

CRYPTO ETL PIPELINE

INTEGRANTES

Ronald Chipana Wariste
Luz Alizon Mamani Mena
Roni Edwin Oyardo Acuña
Ever Alcides Soto Palli



INTRODUCCIÓN

Contexto

Este proyecto se enmarca en el análisis integral del mercado de criptomonedas, un sector financiero que opera continuamente (24/7), se caracteriza por su naturaleza descentralizada y una alta volatilidad.

El objetivo principal fue construir un sistema automatizado que consolide la información proveniente de múltiples fuentes de datos sobre criptomonedas (API, Kaggle).



INGESTA Y LIMPIEZA DE LOS DATOS

Dataset histórico de Kaggle:

Representa una instantánea temporal del mercado de criptomonedas en un momento específico no documentado.

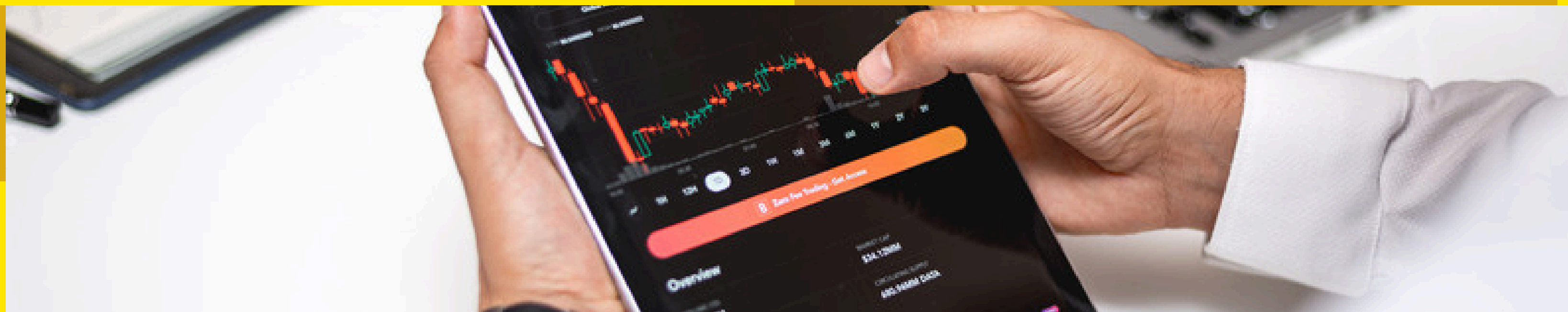
	Rank	Coin Name	Symbol	Price	1h	24h	7d	30d	24h Volume	Circulating Supply	Total Supply	Market Cap
0	1	Bitcoin	BTC	36,456.94	0.40%	-1.70%	1.00%	18.40%	\$22,801,222,945.00	19,549,806	21 Million	\$712,726,163,003.00
1	2	Ethereum	ETH	2,027.60	0.50%	1.40%	1.00%	20.70%	\$26,845,710,464.00	120,249,015	120 Million	\$243,488,187,281.00
2	3	Tether	USDT	1.00	0.10%	-0.30%	-0.10%	-0.10%	\$47,122,466,339.00	88,308,652,879	88.3 Billion	\$88,027,617,310.00
3	4	BNB	BNB	231.63	-0.10%	-12.60%	-8.00%	5.40%	\$3,715,265,116.00	153,856,150	154 Million	\$35,716,332,862.00
4	5	XRP	XRP	0.59	0.10%	-1.90%	-6.90%	12.10%	\$1,339,890,506.00	53,718,306,475	100 Billion	\$31,863,926,051.00

Datos de una API – CoinGecko:

Representa la fuente de datos dinámicos del sistema, proporcionando información actualizada que captura la naturaleza volátil del mercado crypto.

