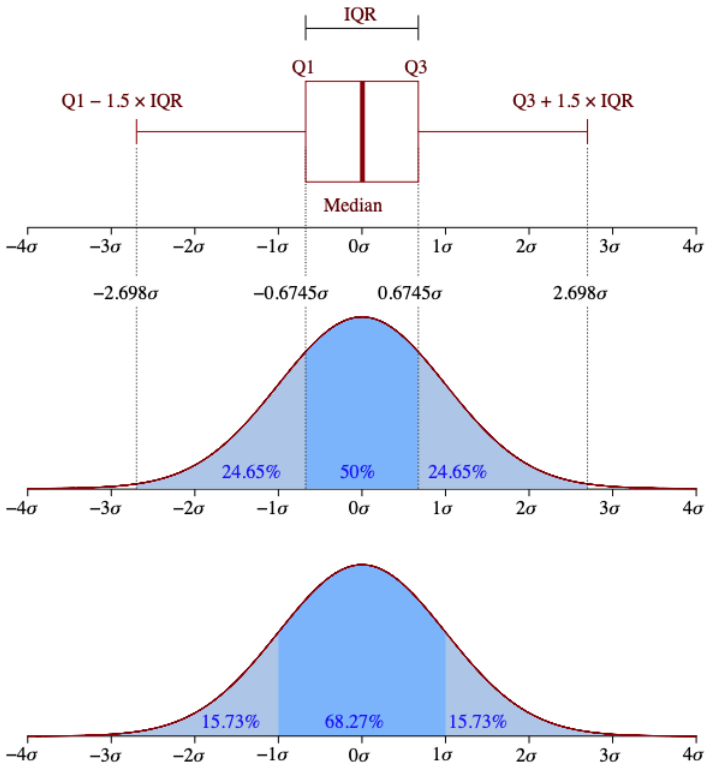
Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Data Science Glosary

Boxplot: A boxplot is a standarized way of displaying a summary of an variable based on five numbers.

$\underline{Q_0}$: The lowest value of the variable.
$\underline{Q_1}$: Is the 25th percentile of the variable.
$\underline{Q_2}$: Is the median of the variable (or 50th percentile).
$\underline{Q_3}$: Is the 75th percentile of the variable.
$\underline{Q_4}$: The highest value of the variable.

IQR (Interquartile range) $= Q_3 - Q_1$



Percentiles: In statistics the $k-th$ percentile also known as *percentile score* or *centile* is a score which below it is the $k-th$ percentage of the distribution.

Compute the k-th: Given N values, to compute the k of a particular value the first to do is sort the values in ascending order where the lowest value you assign the first position and the highest value you assign the last position. Now consider the n position of the particular value and compute: $k = \left(\frac{n}{N}\right) \cdot 100.$

Compute p_k: Given N values and k, first of all you have to order the values and then the position associated to p_k is $i = \frac{k}{100} \cdot (N+1)$. If i is not a integer you cand round.

Boostings: The boosting is a method used in Machine Learning to reduce the error in the predictive analysis of data. The boosting try to improve the performance of a model through the secuential training of various models.

Digital Transformation: Digital transformation is the integration of digital technology into all areas of a business, fundamentally changing how you operate and deliver value to customers. It's also a cultural change that requires organizations to continually challenge the status quo, experiment, and get comfortable with failure.

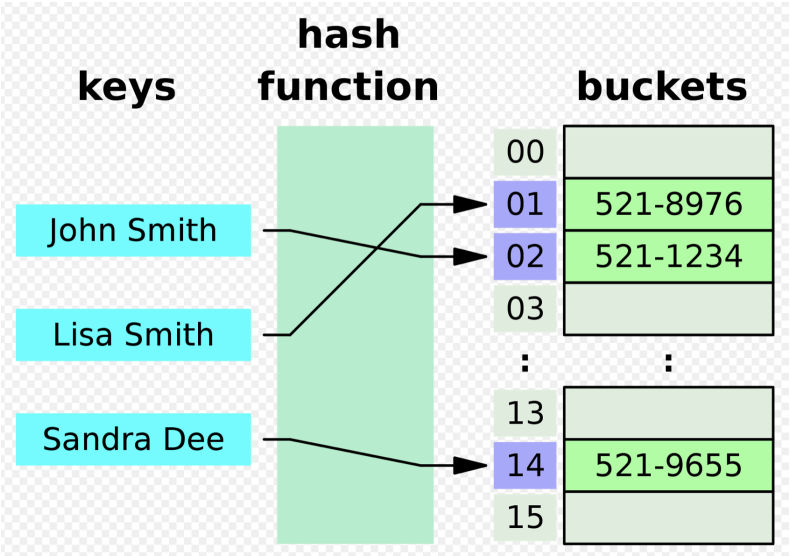
There are three laws of Digital Transformation.

