BLOCKFLOW

INSA Lyon 5IF
Data Engineering - Prof: Riccardo Tommasini

Ana Luisa Girio Berlingieri
Hannah Schulz
Janis Peter

CONTEXT & PROBLEM

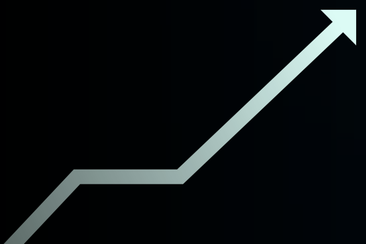
- Blockchain data is large, heterogeneous, and fast-changing
- Exchange, on-chain, and market data differ in structure and frequency
- Joint analysis requires a structured pipeline and temporal alignment

OBJECTIVES

- Build an end-to-end data engineering pipeline
- Integrate Binance, Ethereum, and CoinMarketCap data
- Store data in an analytics-ready data warehouse
- Analyze relationships between price, volume, and gas fees

Questions:

- Price vs Gas Fee: Does ETH price movement correlate with network activity?
- Volume vs Gas Fee: Does trading volume correlate with on-chain congestion?
- Historical Trends: How do correlations evolve over time?



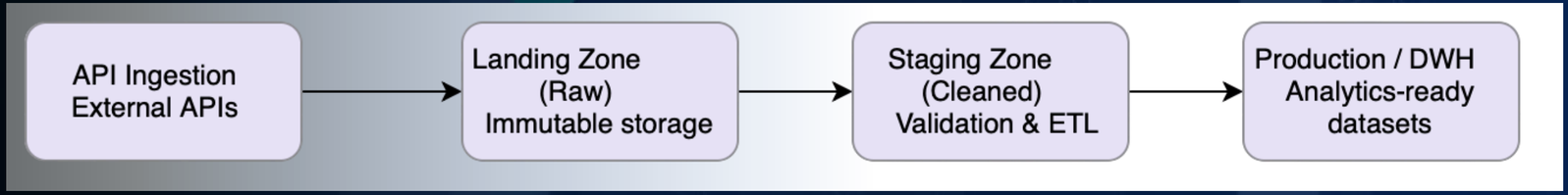
GLOBAL ARCHITECTURE

Pipeline Overview:

- Orchestrated with Apache Airflow
- Four layers:

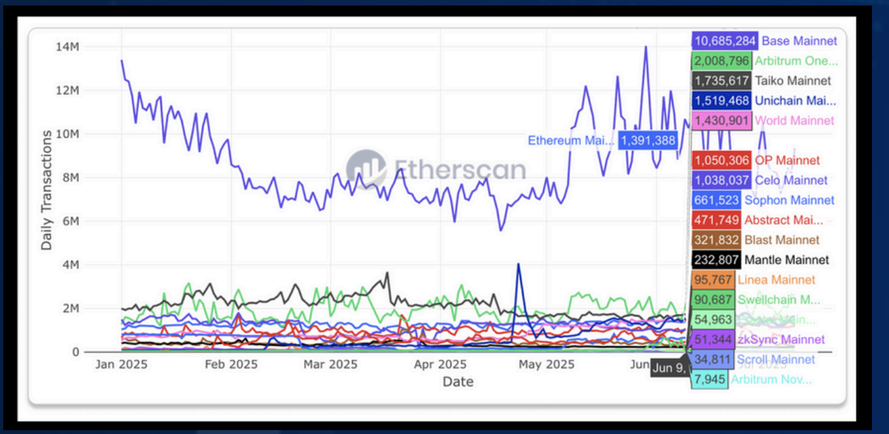
1. API Ingestion
2. Landing Zone
3. Staging Zone
4. Production / Data Warehouse