	symbol	current_price	price_change_percentage_24h	market_cap	total_volume	high_24h	low_24h
0	btc	120467.00	1.79513	2400721293624	7.100968e+10	121044.000	118343.00
1	eth	4475.03	3.14134	540155048597	4.432810e+10	4514.810	4334.97
2	xrp	3.04	3.17250	181914498259	6.804073e+09	3.100	2.94
3	usdt	1.00	0.00503	175814181174	1.326874e+11	1.001	1.00
4	bnb	1087.61	6.00597	151462569006	2.626036e+09	1097.500	1023.46





LIMPIEZA

Datos historicos

Se transformó datos con formatos monetarios, notaciones textuales y símbolos en datasets optimizados para procesamiento computacional, manteniendo simultáneamente la integridad semántica de la información. Los valores nulos se imputaron utilizando la mediana para valores numéricos y la moda para datos categóricos.

Datos de la API

En cuanto a este dataset, se hizo la limpieza. Pero además, se tomó en cuenta solo a las 100 criptomonedas con mayor capitalización de mercado, aplicando el Principio de Pareto, que representa aproximadamente el 85% del valor total del mercado para asegurar eficiencia.

ETIQUETADO Y CLASIFICACIÓN DE TENDENCIAS

Datos atípicos

En la fase de calidad de datos, se identificaron valores atípicos (outliers); sin embargo, se tomó la decisión estratégica de no eliminar estos registros. Etiquetandolos como: "is_outlier"

Clasificación de Tendencia

Se etiquetaron los datos de la API para determinar la tendencia de los precios en las últimas 24 horas.

Resumen de clasificación:

Tendencia moderada alcista: 48 criptomonedas
Tendencia estable: 31 criptomonedas
Tendencia fuerte alcista: 14 criptomonedas
Tendencia moderada bajista: 5 criptomonedas
Tendencia fuerte bajista: 2 criptomonedas



CONSTRUCCIÓN DEL PIPELINE

Google Colaboratory

Antes de implementar la solución en Airflow, se desarrolló un prototipo comprehensivo en Google Colab que permitió validar las hipótesis de procesamiento y refinar los algoritmos en un entorno de iteración rápida.

Migración a Airflow

Presentó desafíos técnicos significativos:

- Complejidad de configuración y el levantamiento del programa:
- Conflictos entre versiones de pandas, numpy y Airflow
- Acceso a sistemas de archivos y variables de entorno
- Configuración óptima de paralelismo y recursos de la computadora.



