SPRINT1

Tecnologías de Procesamiento Big Data

Nicolas Villagrán Prieto

Pablo Díaz Vega

Mario Kroll Merino

Antonio Mora Abós

ÍNDICE

INTRODUCCIÓN

Nuestro proyecto se centra en la necesidad de optimizar la estrategia de Trading de nuestra compañía Broskis Brokers. Nuestro departamento se centrará en las compañías del SP-500 del sector de Tecnologías de la información.

Nuestro proyecto se basa en Scrum para proponer un modo de trabajo ágil. Acorde a la metodología de trabajo de Scrum trabajaremos de manera intensa en Sprints de una semana de duración. Los objetivos de este primer sprint son: obtener un histórico de los datos desde inicios de 2018 hasta finales de 2024 y almacenar dichos datos en distintos formatos, tales como: avro, parquet, csv, json, orc, excel. En concreto, para cada año, los datos correspondientes se guardarán individualmente en cada formato mencionado. Cada año tendremos los datos almacenados en los 6 formatos mencionados. Mediante este enfoque agilizaremos el acceso a los datos en el formato que más útil nos sea dependiendo del momento.

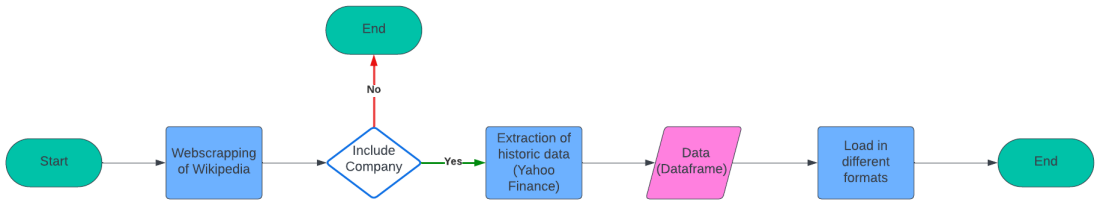
Para la obtención del histórico de datos emplearemos como fuente yahoo finance.

ORGANIZACIÓN DE TAREAS

Hemos designado un scrum master (Mario Kroll) y un product owner(Antonio Mora). Las funcionalidades del scrum máster son: facilitar el proceso Scrum, guiando al equipo en la adopción y mejora continua de las prácticas ágiles. Sus funciones incluyen eliminar obstáculos, fomentar la colaboración y asegurar que el equipo siga los principios de Scrum para lograr entregas exitosas y eficientes. Además, se enfoca en el desarrollo del equipo y la mejora del proceso. Por otro lado, el product owner será el encargado de definir y priorizar las características del producto, asegurando el valor empresarial. Colabora estrechamente con el equipo de desarrollo y actúa como representante del cliente para tomar decisiones que optimicen la entrega de un producto exitoso y alineado con los objetivos del negocio en la metodología ágil Scrum.

TO DO	RESPONSIBLE	DONE
Webscrapping	Pablo Diaz	✓
ETL	Mario Kroll, Nicolas Villagrán	✓
DOCUMENTACIÓN	Antonio Mora	✓

Incluimos una explicación en forma de diagrama de flujo de las tareas que haremos en este sprint.



WEBSCRAPPING

El "web scraping" (es una técnica utilizada para extraer información de sitios web de manera automatizada. Consiste en acceder a la estructura HTML de una página web y extraer datos específicos de interés. Este proceso se realiza mediante el uso de programas o scripts, que simulan la navegación de un usuario en un navegador web, pero en lugar de interactuar manualmente, extraen datos de la página.

Hemos empleado la librería BeautifulSoup mediante un script en Python para quedarnos únicamente con las empresas pertenecientes al sector de tecnologías de la información, que es en el que nos vamos a centrar.

Mediante esta librería hemos obtenido una tabla de Wikipedia. Dicha tabla contiene la siguiente información sobre las distintas emrpesas del SP-500: Símbolo , Seguridad, GICS (sector) , Location, CIK, Founded. El CIK es un número de identificación único asignado a las empresas y entidades registradas en la Securities and Exchange Comission (SEC) en los Estados Unidos.

Una vez hemos extraído la tabla nos quedamos solo con aquellas columnas que vamos a necesitar para llevar a cabo nuestro proyecto: Símbolo, GICS (sector) y CIK. Nos quedaremos únicamente con estos campos, pues son muy relevantes con vistas a en un futuro crear nuestra clave única que nos permita acceder a los datos de manera inequívoca.

ETL

Después de identificar las empresas pertenecientes al sector de tecnologías de la información, utilizamos Yahoo Finance para recopilar datos bursátiles de cada una de ellas. A través de Yahoo Finance, obtenemos diariamente información detallada que incluye la apertura, el máximo, el mínimo, el cierre, el volumen, los dividendos y las divisiones de acciones de cada empresa.

En este punto del Sprint nos encontramos con un gran dilema, y es que los datos tenían doble indexación, de manera que el almacenamiento de los datos no era claro. Para solucionar dicho problema hemos optado por cambiar la dimensión de los datos. En un principio, cuando se obtienen datos de Yahoo Finance se obtiene, para cada empresa, una tabla de datos en el que por cada fila tenemos datos como precio de cierre o apertura. Sin embargo, el problema viene cuando necesitamos guardar esta información para cada empresa en una misma tabla.

Tras barajar opciones para arreglar este problema, nos decantamos por añadir un par de columnas más y eliminar las fechas como índice. De este modo, añadiremos una columna que contendrá el símbolo de la empresa y otra columna que corresponda con la fecha de los datos. Así, la tabla final tendrá dimensiones $n * (365 - (fines\ de\ semana\ o\ festivos))\ x\ 10$, siendo n la cantidad de empresas del sector de Information Technology y 10 la cantidad de columnas. Aquí se proporciona un ejemplo de cómo se representan los datos:

Date	CIK	Symbol	Open	High	Low	Close	Volume	Dividends	Stock Splits