1. Moore's Law: This law says that approximately every 2 years the number of transistors in a microprocessor doubles.

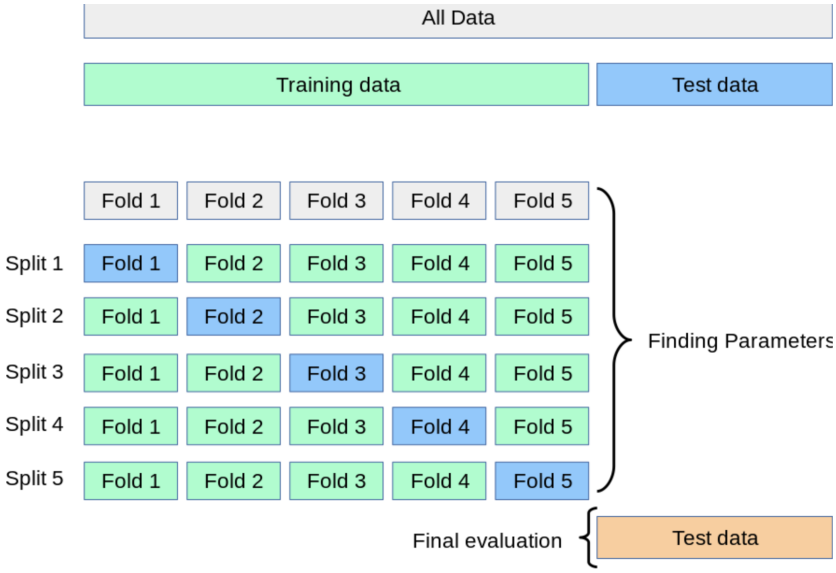
2. Metcalfe's Law: This laws says that the value of a telecommunication network increases proportionaly to the square of the number of user of the system.

3. Bandwidth's Law: Nielsen's Law of Internet Bandwidth simply states that a high-end user's connection speed grows by 50 % per year.

Hash Tables: A hash table, associative array, hashing, hash map, hash table, or chunked table is a data structure that implements the abstract data type called a dictionary (abstract data type). It associates keys or keys with values. The main operation that it supports efficiently is the search: it allows access to the elements (phone and address, for example) stored from a key generated (using the name or account number, for example). It works by transforming the key with a hash function into a hash, a number that identifies the position (box or bucket) where the hash table locates the desired value.



Cross Validation: CrossCross-validation is a technique used to evaluate the results of a statistical analysis and ensure that they are independent of the partition between training and test data.



Pip vs Conda: Conda is a packaging tool and installer that aims to do more than what pip does; handle library dependencies outside of the Python packages as well as the Python packages themselves. Conda also creates a virtual environment, like virtualenv does.

CPU vs GPU: The CPU is the main brain of the computer and is responsible for performing calculations and executing program instructions. The GPU is responsible for processing graphics and visualizations.

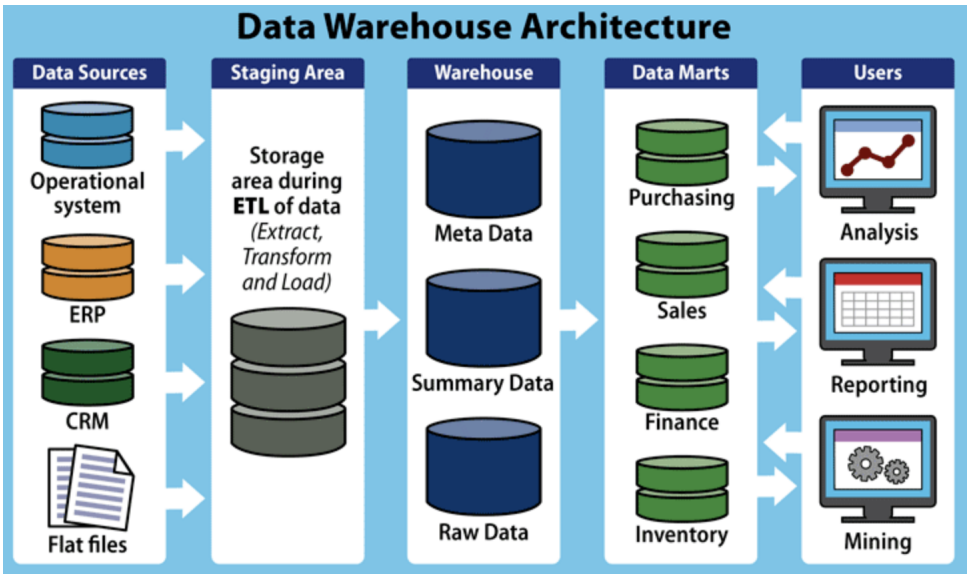
Database engines: A database engine is an underlying element under the system of a database that is used for its operation. These engines build the blocks on which the rest of the elements of the database will be sustained and developed. Also, a database engine is an element that is characterized by all system components, which are responsible for storing and retrieving data.