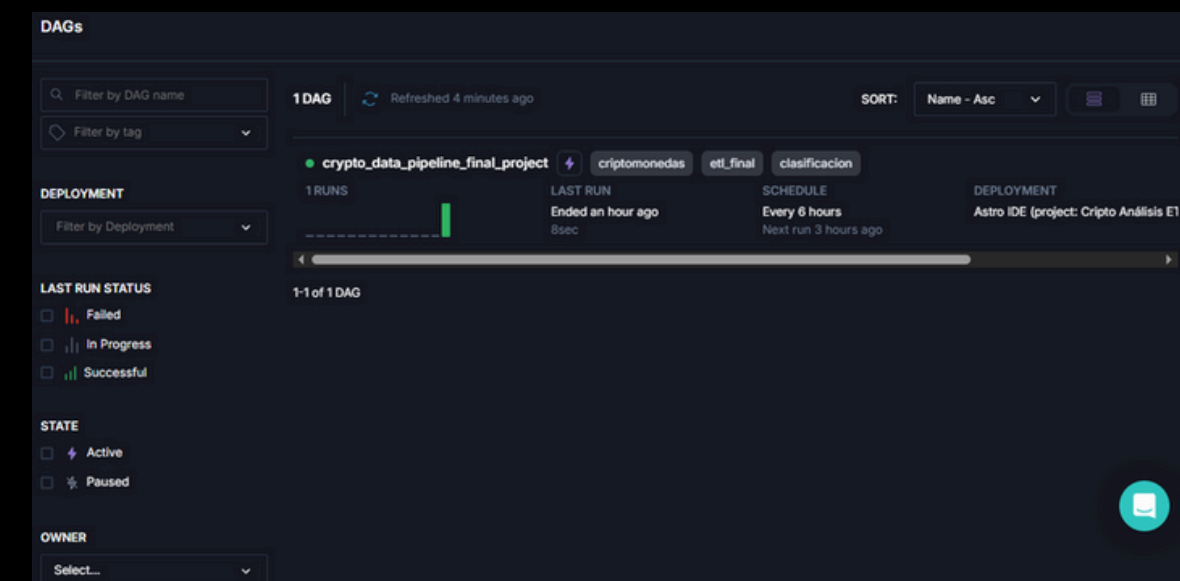
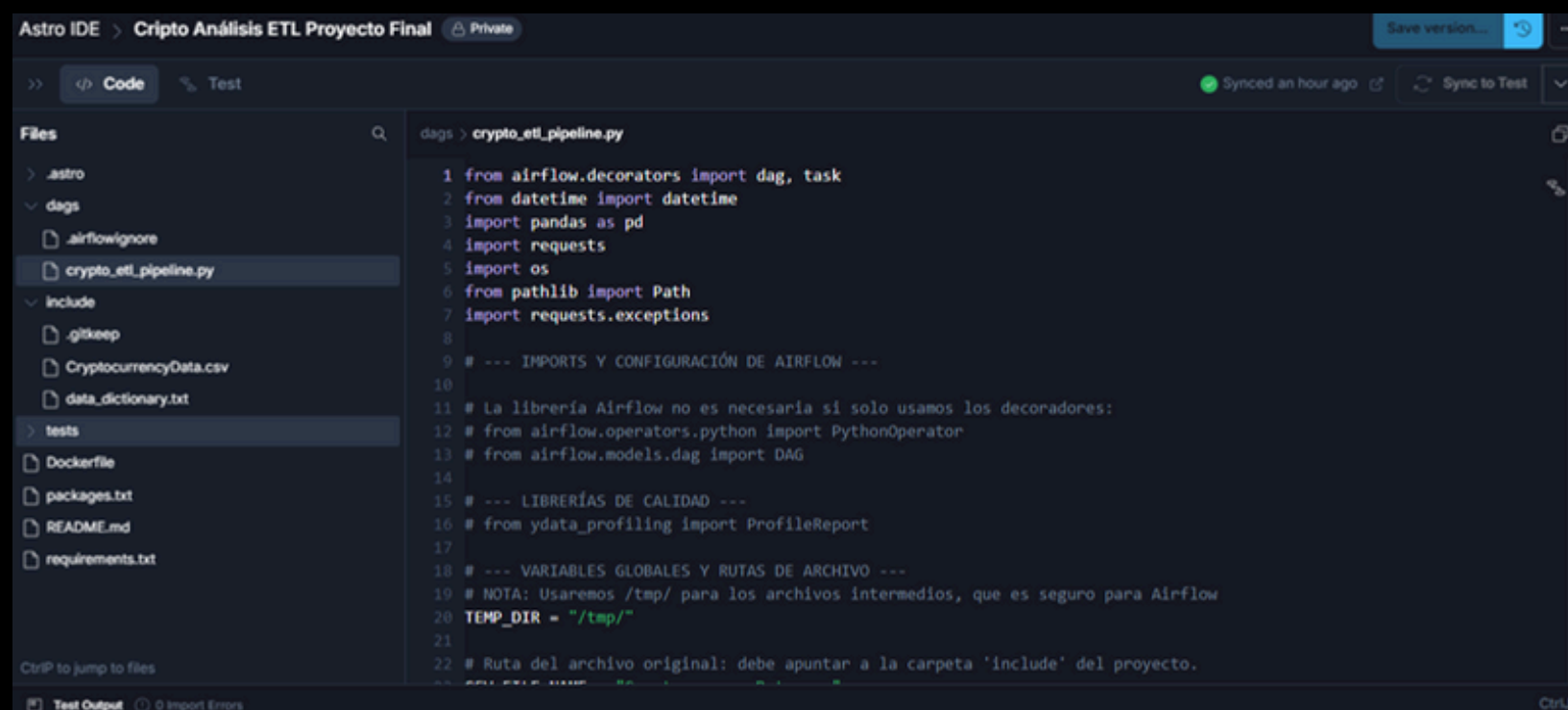
