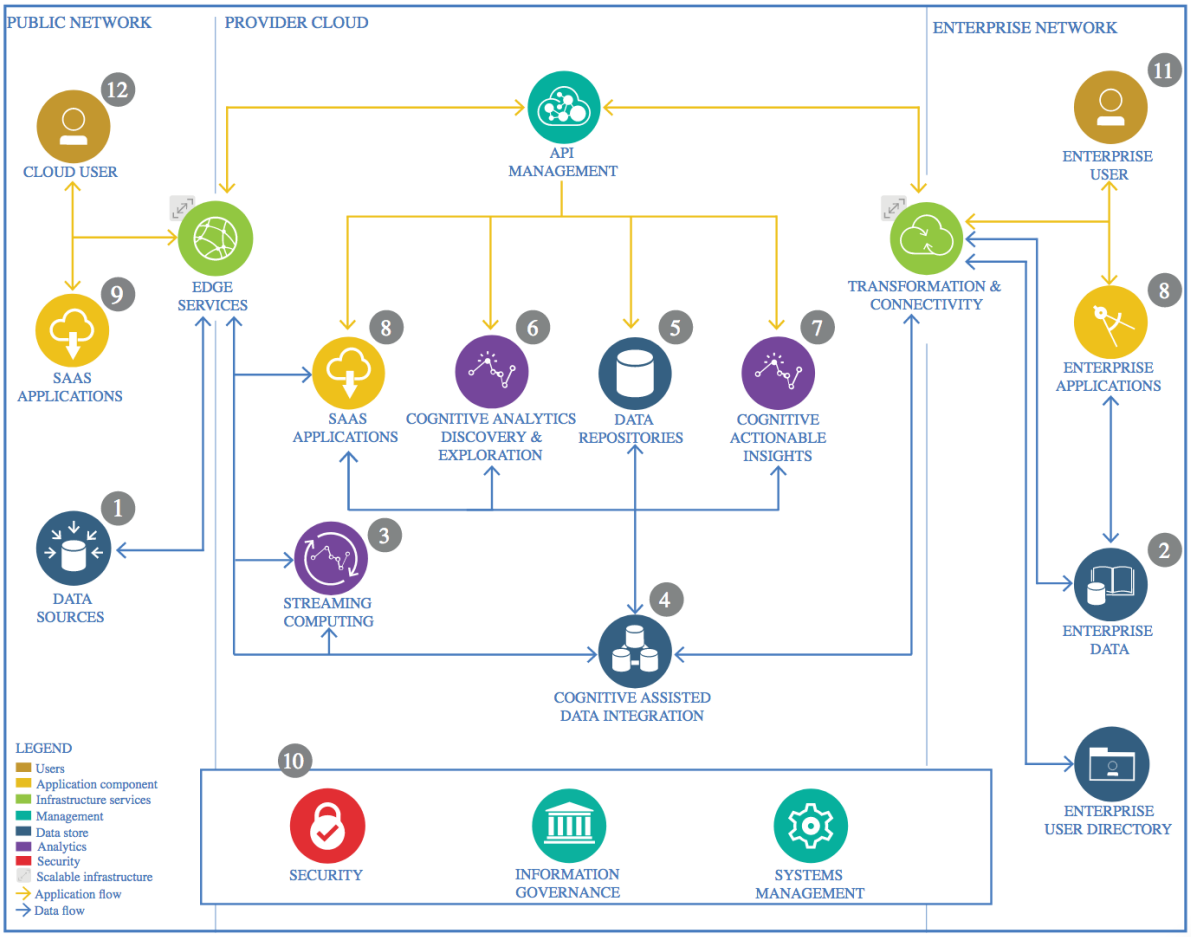
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

For extracting financial data from specific companies for the project, we chose to use Python with the Pandas and datetime libraries. These tools enable efficient and precise web scraping from the Yahoo Finance platform, fetching data from the last 10 years up to the current date. The companies included in this analysis are Apple, Google, IBM, Meta, Amazon, Tesla, and Microsoft.

Data extraction was performed in the StockMaxPredict_WebScraping_LSTM notebook.

1.1.2 Justification

The choice of this technology is based on the robustness and versatility of Python, supported by Pandas' capability for processing tabular data and date manipulation through the datetime library. Additionally, we opted to use Yahoo Finance as the data source due to its recognized reliability and established reputation in the financial domain. This ensures the availability of accurate and real-time data, which is essential for the application's success.

Furthermore, the need to continuously update data for the Deep Learning model has been considered. In this regard, we will implement the addition of new data to the historical dataset as time progresses, thus ensuring that the model is always backed by recent and relevant information for deployment.

1.2 Enterprise Data

1.2.1 Technology Choice

In this project, "business data" is not a separate data source and does not require a specific technology choice. The primary data source focuses on extracting financial data from specific companies, as described in the previous section. Therefore, no additional justification is required for business data, as it is not a separate technological component in this context.

1.3 Streaming analytics

Since the project does not involve data transmission through a network or communication channel, no additional justification is needed for transmission analysis, as it is not relevant in this context. Technologies and approaches have been directed toward other project aspects that are pertinent to the established objectives.

