

Proyecto Integrado V - Línea de Énfasis (Entrega 1).

Julio Cesar Muñoz Rios - 1116436782

Ingeniería de Software y Datos, Institución Universitaria Digital de Antioquia

PREICA2501B020128 – Proyecto Integrado V - Línea de Énfasis

Andrés Felipe Callejas

Mayo de 2025

Automatización de la Recolección Histórica de Datos de NVIDIA: Persistencia, Control de Versiones y Preparación para el Análisis Predictivo

Resumen

Este documento presenta la primera fase de un proyecto orientado a automatizar la recolección continua de datos históricos del comportamiento bursátil de NVIDIA Corporation. Se implementa un sistema que extrae datos desde Yahoo Finance, los limpia, almacena en una base de datos SQLite y los exporta a un archivo CSV versionado, todo bajo un entorno de ejecución automatizada con GitHub Actions. Esta fase busca garantizar la persistencia, trazabilidad y calidad de los datos desde la fuente original, estableciendo la base para futuras etapas centradas en el análisis, modelado y visualización de indicadores clave de desempeño (KPI). La recolección sistemática permitirá respaldar decisiones informadas mediante herramientas de análisis predictivo y tableros interactivos.

Introducción

NVIDIA Corporation es una de las empresas líderes en tecnología de procesamiento gráfico y computación de alto rendimiento. Fundada en 1993, comenzó a cotizar en bolsa en enero del año 1999 bajo el tiker NVDA en el Nasdaq. Su desempeño bursátil se ha convertido en un indicador representativo del comportamiento del sector tecnológico y, en particular, del crecimiento de industrias emergentes como la inteligencia artificial y la computación gráfica. La evolución histórica de su precio de acción, volumen de transacciones y otras variables financieras clave permite calcular indicadores financieros como la volatilidad, el promedio móvil, el índice de fuerza relativa (RSI), entre otros.

La automatización de la recolección de esta información es fundamental para permitir su análisis continuo, confiable y reproducible. En esta primera fase, se enfoca el proyecto en establecer la base de datos histórica y garantizar la calidad del dato desde su origen, sirviendo de soporte a fases futuras orientadas al análisis, predicción y visualización de indicadores clave de desempeño (KPI).

Objetivo General

Automatizar la recolección continua de datos históricos del indicador bursátil de NVIDIA Corporation, garantizando su persistencia, trazabilidad y preparación para futuras fases de análisis, modelado predictivo y visualización de indicadores clave de desempeño.

Objetivos Específicos

1. Obtener de forma automatizada datos históricos bursátiles de NVIDIA desde Yahoo Finance mediante web scraping.
2. Preprocesar y limpiar los datos eliminando inconsistencias, formatos erróneos y valores faltantes.
3. Calcular indicadores financieros básicos como precios de apertura/cierre, volumen, y establecer la estructura para KPIs futuros.
4. Almacenar los datos limpios en una base de datos SQLite y exportarlos en formato CSV bajo un entorno de control de versiones.
5. Documentar el proceso y sentar las bases para futuras fases del proyecto orientadas a entrenar modelos, realizar predicciones y crear una interfaz visual interactiva.

Metodología

La metodología empleada se estructura en cinco pasos:

(1) extracción de datos históricos de NVIDIA desde Yahoo Finance mediante la biblioteca requests y BeautifulSoup; (2) transformación y limpieza de los datos en Pandas, incluyendo la conversión de fechas, eliminación de valores nulos y transformación de cadenas numéricas; (3) almacenamiento persistente de los datos en formato SQLite y CSV, garantizando su reescritura en cada ejecución; (4) automatización del proceso mediante flujos de trabajo programados en GitHub Actions; y (5) registro de logs detallados para asegurar trazabilidad, auditabilidad y depuración del sistema.

Este enfoque garantiza la calidad y disponibilidad de los datos, permitiendo extender el proyecto hacia fases posteriores donde se aplicarán modelos de aprendizaje automático, cálculo de indicadores como medias móviles o RSI, y visualización de KPIs financieros mediante tableros interactivos.