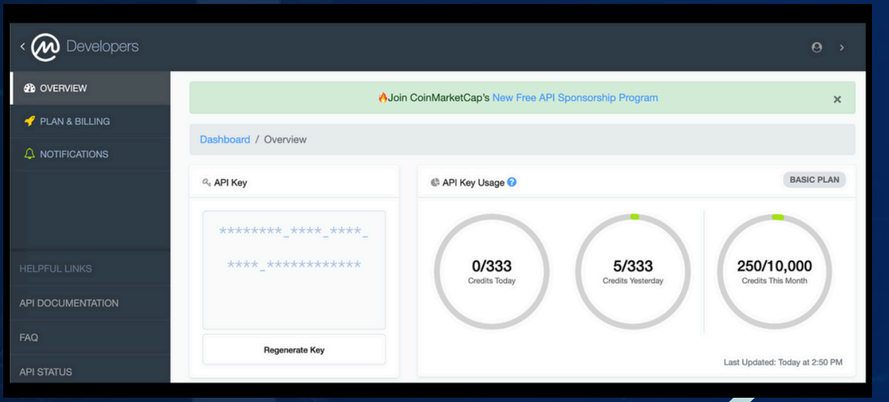
- Dataset-driven dependencies between pipelines



DATA INGESTION & LANDING

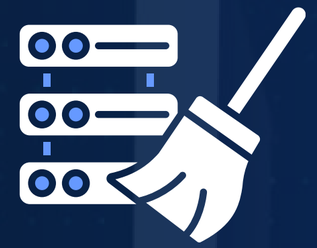
- Binance: OHLCV candlesticks
- Etherscan: Ethereum block data
- CoinMarketCap: market snapshots
- Raw JSON stored in MongoDB and Postgres
- No transformation at this stage





STAGING & DATA WAREHOUSE

- Data cleaning and validation
- Type normalization and hex conversion
- Incremental loading with checkpoints
- Star schema with fact and dimension tables
- Centralized time dimension (UTC)



ANALYTICS PIPELINE

Hourly aggregation of:

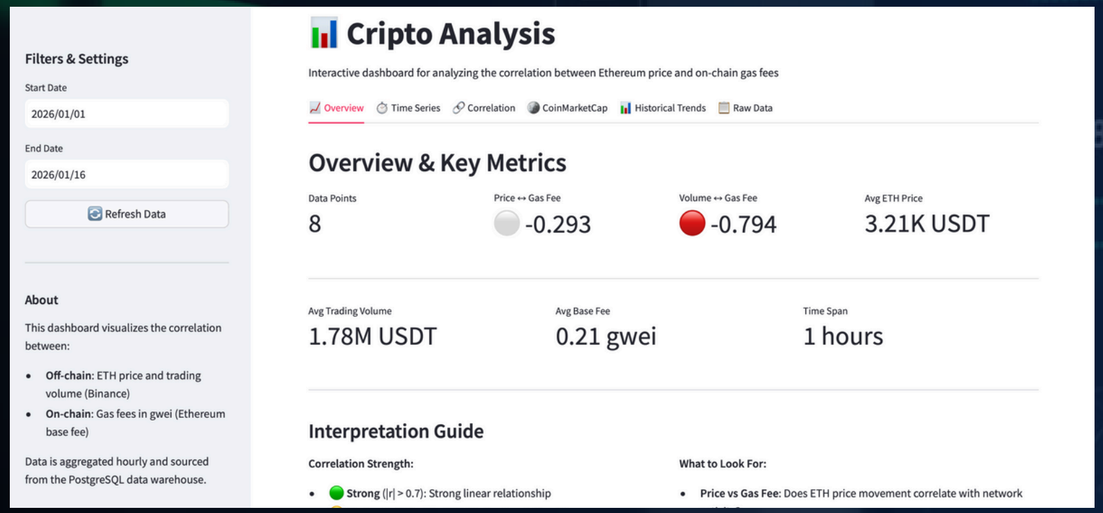
- ETH price
- Trading volume
- Ethereum base gas fee
- Daily integrated analytics across sources
- Pearson correlation analysis

RESULTS

- Weak but positive correlation between volume and gas fees (r ≈ 0.16)
- Gas fees align more with network activity than with price
- Price alone does not explain congestion

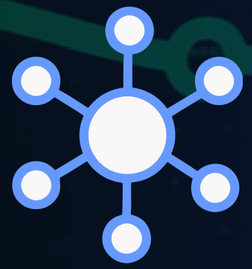
Key insight:

Gas fees are primarily driven by usage, not valuation.



ENGINEERING TAKEAWAYS & CONCLUSION

- Incremental pipelines improve reliability
- Star schema simplifies multi-source analytics
- Temporal alignment is critical in blockchain data
- Airflow enables reproducible analytics workflows



- BlockFlow demonstrates scalable blockchain analytics via data engineering

THANK YOU!