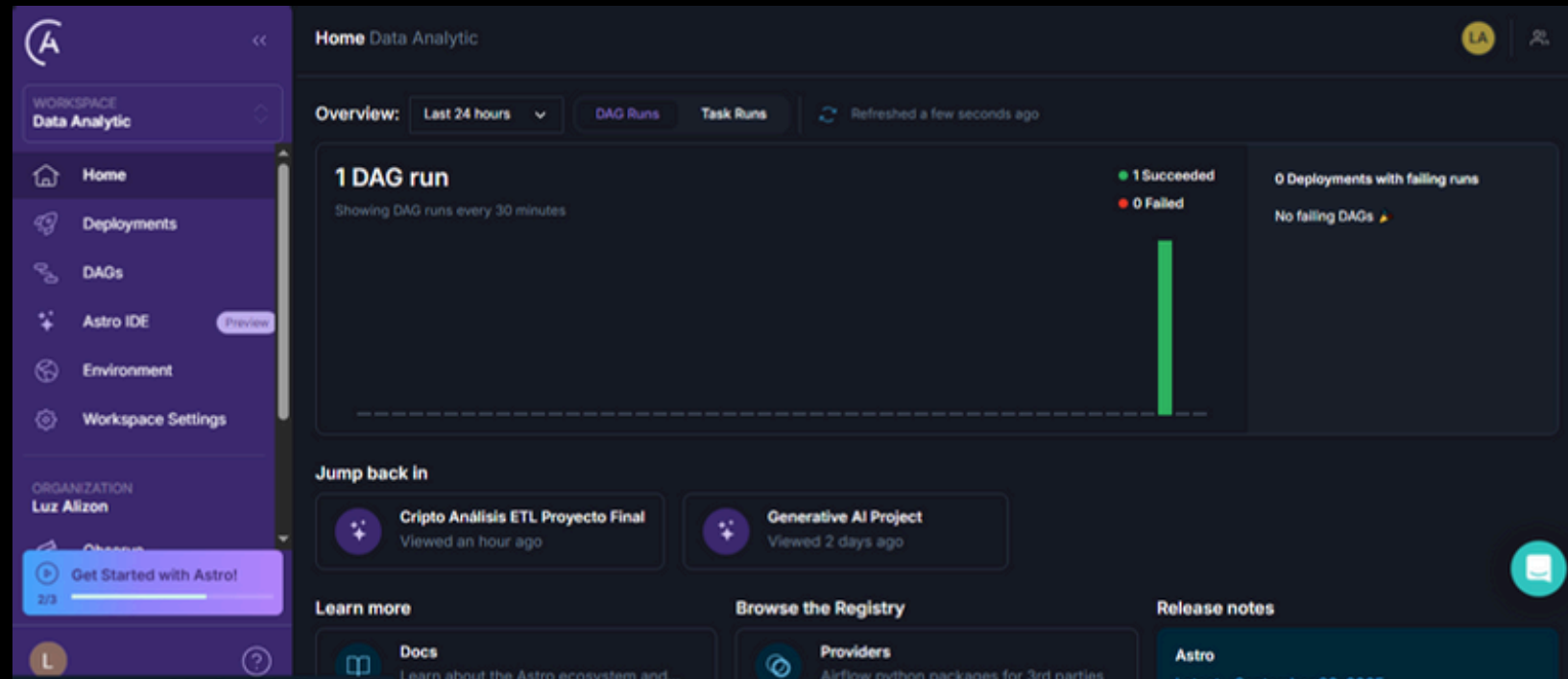
ASTRONOMER



