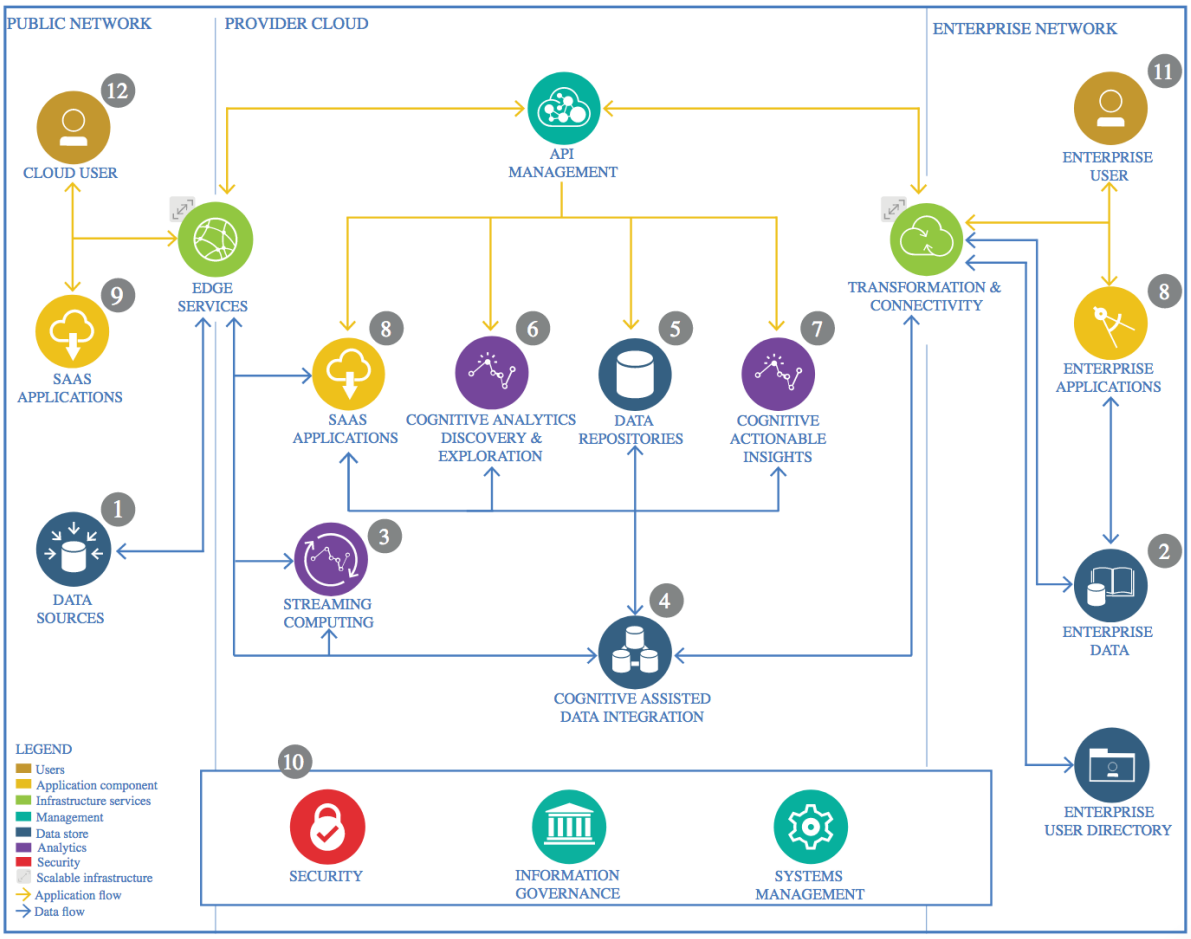
El método ligero de IBM Cloud Garage para la ciencia de datos

Plantilla de documento de decisiones arquitectónicas

1 Descripción general de los componentes arquitectónicos



Arquitectura de referencia de análisis y datos de IBM. Fuente: Corporación IBM

1.1 Fuente de datos

1.1.1 Elección de tecnología

Para la extracción de datos financieros de las empresas específicas para el proyecto, se optó por utilizar Python con las bibliotecas Pandas y datetime. Estas herramientas permiten realizar web scraping de manera eficiente y precisa desde la plataforma de Yahoo Finance, obteniendo datos de los últimos 10 años hasta la fecha actual. Las empresas incluidas en este análisis son Apple, Google, IBM, Meta, Amazon, Tesla y Microsoft.

La extracción de datos se llevó a cabo en el notebook *StockMaxPredict_WebScraping_LSTM*.