1.4 Data Integration

1.4.1 Technology Choice

For data integration, Apache Spark, specifically its PySpark component, was initially considered along with Python data processing libraries like Pandas and NumPy. However, after data extraction, we chose to use only the Pandas library.

This stage was carried out in the StockMaxPredict_ETL_LSTM notebook.

1.4.2 Justification

The initial choice of PySpark was based on the expectation of handling large volumes of data due to the financial nature of the project and the need for scalability and distributed processing. However, after extracting historical data for the last ten years from the seven

selected companies, it was found that the dataset consisted of a total of 18,872 rows and 8 columns, and the raw data, as per the EDA, only exhibited periodicity issues.

Therefore, we opted to exclusively use the Pandas library of Python for the ETL stage, given its efficiency in handling and processing moderately sized tabular data.

1.5 Data Repository

1.5.1 Technology Choice

To manage and store data in this project, CSV files were chosen as the primary format. The processed data after the ETL stage was saved in CSV files, as well as the original data. These files are organized in the "DataStorage" folder within the project repository on GitHub. In addition to data, this folder houses the trained Deep Learning models, each in .keras format, corresponding to the seven analyzed companies.

1.5.2 Justification

The choice to store data in CSV files is based on the simplicity and versatility that this format offers in the context of Machine Learning and Deep Learning. CSV files are easy to manipulate and load into data analysis libraries and tools, simplifying the preprocessing and model training workflow.

Moreover, the choice of GitHub as the storage platform has significant advantages. GitHub provides a collaborative environment and version control, ensuring the integrity and availability of data and models throughout the project. This is especially relevant for team collaboration and project replicability.

1.6 Discovery and Exploration

1.6.1 Technology Choice

For the Data Discovery and Exploration (EDA) process, a combination of Python technologies and libraries was used. These include Pandas, Matplotlib, Seaborn, and SciPy.

The EDA was carried out in the StockMaxPredict_EDAInitial_LSTM notebook, which details the procedures, techniques, and conclusions of the analyses.

1.6.2 Justification

The choice of these technologies is based on their widespread recognition and effectiveness in data analysis. Pandas facilitated data preparation and the conduct of descriptive statistical analyses. Matplotlib and Seaborn provided visualizations that helped identify patterns and trends in the data for each company's stocks. SciPy enriched our EDA by allowing us to perform more advanced statistical analyses, contributing to a deeper understanding of the data's characteristics.

1.7 Actionable Insights

1.7.1 Technology Choice

For generating actionable insights, we employed a combination of Python technologies and libraries, including TensorFlow, Keras, Scikit-Learn, Pandas, NumPy, Matplotlib, and Seaborn. These tools are essential for building multivariate multi-step time series models to predict the highest stock prices of each company.

The complete process of creating, training, and validating the seven LSTM neural network models is documented in the "DeepLearning_Models" folder. In this folder, there are seven individual notebooks, each corresponding to a specific company, detailing all the steps involved in the process.

1.7.2 Justification

The choice of these technologies is based on their robustness and effectiveness in constructing and training Machine Learning and Deep Learning models. TensorFlow, in conjunction with Keras, represents a solid choice for implementing LSTM neural networks in the context of time series. Scikit-Learn provides a wide variety of algorithms and tools for predictive modeling and model evaluation. Pandas and NumPy simplify data manipulation and preparation before use in modeling.

Additionally, during this project stage, Matplotlib and Seaborn were crucial for visualizing and tracking training and prediction results. These libraries allowed for the creation of relevant charts, such as identifying underfitting and overfitting issues, as well as comparing actual values with predicted values, among other key aspects.

The synergy of these technologies provides us with the ability to develop accurate and effective models that generate actionable information. This, in turn, allows us to accurately forecast the highest stock prices of the analyzed companies, which is essential for making informed decisions in the financial domain.

1.8 Applications / Data Products

1.8.1 Technology Choice

For the final product of the project, we selected the technologies Streamlit and Plotly. These allowed us to create an interactive web page with advanced visualization dashboards presenting relevant company data over time, as well as predictions generated by our models. These applications will offer real-time data and interactive filtering capabilities.

1.8.2 Justification

The choice of Streamlit and Plotly is based on their ability to provide investors, traders, portfolio managers, and financial analysts with effective tools for decision-making. These

technologies enable dynamic and accessible data presentation that will help them better understand trends and patterns in stock prices, identify investment opportunities, and optimize their trading strategies. With these applications, we aim to provide an informative and practical experience that facilitates informed decision-making in the financial domain.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

To address security and data management in this project, we primarily used GitHub as our version control and collaboration platform. GitHub provides a secure environment for storing and tracking changes in project data and code.

1.9.2 Justification

GitHub was chosen due to its wide acceptance in the development and Data Science community, as well as its security features such as two-factor authentication, access control, and change auditing, which protect our data and code. GitHub also facilitates documentation and change tracking, ensuring project integrity and transparency, even though we did not use specific data governance tools. This choice reflects best practices in security and data governance for Data Science projects.