

Data architectuur in EA 17



Ir. Ing. Bert Dingemans
bert@data-docent.nl

Inhoudsopgave

Inhoudsopgave.....	2
Inleiding	3
Aanvullende kenmerken	3
Data Architectuur in EA17.....	3
Diagrammen	3
Surveillance Camera Video Ingestion - AWS.....	3
Algorithm Types	5
IoT Agriculture Use Case	11
IoT Agriculture Use Case Example	12
Random Forest - Telecommunication Predicting Customer Churn.....	13
Random Forest - Telecommunication Predicting Customer Churn.....	18
K-Nearest Neighbors - Classifying Flowers	23
Glossaries.....	28
Data Architecture Glossary	28
Data Principle Glossary.....	31
Data Role Glossary.....	34
Data Governance Glossary.....	37
Data Management Glossary	39
Data Management Acronym Glossary	41
Data Quality Glossary.....	46
Data Modeling Glossary.....	47
Data Science Glossary	49
Machine Learning Glossary	51
Image Libraries	53
Business Visualization Images	54
Personally Identifiable Information (PII) Images	57
Internet of Things (IoT) Images.....	58
Data Storage Images.....	60
Principles	61
Universal Data Principles.....	61
Data Architecture Principles.....	62
Data Modeling Principles	63
General Data Protection Regulation (GDPR) Principles	65
Data Role Library.....	66
Data Role Library	66
Tot slot.....	69
Over de auteur.....	69

Inleiding

Sinds medio mei 2024 is versie 17 van Sparx Enterprise Architect beschikbaar. Op dit moment van schrijven, juli 2024, nog als beta versie. In versie 17 zitten een aantal interessante uitbreidingen voor de data architect en de metadata specialist. Dit omdat er een perspectief ontwikkeld is die binnen Sparx Enterprise Architect DataWareHousing genoemd wordt. De naam is wellicht wat eigenaardig gekozen want de inhoud van het perspectief is veel breder. Het gaat namelijk ook in op allerlei zaken zoals data science, data architectuur en data governance rond bedrijfsrollen gerelateerd aan data.

Met deze uitbreiding van de perspectieven is er voor data architecten een mooie aanvulling gekomen op modellering. Daarom heb ik een voorbeeld package uitgewerkt met daarin een aantal voorbeeld diagrammen en vulling van packages op basis van de voorbeelden zoals die standaard in het tool zitten. Hieronder daarom een uitwerking met de aanwezige voorbeeldmodellen.

Aanvullende kenmerken

Naast de voorbeelden zoals hieronder beschreven zijn er nog een aantal aanvullingen in EA 17 die interessant zijn voor de data architect. Die aanvullingen worden hier genoemd en worden verder uitgewerkt in een ander document met een beschrijving van de aanpassingen voor architectuur in het algemeen.

- **ArchiMate 3.2**, de naam in Sparx doet anders vermoeden in het tool maar ArchiMate 3.2 is opgenomen in versie 17
- **Ondersteunen van nieuwe database vormen.** Tot op heden was het alleen mogelijk om relationele databases te modelleren in Sparx. Nu zijn er meerdere nieuwe (NoSQL) databases beschikbaar om uit te werken in modellen denk hierbij aan Hadoop, Snowflake, TeraData, BigTable en andere databases
- Time aware modelleren is uitgebreid en biedt de architect meer ondersteuning in het maken van plateau architecturen en richt zich meer op het bewaren van versies van de architectuurmodellen binnen één repository of in meerdere repositories.

Data Architectuur in EA17

Deze uitwerking is gebaseerd op het datawarehouse perspectief zoals dat is uitgewerkt in versie 17. Het valt daar onder de categorie datawarehouse en dat vind ik persoonlijk wat eigenaardig, vandaar dat ik de term Data Architectuur gebruik. Er zitten ook een aantal hele interessante metadata zaken in.

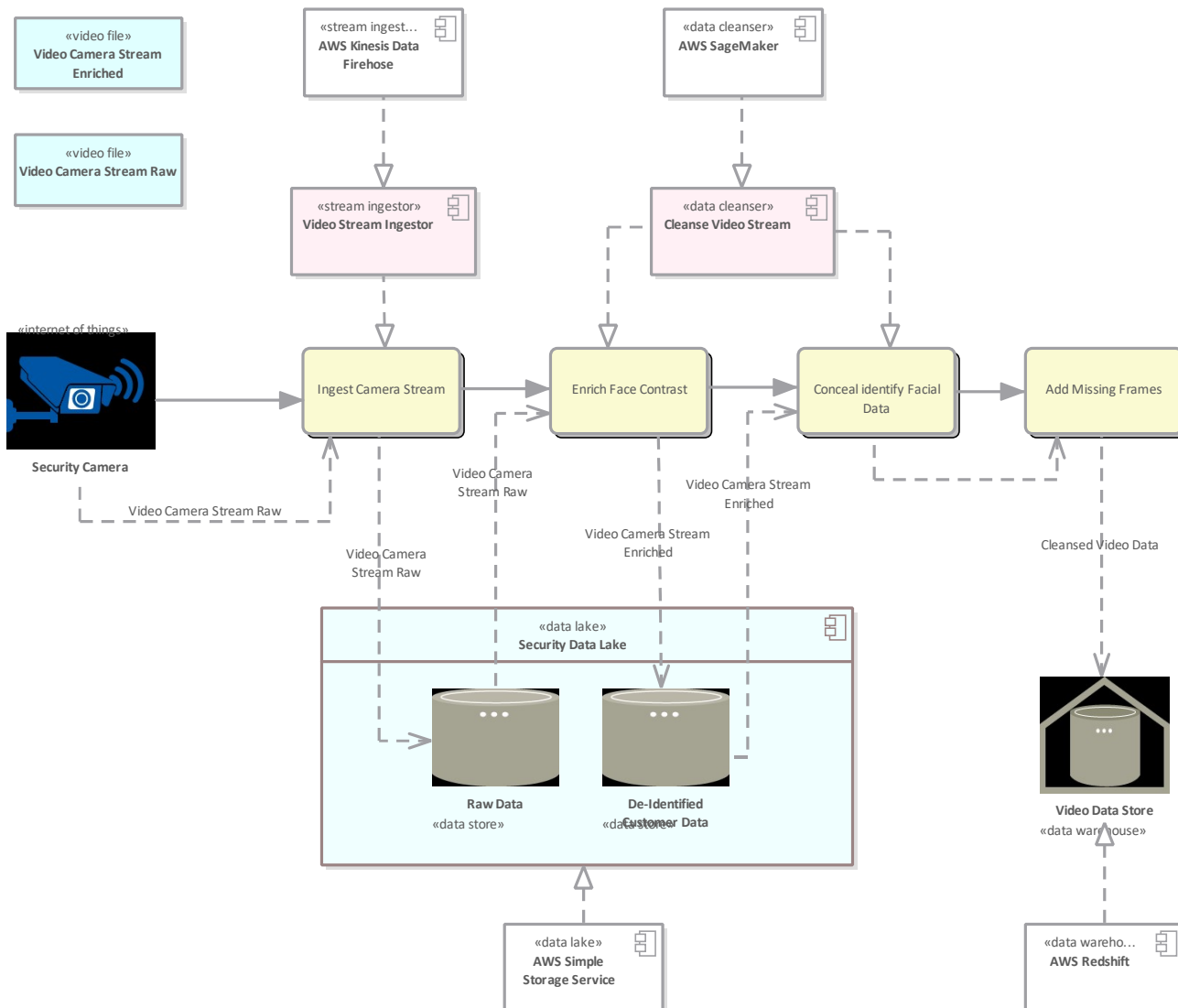
Diagrammen

Onder de diagrammen zijn een aantal interessante voorbeelduitwerkingen te vinden. Vooral interessant is dat er een aantal diagramtypen zijn gemaakt waarmee je feitelijk rond data een aantal metadata zaken kunt gaan uitmodelleren.

De voorbeelden zijn verder niet uitgewerkt en toegelicht, ze bestaan allemaal alleen maar uit de voorbeelden die geleverd worden Sparx in dit perspectief.

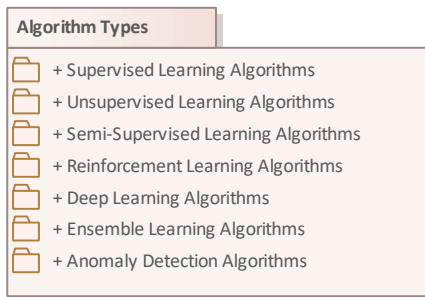
Surveillance Camera Video Ingestion - AWS

Surveillance Camera Video Ingestion AWS



Algorithm Types

Algorithm Types

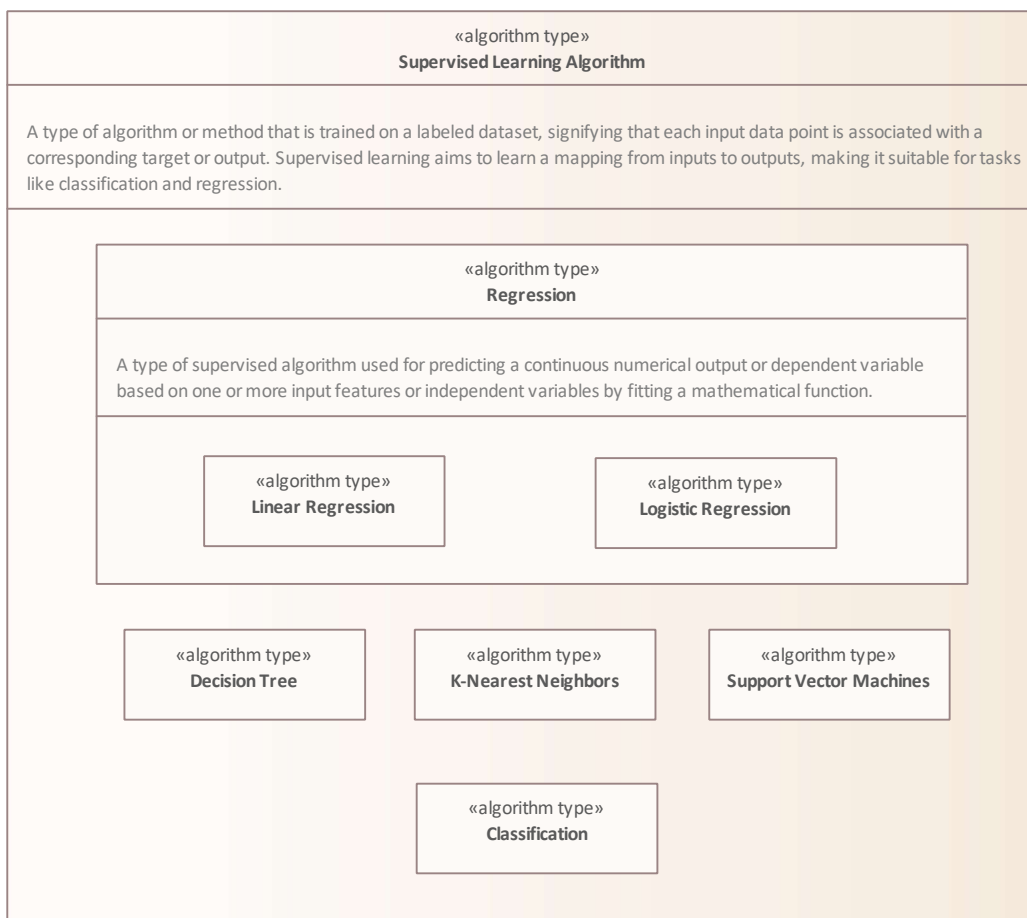


(from Diagrammen)

Algorithm Types

Supervised Learning Algorithms

Supervised Learning Algorithms



Classification

A type of supervised learning algorithm that categorizes a given set of input data into classes based on one or more variables. The primary goal of classification is to create a model that can generalize patterns and relationships within the data, leading to accurate predictions about the class labels when new instances are encountered.

Decision Tree

A type of supervised (or occasionally unsupervised) machine learning algorithm that split data into branches based on feature values, thus forming a tree that is traversed, selecting a branch leading to a conclusion or result.

K-Nearest Neighbors

A type of supervised learning algorithm that makes predictions based on the similarity of data points to their nearest neighbors in the training dataset. It attempts to determine what group a data point belongs to by analyzing the data points around it.

Linear Regression

A type of supervised machine learning algorithm or statistical method used for modeling the relationship between a dependent variable (target or response) and one or more independent variables (features or predictors).

Logistic Regression

A type of supervised machine learning algorithm used for binary and multiclass classification tasks, that is, those that only have two outcomes.

While it is named a regression algorithm, it can also be considered a classification type.

Regression

A type of supervised algorithm used for predicting a continuous numerical output or dependent variable based on one or more input features or independent variables by fitting a mathematical function.