Algunas limitaciones

A pesar de las limitaciones operativas impuestas por la capa gratuita de la plataforma Cloud (Astro/Airflow), se logró construir y validar la funcionalidad principal del pipeline.

- Descarga directa de archivos finales
- Generación de reportes detallados de calidad



El desarrollo demostró un dominio completo sobre la orquestación de Airflow, la conexión a múltiples fuentes de datos y la correcta implementación de la lógica ETL para la puesta en marcha en un entorno local.



AIRFLOW

Una vez guardado lo realizado en Astro, esta herramienta se conecta con “Airflow” para la ejecución del flujo de las tareas.

The screenshot shows the Apache Airflow web interface. The left sidebar contains navigation links: Inicio, Dags, Assets, Navegar, Administración, Docs, and Usuario. The main panel displays a list of DAGs. The 'crypto_data_pipeline_final_project' DAG is selected, showing its status as 'Exitoso' (Successful) and its last execution time as '2025-10-03 21:46:42'.

ID del Dag	Programación	Siguiente Ejecución	Última Ejecución	Etiquetas
crypto_data_pipeline_final_project	0 */6 * * *	2025-10-03 20:00:00	2025-10-03 21:46:42	criptomonedas, etl_final, clasificacion

The screenshot shows the Apache Airflow web interface displaying the execution details of the 'crypto_data_pipeline_final_project' DAG. The DAG is shown as 'Exitoso' (Successful) and the execution time is '2025-10-03 22:49:53'. The tasks listed are: clean_historical_data, ingest_api_data, classify_api_data, download_final_data, and generate_data_quality_reports.

ID de la Tarea	Mapa de Índice	Estado	Fecha Inicial	Fecha Final
clean_historical_data		Exitoso	2025-10-03 22:49:57	2025-10-03 22:50:01
ingest_api_data		Exitoso	2025-10-03 22:49:57	2025-10-03 22:50:01
classify_api_data		Exitoso	2025-10-03 22:49:59	2025-10-03 22:50:01
download_final_data		Exitoso	2025-10-03 22:49:57	2025-10-03 22:50:01
generate_data_quality_reports		Exitoso	2025-10-03 22:49:57	2025-10-03 22:50:01

Task del DAG

Tasks del DAG

1. **ingest_api_data** - Extracción datos API CoinGecko
2. **clean_historical_data** - Limpieza dataset histórico Kaggle
3. **classify_api_data** - Clasificación de tendencias
4. **generate_data_quality_reports** - Generación de reportes de calidad
5. **download_final_data** - Descarga datasets procesados

Flujo de tareas



- **Dominio de Airflow:** El desarrollo demostró un dominio completo sobre la orquestación y la conexión a múltiples fuentes de datos.
- **Estado Final:** Todas las tareas críticas del flujo (ingest_api_data, clean_historical_data, classify_api_data, download_final_data y generate_reports) completaron su ejecución en estado "success", lo que valida la integración y la calidad del código.


El DAG ejecutó su ciclo ETL-L con éxito, confirmando el correcto funcionamiento de la lógica de negocio y la orquestación.

ALMACENAMIENTO

El almacenamiento actual implementa una estrategia dual con archivos CSV temporales procesamiento intermedio. Si bien esta solución es funcional para el prototipo, presenta limitaciones significativas en escalabilidad, eficiencia de consultas y capacidades de análisis histórico.

symbol	current_price	price_change_percentage_24h	market_cap	total_volume	high_24h	low_24h	tenden
btc	121939.0	1.27172	2429777542044	82189108096.0	123855.0	119514.0	Tenden modera alcista
eth	4481.34	-1.50404	540967375699	44321815586.0	4583.89	4443.26	Tenden modera bajista
xrp	3.02	-0.41176	180771266762	6408745224.0	3.09	3.02	Tenden estable
usdt	1.001	0.00306	176342479694	140932593147.0	1.001	1.0	Tenden estable
bnb	1175.74	6.69032	163632980367	4931470866.0	1190.05	1085.89	Tenden fuerte alcista
sol	229.49	-2.00239	125304862242	9608545034.0	236.41	228.01	Tenden modera bajista
usdc	0.999759	0.00441	75397342154	20709803304.0	0.999901	0.999607	Tenden estable
doge	0.253762	-2.6078	38366598059	3128407254.0	0.264548	0.253674	Tenden modera bajista

