Technical Report

Team Members

Zhumanova Zhanel - 22B030544

Zeinolla Dilnaz – 22B0331177

Olzhas Sengaliyev - 23B151050

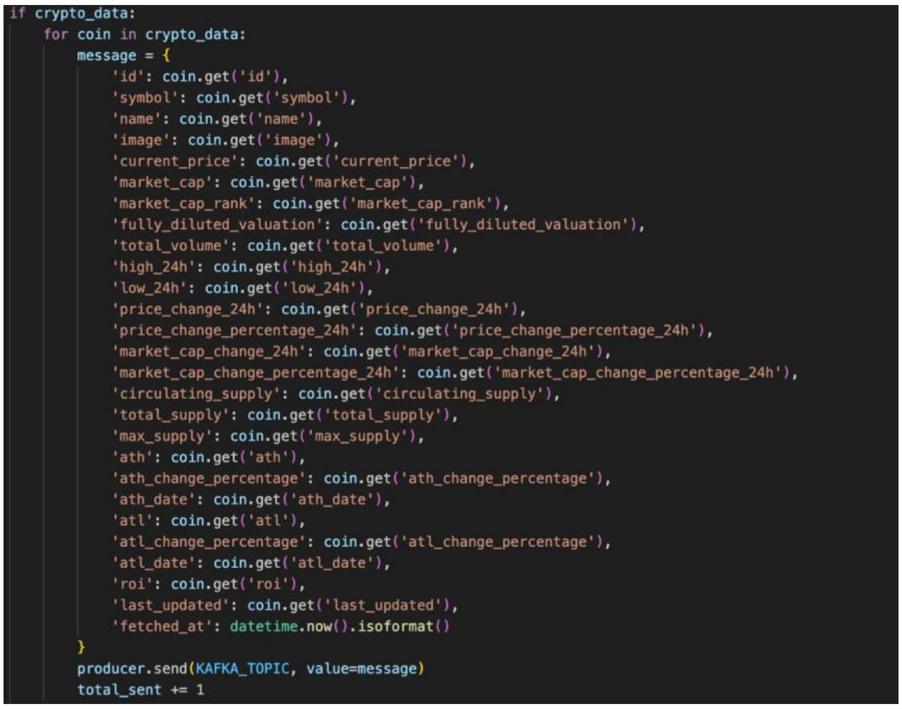
1. API Justification

We selected **CoinGecko API** (endpoint: `https://api.coingecko.com/api/v3/coins/markets`) for our data pipeline. After evaluating several options, we chose cryptocurrency data because it's dynamic , provides rich information (250+ coins with prices, volumes, market caps), has free API access suitable for demo projects, and represents real-world financial data pipelines used in industry.

2. Kafka Topic Schema

Topic: crypto_raw_events | Broker: kafka:29092

Message structure :



```
if crypto_data:
    for coin in crypto_data:
        message = {
            'id': coin.get('id'),
            'symbol': coin.get('symbol'),
            'name': coin.get('name'),
            'image': coin.get('image'),
            'current_price': coin.get('current_price'),
            'market_cap': coin.get('market_cap'),
            'market_cap_rank': coin.get('market_cap_rank'),
            'fully_diluted_valuation': coin.get('fully_diluted_valuation'),
            'total_volume': coin.get('total_volume'),
            'high_24h': coin.get('high_24h'),
            'low_24h': coin.get('low_24h'),
            'price_change_24h': coin.get('price_change_24h'),
            'price_change_percentage_24h': coin.get('price_change_percentage_24h'),
            'market_cap_change_24h': coin.get('market_cap_change_24h'),
            'market_cap_change_percentage_24h': coin.get('market_cap_change_percentage_24h'),
            'circulating_supply': coin.get('circulating_supply'),
            'total_supply': coin.get('total_supply'),
            'max_supply': coin.get('max_supply'),
            'ath': coin.get('ath'),
            'ath_change_percentage': coin.get('ath_change_percentage'),
            'ath_date': coin.get('ath_date'),
            'atl': coin.get('atl'),
            'atl_change_percentage': coin.get('atl_change_percentage'),
            'atl_date': coin.get('atl_date'),
            'roi': coin.get('roi'),
            'last_updated': coin.get('last_updated'),
            'fetched_at': datetime.now().isoformat()
        }
        producer.send(KAFKA_TOPIC, value=message)
        total_sent += 1
```