Supervised Learning Algorithm

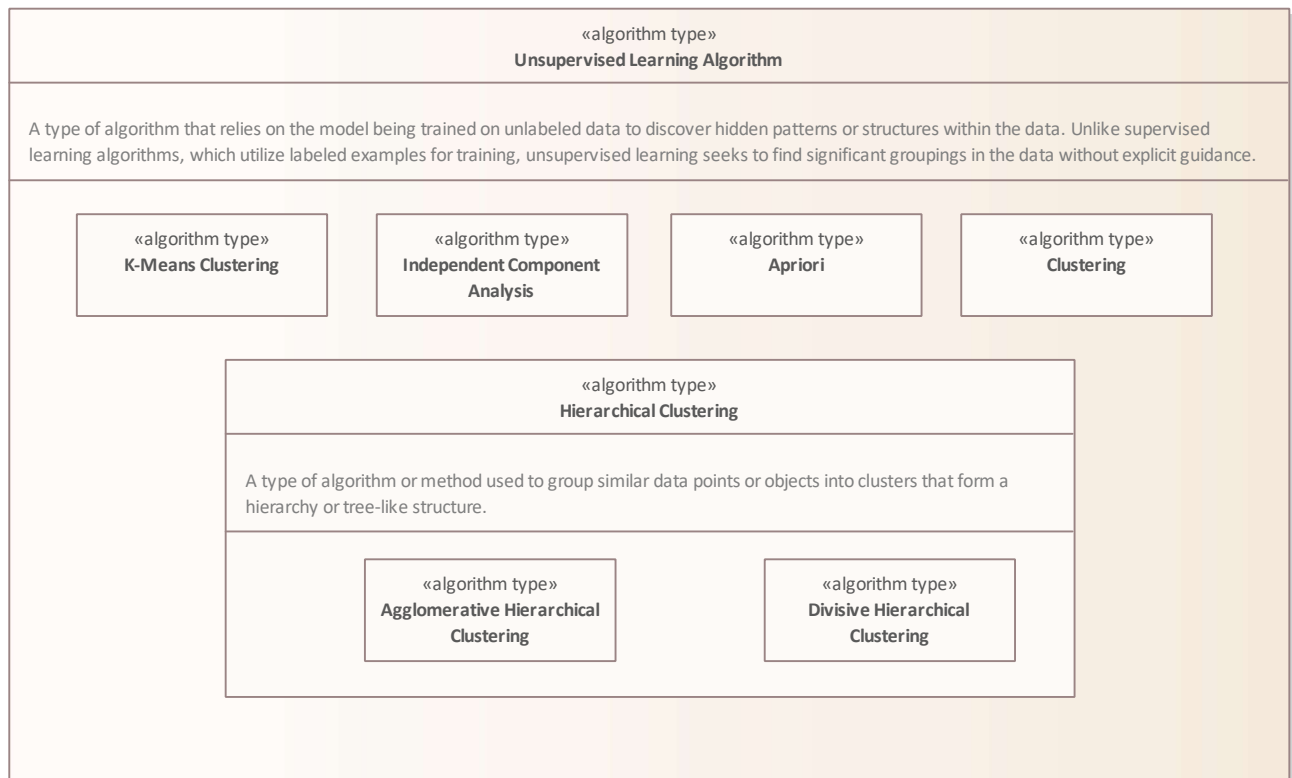
A type of algorithm or method that is trained on a labeled dataset, signifying that each input data point is associated with a corresponding target or output. Supervised learning aims to learn a mapping from inputs to outputs, making it suitable for tasks like classification and regression.

Support Vector Machines

Classification and regression technique that finds a hyperplane that best separates data points into classes.

Unsupervised Learning Algorithms

Unsupervised Learning Algorithms



Agglomerative Hierarchical Clustering

A type of hierarchical clustering algorithm that begins with each data point as its own cluster and then iteratively merges clusters closest to each other, creating a hierarchy or tree structure called a dendrogram.

Apriori

Apriori is an algorithm used for Association Rule Mining that searches for a series of recurring sets of items in the datasets. It then creates an association-based rule and correlations between the itemsets, extending them to larger and larger itemsets where the itemsets appear regularly in the dataset.

Clustering

A type of unsupervised algorithm that groups unlabeled data points into clusters based on their similarity. It is typically used in the exploratory data analysis phase to discover patterns and new information in the data.

Divisive Hierarchical Clustering

A type of hierarchical clustering algorithm that begins with all data points in one cluster and then divides them into smaller clusters in a top-down manner creating a hierarchy or tree structure called a dendrogram.

Hierarchical Clustering

A type of algorithm or method used to group similar data points or objects into clusters that form a hierarchy or tree-like structure.

Independent Component Analysis

A type of algorithm and computational method used in signal processing for separating a multivariate signal into additive sub-components or independent sources.

K-Means Clustering

A type of algorithm used to partition a dataset into distinct groups or clusters based on the similarity of data points. It is a high-speed algorithm that works with large datasets, minimizing the variance within each cluster.

Unsupervised Learning Algorithm

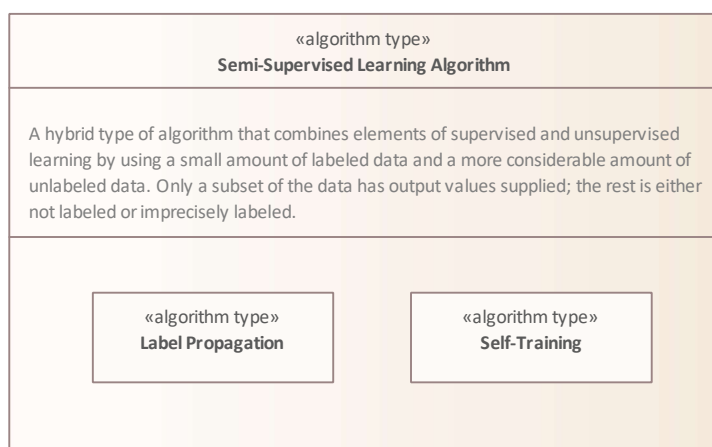
A type of algorithm that relies on the model being trained on unlabeled data to discover hidden patterns or structures within the data. Unlike supervised learning algorithms, which utilize labeled examples for training, unsupervised learning seeks to find significant groupings in the data without explicit guidance.

Dimensionality Reduction Algorithms

Semi-Supervised Learning Algorithms

A hybrid type of algorithm that combines elements of supervised and unsupervised learning by using a small amount of labeled data and a more considerable amount of unlabeled data. Only a subset of the data has output values supplied; the rest is either not labeled or imprecisely labeled.

Semi-Supervised Learning Algorithms



Label Propagation

A type of semi-supervised machine learning algorithm that assigns labels to previously unlabeled data points. It is used for classification tasks, especially when you have a limited amount of labeled data and a more extensive collection of unlabeled data. The algorithm creates a similarity matrix that measures the similarity or affinity between each pair of data points in the dataset.

Self-Training

A type of semi-supervised machine learning algorithm that assigns labels to previously unlabeled data points. It is used for classification tasks, especially when you have a limited amount of labeled data and a more extensive collection of unlabeled data. Parts of the algorithm are applied iteratively until convergence is achieved.

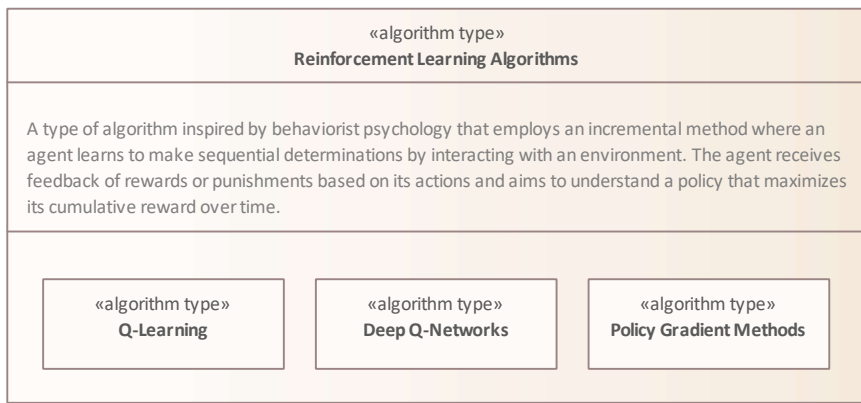
Semi-Supervised Learning Algorithm

A hybrid type of algorithm that combines elements of supervised and unsupervised learning by using a small amount of labeled data and a more considerable amount of unlabeled data. Only a subset of the data has output values supplied; the rest is either not labeled or imprecisely labeled.

Reinforcement Learning Algorithms

A type of algorithm inspired by that employs an incremental method where an agent learns to make sequential determinations by interacting with an environment. The agent receives feedback of rewards or punishments based on its actions and aims to understand a policy that maximizes its cumulative reward over time.

Reinforcement Learning Algorithms



Deep Q-Networks

A type of reinforcement learning algorithm useful in environments with high-dimensional state spaces. The algorithm integrates Q-Learning with deep neural networks to approximate the Q-values for each state-action pair stored in the Q-table.

Policy Gradient Methods

A type of reinforcement learning algorithm or method used to find an optimal policy for solving sequential decision-making problems in circumstances where both the state and action spaces can be continuous or high-dimensional. Policy gradient algorithms learn a parameterized policy that determines the probability distribution over actions.

Q-Learning

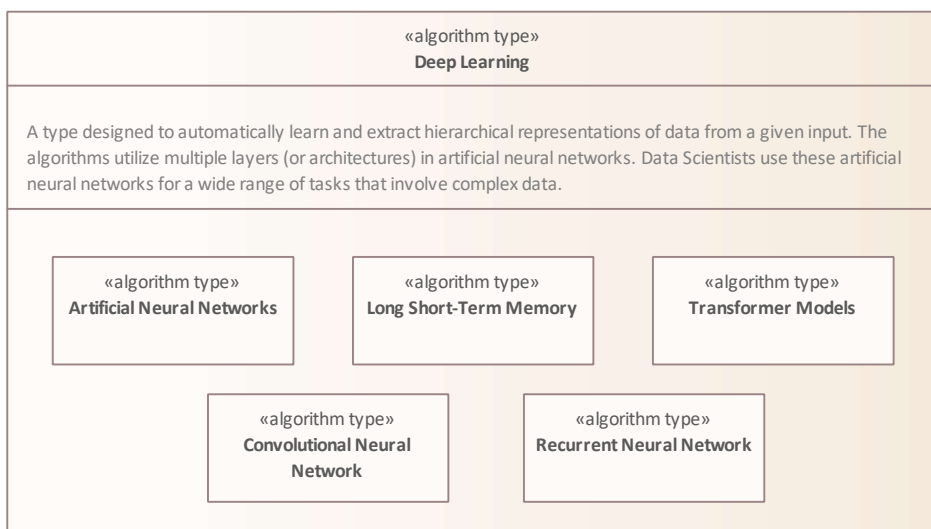
A type of Reinforcement learning algorithm used for solving sequential decision-making problems learning policy that will find the next best action, given a current state. It chooses this action randomly to maximize the cumulative reward, known as the Q-value.

Reinforcement Learning Algorithms

A type of algorithm inspired by behaviorist psychology that employs an incremental method where an agent learns to make sequential determinations by interacting with an environment. The agent receives feedback of rewards or punishments based on its actions and aims to understand a policy that maximizes its cumulative reward over time.

Deep Learning Algorithms

Deep Learning Algorithms



Artificial Neural Networks

A type of deep learning algorithm commonly referred to simply as a neural network. They represent a learning model inspired by the connection of neurons and the structure and functioning of the human brain.

Convolutional Neural Network

A regularized type of feed-forward neural network that learns feature engineering by itself using filters optimization. Specifically designed for tasks involving grid-like data, such as images and videos.

Deep Learning

A type designed to automatically learn and extract hierarchical representations of data from a given input. The algorithms utilize multiple layers (or architectures) in artificial neural networks. Data Scientists use these artificial neural networks for a wide range of tasks that involve complex data.

Long Short-Term Memory

A type of recurrent neural network deep learning algorithm capable of learning order dependence in sequence prediction problems. The algorithms use specialized gates designed to capture and remember information over long data sequences, making them suited for modeling sequences with complex dependencies.

Recurrent Neural Network

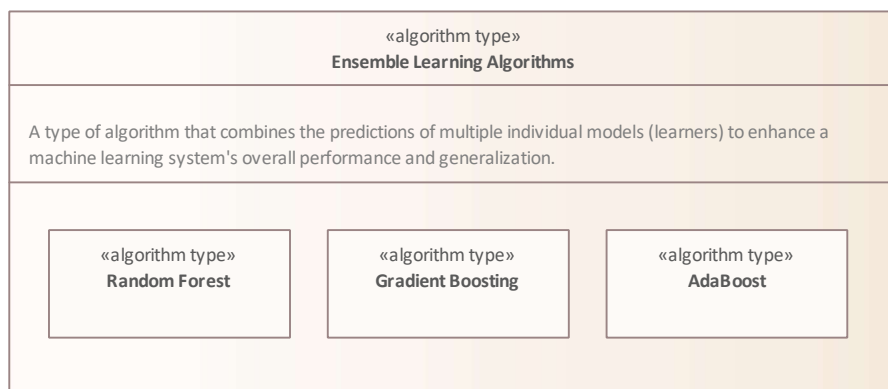
A type of artificial neural network suited for tasks involving sequential data, like time series and text. They are suitable for temporal or ordinal problems, such as image captioning, translation, and processing of natural languages, including speech recognition.

Transformer Models

Transformers have gained popularity for their ability to efficiently capture long-range dependencies in sequences and their parallelizability, making them highly scalable. They solve problems of transduction or transformation of input sequences into output sequences.

Ensemble Learning Algorithms

Ensemble Learning Algorithms



AdaBoost

A type of binary classification ensemble algorithm that focuses on data points misclassified by the current ensemble. It improves the performance of weak learners by iteratively training weak classifiers, assigning weights to data points, and combining them into a strong classifier.

Ensemble Learning Algorithms

A type of algorithm that combines the predictions of multiple individual models (learners) to enhance a machine learning system's overall performance and generalization.

Gradient Boosting

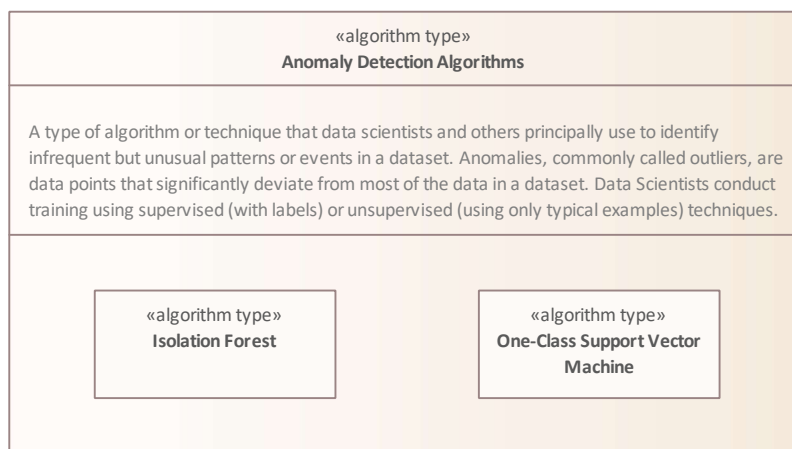
A type of ensemble machine learning algorithm applicable for both classification and regression tasks that combines the predictions from multiple individual models to create a more accurate and robust final prediction. Gradient Boosting sequentially builds an ensemble of decision trees, with each new tree attempting to correct the previous trees' prediction errors (residuals). The algorithm captures complex relationships in the data with high performance.

Random Forest

A type of ensemble machine learning algorithm applicable for both classification and regression tasks. Important properties include predictive accuracy, versatility, and resistance to overfitting. It combines the predictions from multiple individual machine-learning models to produce more robust and precise final predictions. It typically consists of a series of decision trees.

Anomaly Detection Algorithms

Anomaly Detection Algorithms



Anomaly Detection Algorithms

A type of algorithm or technique that data scientists and others principally use to identify infrequent but unusual patterns or events in a dataset. Anomalies, commonly called outliers, are data points that significantly deviate from most of the data in a dataset. Data Scientists conduct training using supervised (with labels) or unsupervised (using only typical examples) techniques.

