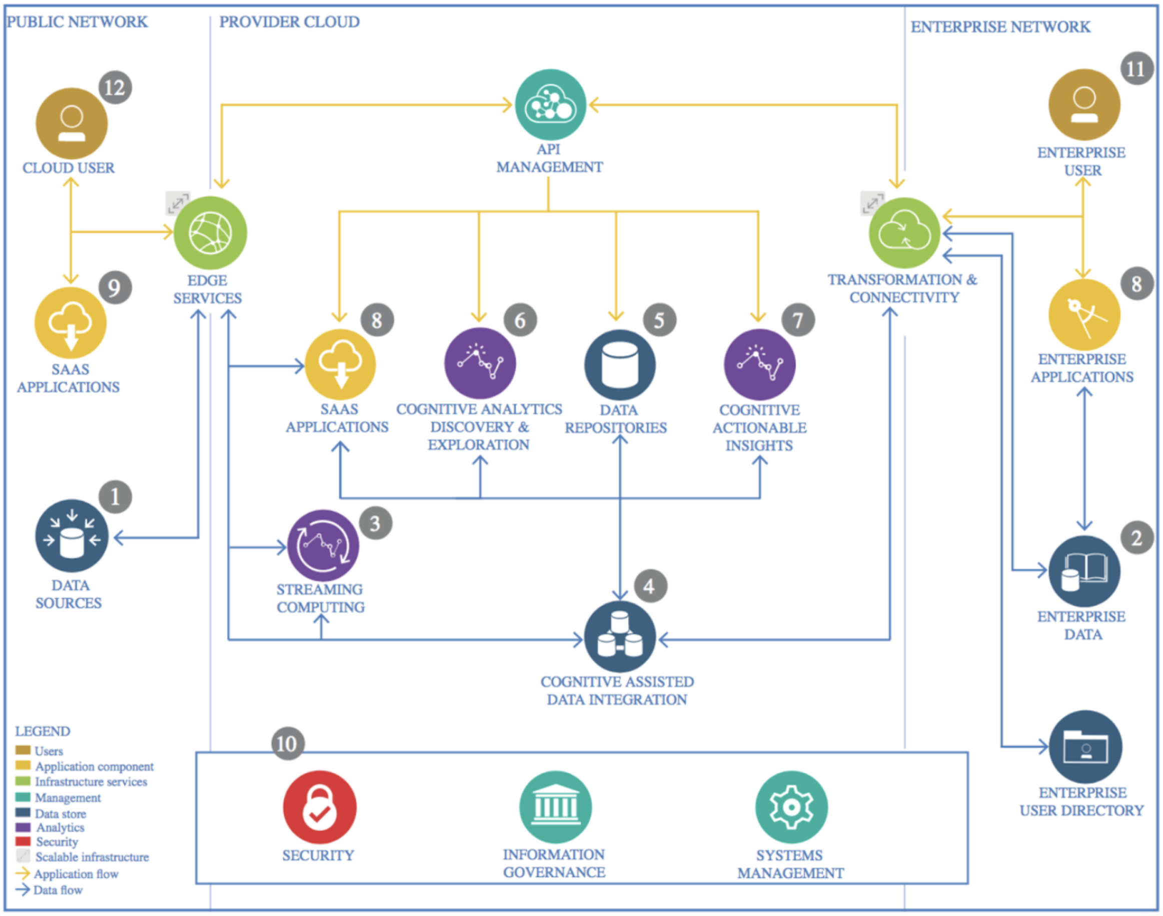
Architectural Decisions Document



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Architectural Components Overview

Data Source and Justification:

The technology choice was an external data source which included Databases in CSV files. The data set Customer demographics and sales has 13.733 rows and 62 columns, The type project store the data in csv is the most adequate.

Enterprise Data and Justification:

This database was provided by IBM and is stored in your cloud so it could be updated automatically.

Streaming analytics and Justification:

This database does not use technology Streaming, this information is collected from customer shopping.

Data Integration and Justification:

The data set Customer demographics and sales from all states of The United States and has 13.733 rows and 62 columns, but after cleaned data set has 36 rows, prepared by category: Gender, State, Credit card type, purchase point, and order type, and category by principals products.

The database object of this study presents the characteristics of segmenting the market into two categories, one is by-products and the other by quality or characterization of the customers, so it will be coded in such a way that it shows us the information grouped or unifying both categories. To answer which market segmentation the product should be aimed at and which product should put more emphasis on.

The technology used was a cross-validation table, group by different categories.

Data Repository and Justification:

The persistent storage data I used github:

https://github.com/Moly-malibu/Customer-demographics-and-sales-in-United-States.git

Discovery, Exploration and Justification:

Understanding the business consists of identifying the variables and clarifying the business problems through metrics that must be: Specific, Measurable, Achievable, Relevant, Limited in time to achieve the objectives of the model. Determining sales forecasts or the probability that an order is fraudulent are among many the main objectives in the analyzes, as well as determining product demand if there is a desire for the product or service, the size of the market; determine how many people would be interested in your offer, including a study of economic indicators such as income range and employment rate, highlighting whether or not there is market saturation, whether there is product diversity for consumers, what prices from potential customers pay for these alternatives, for this Data Science has differents tools to help in this process that we will apply in this project. Such as, Use correlation, covariance, skewness and kurtosis to analyze what productos is the most important for the customers and determine if the company can supply this type of product.

Actionable Insights and Justification:

The different tools used in data science or machine learning can be used in Linear Regression to get answers such as How much or how many? or classification model for which category, also Clustering to select which group or Anomaly detection if this is weird. In this analysis we used Linear Regression.

Regression analysis is supervised machine learning, this model helps to see the relationship between a certain number of features and target variable. Also, Logistic Regression the probability of some events is represented as a linear function of combination of predictor variables, but no required linear relationship between dependent and independent variables. These variables can be represented in the binary values (0 0r one, False or True). With sigmoid function we can do the value of range 0 to 1.

Prepared categorical data:

Normalization is the process creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency. Train a dataset that uses samples to create the model, and validation that is used to qualify performance. Features and target. Steps to create the model, in case I did Logistic Regression.

Applications / Data Products and Justification:

With the Technologic use as Logistic regression, we analysis different features in this dataset, the principal product the customer expense their money are the lotions or cosmetics, even in comparison that beer, cigarettes, and medicine if went be compared way to pay, we can see that Diners Club is the credit card with the principal move in sales of Lotions, followed by American Express.

Compared the Lotion versus Cheese the customers have the same level to buy of preference at the time they acquire this product.

If we compare the Lotion by generations, and see who spends more in these items, and is the generation Z(born after 1997) -that will be the most target to sell this type of products-. The genders are Miss and Mr. In the order type, medium value is the quantity of money the customer expend. the preference most representative to buy Lotion is directly in the stores (desktop)

Security, Information Governance and Systems Management and Justification:

For this project, access to information is free, but when it comes to companies, access and security of information must be restricted according to the sensitivity of each particular case regarding the projects.