'api_cryptocurrency_data.csv'

1 to 10 of 4150 entries Filter 						
symbol	current_price	1h	24h	7d	30d	24h_volume
BTC	36456.94	0.004	-0.017	0.01	0.184	22801222945.0
ETH	2027.6	0.005	0.013999999999999999	0.01	0.207	26845710464.0
USDT	1.0	0.001	-0.003	-0.001	-0.001	47122466339.0
BNB	231.63	-0.001	-0.126	-0.08	0.054000000000000006	3715265116.0
XRP	0.59	0.001	-0.019	-0.069	0.121	1339890506.0
USDC	1.0	0.0	-0.004	0.0	0.0	16458242430.0
SOL	54.97	0.006	-0.006	-0.114	0.879	2374980324.0
STETH	2023.81	0.004	0.015	0.009000000000000001	0.207	17625537.0
ADA	0.37	0.001	-0.016	0.006999999999999999	0.401	373277046.0
DOGE	0.07	0.001	-0.023	0.001	0.149	957225658.0

'cleaned_cryptocurrency_data.csv'



JUSTIFICACIÓN ARQUITECTÓNICA

MICRO – BATCH

- Procesamiento complejo de datos históricos
- Limpieza y validación comprehensiva
- Generación de reportes de calidad
- Clasificación por lotes completos