Isolation Forest

A type of unsupervised machine learning algorithm used for detecting abnormal data points in a dataset - referred to as anomaly detection. Isolation Forests are highly effective for detecting outliers in high-dimensional data because the outliers are typically isolated and separated from normal data points.

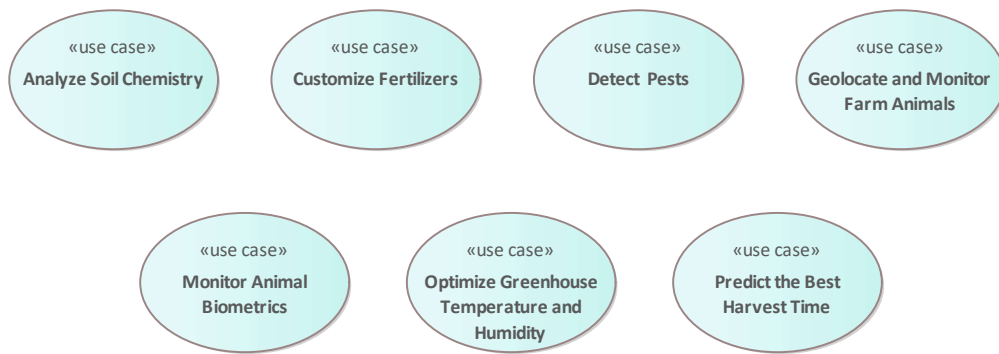
One-Class Support Vector Machine

A type of unsupervised learning algorithm used to detect anomalies or outliers - data points that deviate significantly from most of the data. Unlike the standard supervised SVM, the one-class SVM does not include target labels for the model training process. The algorithm is adapted for the one-class classification problem, learns the boundary for the primary data points, and then identifies the data points outside the border as anomalies.

IoT Agriculture Use Case

Agricultural systems, including crop production, livestock, poultry, bees, and other animal farming, benefit from information from IoT devices to improve efficiency and increase profitability.

IoT Agriculture Use Case



Analyze Soil Chemistry

IoT devices and sensors can monitor soil chemistry to report on pH and mineral levels. The chemistry of the soil will influence the supply of nutrients required for plant growth and production. The soil composition is critical to soil degradation and water transport processes.

Customize Fertilizers

IoT devices measure soil nutrients, and soil scientists use this information to customize fertilizers constituents and quantities applicable to the plant's growth stage resulting in optimized yields and reduced costs.

Detect Pests

Farmers use IoT devices such as noise sensors to detect swarms of insects and thermographic sensors that measure the differences in surface temperature of the plant leaves and canopy to determine the presence of plant pathogens.

Geolocate and Monitor Farm Animals

Animal location monitoring provides data scientists and researchers with the information required to evaluate such factors as the movements of animals in paddocks and farms, the spatial heterogeneity of field occupancy by livestock, poultry, bees, pasture utilization, animal performance and behavior, and herd socialization.

Monitor Animal Biometrics

Farm managers and data scientists use Livestock biometrics to identify farm animals using a pattern recognition system based on the physical attributes of an animal. Analysts use the data for livestock identification, traceability, and health and welfare assessments to enhance efficiency, productivity, and sustainability on your farm.

Optimize Greenhouse Temperature and Humidity

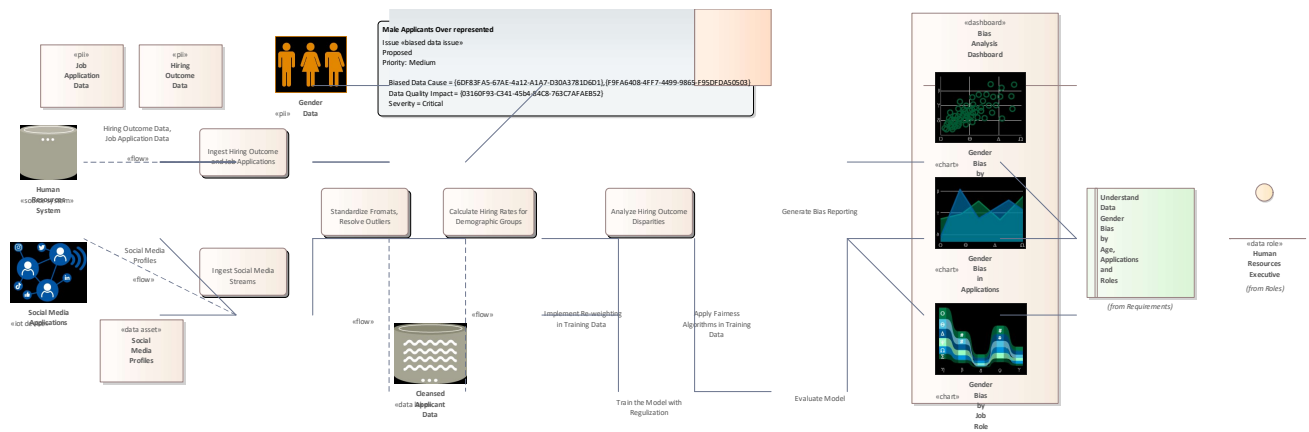
Yields from greenhouses can be optimized by using IoT pressure, temperature, light, and humidity sensors. The data collected allows automatic adjustment of temperature, light, and other conditions, and irrigation to maximize the yields and efficiency of the greenhouse.

Predict the Best Harvest Time

Farm managers use IoT sensors to predict optimal crop harvest times using devices that measure weather conditions such as temperature and humidity and monitor soil properties.

IoT Agriculture Use Case Example

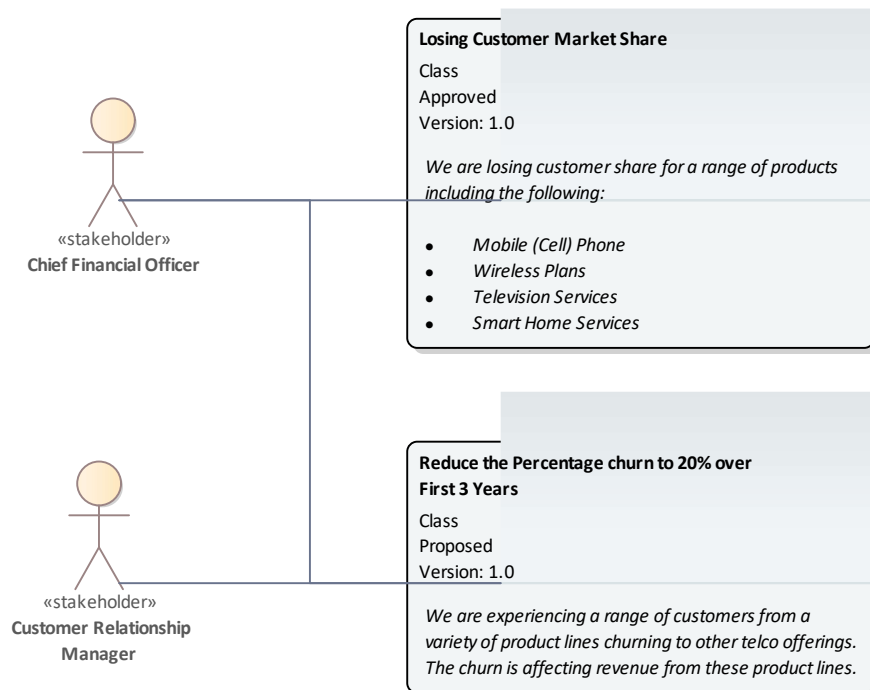
Gender Bias in Human Resources Hiring



Random Forest - Telecommunication Predicting Customer Churn

Strategy

Motivation



Chief Financial Officer

Customer Relationship Manager

Losing Customer Market Share

We are losing customer share for a range of products including the following:

- Mobile (Cell) Phone
- Wireless Plans

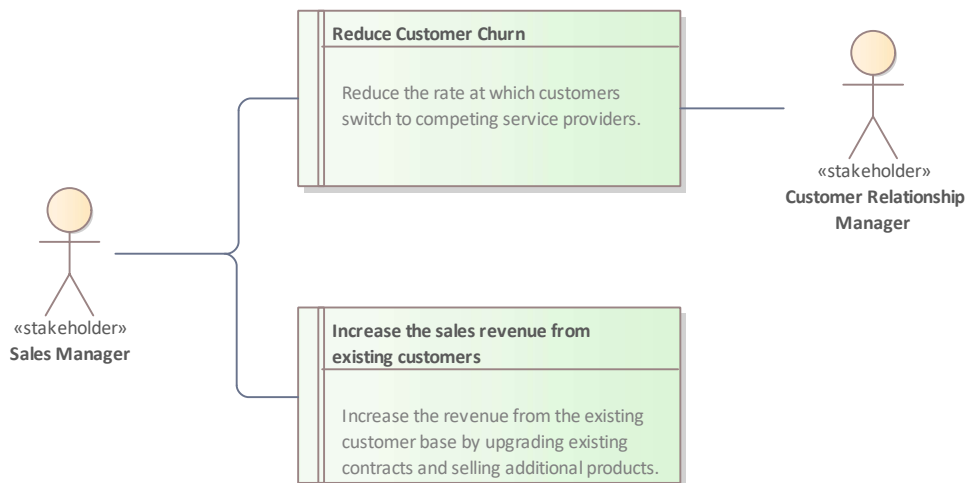
- Television Services
- Smart Home Services

Reduce the Percentage churn to 20% over First 3 Years

We are experiencing a range of customers from a variety of product lines churning to other telco offerings. The churn is affecting revenue from these product lines.

Business

Business



Customer Relationship Manager

Increase the sales revenue from existing customers

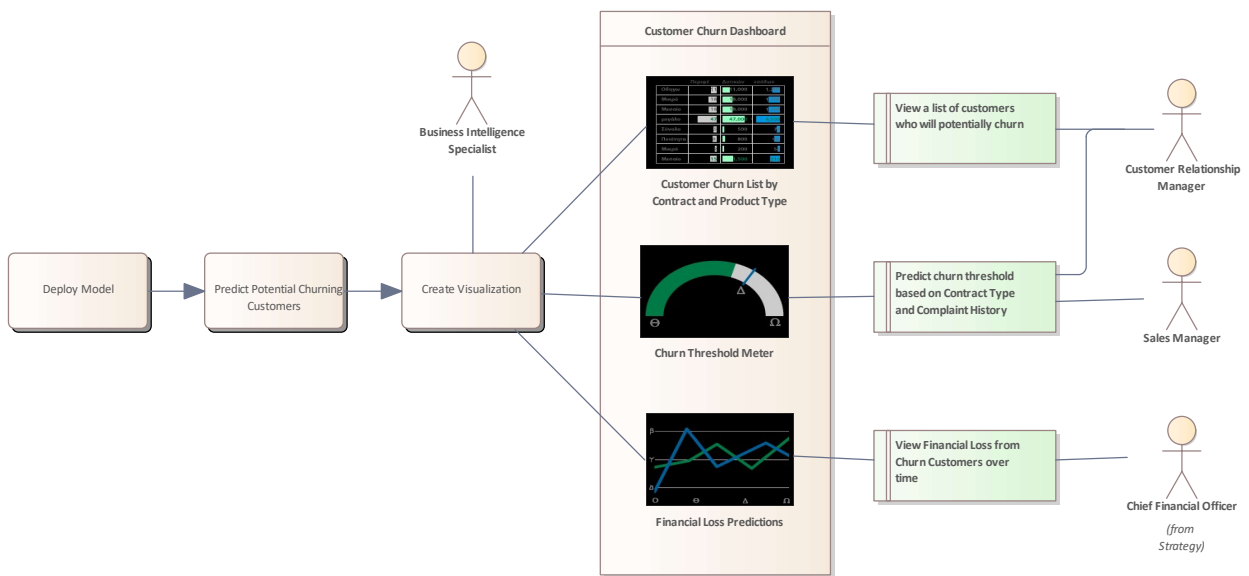
Increase the revenue from the existing customer base by upgrading existing contracts and selling additional products.

Reduce Customer Churn

Reduce the rate at which customers switch to competing service providers.

Sales Manager

Dashboard Visualization

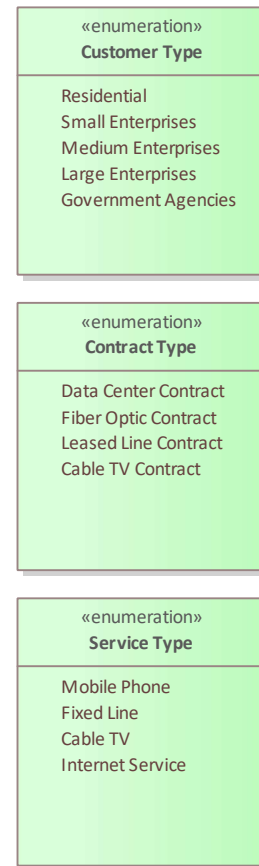
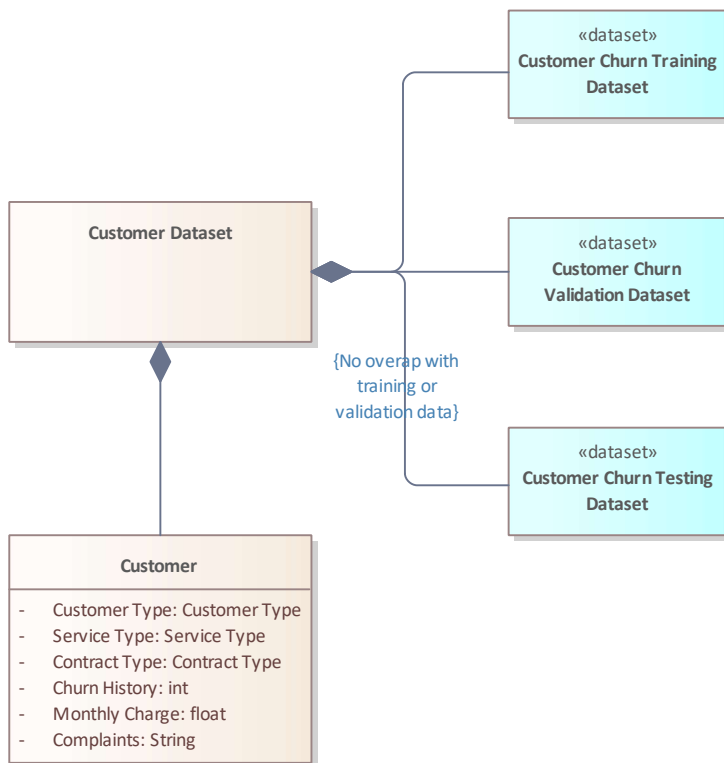


