COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Sprint 3

Grupo 2

Isabel V. Morell Maudes
Teresa X. Garvía Gallego
María Carreño Nin de Cardona
Raquel Fernández Esquinas

Tecnologías de Procesamiento Big Data
3ºA Grado en Ingeniería Matemática e Inteligencia Artificial

Índice

Introducción

Este sprint supone la tercera práctica del proyecto final de la asignatura Tecnologías de Procesamiento Big Dara sobre las criptomonedas. En esta fase del proyecto, nos centraremos en la implementación de una estrategia de almacenamiento de datos históricos basada en tres capas dentro de Amazon S3: Bronce, Plata y Oro.

El objetivo principal de este sprint es desarrollar un sistema automatizado utilizando AWS Glue y Apache Spark para mejorar la organización y transformación de los datos, permitiendo una gestión eficiente y estructurada para su posterior análisis técnico.

Para lograrlo, se implementará un proceso en el que:

- Los datos históricos en formato CSV almacenados en la capa Bronce serán leídos y transformados en formato Parquet en la capa Plata.

- Se calcularán diferentes índices clave de análisis técnico (KPIs) como:
    - Moving Average Simple (SMA)
    - Moving Average Exponential (EMA)
    - Relative Strength Index (RSI)
    - Moving Average Convergence Divergence (MACD)

- Los resultados de estos cálculos se almacenarán en la capa Oro, facilitado su uso en futuras fases del proyecto.

Además, para mejorar la gobernanza y el acceso a los metadatos de los datos almacenados en S3, se desarrollará un script en Python que automatice la creación y configuración de un crawler utilizando AWS Glue Data Catalog y AWS Glue Crawler. Esto garantizará una correcta estructuración y consulta de los datos en el ecosistema de AWS.

Dado que la seguridad es un aspecto fundamental en el proyecto, se garantizará que el código subido no contenga credenciales ni información sensible.

Toda la información, incluyendo los archivos y los datos relacionados con este sprint, se encuentra disponible en la rama de desarrollo del repositorio de GitHub: https://github.com/tgarviagallego/Proyecto_BigData.git

Metodología

El objetivo principal de este Sprint es diseñar e implementar una estructura de almacenamiento de datos históricos basada en tres capas, garantizando la correcta separación y transformación de los datos para su posterior análisis. Para ello, se han creado las capas Silver y Golden a partir de la capa Bronze, la cual fue desarrollada en sprints anteriores. Cada una de estas capas ha sido almacenada en diferentes buckets de Amazon S3 con el fin de mejorar la organización, eficiencia y accesibilidad de la información.

Dado que el bucket obtenido en el Sprint 2 corresponde a la capa Bronze, el primer paso consistió en transformar estos datos para generar la capa Silver. Esta transformación tiene como objetivo mejorar el rendimiento de las consultas y la compatibilidad con herramientas analíticas mediante el uso del formato Parquet, que proporciona mayor eficiencia en la compresión y acceso a la información. Por útlimo, la creación de la capa oro a partir de esa capa plata

Para llevar a cabo todas estas implementaciones, usamos diferentes herramientas que permitieron garantizar la eficiencia y escalabilidad del proceso. Entre ellas destacan:

- **Lenguaje de programación:** Python, utilizado para escribir los scripts de procesamiento de datos en AWS Glue.
- **Frameworks y librerías:** PySpark, empleado para el procesamiento distribuido y eficiente de grandes volúmenes de datos.
- **Servicios en la nube:** AWS Glue, encargado de la ejecución de los scripts de transformación, y Amazon S3, utilizado para almacenar los datos en sus respectivas capas.
- **Formatos de almacenamiento:** CSV para la capa Bronze y Parquet para las capas Silver y Golden, aprovechando las ventajas que este último ofrece en términos de compresión y velocidad de consulta.

La arquitectura del sistema se basa en un modelo de tres capas diseñado para optimizar el almacenamiento y procesamiento de los datos:

1. **Capa Bronze:** Es la capa donde se almacenan los datos crudos en formato CSV sin modificaciones. Aquí se encuentran los datos tal como fueron obtenidos desde su fuente original.
2. **Capa Silver:** A partir de los datos en la capa Bronze, se lleva a cabo un proceso de limpieza y transformación que incluye la conversión de formato de CSV a Parquet. Esto permite mejorar la eficiencia en la lectura y manipulación de los datos.
3. **Capa Golden:** En esta última capa, a los datos almacenados en la capa Silver se le añaden indicadpres técnicos clave que facilitan el análisis y la toma de decisiones.

