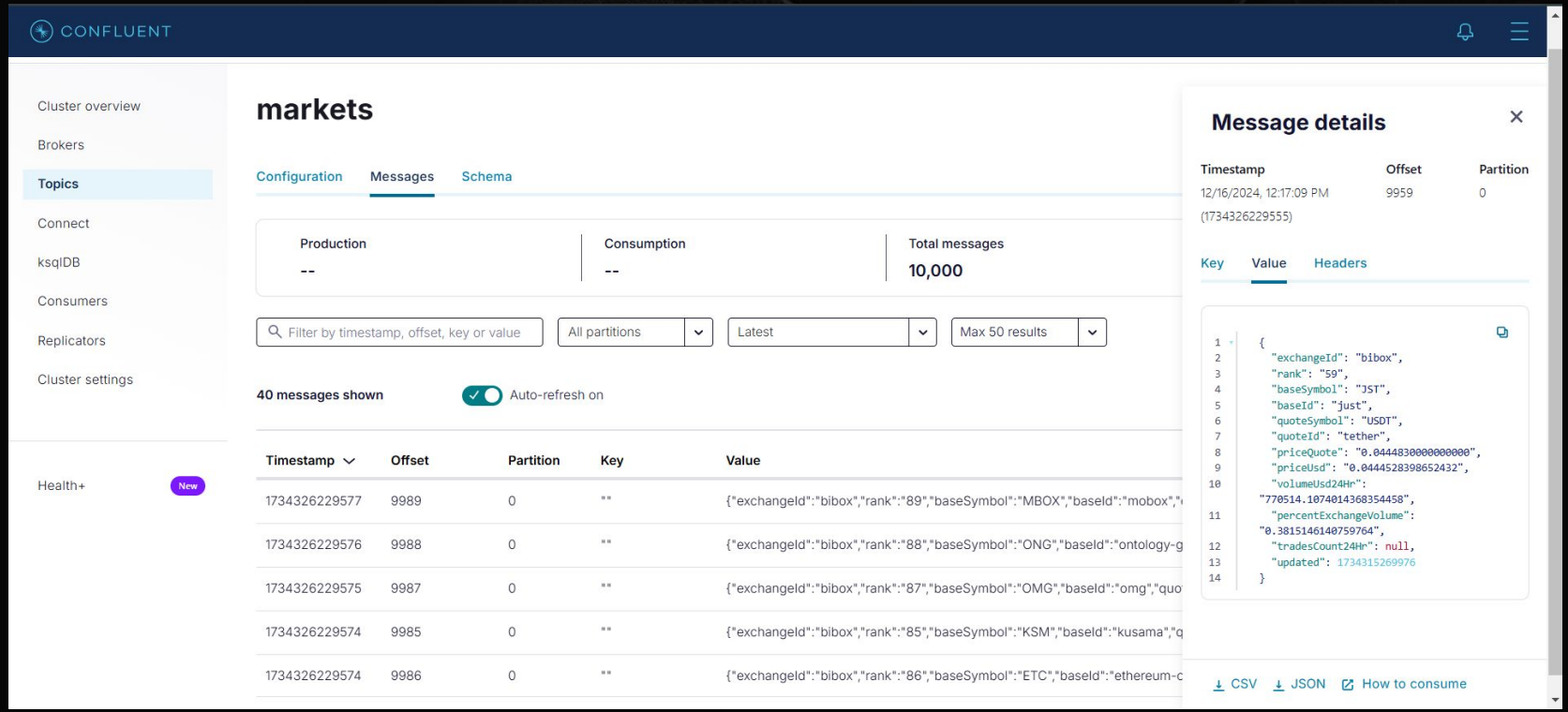


Keterangan:

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

ibimbing

Topic pada Confluent



ibimbing

Spark streaming dan Batch

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

Spark 3.3.2

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)