Business Intelligence Specialist

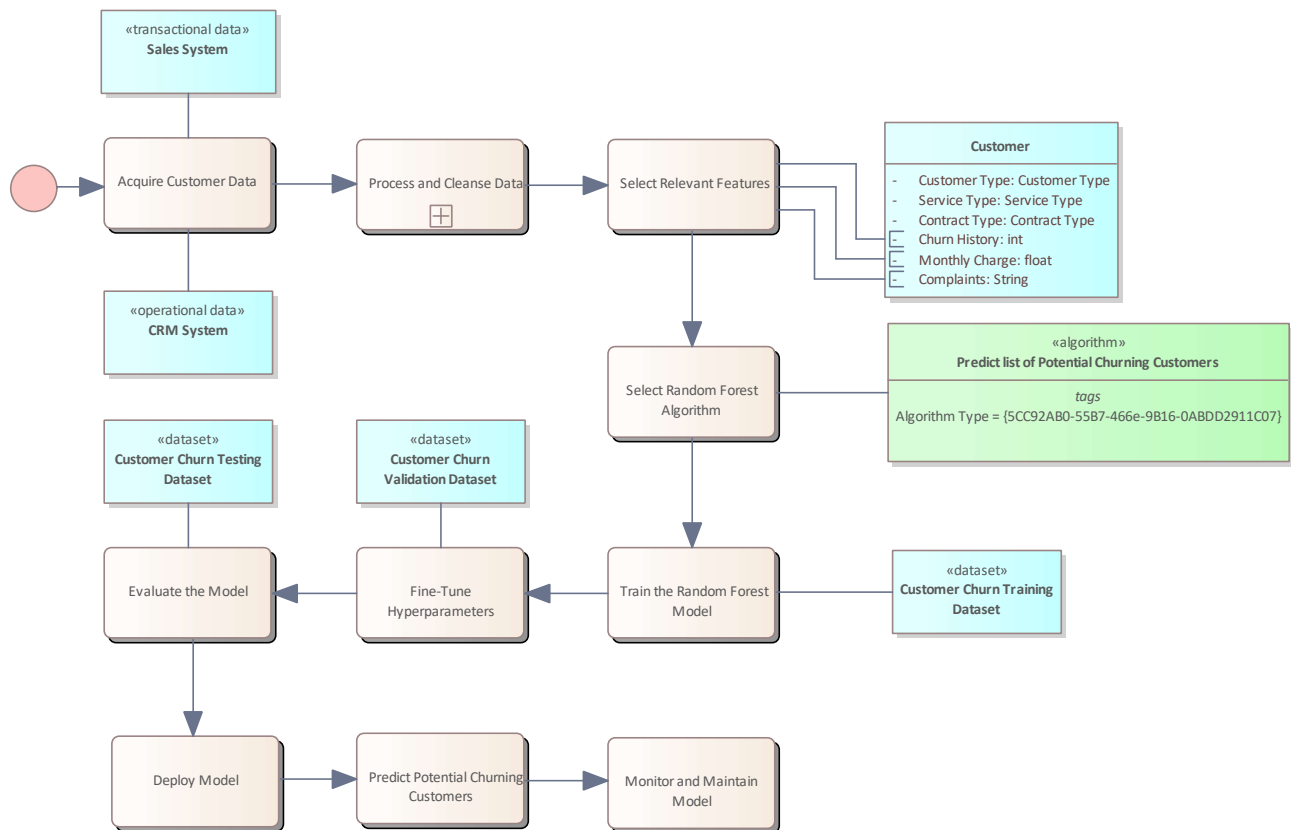
Business Intelligence Specialists are responsible for designing, developing, and maintaining business intelligence solutions, such as data warehouses, data marts, and reporting systems. They enable users to access, visualize and analyze data easily. The visualizations and analysis support meaningful and actionable insights that stakeholders can use to support decision-making and drive business strategies.

Data

Data



Data Science



Acquire Customer Data

Collect data from transactional systems including the sales database and the Customer Relationship Management System. Divide the data set into three subsets: Training, Validation and Testing Datasets.

CRM System

Customer

Customer Churn Testing Dataset

Customer Churn Training Dataset

Customer Churn Validation Dataset

Deploy Model

Evaluate the Model

Fine-Tune Hyperparameters

Tune or optimize the Random Forest hyperparameters. You can use approaches like cross-validation to find the most suitable hyperparameters to optimize. These could include the number of trees in the forest, the maximum depth of trees, and the minimum number of samples required to split a node.

Monitor and Maintain Model

Predict list of Potential Churning Customers

Predict Potential Churning Customers

Process and Cleanse Data

Analyze the dataset and perform any necessary preprocessing of the data. The processes may include data sampling, handling missing data, Removing duplicates, converting categorical Variables into Numerical Representations, and scaling numerical features.

Sales System

Select Random Forest Algorithm

Choose the Random Forest Algorithms as the basis for a predictive model due to its ability to work with both numerical and categorical data points. It also has the advantage of being able to handle complex relationships in the data.

Select Relevant Features

Select the features that could possibly contribute to a customer's decision to churn.

Start

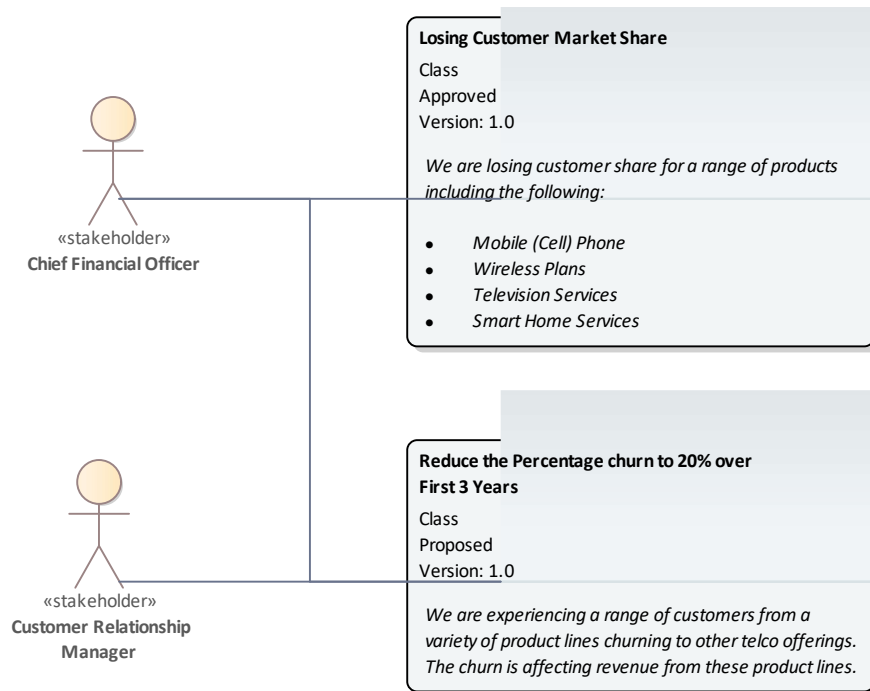
Train the Random Forest Model

Using the training dataset to train the Random Forest model, specifying that the target variable is whether a customer churned or not. The model will learn to predict churn based on the features defined in the 'Select Relevant Features ' activity.

Random Forest - Telecommunication Predicting Customer Churn

Strategy

Motivation



Chief Financial Officer

Customer Relationship Manager

Losing Customer Market Share

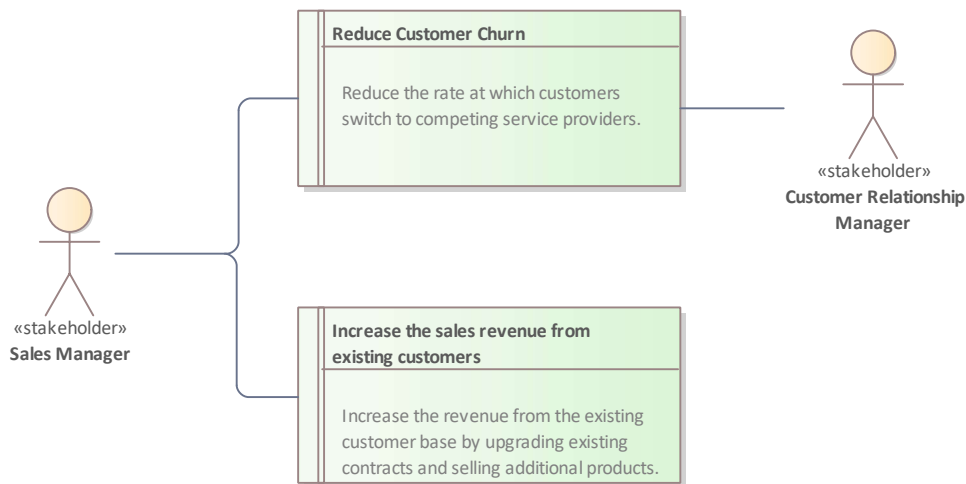
We are losing customer share for a range of products including the following:

- Mobile (Cell) Phone
- Wireless Plans
- Television Services
- Smart Home Services

Reduce the Percentage churn to 20% over First 3 Years

We are experiencing a range of customers from a variety of product lines churning to other telco offerings. The churn is affecting revenue from these product lines.

Business



Customer Relationship Manager

Increase the sales revenue from existing customers

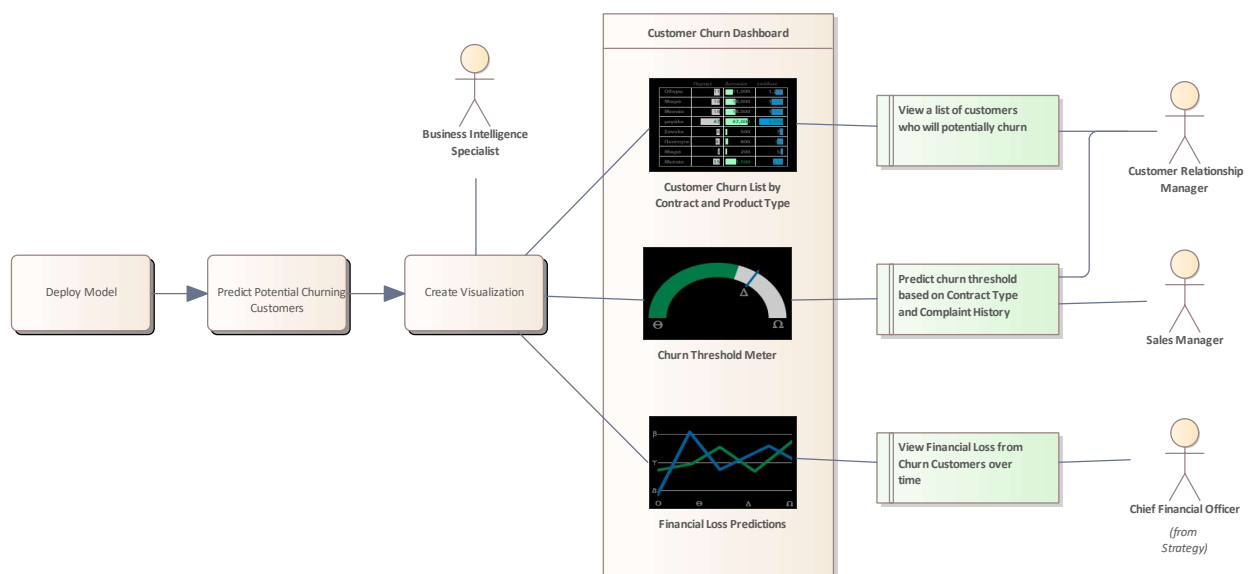
Increase the revenue from the existing customer base by upgrading existing contracts and selling additional products.

Reduce Customer Churn

Reduce the rate at which customers switch to competing service providers.

Sales Manager

Dashboard Visualization

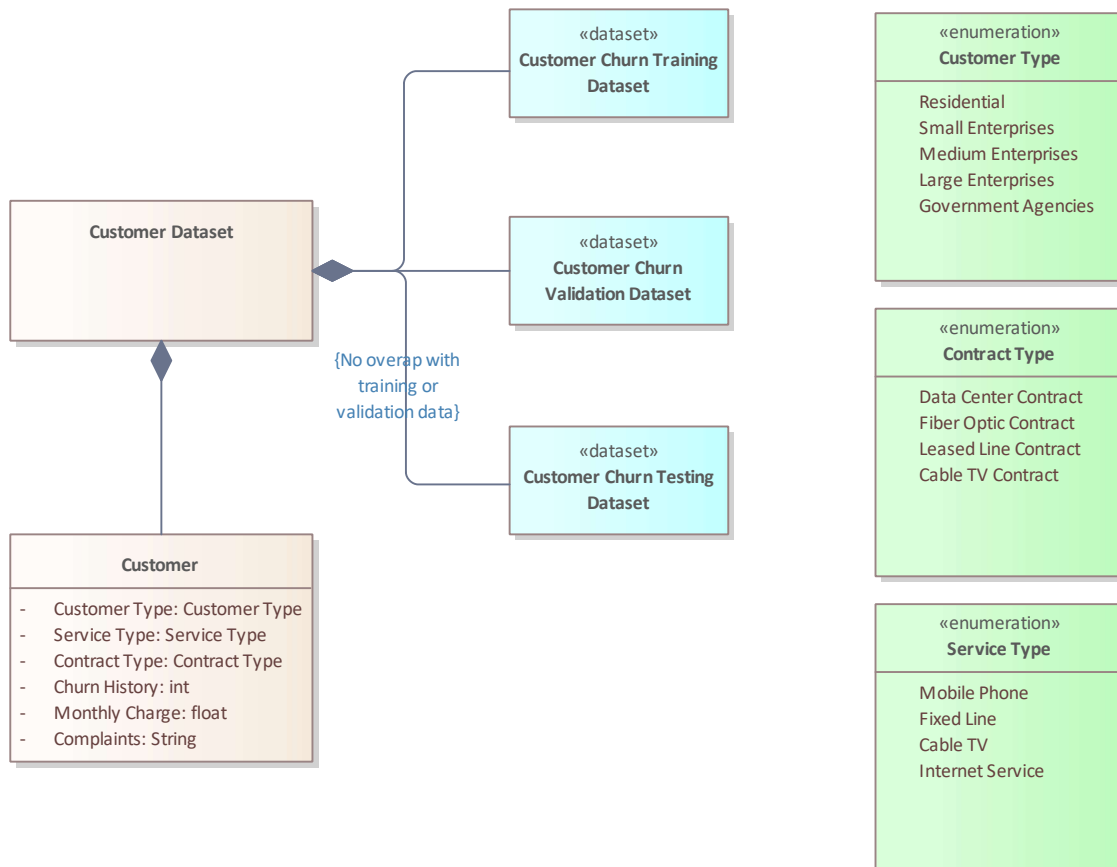


Business Intelligence Specialist

Business Intelligence Specialists are responsible for designing, developing, and maintaining business intelligence solutions, such as data warehouses, data marts, and reporting systems. They enable users to access, visualize and analyze data easily. The visualizations and analysis support meaningful and actionable insights that stakeholders can use to support decision-making and drive business strategies.