Major database engines:

1. Microsoft SQL Server.
2. MySQL.
3. SQLite.
4. Oracle.
5. ODBC.
6. PostgreSQL.

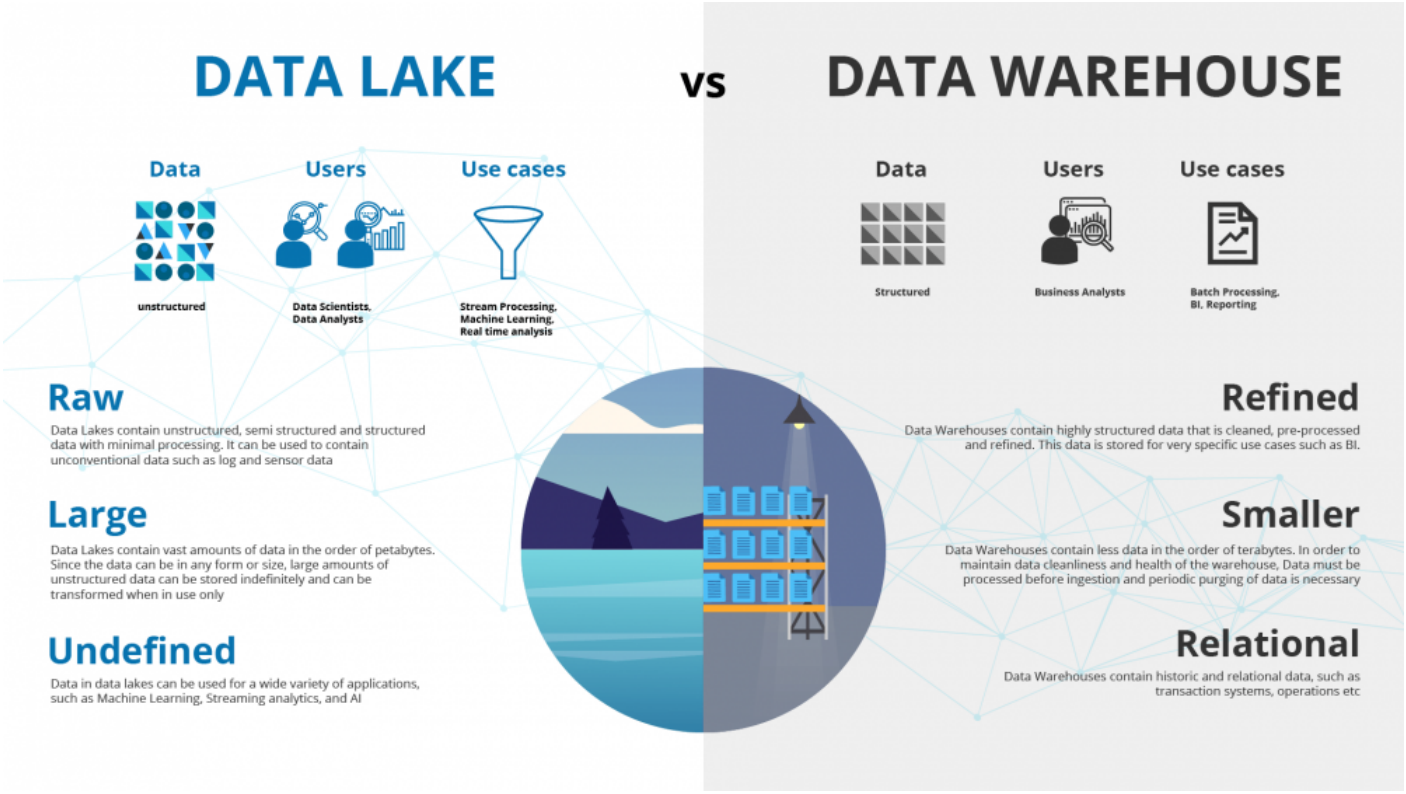
Data Warehouse: The term "Data Warehouserefers to the process that consists of collecting and manipulating data from various sources, in order to recover valuable information for a company.