3. Data Cleaning Rules

We implemented six Pandas-based cleaning operations:

1. Drop incomplete records- Remove entries missing critical fields

2. Fill missing numerics - Replace NaN values with 0 for fields like `market_cap`, `total_volume`

3. Standardize text - Normalize coin IDs to lowercase, symbols to UPPERCASE

4. Convert timestamps- Parse timestamp strings to datetime objects for `last_updated`, `ath_date`, `atl_date`

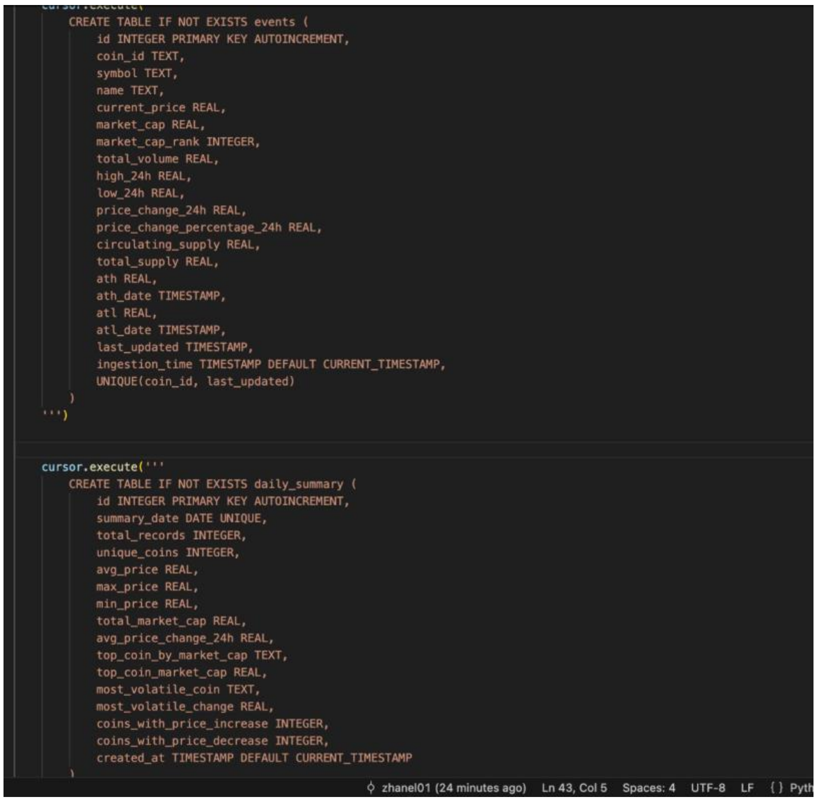
5. Remove duplicates - Sort by timestamp and keep latest version of each `(coin_id, last_updated)` pair

6. Validate prices - Filter negative prices, prices over $1B, and negative market cap ranks

4. SQLite Database Schema

Table 1: events (stores cleaned cryptocurrency data)