Para garantizar una implementación eficaz, se siguieron los siguientes pasos. Inicialmente realizamos la creación de la capa Silver a partir de la capa Bronze. Para ello, desarrollamos un script en Python utilizando PySpark dentro de AWS Glue. Este script nos permitió hacer una lectura de los archivos CSV originales, su conversión al formato Parquet y su almacenamiento en un bucket de S3 específico para esta capa. La conversión a Parquet la hicimos para poder cumplir con los requistos de la capa silver, que tiene como objetivo mejorar la eficiencia en la lectura y consulta de datos, aprovechando la compresión y estructura optimizada que ofrece este formato.

Una vez que los datos quedaron almacenados en el bucket de la capa Silver, generamos una serie de indicadores técnicos a partir de estos datos. Implementamos una serie de cálculos para obtener métricas clave utilizadas en análisis financiero, como la Media Móvil Simple (SMA), la Media Móvil Exponencial (EMA), el Índice de Fuerza Relativa (RSI) y la Convergencia/Divergencia de Medias Móviles (MACD). Estos indicadores permiten realizar un análisis técnico detallado de la evolución de los datos almacenados y su comportamiento a lo largo del tiempo.

Para calcular estos indicadores, usamos PySpark debido a su capacidad de procesamiento distribuido, lo que nos permitió trabajar eficientemente con grandes volúmenes de datos. Una vez calculados, los resultados fueron almacenados en la capa Golden, que representa la versión final y enriquecida de los datos. Los datos en esta capa se guardaron en formato Parquet dentro de un bucket S3 dedicado, asegurando así su disponibilidad para futuros análisis y estudios técnicos.

Para garantizar la fiabilidad y calidad de la solución implementada, llevamos a cabo diversas pruebas en cada una de las fases del proceso. En primer lugar, realizamos pruebas de integración para verificar la correcta lectura, transformación y almacenamiento de los datos en cada una de las capas. Comprobamos que los datos originales de la capa Bronze coincidieran con los almacenados en la capa Silver tras la conversión de formato, asegurando que no hubiera pérdidas ni modificaciones inesperadas.

Por último, realizamos pruebas de validación de datos para garantizar la integridad y coherencia de la información almacenada en cada una de las capas. Verificamos que los indicadores técnicos calculados fueran correctos y consistentes con los datos originales, asegurando así la precisión de la información almacenada en la capa Golden.

La implementación de esta solución ha permitido estructurar de manera eficiente el almacenamiento y procesamiento de los datos históricos en tres capas diferenciadas. Gracias al uso de AWS Glue y PySpark, logramos realizar una transformación eficiente de los datos, mejorando su accesibilidad y optimizando los cálculos analíticos. La integración de indicadores técnicos en la capa Golden proporciona una base sólida para la toma de decisiones y el análisis avanzado de criptomonedas, asegurando así que la información procesada sea confiable y útil para estudios financieros y de mercado.

Resultados

A partir de los resultados del sprint anterior, hemos creado las capas plata y oro que se describen en la sección anterior. Los datos de los que partíamos eran los siguientes, visualizados con Amazon Athena tras pasarlos por el crawler:

Resultados de la consulta | Estado de la consulta

Completado — Tiempo en cola: 77 ms — Tiempo de ejecución: 905 ms — Datos analizados: 2.88 KB

Resultados (5) — Copiar — Descargar resultados en formato CSV

Filas de búsqueda — 1

#	datetime	symbol	open	high	low	close	volume	partition_0	partition_1
1	2024-02-01 01:00:00	CRYPTO:XRPUSD	0.5033	0.5102	0.4899	0.5055	0.0	XRP	2024
2	2024-02-02 01:00:00	CRYPTO:XRPUSD	0.5055	0.5138	0.4988	0.5104	0.0	XRP	2024
3	2024-02-03 01:00:00	CRYPTO:XRPUSD	0.5104	0.5265	0.5058	0.5184	0.0	XRP	2024
4	2024-02-04 01:00:00	CRYPTO:XRPUSD	0.5184	0.519	0.5007	0.503	0.0	XRP	2024
5	2024-02-05 01:00:00	CRYPTO:XRPUSD	0.503	0.5138	0.4969	0.5066	0.0	XRP	2024