Data

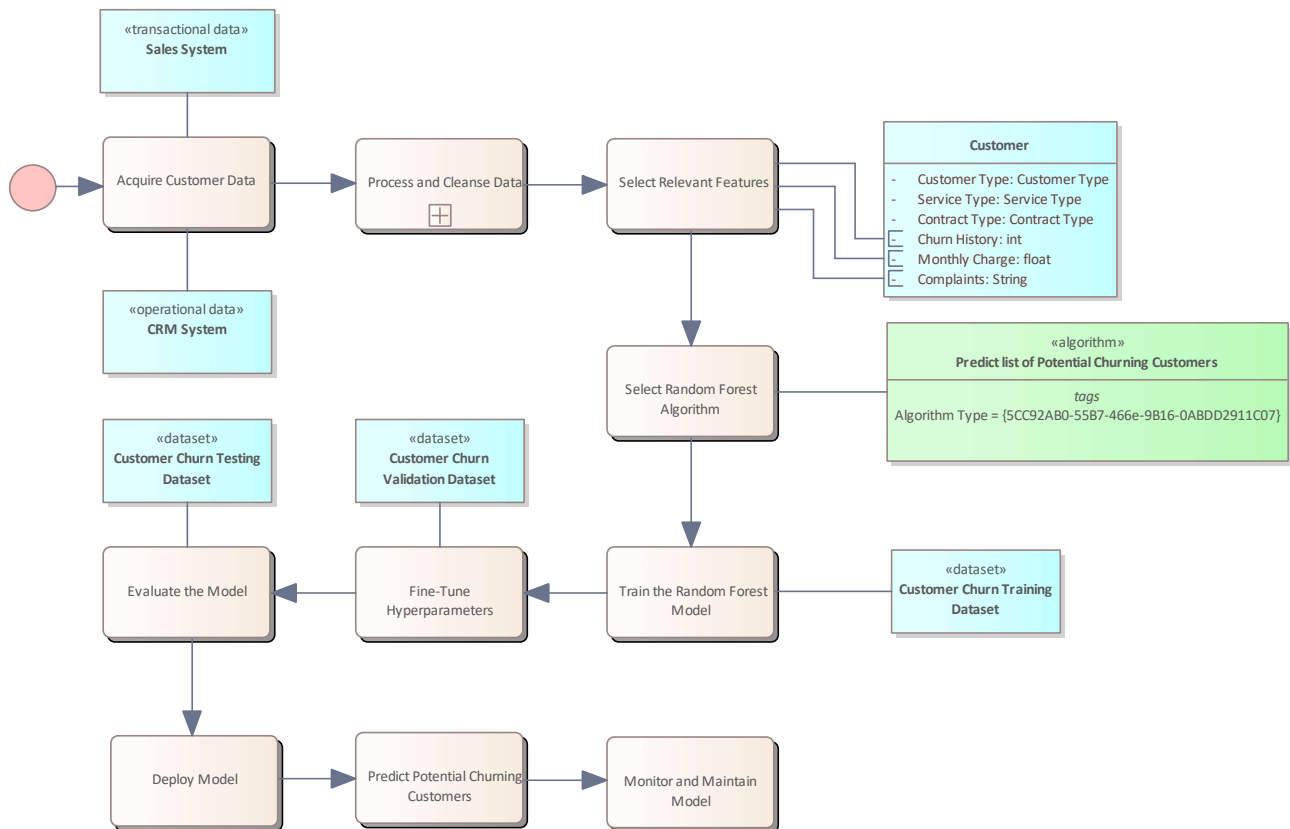
Data



Application

Data Science

Data Science



Acquire Customer Data

Collect data from transactional systems including the sales database and the Customer Relationship Management System. Divide the data set into three subsets: Training, Validation and Testing Datasets.

Fine-Tune Hyperparameters

Tune or optimize the Random Forest hyperparameters. You can use approaches like cross-validation to find the most suitable hyperparameters to optimize. These could include the number of trees in the forest, the maximum depth of trees, and the minimum number of samples required to split a node.

Process and Cleanse Data

Analyze the dataset and perform any necessary preprocessing of the data. The processes may include data sampling, handling missing data, Removing duplicates, converting categorical Variables into Numerical Representations, and scaling numerical features.

Sales System

Select Random Forest Algorithm

Choose the Random Forest Algorithms as the basis for a predictive model due to its ability to work with both numerical and categorical data points. It also has the advantage of being able to handle complex relationships in the data.

Select Relevant Features

Select the features that could possibly contribute to a customer's decision to churn.

Start

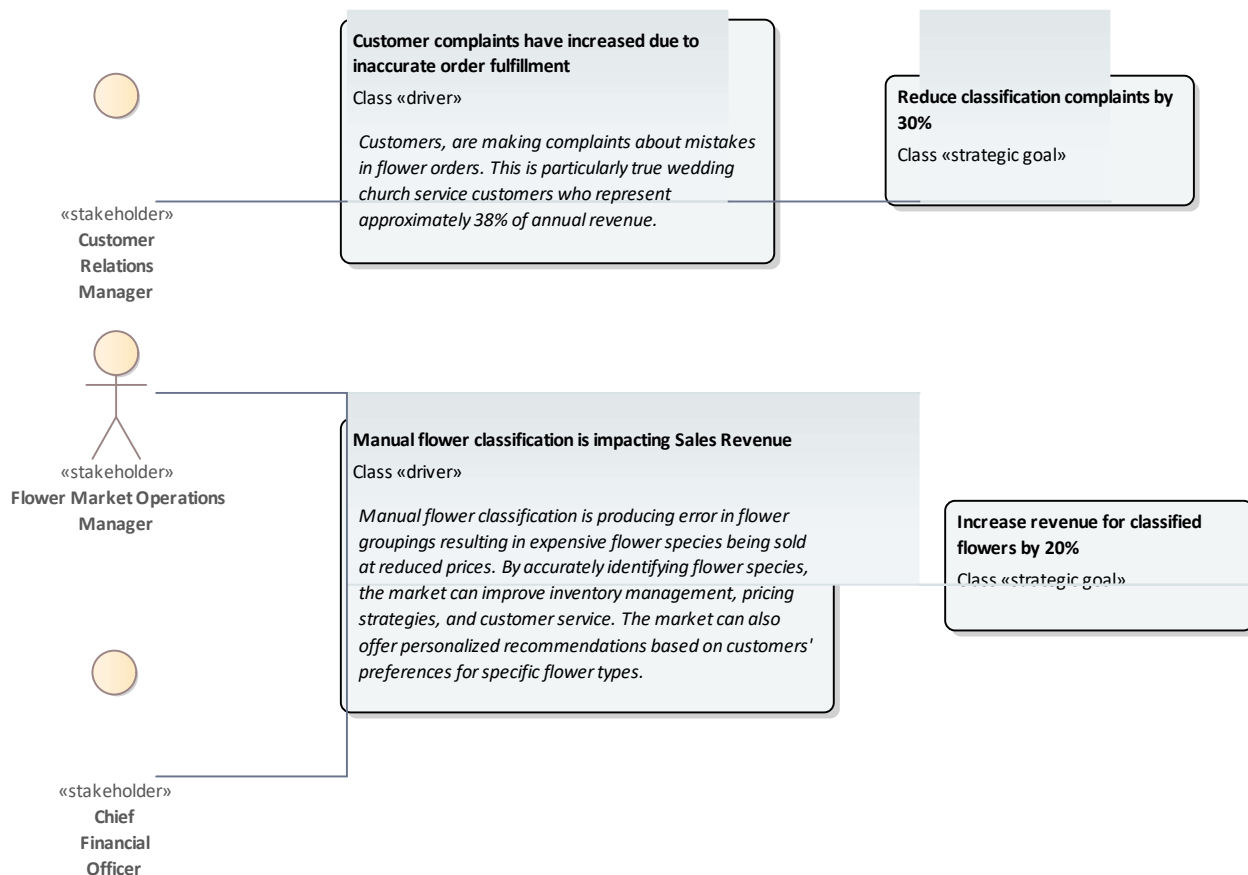
Train the Random Forest Model

Using the training dataset to train the Random Forest model, specifying that the target variable is whether a customer churned or not. The model will learn to predict churn based on the features defined in the 'Select Relevant Features' activity.

K-Nearest Neighbors - Classifying Flowers

Strategy

Strategy



Chief Financial Officer

Customer complaints have increased due to inaccurate order fulfillment

Customers, are making complaints about mistakes in flower orders. This is particularly true wedding church service customers who represent approximately 38% of annual revenue.

Customer Relations Manager

Flower Market Operations Manager

Increase revenue for classified flowers by 20%

Manual flower classification is impacting Sales Revenue

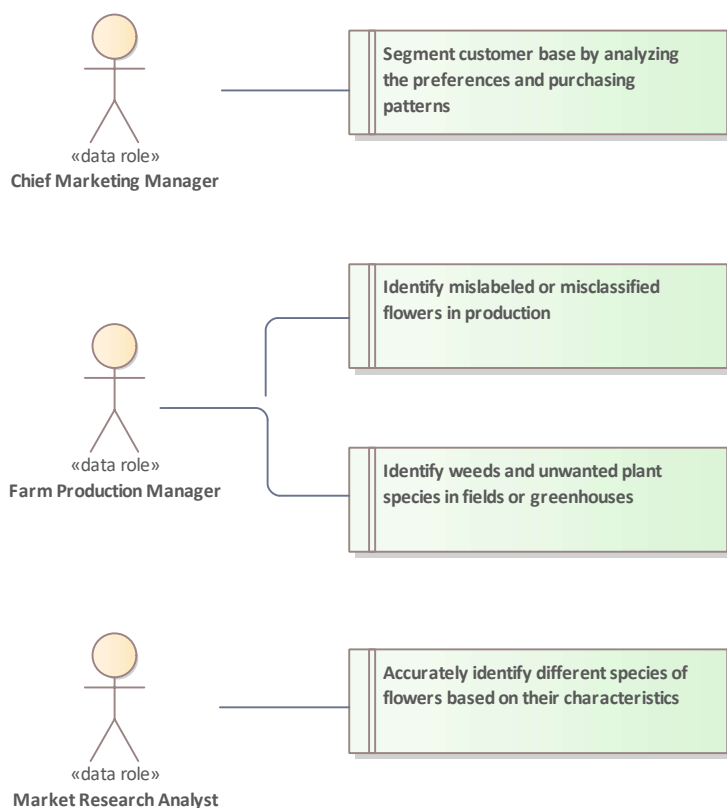
Manual flower classification is producing error in flower groupings resulting in expensive flower species being sold at reduced prices. By accurately identifying flower species, the market can improve inventory

management, pricing strategies, and customer service. The market can also offer personalized recommendations based on customers' preferences for specific flower types.

Reduce classification complaints by 30%

Business

Business



Accurately identify different species of flowers based on their characteristics

Chief Marketing Manager

Farm Production Manager

Identify mislabeled or misclassified flowers in production

Ensuring the accuracy of species identification is crucial, especially for businesses that deal with rare or endangered plant species.

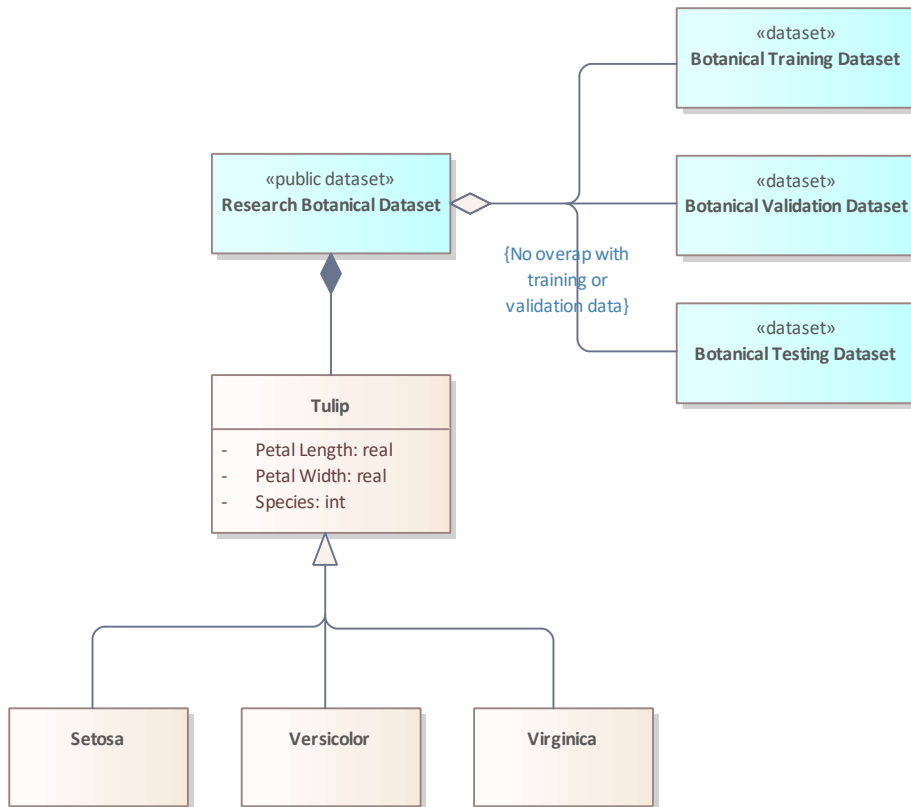
Identify weeds and unwanted plant species in fields or greenhouses