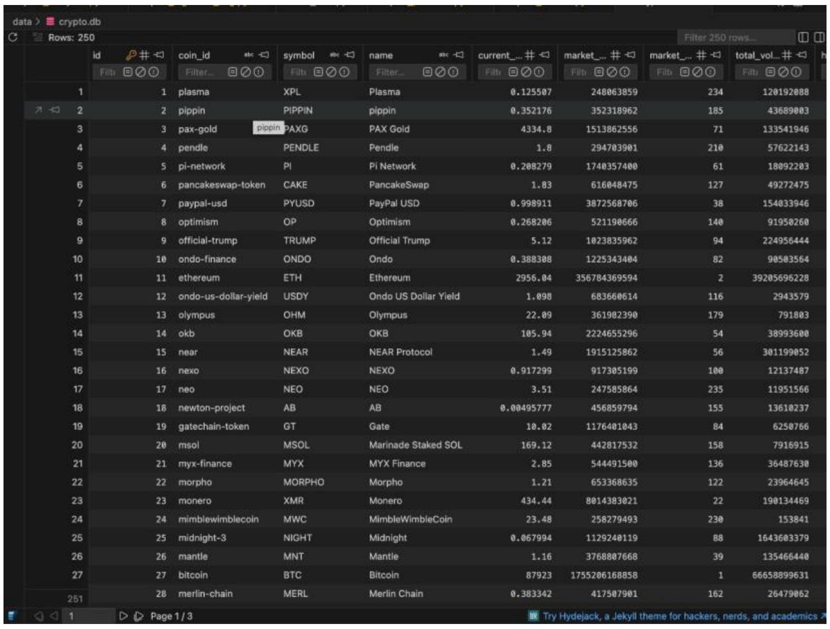
Table 2: daily_summary



```
    CREATE TABLE IF NOT EXISTS events (
        id INTEGER PRIMARY KEY AUTOINCREMENT,
        coin_id TEXT,
        symbol TEXT,
        name TEXT,
        current_price REAL,
        market_cap REAL,
        market_cap_rank INTEGER,
        total_volume REAL,
        high_24h REAL,
        low_24h REAL,
        price_change_24h REAL,
        price_change_percentage_24h REAL,
        circulating_supply REAL,
        total_supply REAL,
        ath REAL,
        ath_date TIMESTAMP,
        atl REAL,
        atl_date TIMESTAMP,
        last_updated TIMESTAMP,
        ingestion_time TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
        UNIQUE(coin_id, last_updated)
    )
''')

cursor.execute('''
    CREATE TABLE IF NOT EXISTS daily_summary (
        id INTEGER PRIMARY KEY AUTOINCREMENT,
        summary_date DATE UNIQUE,
        total_records INTEGER,
        unique_coins INTEGER,
        avg_price REAL,
        max_price REAL,
        min_price REAL,
        total_market_cap REAL,
        avg_price_change_24h REAL,
        top_coin_by_market_cap TEXT,
        top_coin_market_cap REAL,
        most_volatile_coin TEXT,
        most_volatile_change REAL,
        coins_with_price_increase INTEGER,
        coins_with_price_decrease INTEGER,
        created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
    )
```

```
zhanel01 (24 minutes ago)   Ln 43, Col 5   Spaces: 4   UTF-8   LF   { } Pyth
```



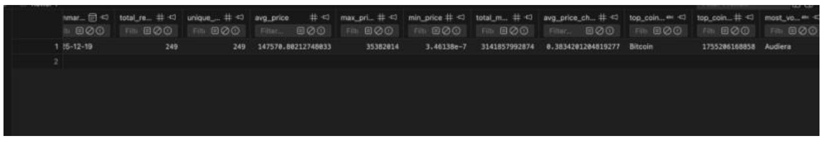
data > crypto.db

Rows: 250

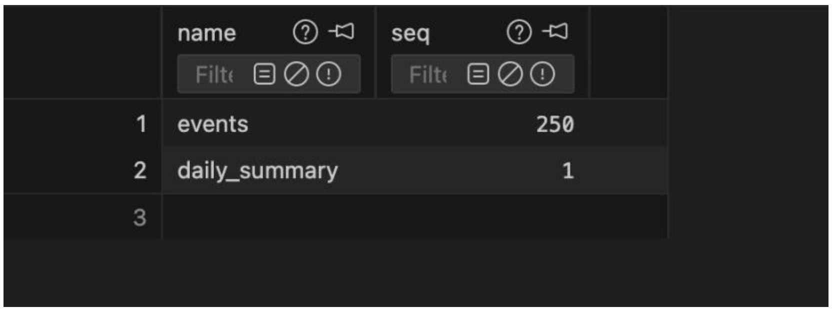
	id	coin_id	symbol	name	current_...	market_...	market_...	total_vol...
1	1	plasma	XPL	Plasma	0.125507	248063859	234	120192088
2	2	pippin	PIPPIN	pippin	0.352176	352318962	185	43689003
3	3	pax-gold	PAXG	PAX Gold	4334.8	1513862556	71	133541946
4	4	pendle	PENDLE	Pendle	1.8	294703901	210	57622143
5	5	pi-network	PI	Pi Network	0.208279	1740357400	61	18092203
6	6	pancakeswap-token	CAKE	PancakeSwap	1.83	616048475	127	49272475
7	7	paypal-usd	PYUSD	PayPal USD	0.998911	3872568706	38	154033946
8	8	optimism	OP	Optimism	0.268206	521190666	140	91950260
9	9	official-trump	TRUMP	Official Trump	5.12	1023835962	94	224956444
10	10	ondo-finance	ONDO	Ondo	0.388308	1225343404	82	90503564
11	11	ethereum	ETH	Ethereum	2956.04	356784369594	2	39205696228
12	12	ondo-us-dollar-yield	USDY	Ondo US Dollar Yield	1.098	683660614	116	2943579
13	13	olympus	OHM	Olympus	22.09	361982390	179	791803
14	14	okb	OKB	OKB	105.94	2224655296	54	38993600
15	15	near	NEAR	NEAR Protocol	1.49	1915125862	56	301199052
16	16	nexo	NEXO	NEXO	0.917299	917305199	100	12137487
17	17	neo	NEO	NEO	3.51	247585864	235	11951566
18	18	newton-project	AB	AB	0.00495777	456859794	155	13610237
19	19	gatechain-token	GT	Gate	10.02	1176401043	84	6250766
20	20	msol	MSOL	Marinade Staked SOL	169.12	442817532	158	7916915
21	21	myx-finance	MYX	MYX Finance	2.85	544491500	136	36487630
22	22	morpho	MORPHO	Morpho	1.21	653368635	122	23964645
23	23	monero	XMR	Monero	434.44	8014383021	22	190134469
24	24	mimblewimblecoin	MWC	MimbleWimbleCoin	23.48	258279493	230	153841
25	25	midnight-3	NIGHT	Midnight	0.067994	1129240119	88	1643603379
26	26	mantle	MNT	Mantle	1.16	3768807668	39	135466440
27	27	bitcoin	BTC	Bitcoin	87923	1755206168858	1	66658899631
251	28	merlin-chain	MERL	Merlin Chain	0.383342	417507901	162	26479062

