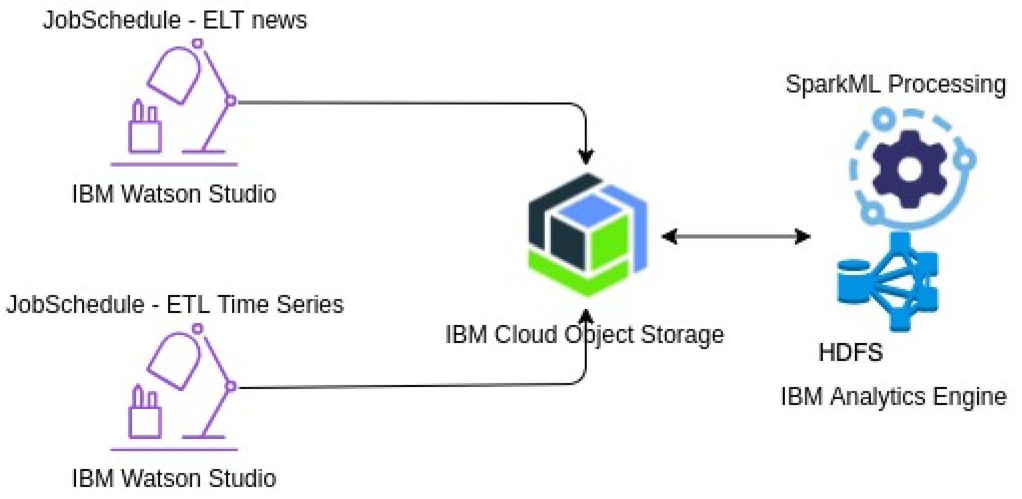
The Lightweight IBM Cloud Garage Method for Data Science

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

The project will have two external data source. For the news used in the project, it is use the economic site called Infomoney, and the data source for the time series Itaú stock price will be B3 (Brasil Bolsa Balcão S.A).

1.1.2 Justification

The Web API is the simple and fast way to get information from external data set. There's a high power integration, optimization of internal management, feasibility of partnerships and customization of services.

1.2 Enterprise Data

1.2.1 Technology Choice

External data categories.

1.2.2 Justification

Because we don’t generate the data so consuming the data from a Website DB and storage it in our Cloud DB.

1.3 Streaming analytics

1.3.1 Technology Choice

IBM Analytics Engine using Apache Spark for big data processing.

1.3.2 Justification

The aim of the application is to predict the Closing price of the stoke market, then its daily and don’t need the real-time streaming, the batch analytics is enough. The IBM AE allows to storage the data in hdfs and processing with Apache Spark, the spark is a cluster computing system that is faster analytics than apache Hadoop, ease to use and multilingual. In this project the script that will be processing in Pyspark, this is a Python API to support Apache Spark.

1.4 Data Integration

1.4.1 Technology Choice

To extract the news is used Crawlers, but the first step is the google search API to get the links of the news and Scrapy for extracting the data from the website.
To get the continuous variables of the time series, the project is using a financial API called Yahoo Finance API.

1.4.2 Justification

The google search API was choice due it facility and stable, this API allows to get 8 links with the free credential, but there isn’t more than 8 news of itaú per day, so it’s fittable in the project.
The Scrapy is a open source and collaborative framework for extracting the data from websites. In a fast, simple to use way as long as you know which elements you want extract.
The Yahoo finance API is free and stable and has all the necessary time series.

1.5 Data Repository

1.5.1 Technology Choice

IBM Cloud Object Storage and hdfs on Analytics Engine Cluster.

1.5.2 Justification

The Hadoop file system is faster than MongoDB using Apache Spark, so the goal for this technology is high performance due the parallel processing, the ibm COS is used only to storage the ETL data, so there’s many small data that is gather and storage on hdfs.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter Lab running locally and IBM Watson studio.

1.6.2 Justification

The jupyter lab is my favorite exploration data tool, because support python to analyzes and creating data visualization easily and fast. The IBM Watson Studio (Jupyter notebook on Cloud) is very easily to connect to IBM Cloud object storage, so it was used to test a fast and cheap ETL.

1.7 Actionable Insights

1.7.1 Technology Choice

Machine Learning RNN NLP and Linear Regression Model.

1.7.2 Justification

The Recurrent Neural network is the most recommended Classifier in Natural Processing Language, it’s use to classifier Itau News in 3 scenarios Positive, Negative and Neutral.
The Linear Regression models was the best model to predict the Closing price of the stock market in our analysis.

1.8 Applications / Data Products

1.8.1 Technology Choice

Access to Processed data from hdfs through API.

1.8.2 Justification

The Machine Learning models and the data sets are in cloud, so the processed data need to be consumed from the cloud, so the easy way is through API, it will give the possibility of the stakeholder to use any Business Intelligence Tool of your preference.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

IBM IAM rules.

1.9.2 Justification

With IBM IAM rules we can control easily who has the access to the processed data, ensuring their security. It’s very easy to configure this, to obtain the data that we can get access using IBM HMAC credentials.