1.1.2 Justificación

La elección de esta tecnología se fundamenta en la robustez y versatilidad de Python, respaldada por la capacidad de Pandas para el procesamiento de datos tabulares y la manipulación de fechas a través de la librería datetime. Además, se ha optado por utilizar Yahoo Finance como fuente de datos debido a su reconocida confiabilidad y su establecido prestigio en el ámbito financiero. Esto asegura la disponibilidad de datos precisos y actualizados en tiempo real, esenciales para el éxito de la aplicación.

Asimismo, se ha considerado la necesidad de mantener los datos continuamente actualizados para el modelo de Deep Learning. En este sentido, se implementará la adición de nuevos datos al conjunto histórico extraído a medida que el tiempo avance, asegurando así que el modelo esté siempre respaldado por información reciente y relevante para el Deployment.

1.2 Datos empresariales

En este proyecto, los "datos empresariales" no son una fuente de datos separada ni requieren una elección específica de tecnología. La principal fuente de datos se centra en la extracción de datos financieros de las acciones de empresas específicas, como se describe en la sección anterior. Por lo tanto, no se requiere una justificación adicional para los datos empresariales, ya que no se trata de un componente tecnológico separado en este contexto.

1.3 Análisis de transmisión

Dado que el proyecto no involucra la transmisión de datos a través de una red o canal de comunicación, no se requiere una justificación adicional para el análisis de transmisión, ya que no es relevante en este contexto. Las tecnologías y enfoques se han centrado en otros aspectos del proyecto que son pertinentes para los objetivos establecidos.

1.4 Integración de datos

1.4.1 Elección de tecnología

Para la integración de datos, inicialmente se consideró Apache Spark, específicamente su componente PySpark, junto con las bibliotecas de procesamiento de datos en Python, como Pandas y NumPy. Pero luego de la extracción de datos se optó solo por la librería Pandas.

Esta etapa se llevó a cabo en el notebook *StockMaxPredict_ETL_LSTM*.

1.4.2 Justificación

La elección inicial de PySpark se basó en la expectativa de manejar grandes volúmenes de datos debido a la naturaleza financiera del proyecto y la necesidad de escalabilidad y procesamiento distribuido. Sin embargo, tras la extracción de los datos históricos de los últimos diez años de las 7 empresas seleccionadas, se constató que el conjunto de datos presentaba un total de 18,872 filas y 8 columnas, además de que los datos sin procesar según el EDA solo presentaban problemas de periodicidad.

Por lo tanto, se optó por utilizar exclusivamente la biblioteca Pandas de Python para la etapa de ETL, dada su eficacia en el manejo y procesamiento de datos tabulares de tamaño moderado.

1.5 Repositorio de datos

1.5.1 Elección de tecnología

Para gestionar y almacenar los datos en este proyecto, se optó por utilizar archivos CSV como formato principal. Los datos procesados después de la etapa ETL se guardaron en archivos CSV, al igual que los datos originales. Estos archivos se encuentran organizados en la carpeta "DataStorage" dentro del repositorio del proyecto en GitHub. Además de los datos, esta carpeta alberga los modelos de Deep Learning entrenados, cada uno de ellos en formato .keras, correspondiente a las siete empresas analizadas.

1.5.2 Justificación

La elección de almacenar los datos en archivos CSV se basa en la simplicidad y versatilidad que este formato ofrece en el contexto de Machine Learning y Deep Learning. Los archivos CSV son fáciles de manipular y cargar en bibliotecas y herramientas de análisis de datos, lo que simplifica el flujo de trabajo de preprocesamiento y entrenamiento de modelos.

Además, la elección de GitHub como plataforma de almacenamiento tiene ventajas significativas. GitHub proporciona un entorno colaborativo y de control de versiones, lo que garantiza la integridad y disponibilidad de los datos y modelos a lo largo del proyecto. Esto es especialmente relevante para la colaboración en equipo y la replicabilidad del proyecto.

1.6 Descubrimiento y exploración

1.6.1 Elección de tecnología

Para el proceso de Descubrimiento y Exploración de Datos (EDA), se empleó una combinación de tecnologías y bibliotecas de Python. Estas incluyen Pandas, Matplotlib, Seaborn y SciPy.

La EDA se llevó a cabo en el notebook *StockMaxPredict_EDAInitial_LSTM*, donde se detallan los procedimientos, las técnicas y conclusiones de los análisis.

1.6.2 Justificación

La elección de estas tecnologías se basa en su amplio reconocimiento y eficacia en el análisis de datos. Pandas facilitó la preparación de los datos y la realización de análisis estadísticos descriptivos. Matplotlib y Seaborn proporcionaron visualizaciones visuales que ayudaron a identificar patrones y tendencias en los datos de las acciones de cada empresa. SciPy enriqueció nuestro EDA al permitirnos realizar análisis estadísticos más avanzados, lo que contribuyó a una comprensión más profunda de las características de los datos.