Page 1 / 3

Try Hydejack, a Jekyll theme for hackers, nerds, and academics



	nmar...	total_re...	unique_...	avg_price	max_pri...	min_price	total_m...	avg_price_ch...	top_coin...	top_coin...	most_vo...
1	25-12-19	249	249	147570.80212748033	35382014	3.46138e-7	3141857992874	0.3834201204819277	Bitcoin	1755206168858	Audiera
2											

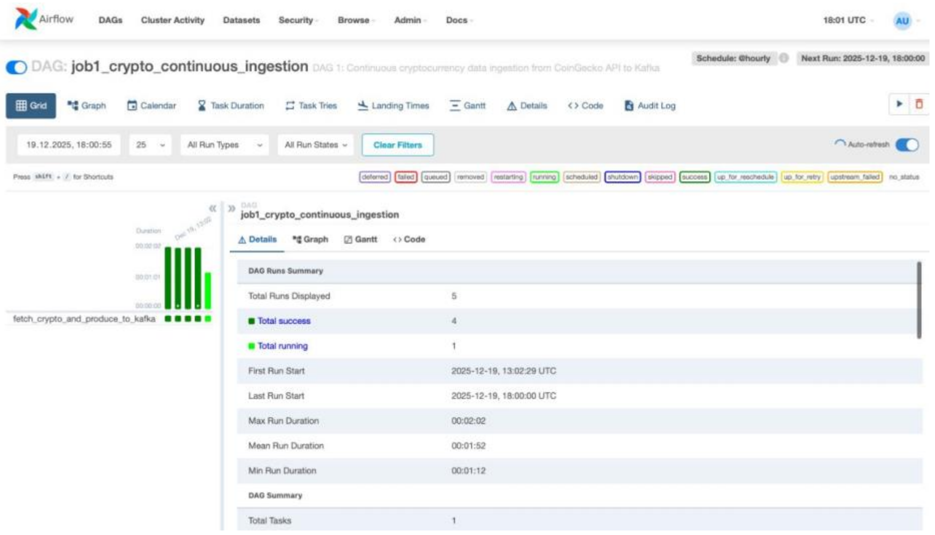


	name	seq
1	events	250
2	daily_summary	1
3		

5. Pipeline Architecture & DAGs

DAG 1: job1_crypto_continuous_ingestion (1-2 minute)

- Fetches crypto data for 2 minutes, sends around 250 messages to Kafka



DAG: job1_crypto_continuous_ingestion DAG 1: Continuous cryptocurrency data ingestion from CoinGecko API to Kafka