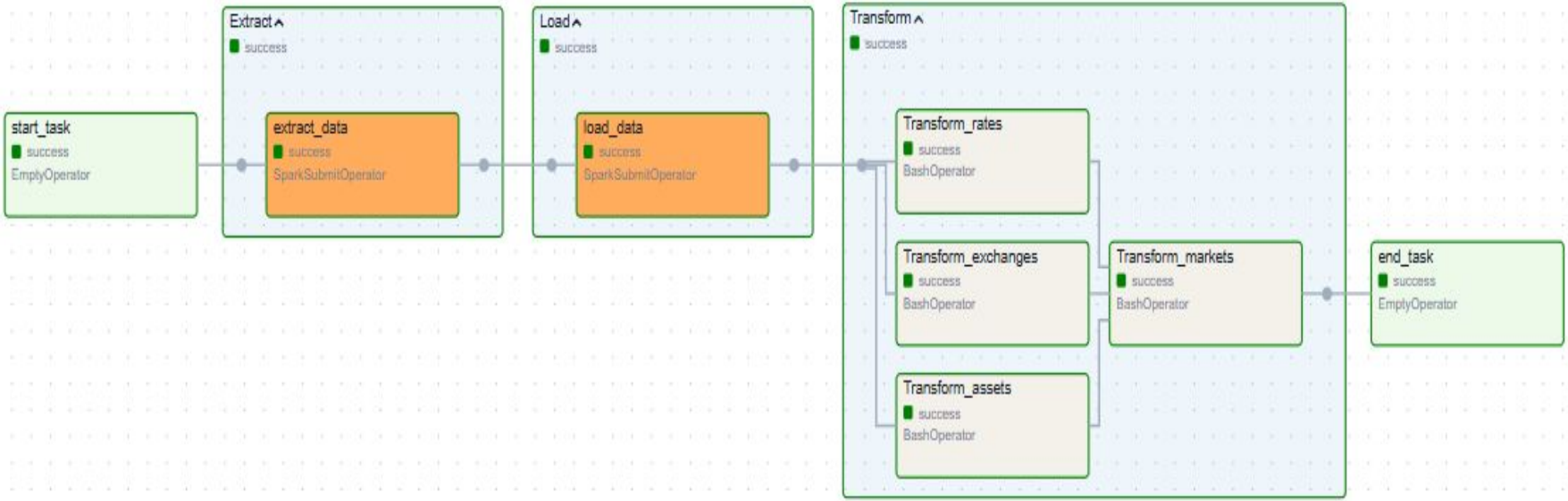
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

Completed Applications (0)

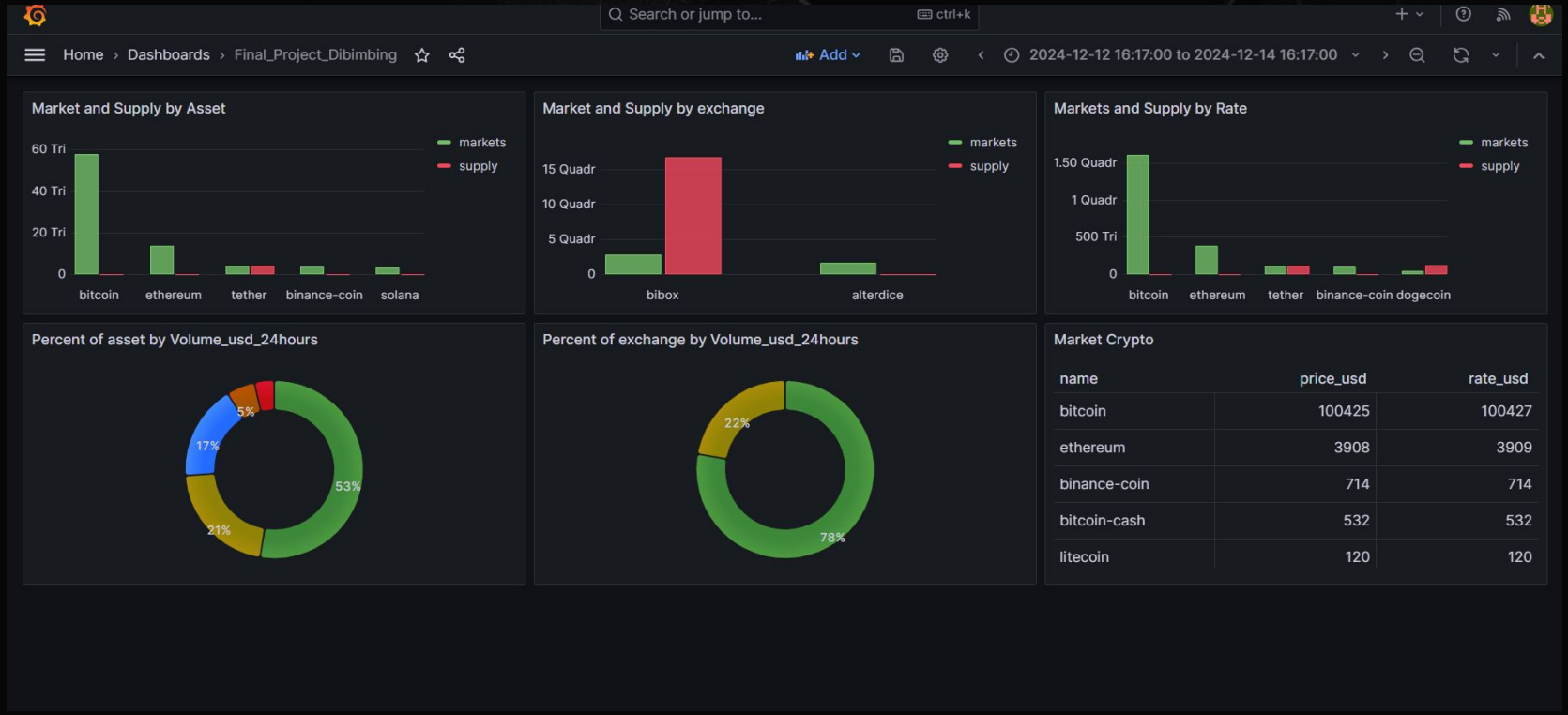
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