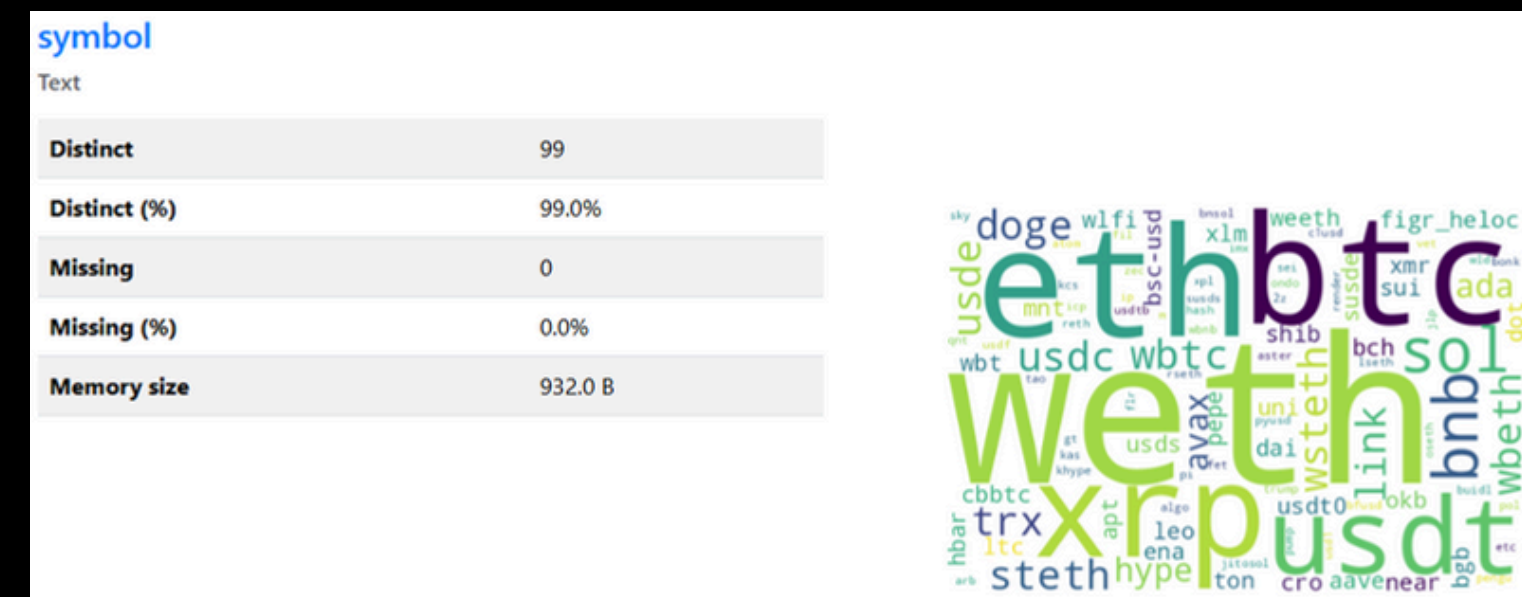
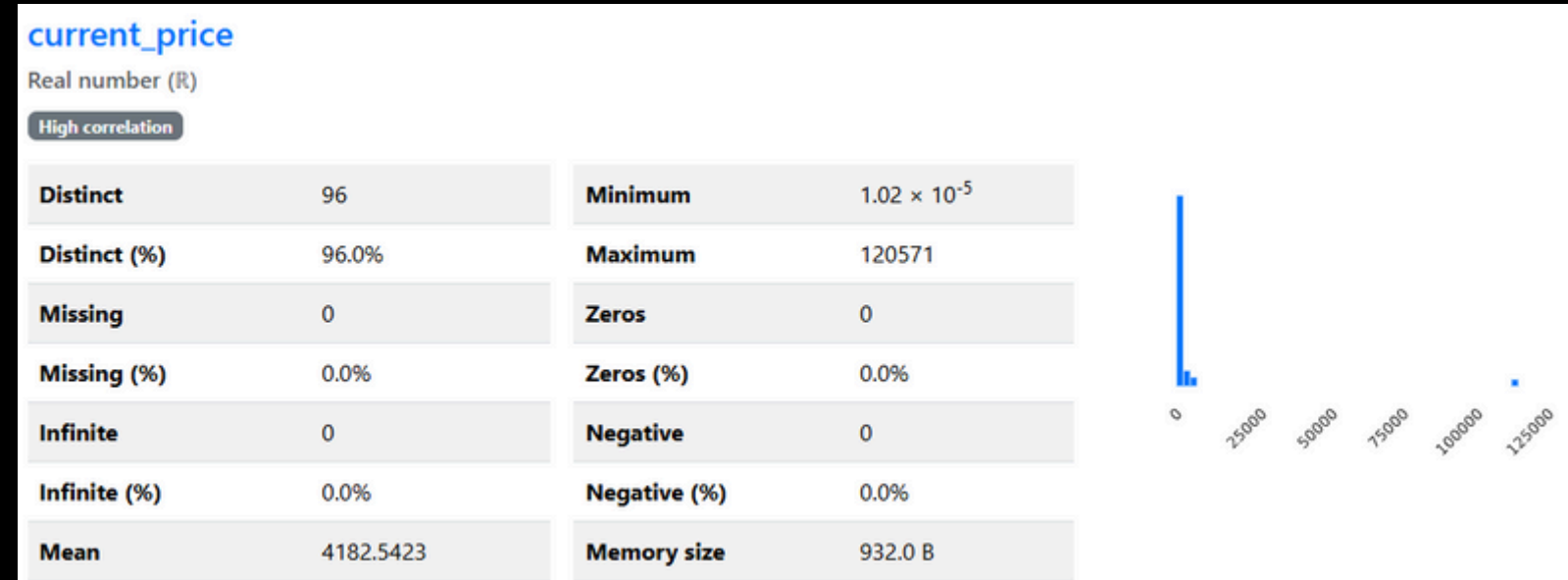
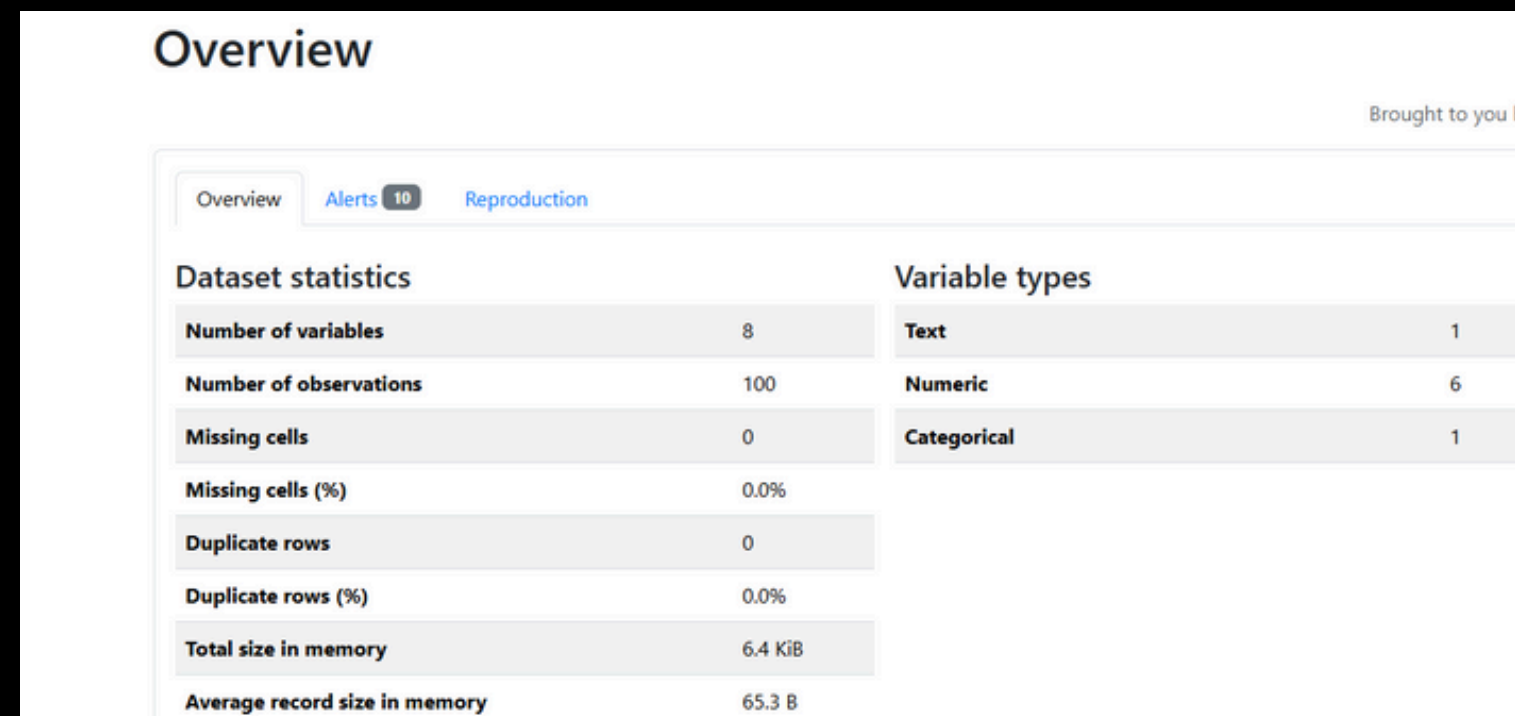
CUASI – STREAMING

- Extracción API cada 6 horas (frecuencia aumentable)
- Clasificación inmediata tras recepción de datos
- Alertas potenciales basadas en resultados

Híbrido

REPORTES

Reporte de calidad elaborado por ProfileReport, mostrando algunos datos estadísticos.



LIMITACIÓN Y MEJORAS FUTURAS

Conclusión

A pesar de las limitaciones operativas impuestas por la capa gratuita de la plataforma Cloud (Astro/Airflow), que restringieron la descarga directa de archivos y la generación de reportes avanzados (ProfileReport), se logró construir y validar la funcionalidad principal del pipeline.

Mejoras futuras y recomendaciones

- Migración a una arquitectura de Streaming utilizando Spark Streaming y Kafka.
- Migrar a soluciones más robustas y escalables: un Data Lake (S3/GCS) para los datos sin procesar o PostgreSQL.
- Implementar un "Data Contract".



GRACIAS