Schedule: @hourly

Next Run: 2025-12-19, 18:00:00

19.12.2025, 18:00:55 | 25

fetch_crypto_and_produce_to_kafka

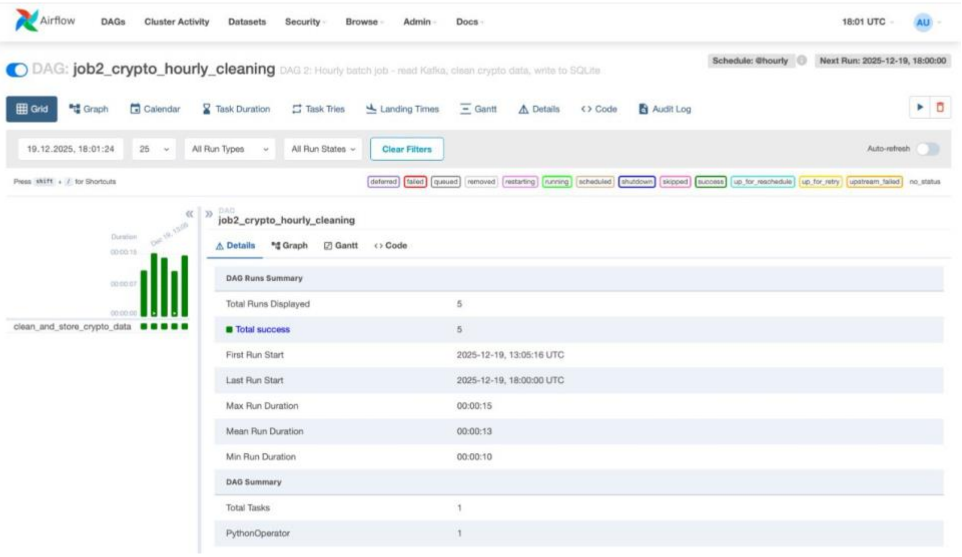
DAG
job1_crypto_continuous_ingestion

Details | Graph | Gantt | Code

DAG Runs Summary	
Total Runs Displayed	5
Total success	4
Total running	1
First Run Start	2025-12-19, 13:02:29 UTC
Last Run Start	2025-12-19, 18:00:00 UTC
Max Run Duration	00:02:02
Mean Run Duration	00:01:52
Min Run Duration	00:01:12
DAG Summary	
Total Tasks	1

DAG 2: job2_crypto_hourly_cleaning (hourly)

- Consumes Kafka messages, applies cleaning rules, inserts into events table



DAG: job2_crypto_hourly_cleaning DAG 2: Hourly batch job - read Kafka, clean crypto data, write to SQLite