Como podemos observar tenemos dos particiones, que surgen de la distribución de nuestros datos en el bucket de S3 llamado trading-view-data. Este bucket se estructuraba dividiendo los datos por tipo de criptomoneda y dentro de cada tipo de criptomoneda cada uno de los 4 años almacenados. Como consideramos que la columna volume no nos aporta información, decidimos eliminarla al pasar los datos a la capa plata. Los tipos de las columnas de nuestra tabla son los siguientes:

Resultados de la consulta | Estado de la consulta

Completado — Tiempo en cola: 149 ms — Tiempo de ejecución: 733 ms — Datos analizados: -

```
datetime                string
symbol                  string
open                    double
high                    double
low                     double
close                   double
volume                  double
partition_0             string
partition_1             string

# Partition Information
# col_name              data_type             comment

partition_0             string
partition_1             string
```

Por lo tanto, nuestros datos cargados en la capa bronce están correctamente tipificados y podemos operar con ellos.

Tras realizar el proceso de limpieza y conversión de los datos de CSV a Parquet, además de la eliminación de la columna volumen, los datos quedan:

Resultados de la consulta | Estado de la consulta

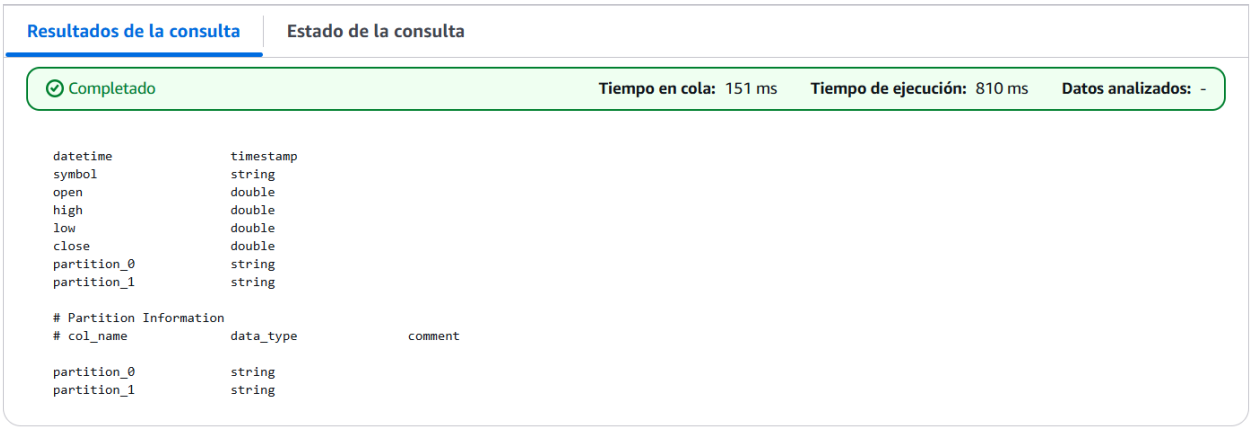
Completado — Tiempo en cola: 59 ms — Tiempo de ejecución: 924 ms — Datos analizados: 2.74 KB

Resultados (5) — Copiar — Descargar resultados en formato CSV

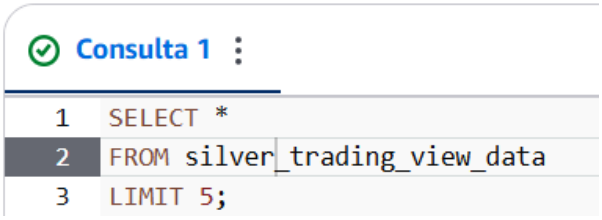
Filas de búsqueda — 1

#	datetime	symbol	open	high	low	close	partition_0	partition_1
1	2023-05-01 02:00:00.000	CRYPTO:XLMUSD	0.09444907	0.09465507	0.09176387	0.0926767	XLM	2023
2	2023-05-02 02:00:00.000	CRYPTO:XLMUSD	0.0926787	0.09381856	0.0918753	0.0936018	XLM	2023
3	2023-05-03 02:00:00.000	CRYPTO:XLMUSD	0.0936018	0.09385178	0.09181614	0.09358925	XLM	2023
4	2023-05-04 02:00:00.000	CRYPTO:XLMUSD	0.093592	0.09418228	0.09276006	0.09358667	XLM	2023
5	2023-05-05 02:00:00.000	CRYPTO:XLMUSD	0.09358617	0.09459927	0.09237582	0.09445629	XLM	2023