ibimbing

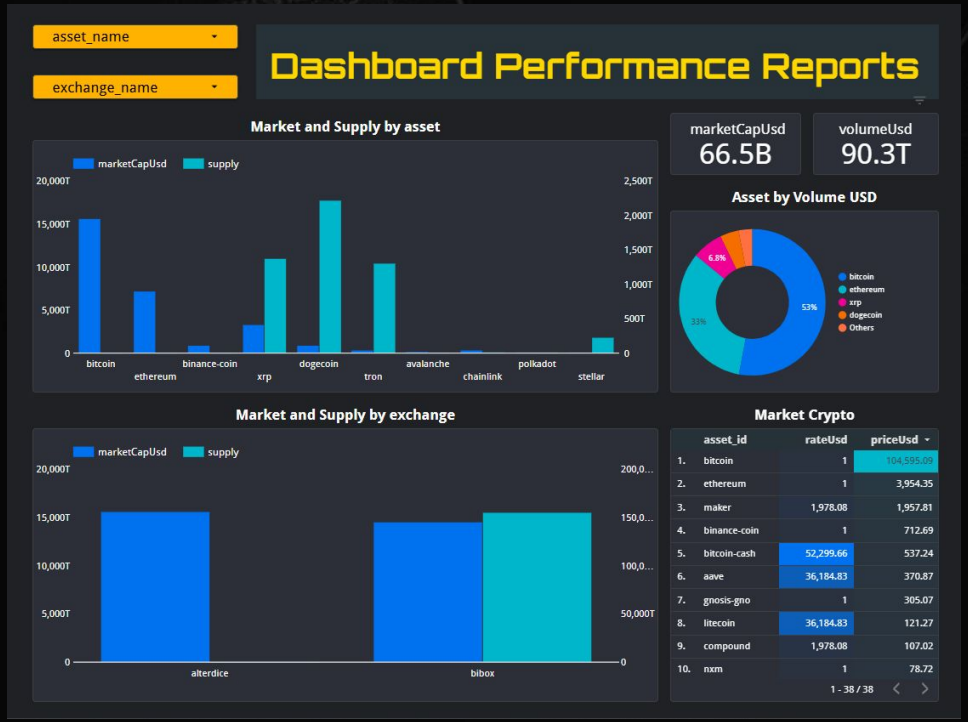
DAGs di Airflow



Dashboard Grafana



Dashboard Looker Studio



https://lookerstudio.google.com/reporting/210d6330-cdad-4e44-a4e3-1708390aa63c

ibimbing

Data Modeling (Business)

ibimbing

Assets

PK	id	varchar
	Rank	varchar
	Symbol	varchar
	Name	varchar
	Explorer	varchar

Rates

PK	id	varchar
	Symbol	varchar
	Currency_Symbol	varchar
	Type	varchar

Exchanges

PK	exchange_id	varchar
	Nama	varchar
	Rank	integer
	Socket	varchar
	Exchange_url	varchar
	updated	datetime

Markets

PK	exchange_id	varchar
FK	base_id	varchar
FK	quote_id	varchar
	Supply	integer
	Market_Cap_usd	integer
	Price_Quote	integer
	Price_usd	integer
	Rate_usd	integer
	Volume_usd	integer
	Change_Percent	integer
	Vwap	integer
	Percent_Exchange_Volume	integer
	Percent_Total_Volume	integer
	Trading_Pairs	Integer
	Updated	datetime

Conclusion & Recommendation

ibimbing

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 applikasi 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.

ibimbing

Terima Kasih

Contact

SCAN ME