A Data Warehouse is a platform used to collect and analyze data from multiple heterogeneous sources.



Feature Store: A feature store is our feature data warehouse, our central storage unit of documented, curated, and access-controlled features that can be used across many different models.

Datalake: A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data without modifying it and without having to structure it first. You can also run different types of analytics: from dashboards and visualizations to big data processing, real-time analytics, and machine learning to make better decisions.



Encapsulation: Encapsulation is a mechanism for gathering data and methods within a structure by hiding the implementation of the object, that is, preventing access to the data by any means other than the proposed services.

Modularization: Modularization is a practice of organizing a code base into loosely coupled parts and independent elements. Each part is a module. Each module is independent and has a clear purpose.

Serverless: Serverless means without a server, it is a solution that allows you to create and run applications quickly and with a lower total cost of ownership, since it is not necessary to provision and manage infrastructure. Obviously, there are servers behind to run the applications, but the cloud provider takes care of the administration, therefore, on our side we stop worrying about managing servers, operating systems, software and other resources, and we only focus on the application code.

Target Leakeage: Target leakage, sometimes called data leakage, is one of the most difficult problems when developing a machine learning model. It happens when you train your algorithm on a dataset that includes information that would not be available at the time of prediction when you apply that model to data you collect in the future. Since it already knows the actual outcomes, the model's results will be unrealistically accurate for the training data, like bringing an answer sheet into an exam.

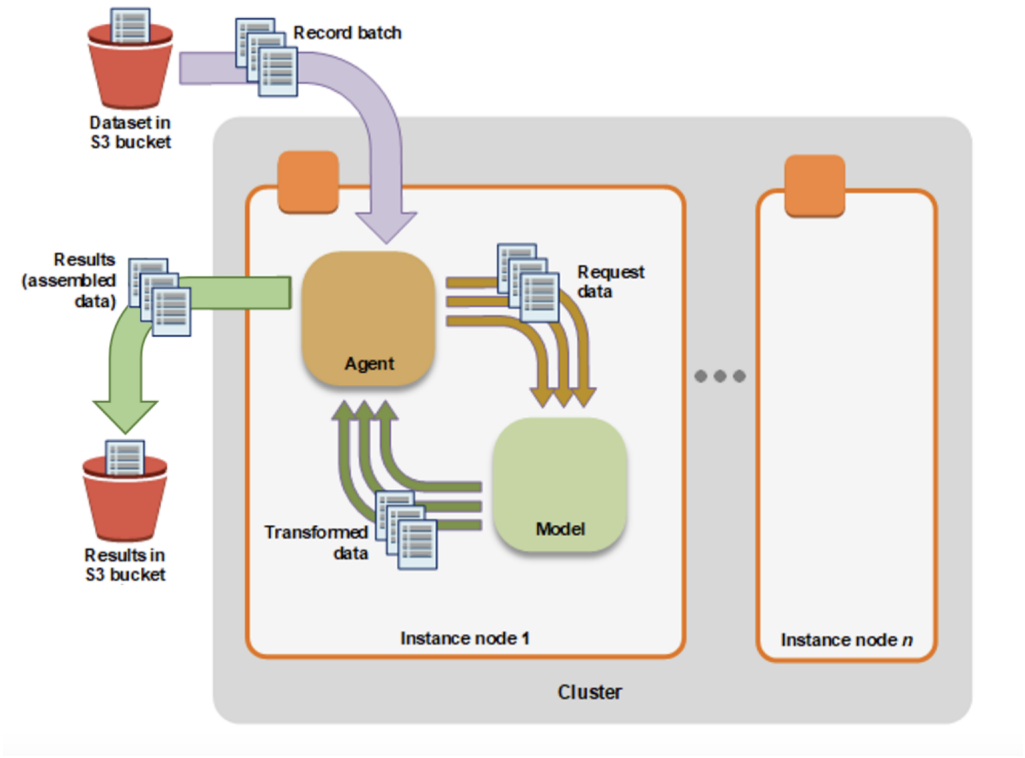
Cache storage: A cache is a high-speed data storage layer that stores a subset of data, typically transient, so that future requests for that data are served faster than if the data must be accessed from the storage location.