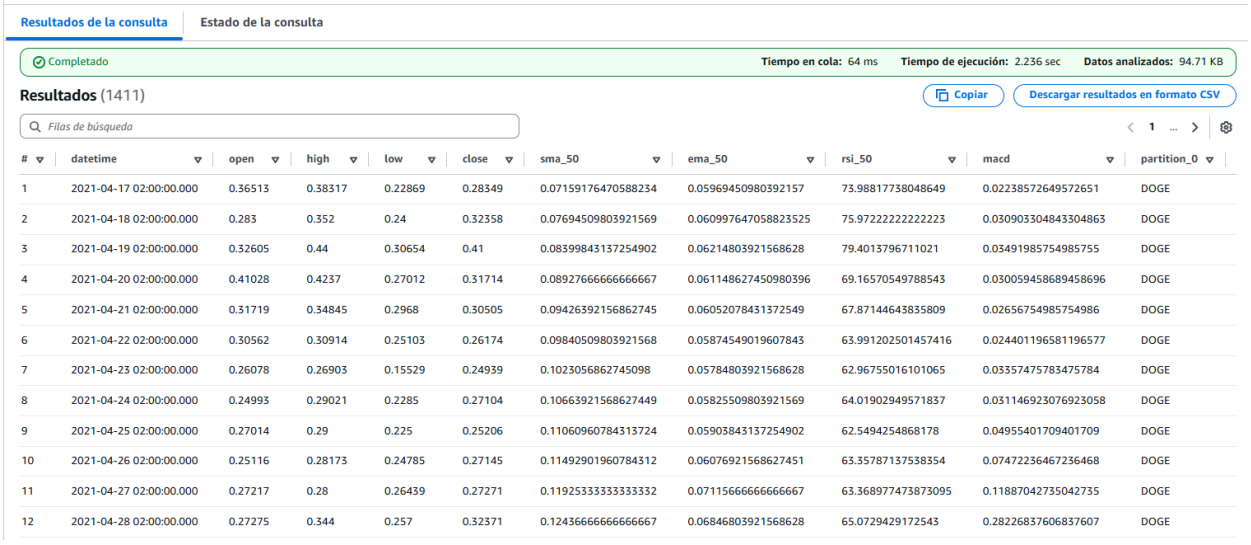
Estos datos se han obtenido ejecutando el ETL job de AWS Glue que tomaba los datos de la capa bronce y, tras realizar una serie de operaciones, guardaba los datos en el nuevo formato en la capa plata. Los tipos de las columnas son:



Que como vemos coinciden con los de la capa bronce. Parece que los datos se han guardado bien y que no hemos realizado ninguna conversión sin ser conscientes de ello.
Las consultas ejecutadas en Amazon Athena para la obtención de los resultados anteriores son del estilo de:



Cambiando el nombre de la tabla.
A continuación, ejecutamos el ETL job que nos calculará los indicadores KPI de estos datos y nos los guardará en la capa oro. De nuevo, para poder visualizarlos con Amazon Athena ejecutamos el crawler correspondiente primero para crear tablas en la database. Nuestra nueva tabla queda:



#	datetime	open	high	low	close	sma_50	ema_50	rsi_50	macd	partition_0
1	2021-04-17 02:00:00.000	0.36513	0.38317	0.22869	0.28349	0.07159176470588234	0.059694509803921 57	73.98817738048649	0.02238572649572651	DOGE
2	2021-04-18 02:00:00.000	0.283	0.352	0.24	0.32358	0.07694509803921569	0.060997647058823525	75.97222222222223	0.030903304843304863	DOGE
3	2021-04-19 02:00:00.000	0.32605	0.44	0.30654	0.41	0.08399843137254902	0.06214803921568628	79.4013796711021	0.03491985754985755	DOGE
4	2021-04-20 02:00:00.000	0.41028	0.4237	0.27012	0.31714	0.08927666666666667	0.061148627450980396	69.16570549788543	0.030059458689458696	DOGE
5	2021-04-21 02:00:00.000	0.31719	0.34845	0.2968	0.30505	0.09426392156862745	0.06052078431372549	67.87144643835809	0.02656754985754986	DOGE
6	2021-04-22 02:00:00.000	0.30562	0.30914	0.25103	0.26174	0.09840509803921568	0.05874549019607843	63.991202501457416	0.024401196581196577	DOGE
7	2021-04-23 02:00:00.000	0.26078	0.26903	0.15529	0.24939	0.1023056862745098	0.05784803921568628	62.96755016101065	0.03357475783475784	DOGE
8	2021-04-24 02:00:00.000	0.24993	0.29021	0.2285	0.27104	0.10663921568627449	0.05825509803921569	64.01902949571837	0.031146923076923058	DOGE
9	2021-04-25 02:00:00.000	0.27014	0.29	0.225	0.25206	0.11060960784313724	0.05903843137254902	62.5494254868178	0.04955401709401709	DOGE
10	2021-04-26 02:00:00.000	0.25116	0.28173	0.24785	0.27145	0.11492901960784312	0.06076921568627451	63.35787137538354	0.07472236467236468	DOGE
11	2021-04-27 02:00:00.000	0.27217	0.28	0.26439	0.27271	0.11925333333333332	0.07115666666666667	63.368977473873095	0.11887042735042735	DOGE
12	2021-04-28 02:00:00.000	0.27275	0.344	0.257	0.32371	0.12436666666666667	0.06846803921568628	65.0729429172543	0.28226837606837607	DOGE

Como podemos observar, las columnas de los KPIs se han creado de forma correcta.
Las columnas SMA_50 y EMA_50 muestran tendencias suavizadas del precio en base a los últimos 50 periodos. Como los valores de ambas medias son similares, parece que hay estabilidad en la tendencia a mediano plazo. El RSI_50 mide la fortaleza del mercado en un rango de 0 a 100. Vemos que tenemos valores cercanos a 70, incluso un poco superiores. Los valores superiores a 70 indican sobrecompra mientras que los valores inferiores a 30 indican sobreventa. Los valores RSI que tenemos podrían indicar una tendencia en aumento con probabilidad de sobrecompra. El MACD ayuda a identificar cambios de tendencia y señales de compra o venta. Como nuestros valores de MACD son positivos, indican una posible tendencia alcista.
Para garantizar la correcta interpretación de los KPIs, fue necesario eliminar los primeros 50 registros, ya que no contenían suficiente información para calcular las medias móviles de manera precisa. Esta reducción representó aproximadamente un 3-4% del total de los datos, sin afectar significativamente el análisis.
Los valores obtenidos en los KPIs respaldan la hipótesis de una tendencia alcista en el mercado. La combinación de un RSI elevado, un MACD positivo y la estabilidad de las medias móviles refuerzan esta conclusión.

Conclusión

La división de los datos en tres capas nos va a permitir gestionar, transformar y optimizar el uso de los datos a lo largo de su ciclo de vida. Tener los datos en la capa bronce nos va a permitir tener una copia del histórico por si acaso se necesitará reprocesar los datos o por temas de auditoría. La capa plata nos va a permitir tener unos datos limpios y procesados a partir de los cuales podremos calcular los diferentes KPIs que incluimos en la capa oro. Estos KPIs, al estar en la capa oro, garantizan que la información utilizada para la toma de decisiones sea consistente, confiable y esté alineada con las reglas de negocio establecidas**.** Mantener estas capas separadas nos da flexibilidad y una mayor gobernanza y calidad de datos.

Como hemos mencionado, los KPIs calculados nos proporcionan tendencias suavizadas del precio, la fortaleza del mercado y los cambios de tendencia y señales de compra o venta. Tener almacenadas todas estas medidas en la capa oro nos permitirá estudiarlas en mayor profundidad y con más detenimiento, así como realizar una comparación entre las diferentes criptomonedas.

Todo este proceso lo hemos realizado a través de los ETL Jobs de AWS Glue lo que ha favorecido la automatización, el escalado y la optimización de la transformación de los datos. Además, gracias a otros servicios como S3 y Athena hemos logrado mejorar la eficiencia en la limpieza y estructuración de los datos, así como comprobar el correcto funcionamiento de todas las capas.