Market Research Analyst

Segment customer base by analyzing the preferences and purchasing patterns

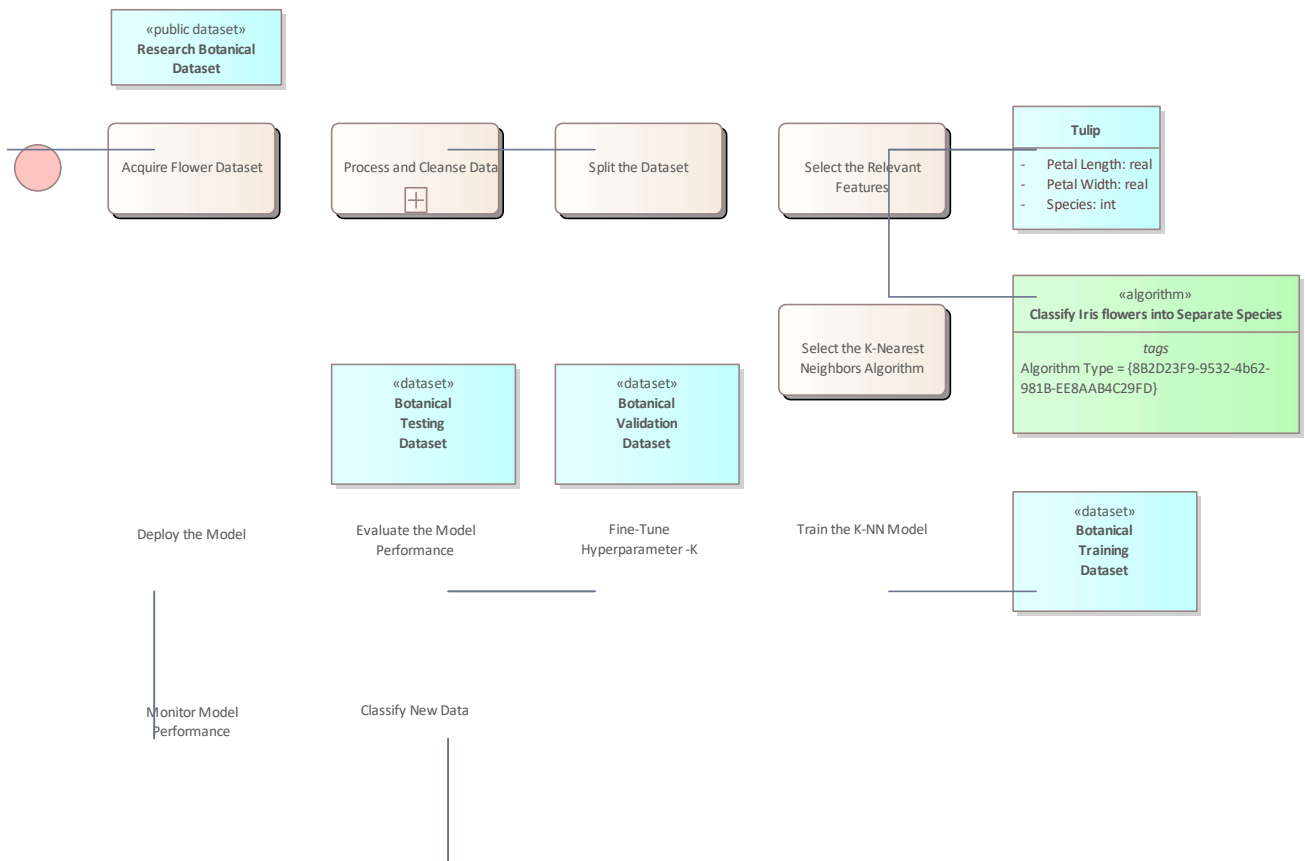
Data

K-Nearest Neighbors - Botany Classifying Flowers



Data Science

Data Science



Acquire Flower Dataset

Acquire a dataset that includes measurements of iris flowers and their associated species labels.

Botanical Testing Dataset

Botanical Training Dataset

Botanical Validation Dataset

Classify Iris flowers into Separate Species

The algorithm will classify the rows in the dataset into one of the three species - namely iris flowers into one of three species:

- Setosa
- Versicolor
- Virginica

Classify New Data

Use the model to make predications by classifying the rows of flowers in a new dataset.

Deploy the Model

End

Evaluate the Model Performance

You can evaluate the model's performance using the testing dataset. Standard evaluation metrics for classification tasks include accuracy, precision, recall, and F1-score. You could also consider producing a confusion matrix to determine the model's efficacy in classifying each iris species.

Fine-Tune Hyperparameter -K

Select the optimal value of the hyperparameter "K," which designates the number of nearest neighbors to consider when making a prediction by using a technique like cross-validation.

Monitor Model Performance

Continuously monitor the model's performance in situ in your production environments. You may need to re-train the model occasionally with new data to keep it current.

Process and Cleanse Data

Analyze the dataset and perform any necessary preprocessing of the data. The processes may include data sampling, handling missing data, Removing duplicates, converting categorical Variables into Numerical Representations, and scaling numerical features.

Research Botanical Dataset

Select the K-Nearest Neighbors Algorithm

The K-NN algorithm is a non-parametric supervised learning method. It finds the K training samples closest to the new data point and then makes a prediction based on the majority class among those K neighbors. Where K specifies how many neighbors will be checked to determine the classification of a specific query point.

Select the Relevant Features

Choose the features from the flower data, in this case use the following:

- Petal Length
- Petal Width

These two features will be used to determine the species of iris flowers.

Split the Dataset

Split the dataset into three distinct subsets of data namely:

- Training
- Testing
- Validation

Start

Train the K-NN Model

Train the K-NN model using the training data set, which should be labeled.

Tulip

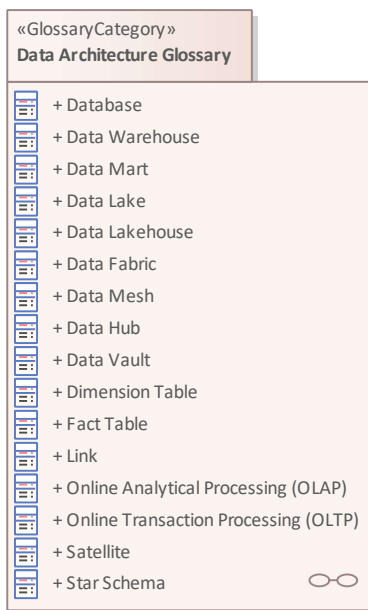
Glossaries

Er zitten een aantal glossary elementen in de voorbeeld uitwerking met daarin een glossary model met definities van een groot aantal zaken die relevant zijn binnen data management, metadata, data architectuur en data science.

Binnen EA is dit een handige functionaliteit maar wellicht ook voor het uitwerken van een data architectuur of een metadata uitwerking is dit zeer interessant.

Data Architecture Glossary

Data Architecture Glossary Category



(from Glossaries)

Data Architecture Glossary

Data Architecture Glossary Entries

<div>«GlossaryEntry» Database</div> <div>A Database is a structured and organized collection of data that is stored and accessed electronically . Data roles store, retrieve, manipulate, and manage the data using a database management system.</div>	<div>«GlossaryEntry» Data Warehouse</div> <div>A Data Warehouse is a structured and optimized storage system for collecting, storing, and managing data from various sources within an organization. It is a central repository for historical, current, transformed, integrated, and organized data for querying, reporting, and analytical purposes. The primary goal of a Data Warehouse is to support business intelligence (BI), data analysis, and data science activities by providing a consistent and reliable data source for decision-making and drawing insights.</div>	<div>«GlossaryEntry» Data Mart</div> <div>A Data Mart is a discreet subset of a data warehouse and is typically tailored and optimized for distinct groups of users, business departments, or functions within an organization. It is a focused data repository designed to meet a group of users' unique reporting and analysis needs. Data marts improve performance, simplify access, and enhance the relevance of data for those who require specific insights.</div>	<div>«GlossaryEntry» Data Lake</div> <div>A Data Lake is a repository or storage system comprising large amounts of raw, structured, semi-structured, and unstructured data. It can accommodate data in its native format and at scale, making it a flexible and complete solution for storing and analyzing diverse types of data. Contrary to conventional data storage systems that require data to be structured and organized in advance, a Data Lake allows data to be ingested and stored as-is, with data analysts and scientists determining its structure at the time of analysis.</div>
<div>«GlossaryEntry» Data Lakehouse</div> <div>A Data Lakehouse is a hybrid repository that combines the advantages of both Data Lakes and Data Warehouses and addresses some of the limitations and challenges of each system. A Data Lakehouse can store structured, unstructured, and semi-structured data. They aim to deliver the best of both approaches and provide a harmonious and scalable solution for managing and analyzing data in various formats while producing the performance and structure required for efficient querying and analytics.</div>	<div>«GlossaryEntry» Data Fabric</div> <div>A Data Fabric is an architecture framework and set of data services designed to provide a unified and consistent way to manage, integrate, access, and share data across an organization's various systems, applications, and environments. It addresses the challenges of data silos, data fragmentation and standardizes data management practices across cloud, on-premises, and edge devices.</div>	<div>«GlossaryEntry» Data Mesh</div> <div>A Data Mesh operates a federated data architecture where data is distributed across various storage systems and domains that provide access to their data products through well-defined APIs and interfaces. With data mesh, the responsibility for analytical data lies with the domain teams, not the central data team, who provide a domain-agnostic data platform.</div>	<div>«GlossaryEntry» Data Hub</div> <div>A Data Hub is a centralized and integrated repository for storing, managing, and sharing data from different sources within an organization. It delivers a unified view of an organization's data assets, enabling efficient data management, analysis, and collaboration between various data roles.</div>
<div>«GlossaryEntry» Data Vault</div> <div>Data Vault is a data modeling and architecture method used in Data Warehouse Architecture and business intelligence. It provides a structured and scalable approach for handling and organizing data in complex and volatile environments. The Data Vault method seeks to improve flexibility, scalability, and maintainability in data warehouse initiatives, mainly when there are many data sources and data structures subject to frequent changes.</div>	<div>«GlossaryEntry» Dimension Table</div> <div>A Dimension Table is a critical entity of a dimensional model used in Data Warehouse Architecture. It contains descriptive attributes that provide context and descriptive information about the business entities being analyzed. Dimension tables provide the context and details required to analyze the measures stored in fact tables. They allow data analysts and others to analyze data across several dimensions, enabling meaningful analysis, reporting, and querying in Data Warehouse Architecture systems and business intelligence environments.</div>	<div>«GlossaryEntry» Fact Table</div> <div>A Fact Table is a central component that holds quantitative and numerical data related to the business processes or events being analyzed in a star schema. The table is part of a relational database design used in Data Warehouse Architecture, structured primarily for online analytical processing (OLAP). Fact tables contain descriptive attributes or context related to the measures in the fact table. Foreign keys establish relationships with one or more dimension tables. Data analysts and others use these keys to join and retrieve descriptive information from the dimension tables that provide context to the measures stored in the fact table.</div>	<div>«GlossaryEntry» Link</div> <div>A Link is a connection or relationship that exists between two entities in a data model or schema. Foreign key relationships are links where one or more fields or columns in one table refer to a primary key field in another table.</div>
<div>«GlossaryEntry» Online Analytical Processing (OLAP)</div> <div>Online Analytical Processing (OLAP) is a class of technology used for organizing and analyzing large volumes of data to support business intelligence (BI) and decision-making processes. OLAP systems allow users to explore and analyze multidimensional data interactively, gain insights, identify tendencies, and make informed decisions. Data scientists, analysts, and others use OLAP to efficiently process complex queries and provide an intuitive interface for data exploration.</div>	<div>«GlossaryEntry» Online Transaction Processing (OLTP)</div> <div>Online Transaction Processing (OLTP) is a class of technology used for managing and processing high volumes of operational and transactional data in real-time. OLTP systems can handle routine business operations, such as capturing, updating, and retrieving data related to daily transactions. These transactions include processing customer orders, updating inventory levels, recording financial transactions, and more.</div>	<div>«GlossaryEntry» Satellite</div> <div>Satellite data modeling is a method that creates additional tables, often referred to as satellite tables or simply satellites, to define and store additional or detailed attributes related to a particular dimension. Data Modelers use the tables to provide supplementary descriptive details or historical changes that might not be appropriate to include in the main dimension table</div>	<div>«GlossaryEntry» Star Schema</div> <div>A Star Schema is a data modeling method used in Data Warehouse Architecture that manages data into a central fact table surrounded by multiple dimension tables. The fact table is placed in the center and is surrounded by several dimension tables in a radial pattern resembling the points of a star. The fact table contains the quantitative data or measures, while the dimension tables store descriptive attributes that provide context to the fact table's measures. Each dimension table represents a specific aspect of the business. The star schemas' straightforward design facilitates data retrieval and analysis in Data Warehouse Architecture systems and business intelligence environments.</div>

Data Fabric

A Data Fabric is an architecture framework and set of data services designed to provide a unified and consistent way to manage, integrate, access, and share data across an organization's various systems, applications, and environments. It addresses the challenges of data silos, data fragmentation and standardizes data management practices across cloud, on-premises, and edge devices.

Data Hub

A Data Hub is a centralized and integrated repository for storing, managing, and sharing data from different sources within an organization. It delivers a unified view of an organization's data assets, enabling efficient data management, analysis, and collaboration between various data roles.

Data Lake

A Data Lake is a repository or storage system comprising large amounts of raw, structured, semi-structured, and unstructured data. It can accommodate data in its native format and at scale, making it a flexible and complete solution for storing and analyzing diverse types of data. Contrary to conventional data storage systems that require data to be structured and organized in advance, a Data Lake allows data to be ingested and stored as-is, with data analysts and scientists determining its structure at the time of analysis.

Data Lakehouse

A Data Lakehouse is a hybrid repository that combines the advantages of both Data Lakes and Data Warehouses and addresses some of the limitations and challenges of each system. A Data Lakehouse can store structured, unstructured, and semi-structured data. They aim to deliver the best of both approaches and provide a harmonious and scalable solution for managing and analyzing data in various formats while producing the performance and structure required for efficient querying and analytics.

Data Mart

A Data Mart is a discreet subset of a data warehouse and is typically tailored and optimized for distinct groups of users, business departments, or functions within an organization. It is a focused data repository designed to meet a group of users' unique reporting and analysis needs. Data marts improve performance, simplify access, and enhance the relevance of data for those who require specific insights.

Data Mesh

A Data Mesh operates a federated data architecture where data is distributed across various storage systems and domains that provide access to their data products through well-defined APIs and interfaces. With data mesh, the responsibility for analytical data lies with the domain teams, not the central data team, who provide a domain-agnostic data platform.