Scaffolding: Scaffolding consists of generating code from predefined templates and a specification provided by the developer. It is typically used to generate boilerplate code that can be easily specified and generated from a template.

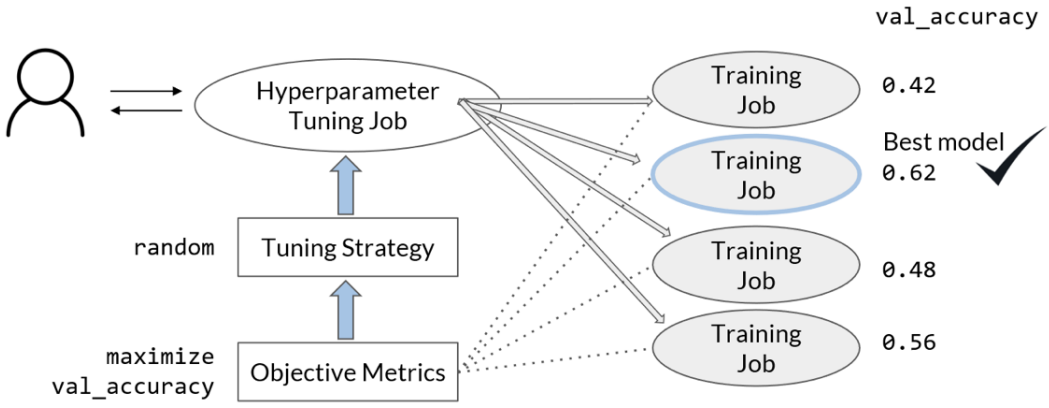
Continuous Integration and Distribution (CI/CD): CI/CD is a method of distributing applications to customers frequently by using automation in the application development stages.

AWS Quicksight: Amazon QuickSight is a scalable, serverless, embeddable, machine learning-powered business intelligence (BI) service built for the cloud. QuickSight enables you to easily create and publish interactive business intelligence dashboards that include insights powered by machine learning.

Batch Transform: Amazon SageMaker batch transform manages all the compute resources required to get the predictions, this includes launching the compute instances and deleting them once the batch transform process has finished. The batch transform manages the interactions between the dataset and the model through an object inside the instance node called an agent.



Hyperparameters Job: The hyperparameter tuning job will launch training jobs to find an optimal configuration of hyperparameters. These training jobs should be configured using the SageMaker CreateHyperParameterTuningJob API.



Data Pipeline: A data pipeline is a series of data ingestion and processing steps that represent the flow of data from a selected single source or multiple sources, over to a target placeholder. The target can be specified either as a data platform or an input to the next pipeline, as the beginning of the next processing steps. Generally, each time we want to process data between points A and B, there is always some kind of data pipeline behind the scenes. This may also involve multiple points, which can also be understood as separate systems.

Notebooks Jobs: With SageMaker notebook jobs, you can now run your notebooks as is or in a parameterized fashion with just a few simple clicks from the SageMaker Studio or SageMaker Studio Lab interface. You can run these notebooks on a schedule or immediately. There's no need for the end-user to modify their existing notebook code. When the job is complete, you can view the populated notebook cells, including any visualizations.

AWS Lambda: AWS Lambda is a compute service that lets you run code without provisioning or managing servers.

Lambda runs the code on a highly available compute infrastructure and performs all compute resource management tasks, including server and operating system maintenance, capacity provisioning and autoscaling, and logging functions . With Lambda, all you have to do is supply your code in one of the Lambda-supported language runtimes.

Organize your code into Lambda Functions. The Lambda service runs the function only when needed and scales automatically. You only pay for the computing time you consume; there is no charge when the code is not running.

Step Functions: AWS Step Functions is a serverless orchestration service that allows you to integrate with AWS Lambda Functions and other AWS Services to build business-critical applications. Through the Step Functions graphical console, you can view your application's workflow as a series of event-based steps.

Endpoint: The VPC Endpoint option is defined as an Amazon Web Service tool that enables private connection between the VPC Endpoints (VPC) system virtual private cloud with AWS-supported services and other Endpoint services.
