Unidad 1

Actividad:

Documentación del proceso de extracción de datos de la página Yahoo Finanzas.

Julio César Cárdenas Veloth

Módulo – Proyecto integrado V

Grupo PREICA2501B020128

Profesor

Andrés Felipe Callejas

Institución Universitaria Digital de Antioquia

Ingeniería de Software y Datos

Medellín

2025

1 RESUMEN

Para este primer desarrollo del proyecto integrado, se ha elaborado un script en Python el cual accede automáticamente a la página de Yahoo Finance del índice de Shanghái (https://es.finance.yahoo.com/quote/000001.SS/history/), acepta cookies y descarga la tabla de índices históricos de la bolsa para el periodo entre el 9 de mayo de 2024 y el 9 de mayo de 2025, de igual forma se realiza una limpieza y estandarización de los nombres de columna, se transforma las fechas de formato español a ISO y se ajusta los valores numéricos para que los precios queden como decimales y los volúmenes como enteros. El script esta diseñado para insertar en la base de datos SQLite solo los registros que no estuvieran previamente almacenados (los deltas), adicionalmente, genera un archivo CSV con la serie completa, con esto se garantiza un repositorio estructurado, eficiente y siempre actualizado para el cálculo y seguimiento de los indicadores claves de la bolsa de Shanghái, y que además puede ser ejecutado desde GitHub.

.

2 INTRODUCCIÓN

El contar con datos bursátiles precisos y actualizados, genera confianza y es clave para elaborar indicadores de desempeño (KPI) que guíen las estrategias de inversión y la gestión de riesgos en la bolsa de Shanghái. En el contexto de Big Data, donde el volumen y la velocidad de la información crecen exponencialmente, contar con un proceso de automatización robusto se convierte en un diferenciador competitivo: el script desarrollado no solo captura de forma continua los precios y volúmenes históricos desde Yahoo Finance, sino que también los integra de manera incremental en una base de datos SQLite y en un CSV. Esta infraestructura automatizada permite procesar grandes volúmenes de datos con mínima intervención manual, asegurando la integridad y actualización permanente de la información. Con esto, analistas e interesados disponen de una fuente unificada y confiable para calcular métricas avanzadas (como volatilidad, medias móviles y ratios de volumen), y pueden responder ágilmente a las dinámicas del mercado basándose en análisis de datos a gran escala.

3 METODOLOGIA

Metodología para la descarga automatizada de datos desde Yahoo Finanzas

3.1 Definición del alcance y fuente de datos

Se seleccionó como objeto de estudio el índice de Shanghái (https://es.finance.yahoo.com/quote/000001.SS/history/) y su histórico de precios y volúmenes, disponible en la URL indicada:

El período de interés abarca del 9 de mayo de 2024 al 9 de mayo de 2025 para el momento de la realización de esta documentación..

3.2 Configuración del entorno de Big Data y automatización

Se emplea Python como lenguaje de scripting, con entornos virtuales para gestionar dependencias.

Se instalan bibliotecas clave:

Selenium para controlar un navegador en modo “headless” y reaccionar ante elementos dinámicos (botón de cookies, carga de tablas).

BeautifulSoup para extraer el HTML resultante.

Pandas para estructurar y transformar los datos tabulares.

SQLite como gestor de base de datos ligero y escalable para series temporales.

3.3 Extracción

La función de extracción se automatizó mediante un script de Python con Selenium, el funcionamiento de explica a continuación.

El script abre la página en un navegador controlado por Selenium, asegurando la ejecución de JavaScript y la carga completa del contenido.

Se detecta y pulsa automáticamente el botón de “Aceptar cookies” si aparece, garantizando acceso sin interferencias.

Se espera a que el elemento <table> con el histórico esté presente en el DOM y se realiza la captura de cada uno de los campos.

3.4 Transformación

Se captura el HTML con la tabla completa y se envía a BeautifulSoup para extraer cabeceras y filas.

Las cabeceras se limpian de etiquetas anidadas y caracteres especiales, normalizándose a ASCII y reemplazando espacios por guiones bajos.

Las fechas en español (p. ej. “30 abr 2025”) se descomponen en día, mes y año; el mes se traduce a su número correspondiente y la fecha se convierte a formato ISO (YYYY-MM-DD).

Los valores numéricos eliminan separadores de miles, convierten coma decimal a punto y se escriben adecuadamente: precios como float y volúmenes como int.

3.5 Carga incremental

Se conecta a la base SQLite (historical.db) y se obtienen las fechas ya almacenadas.

Solo los registros con fecha nueva se insertan en modo append, evitando recargar datos anteriores y garantizando un delta incremental.

Se genera además un CSV (historical.csv) con todo el conjunto de datos, listo para uso en otros entornos.

3.6 Automatización y orquestación

El proceso se puede programar con GitHub Actions para ejecutarse periódicamente sin intervención humana.

El uso de un logger configurado en nivel DEBUG/INFO permite auditar cada ejecución, detectar errores de carga o parseo y verificar el número de registros añadidos.

4 Conclusiones y recomendaciones

En conclusión, la automatización de la captura y procesamiento de datos bursátiles convierte un flujo tedioso y propenso a errores en un servicio continuo y fiable, alineado con los principios de Big Data donde la velocidad y el volumen de información exigen soluciones robustas. Esta infraestructura, basada en herramientas de scraping dinámico, transformaciones estandarizadas y almacenamiento incremental, facilita el escalado hacia volúmenes mayores sin necesidad de intervención manual. Al centralizar los datos históricos en repositorios estructurados como SQLite y CSV, se crea una base sólida para alimentar pipelines analíticos y generar KPI en tiempo real. De este modo, las organizaciones pueden responder de forma ágil a los cambios del mercado, maximizar la eficiencia operativa y sostener un crecimiento orientado a datos.

5 Bibliografía

Yahoo Finanzas. (s. f.). Historial de cotizaciones de 000001.SS – Índice de Shanghái Composite. Recuperado 11 de mayo de 2025, de https://es.finance.yahoo.com/quote/000001.SS/history/