Data Vault

Data Vault is a data modeling and architecture method used in Data Warehouse Architecture and business intelligence. It provides a structured and scalable approach for handling and organizing data in complex and volatile environments. The Data Vault method seeks to improve flexibility, scalability, and maintainability in data warehouse initiatives, mainly when there are many data sources and data structures subject to frequent changes.

Data Warehouse

A Data Warehouse is a structured and optimized storage system for collecting, storing, and managing data from various sources within an organization. It is a central repository for historical, current, transformed, integrated, and organized data for querying, reporting, and analytical purposes. The primary goal of a Data Warehouse is to support business intelligence (BI), data analysis, and data science activities by providing a consistent and reliable data source for decision-making and drawing insights.

Database

A Database is a structured and organized collection of data that is stored and accessed electronically. Data roles store, retrieve, manipulate, and manage the data using a database management system.

Dimension Table

A Dimension Table is a critical entity of a dimensional model used in Data Warehouse Architecture. It contains descriptive attributes that provide context and descriptive information about the business entities being analyzed.

Dimension tables provide the context and details required to analyze the measures stored in fact tables. They allow data analysts and others to analyze data across several dimensions, enabling meaningful analysis, reporting, and querying in Data Warehouse Architecture systems and business intelligence environments.

Fact Table

A Fact Table is a central component that holds quantitative and numerical data related to the business processes or events being analyzed in a star schema. The table is part of a relational database design used in Data Warehouse Architecture, structured primarily for online analytical processing (OLAP).

Fact tables contain descriptive attributes or context related to the measures in the fact table. Foreign keys establish relationships with one or more dimension tables. Data analysts and others use these keys to join and retrieve descriptive information from the dimension tables that provide context to the measures stored in the fact table.

Link

A Link is a connection or relationship that exists between two entities in a data model or schema. Foreign key relationships are links where one or more fields or columns in one table refer to a primary key field in another table.

Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) is a class of technology used for organizing and analyzing large volumes of data to support business intelligence (BI) and decision-making processes. OLAP systems allow users to explore and analyze multidimensional data interactively, gain insights, identify tendencies, and make informed decisions. Data scientists, analysts, and others use OLAP to efficiently process complex queries and provide an intuitive interface for data exploration.

Online Transaction Processing (OLTP)

Online Transaction Processing (OLTP) is a class of technology used for managing and processing high volumes of operational and transactional data in real-time. OLTP systems can handle routine business operations, such as capturing, updating, and retrieving data related to daily transactions. These transactions include processing customer orders, updating inventory levels, recording financial transactions, and more.

Satellite

Satellite data modeling is a method that creates additional tables, often referred to as satellite tables or simply satellites, to define and store additional or detailed attributes related to a particular dimension. Data Modelers use the tables to provide supplementary descriptive details or historical changes that might not be appropriate to include in the main dimension table

Star Schema

A Star Schema is a data modeling method used in Data Warehouse Architecture that manages data into a central fact table surrounded by multiple dimension tables. The fact table is placed in the center and is surrounded by several dimension tables in a radial pattern resembling the points of a star.

The fact table contains the quantitative data or measures, while the dimension tables store descriptive attributes that provide context to the fact table's measures. Each dimension table represents a specific aspect of the business.

The star schemas' straightforward design facilitates data retrieval and analysis in Data Warehouse Architecture systems and business intelligence environments.

Data Principle Glossary

Data Principle Glossary



Accountability

The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

Accuracy

You must keep personal data accurate and up to date.

Adaptable and Flexible

Refers to the capacity of a data system or architecture to respond effectively to changes in business processes, requirements, data sources, and technology environments. Flexibility allows a system to accommodate variations in data formats, data models, and data processing workflows. An adaptable data architecture can evolve and change to meet new opportunities and challenges without significant refurbishment or disruptions.

Automated Pipelines

Automate pipelines to streamline the process of data ingestion, data integration, data transformation, and data analysis. Efficiently manage and process large volumes of data, derive meaningful insights, make informed decisions, and automate business processes.

Business Focused

Data models should focus on essential aspects of the business domain and abstract away unnecessary details. Regularly validate and verify the data model with business stakeholders to ensure it accurately represents the real-world domain and is fit for purpose.

Confidentiality Classification

Include a field to model data with a confidentiality level for documents, communications, locations, and facilities to ensure this data is distinguishable from other information. Include a reference entity that lists the confidentiality level.

Data Flows are Optimized

The movement of data between environments and parts of a pipeline should be optimized to reduce costs and processing time and increase data currency.

Data has a Custodian

The operational aspects of data are managed and safeguarded by the data custodian on behalf of the data owner.

Data has an Owner

The data owner is typically a business stakeholder or department within an organization with ultimate accountability and responsibility for a specific dataset.

Data is an Asset

Data is an important and valuable organizational asset and must be managed through its lifecycle from creation, enrichment, storage, and visualization to destruction.

Data is Available

Data classified for users will be available and accessible to the roles that need the data to perform their business or technical functions.

Data is Curated

Data is collected, organized, validated, preserved, and managed through its lifecycle to ensure its quality, usability, and long-term value, from acquisition or creation to archiving or disposal.

Data is Defined

Data is defined consistently throughout the organization, and the definitions are comprehensible and accessible to all users.

Data is Secure

Data classified as confidential, sensitive, or personal will be secured and protected in transit, at rest, and in use from unauthorized access or dissemination.

Data is Shared

Data is used and consumed by various organizational functions and roles that need access to specific data to perform their roles.

Data minimisation

You should collect and process only as much data as absolutely necessary for the purposes specified.

Data Normalization

Apply normalization methods to reduce data redundancy and improve data integrity. Normalization increases efficiency and consistency.

Design Flexibility

Design the data model to be flexible enough to accommodate future business and technical changes and requirements. Avoid inflexible and brittle structures that impede scalability and adaptability.

Driven by Artificial Intelligence

Refers to incorporating and operating artificial intelligence (AI) and machine learning (ML) technologies within the design and management of a data architecture or system. Uses include predictive analysis, anomaly detection, cost optimization, and personalization.

Entity Relationships

Define relationships between entities accurately. Depending on the business rules and requirements, these relationships can be one-to-one, one-to-many, or many-to-many.

Governed and Managed

Refers to the structured and regulated approach to overseeing, organizing, and maintaining an organization's data assets to ensure that all personnel and systems use data securely, effectively, and in compliance with regulatory requirements.

Integrity and confidentiality (security)

Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).

Lawfulness, fairness and transparency

Processing must be lawful, fair, and transparent to the data subject.

Level of Abstraction

Represent entities, relationships, and attributes at an appropriate level of granularity. Conceptual Data Models typically do not model attributes and naming is business focused. Logical Data Models add attributes with generic data types. Physical Data Models often change names to suit an implementation and add platform-specific data types.

Model Clarity

Ensure that the data model is straightforward and easy to understand by all stakeholders, including business users, developers, and data professionals. Use clear and consistent naming conventions for entities, attributes, and relationships.

Model Simplicity

Ensure the data model is as simple as possible without sacrificing its ability to represent the real-world domain accurately. Avoid unnecessary complexity that attenuates model comprehensibility.

Personally Identifiable Information

Include a field to model PII data to ensure this data is distinguishable from other information. Include a reference entity that lists the privacy type and another that lists the sensitivity level.

Purpose limitation

You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.

Robust and Resilient

Refers to designing data systems capable of withstanding disruptions, maintaining data integrity, ensuring high availability, and guaranteeing data operations function effectively despite unanticipated challenges.

Scalable

Refers to a system's ability or architecture to manage increasing amounts of data or expanding workloads without sacrificing performance, reliability, or maintainability.

Storage limitation

You may only store personally identifying data for as long as necessary for the specified purpose.

Data Role Glossary

Data Role Glossary



Business Intelligence Specialist

Business Intelligence Specialists are responsible for designing, developing, and maintaining business intelligence solutions, such as data warehouses, data marts, and reporting systems. They enable users to access, visualize and analyze data easily. The visualizations and analysis support meaningful and actionable insights that stakeholders can use to support decision-making and drive business strategies.

Chief Data Officer

Chief Data Officers (CDOs) are senior executives responsible for managing and leveraging an organization's data assets to drive strategic decision-making, improve operational efficiency, ensure regulatory compliance, and foster data-driven innovation. The CDO plays a critical role in developing and implementing data strategies, ensuring data quality and governance, and promoting a data-centric culture within the organization.

Chief Information Officer

Chief Information Officers (CIOs) are senior-level executives responsible for overseeing and managing an organization's information technology (IT) systems and strategies. The CIO typically reports directly to the CEO or another top-level executive. The role aligns the organization's IT initiatives and resources with its high-level business goals and objectives. This alignment involves developing and implementing IT strategies that support the organization's vision, mission, and long-term plans.

Data Analyst

Data Analysts have the knowledge and skills to change raw data into information and insights that benefit data consumers. They are typically given a specific requirement or problem that needs to be solved and use their knowledge of the structure and content of available data to provide answers to the business.

Data Custodian

Data custodians are responsible for operating and managing technology, including systems that collect, store, process, manage, and provide access to the organization's data. They are commonly associated with the technology services and functions of the organization but may also include systems administrators working within one or more functional areas.

Data Engineer

Data Engineers are responsible for designing, developing, and maintaining the infrastructure, systems, and pipelines required for the efficient and reliable processing, storage, and retrieval of large volumes of data. They work closely with data scientists, analysts, and other stakeholders to ensure data is collected, transformed, and accessible for analysis and decision-making.

Data Governance Manager

A Data Governance Manager establishes and enforces the organization's data governance policies and procedures. The role defines data ownership, establishes data quality standards, and ensures compliance with data protection regulations. They collaborate with other stakeholders to develop data governance frameworks and oversee data stewardship activities, including developing, implementing, and managing data governance initiatives.

Data Modeler

Data modelers create graphical representations of business or information systems. The models are used both to communicate and to define the data entities and their relationships, attributes, and data types. There are conceptual, logical, and physical data modelers. Business requirements and reporting needs drive the creation of conceptual models, and platform and system constraints drive the creation of physical data models.

Data Owner

Data Owners are typically directors or managers with the authority to determine business definitions of data, grant access to data, and approve secure data usage for the functional areas within their jurisdiction of authority. By understanding the organization's information needs, data owners can anticipate how data can be used to meet the organization's strategic goals.

Data Protection Officer

A Data Protection Officer is an individual designated by an organization to oversee data protection and privacy matters within the organization and information shared with external entities. The role is typically associated with compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union, and analogous authorities in other jurisdictions concerned with privacy and human rights law.

Data Quality Analyst

Data Quality Analysts focus on assessing and improving the quality of data within an organization. They develop and implement data quality metrics, conduct data quality assessments, and collaborate with data owners to address data quality issues.

Data Scientist

Data scientists are advanced analysts responsible for collecting, analyzing, and interpreting data to help drive organizational decision-making. They use sophisticated analytics techniques, such as machine learning and predictive modeling, to gain hidden insights and create predictive models. They can work with structured, semi-structured or unstructured data sets.

Data Stakeholder

Data Stakeholders are any individual or group interested in or involved in the management, access, quality, and use of data within an organization or a specific project. Data stakeholders can include various roles and departments across an organization, each with different perspectives and responsibilities related to data.

Data Steward

Data stewards are responsible for implementing data policies and managing one or more types of organizational data. They authorize and monitor the secure use of data within their assigned business areas. They ensure appropriate access, accuracy, classification, privacy, and data security is maintained at all times.

Data User

Data users are authorized individuals with access to organization data to perform their assigned duties or functions. When users are given access to data, they assume responsibility for the appropriate use, management, and application of privacy and security standards for the data they are authorized to use.

Database Administrator

Database Administrators (DBAs) are responsible for managing and maintaining an organization's databases. They ensure data integrity, security, and performance by monitoring database performance, optimizing queries, and managing backups.

Executive Sponsor

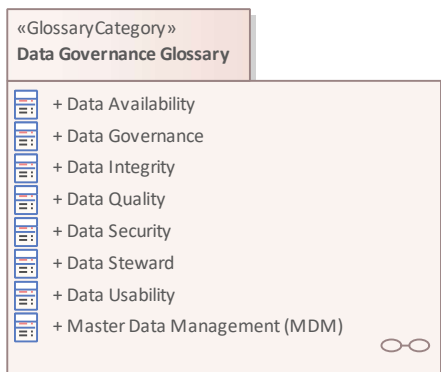
An Executive Sponsor is a senior leadership member with planning and policy responsibility and accountability for major administrative data systems within their functional areas. By understanding the organization's functions, they can anticipate how it will use data to meet internal and external organizational needs.

Information Architect

Information Architects focus on organizing, structuring, and designing information systems to facilitate efficient access, retrieval, and use of information. They are responsible for analyzing an organization's information needs, understanding user requirements, and creating secure and compliant information architectures that meet those needs. They work closely with stakeholders, such as business leaders, content creators, and technology teams, to develop solutions that enhance information management and usability.

Data Governance Glossary

Data Governance Glossary Category



Data Governance Glossary

Data Governance Glossary Entries

<p>«GlossaryEntry» Data Availability</p> <p>Data Availability measures the accessibility and readiness of data for authorized users or systems when required. It ensures that data necessary for business operations, analysis, and decision-making is accessible without delays or disruptions.</p>	<p>«GlossaryEntry» Data Governance</p> <p>Data Governance is setting standards and managing the availability, usability, integrity, and security of data used within an organization. It involves devising and enforcing policies, standards, and procedures to ensure data assets are correctly managed and appropriately utilized. The governance includes how data is sourced, collected, stored, processed, and ultimately archived or destroyed.</p>	<p>«GlossaryEntry» Data Integrity</p> <p>Data Integrity concerns data's accuracy, consistency, and reliability throughout its entire lifecycle, from source to disposal. It ensures that the data remains unchanged, complete, and consistent from its point of creation or capture through analysis, processing, and utilization.</p>	<p>«GlossaryEntry» Data Quality</p> <p>Data Quality concerns data accuracy, completeness, trustworthiness, and consistency across an organization's datasets. High-quality data is critical for decisions, accurate analysis, and achieving an organization's goals and objectives.</p>
<p>«GlossaryEntry» Data Security</p> <p>Data Security concerns protecting an organization's digital data in any form or location from unauthorized access, corruption, theft, or damage throughout the data lifecycle. Security includes constant monitoring, safeguarding private data, verifying the identity of users, and restricting access to data based on their permissions or roles. Security policies and procedures must be devised and implemented to keep the organization's data secure.</p>	<p>«GlossaryEntry» Data Steward</p> <p>A Data Steward is a data governance role responsible for the oversight and management of the organization's data assets within distinct domains or areas. The role's primary function is to ensure data is sourced, processed, accessed, and managed according to policies.</p>	<p>«GlossaryEntry» Data Usability</p> <p>Data Usability concerns the ease with which users can access, understand, process, and utilize data effectively to meet specific business, analytical, or technical needs. It comprises several factors that can empower individuals and teams to derive valuable insights, make informed business and technical decisions, and drive innovation.</p>	<p>«GlossaryEntry» Master Data Management (MDM)</p> <p>Master Data Management (MDM) is a comprehensive method organizations use to link, manage, and synchronize key business data entities within an organization. The master data is typically stored in a centralized repository, enabling controlled access to users and systems.</p>

Data Availability

Data Availability measures the accessibility and readiness of data for authorized users or systems when required. It ensures that data necessary for business operations, analysis, and decision-making is accessible without delays or disruptions.