Schedule: @hourly

Next Run: 2025-12-19, 18:00:00

19.12.2025, 18:01:24 | 25

clean_and_store_crypto_data

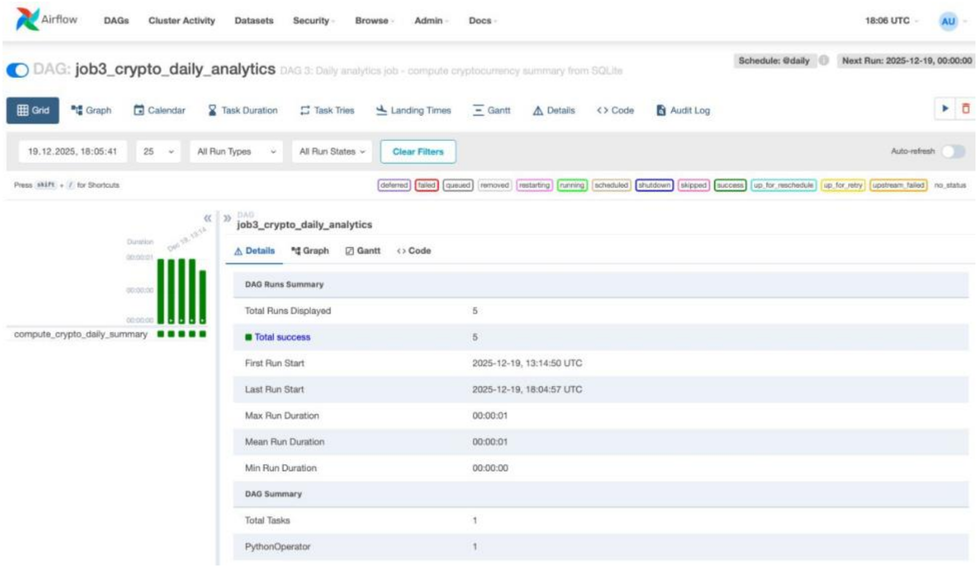
DAG
job2_crypto_hourly_cleaning

Details | Graph | Gantt | Code

DAG Runs Summary	
Total Runs Displayed	5
Total success	5
First Run Start	2025-12-19, 13:05:16 UTC
Last Run Start	2025-12-19, 18:00:00 UTC
Max Run Duration	00:00:15
Mean Run Duration	00:00:13
Min Run Duration	00:00:10
DAG Summary	
Total Tasks	1
PythonOperator	1

DAG 3: job3_crypto_daily_analytics (daily)

- Reads previous day's records, computes statistics, stores in daily_summary



DAG: job3_crypto_daily_analytics DAG 3: Daily analytics job - compute cryptocurrency summary from SQLite

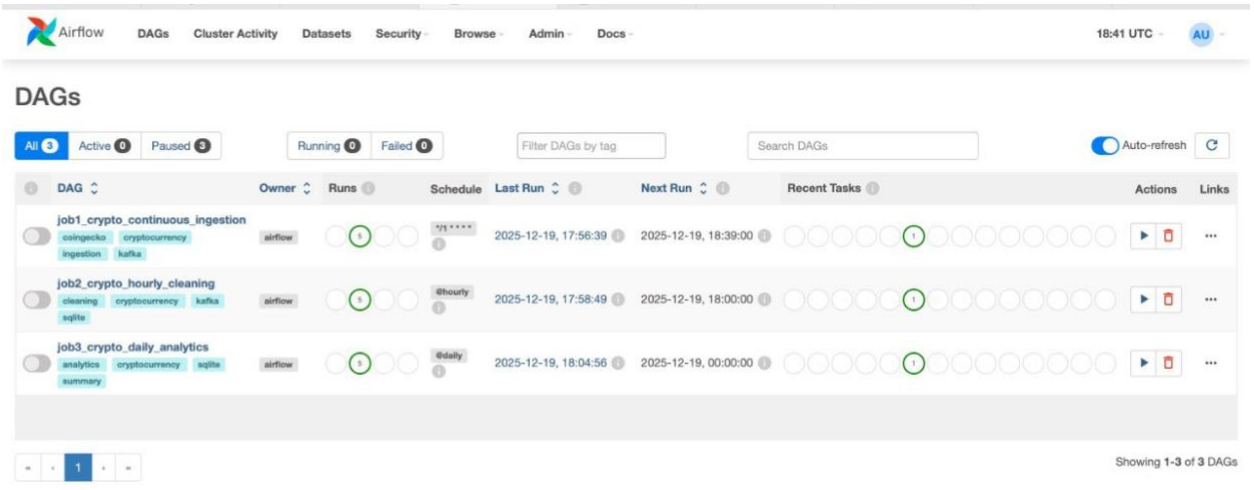
Schedule: @daily | Next Run: 2025-12-19, 00:00:00

19.12.2025, 18:05:41 | 25 | All Run Types | All Run States | Clear Filters

DAG
job3_crypto_daily_analytics

DAG Runs Summary	
Total Runs Displayed	5
🟩 Total success	5
First Run Start	2025-12-19, 13:14:50 UTC
Last Run Start	2025-12-19, 18:04:57 UTC
Max Run Duration	00:00:01
Mean Run Duration	00:00:01
Min Run Duration	00:00:00
DAG Summary	
Total Tasks	1
PythonOperator	1

6. Screenshots & Verification



DAGs

All 3 | Active 0 | Paused 3 | Running 0 | Failed 0 | Filter DAGs by tag | Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
job1_crypto_continuous_ingestion coingecko cryptocurrency ingestion kafka	airflow	5	*/1 * * * *	2025-12-19, 17:56:39	2025-12-19, 18:39:00	1		...
job2_crypto_hourly_cleaning cleaning cryptocurrency kafka sqlite	airflow	5	@hourly	2025-12-19, 17:58:49	2025-12-19, 18:00:00	1		...
job3_crypto_daily_analytics analytics cryptocurrency sqlite summary	airflow	5	@daily	2025-12-19, 18:04:56	2025-12-19, 00:00:00	1		...

Showing 1-3 of 3 DAGs