1.7 Información procesable

1.7.1 Elección de tecnología

Para la generación de Información Procesable, emplearemos una combinación de tecnologías y bibliotecas de Python. Esto incluye TensorFlow, Keras, Scikit-Learn, Pandas, NumPy, Matplotlib y Seaborn. Estas herramientas son esenciales para construir modelos de series de tiempo multivariados multi-step para predecir los precios más altos de las acciones de cada empresa.

El proceso completo de creación, entrenamiento y validación de los 7 modelos de redes LSTM se encuentra documentado en la carpeta "DeepLearning_Models". En dicha carpeta, se disponen de 7 notebooks individuales, cada uno correspondiente a una empresa específica, donde se detallan minuciosamente todos los pasos involucrados en el proceso.

También se realizó un experimento utilizando varios algoritmos de regresión de Machine Learning con Apache Spark, con el propósito de pronosticar la variable 'High' de acuerdo con los requisitos del proyecto. El objetivo principal de este experimento fue demostrar la importancia de utilizar un modelo de Deep Learning basado en redes LSTM para realizar predicciones sobre los valores máximos de las acciones de las empresas. Los detalles de este experimento se encuentran documentados en el notebook denominado StockMaxPredict_MLExperimentation_LSTM.

1.7.2 Justificación

La elección de estas tecnologías se basa en su robustez y eficacia en la construcción y entrenamiento de modelos de Machine Learning y Deep Learning. TensorFlow, en conjunto con Keras, representa una elección sólida para la implementación de redes neuronales LSTM en el contexto de series temporales. Scikit-Learn proporciona una amplia variedad de algoritmos y herramientas para el modelado predictivo y la evaluación de modelos. Por su parte, Pandas y NumPy simplifican la manipulación y preparación de los datos antes de su utilización en el modelado.

Además, durante esta etapa del proyecto, Matplotlib y Seaborn fueron fundamentales para la visualización y el seguimiento de los resultados del entrenamiento y la predicción. Estas bibliotecas permitieron la creación de gráficos relevantes, como la identificación de problemas de underfitting y overfitting, así como la comparación de los valores reales con los valores predichos, entre otros aspectos clave.

La sinergia de estas tecnologías nos brinda la capacidad de desarrollar modelos precisos y eficaces que generan información procesable. Esto, a su vez, nos permite pronosticar con precisión los precios máximos de las acciones de las empresas analizadas, lo que resulta esencial para la toma de decisiones informadas en el ámbito financiero.

1.8 Aplicaciones/Productos de datos

1.8.1 Elección de tecnología

Para el producto final del Proyecto se seleccionaron las tecnologías de Streamlit y Plotly. Estas nos permitieron crear una página web interactiva con paneles de visualización avanzados que presentan datos relevantes de las empresas a lo largo del tiempo, así como las predicciones generadas por nuestros modelos. Estas aplicaciones ofrecerán datos en tiempo real y funcionalidades de filtrado interactivo.

1.8.2 Justificación

La elección de Streamlit y Plotly se fundamenta en su capacidad para proporcionar a inversores, traders, gestores de carteras y analistas financieros herramientas efectivas para la toma de decisiones. Estas tecnologías permiten una presentación de datos dinámica y accesible que les ayudará a comprender mejor las tendencias y patrones en los precios de las acciones, a identificar oportunidades de inversión y a optimizar sus estrategias comerciales. Con estas aplicaciones, buscamos brindar una experiencia informativa y práctica que facilite la toma de decisiones informadas en el ámbito financiero.

1.9 Seguridad, Gobernanza de la Información y Gestión de Sistemas

1.9.1 Elección de tecnología

Para abordar la seguridad y gestión de datos en este proyecto, hemos utilizado principalmente GitHub como nuestra plataforma de gestión de versiones y colaboración. GitHub proporciona un entorno seguro para el almacenamiento y seguimiento de los cambios en los archivos de datos y código del proyecto.

1.9.2 Justificación

Se elegió GitHub debido a su amplia aceptación en la comunidad de desarrollo y Data Science, así como por sus características de seguridad, como la autenticación de dos factores, control de acceso y auditorías de cambios, que protegen nuestros datos y código. GitHub también facilita la documentación y seguimiento de cambios, lo que garantiza la integridad y transparencia del proyecto, aunque no hayamos usado herramientas específicas de gobernanza de datos. Esta elección refleja las mejores prácticas de seguridad y gobernanza de datos en proyectos de Data Science.