Data Governance

Data Governance is setting standards and managing the availability, usability, integrity, and security of data used within an organization. It involves devising and enforcing policies, standards, and procedures to ensure data assets are correctly managed and appropriately utilized. The governance includes how data is sourced, collected, stored, processed, and ultimately archived or destroyed.

Data Integrity

Data Integrity concerns data's accuracy, consistency, and reliability throughout its entire lifecycle, from source to disposal. It ensures that the data remains unchanged, complete, and consistent from its point of creation or capture through analysis, processing, and utilization.

Data Quality

Data Quality concerns data accuracy, completeness, trustworthiness, and consistency across an organization's datasets. High-quality data is critical for decisions, accurate analysis, and achieving an organization's goals and objectives.

Data Security

Data Security concerns protecting an organization's digital data in any form or location from unauthorized access, corruption, theft, or damage throughout the data lifecycle. Security includes constant monitoring, safeguarding private data, verifying the identity of users, and restricting access to data based on their permissions or roles. Security policies and procedures must be devised and implemented to keep the organization's data secure.

Data Steward

A Data Steward is a data governance role responsible for the oversight and management of the organization's data assets within distinct domains or areas. The role's primary function is to ensure data is sourced, processed, accessed, and managed according to policies.

Data Usability

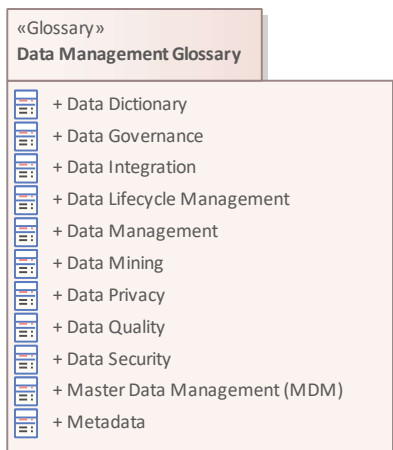
Data Usability concerns the ease with which users can access, understand, process, and utilize data effectively to meet specific business, analytical, or technical needs. It comprises several factors that can empower individuals and teams to derive valuable insights, make informed business and technical decisions, and drive innovation.

Master Data Management (MDM)

Master Data Management (MDM) is a comprehensive method organizations use to link, manage, and synchronize key business data entities within an organization. The master data is typically stored in a centralized repository, enabling controlled access to users and systems.

Data Management Glossary

Data Management Glossary Category



(from Glossaries)

Data Management Glossary

Data Management Glossary Entries

<p>«GlossaryEntry» Data Dictionary</p> <p>A Data Dictionary is a centralized repository or record that comprises precise descriptions and metadata concerning the data used within an organization. It is a comprehensive reference providing information about the structure, definitions, relationships, and attributes of the data elements stored in databases, data warehouses, data lakes, or other information systems.</p>	<p>«GlossaryEntry» Data Integration</p> <p>Data Integration involves merging data from diverse and heterogeneous sources into a unified and coherent view. It consists of retrieving data from multiple sources, such as databases, applications, cloud services, and Internet of things streams, and transforming that data into a format that can be analyzed, understood, and used for different purposes.</p>	<p>«GlossaryEntry» Data Governance</p> <p>Data Governance is setting standards and managing the availability, usability, integrity, and security of data used within an organization. It involves devising and enforcing policies, standards, and procedures to ensure data assets are correctly managed and appropriately utilized. The governance includes how data is sourced, collected, stored, processed, and ultimately archived or destroyed.</p>	<p>«GlossaryEntry» Data Lifecycle Management</p> <p>Data Lifecycle Management concerns safeguarding and managing the flow of data throughout its existence, from creation or ingestion, storage, and usage to archiving or deletion.</p>
<p>«GlossaryEntry» Data Management</p> <p>Data Management constitutes the procedures, techniques, and technologies used to collect, store, organize, secure, process, and retrieve data securely and efficiently. It comprises activities to ensure data is accurate, available, reliable, and accessible to data stakeholders within an organization.</p>	<p>«GlossaryEntry» Data Mining</p> <p>The Data Mining discipline and process discovers patterns, trends, relationships, and insights within large volumes of data. Data mining uses different techniques, algorithms, and tools to extract valuable information from data, which stakeholders can use for decision-making, prediction, and knowledge discovery.</p>	<p>«GlossaryEntry» Data Privacy</p> <p>Data Privacy concerns the proper management, protection, and legal use of an organization's or an individual's information. It involves safeguarding sensitive data and ensuring that personal or confidential information is not disclosed, accessed, or misused without the data owner's explicit consent or in violation of relevant laws and regulations.</p>	<p>«GlossaryEntry» Data Quality</p> <p>Data Quality concerns data accuracy, completeness, trustworthiness, and consistency across an organization's datasets. High-quality data is critical for decisions, accurate analysis, and achieving an organization's goals and objectives.</p>
<p>«GlossaryEntry» Data Security</p> <p>Data Security concerns protecting an organization's digital data in any form or location from unauthorized access, corruption, theft, or damage throughout the data lifecycle. Security includes constant monitoring, safeguarding private data, verifying the identity of users, and restricting access to data based on their permissions or roles. Security policies and procedures must be devised and implemented to keep the organization's data secure.</p>	<p>«GlossaryEntry» Master Data Management (MDM)</p> <p>Master Data Management (MDM) is a comprehensive method organizations use to link, manage, and synchronize key business data entities within an organization. The master data is typically stored in a centralized repository, enabling controlled access to users and systems.</p>	<p>«GlossaryEntry» Metadata</p> <p>Metadata concerns descriptive information or data about other data. It provides context, details, and attributes that help understand, manage, or organize a data set. Metadata acts as a "data about data" that provides insights into the content, structure, and characteristics of the data it describes.</p>	

Data Dictionary

A Data Dictionary is a centralized repository or record that comprises precise descriptions and metadata concerning the data used within an organization. It is a comprehensive reference providing information about the structure, definitions, relationships, and attributes of the data elements stored in databases, data warehouses, data lakes, or other information systems.

Data Governance

Data Governance is setting standards and managing the availability, usability, integrity, and security of data used within an organization. It involves devising and enforcing policies, standards, and procedures to ensure data assets are correctly managed and appropriately utilized. The governance includes how data is sourced, collected, stored, processed, and ultimately archived or destroyed.

Data Integration

Data Integration involves merging data from diverse and heterogeneous sources into a unified and coherent view. It consists of retrieving data from multiple sources, such as databases, applications, cloud services, and Internet of things streams, and transforming that data into a format that can be analyzed, understood, and used for different purposes.

Data Lifecycle Management

Data Lifecycle Management concerns safeguarding and managing the flow of data throughout its existence, from creation or ingestion, storage, and usage to archiving or deletion.

Data Management

Data Management constitutes the procedures, techniques, and technologies used to collect, store, organize, secure, process, and retrieve data securely and efficiently. It comprises activities to ensure data is accurate, available, reliable, and accessible to data stakeholders within an organization.

Data Mining

The Data Mining discipline and process discovers patterns, trends, relationships, and insights within large volumes of data. Data mining uses different techniques, algorithms, and tools to extract valuable information from data, which stakeholders can use for decision-making, prediction, and knowledge discovery.

Data Privacy

Data Privacy concerns the proper management, protection, and legal use of an organization's or an individual's information. It involves safeguarding sensitive data and ensuring that personal or confidential information is not disclosed, accessed, or misused without the data owner's explicit consent or in violation of relevant laws and regulations.

Data Quality

Data Quality concerns data accuracy, completeness, trustworthiness, and consistency across an organization's datasets. High-quality data is critical for decisions, accurate analysis, and achieving an organization's goals and objectives.

Data Security

Data Security concerns protecting an organization's digital data in any form or location from unauthorized access, corruption, theft, or damage throughout the data lifecycle. Security includes constant monitoring, safeguarding private data, verifying the identity of users, and restricting access to data based on their permissions or roles. Security policies and procedures must be devised and implemented to keep the organization's data secure.

Master Data Management (MDM)

Master Data Management (MDM) is a comprehensive method organizations use to link, manage, and synchronize key business data entities within an organization. The master data is typically stored in a centralized repository, enabling controlled access to users and systems.

Metadata

Metadata concerns descriptive information or data about other data. It provides context, details, and attributes that help understand, manage, or organize a data set. Metadata acts as a "data about data" that provides insights into the content, structure, and characteristics of the data it describes.

Data Management Acronym Glossary

Data Management Acronym Glossary Category

Data Management Acronym Glossary	
+ ACID	
+ ADLS	
+ AI	
+ API	
+ BI	
+ CAP	
+ CDC	
+ CRUD	
+ CSV	
+ DBMS	
+ DDL	
+ DL	
+ DLP	
+ DM	
+ DML	
+ DQ	
+ DW	
+ DWH	
+ EDL	
+ EDW	
+ ELT	
+ ERD	
+ ETL	
+ ETLT	
+ FOLAP	
+ GCS	
+ GDPR	
+ HDFS	
+ HIPAA	
+ HOLAP	
+ IoT	
+ JSON	
+ MDM	
+ ML	
+ MOLAP	
+ NoSQL	
+ ODS	
+ OLAP	
+ OLTP	
+ PII	
+ RDBMS	
+ ROLAP	
+ S3	
+ SCD	
+ SNOWFLAKE	
+ SQL	
+ STAR	
+ XML	

(from Glossaries)

Data Management Acronym Glossary

Data Management Acronym Glossary Entries

«GlossaryEntry» MOLAP	«GlossaryEntry» ACID	«GlossaryEntry» ADLS	«GlossaryEntry» AI	«GlossaryEntry» API
Multidimensional Online Analytical Processing	Atomicity, Consistency, Isolation, Durability	Azure Data Lake Storage	Artificial Intelligence	Application Programming Interface
«GlossaryEntry» BI	«GlossaryEntry» CAP	«GlossaryEntry» CDC	«GlossaryEntry» CRUD	«GlossaryEntry» CSV
Business Intelligence	Consistency, Availability, Partition Tolerance	Change Data Capture	Create, Read, Update, Delete	Comma-Separated Values
«GlossaryEntry» DBMS	«GlossaryEntry» DDL	«GlossaryEntry» DL	«GlossaryEntry» DLP	«GlossaryEntry» DM
Database Management System	Data Definition Language	Data Lake	Data Lake Platform	Data Mart
«GlossaryEntry» DML	«GlossaryEntry» DQ	«GlossaryEntry» DW	«GlossaryEntry» DWH	«GlossaryEntry» EDL
Data Manipulation Language	Data Quality	Data Warehouse	Data Warehouse	Enterprise Data Lake
«GlossaryEntry» EDW	«GlossaryEntry» ELT	«GlossaryEntry» ERD	«GlossaryEntry» ETL	«GlossaryEntry» ETLT
Enterprise Data Warehouse	Extract, Load, Transform	Entity-Relationship Diagram	Extract, Transform, Load	Extract, Transform, Load, Transform
«GlossaryEntry» FOLAP	«GlossaryEntry» GCS	«GlossaryEntry» GDPR	«GlossaryEntry» HDFS	«GlossaryEntry» HIPAA
Federated Online Analytical Processing	Google Cloud Storage	General Data Protection Regulation	Hadoop Distributed File System	Health Insurance Portability and Accountability Act
«GlossaryEntry» HOLAP	«GlossaryEntry» IoT	«GlossaryEntry» JSON	«GlossaryEntry» MDM	«GlossaryEntry» ML
Hybrid Online Analytical Processing	Internet of Things	JavaScript Object Notation	Master Data Management	Machine Learning
«GlossaryEntry» NoSQL	«GlossaryEntry» ODS	«GlossaryEntry» OLAP	«GlossaryEntry» OLTP	«GlossaryEntry» PII
Not Only SQL	Operational Data Store	Online Analytical Processing	Online Transaction Processing	Personally Identifiable Information
«GlossaryEntry» RDBMS	«GlossaryEntry» ROLAP	«GlossaryEntry» S3	«GlossaryEntry» SCD	«GlossaryEntry» SNOWFLAKE
Relational Database Management System	Relational Online Analytical Processing	Simple Storage Service (Amazon Web Services)	Slowly Changing Dimension	Normalized Reusable Data Model
«GlossaryEntry» SQL	«GlossaryEntry» STAR	«GlossaryEntry» XML		
Structured Query Language	Standardized Architecture for Reusable Data	Extensible Markup Language		

ACID

Atomicity, Consistency, Isolation, Durability

ADLS

Azure Data Lake Storage

AI

Artificial Intelligence

API

Application Programming Interface

BI

Business Intelligence

CAP

Consistency, Availability, Partition Tolerance

CDC

Change Data Capture

CRUD

Create, Read, Update, Delete

CSV

Comma-Separated Values

DBMS

Database Management System

DDL

Data Definition Language

DL

Data Lake

DLP

Data Lake Platform

DM

Data Mart

DML

Data Manipulation Language

DQ

Data Quality

DW

Data Warehouse

DWH

Data Warehouse

EDL

Enterprise Data Lake

EDW

Enterprise Data Warehouse

ELT

Extract, Load, Transform

ERD

Entity-Relationship Diagram

ETL

Extract, Transform, Load

ETLT

Extract, Transform, Load, Transform

FOLAP

Federated Online Analytical Processing

GCS

Google Cloud Storage

GDPR

General Data Protection Regulation

HDFS

Hadoop Distributed File System

HIPAA

Health Insurance Portability and Accountability Act

HOLAP

Hybrid Online Analytical Processing

IoT

Internet of Things

JSON

JavaScript Object Notation

MDM

Master Data Management

ML

Machine Learning

MOLAP

Multidimensional Online Analytical Processing

NoSQL

Not Only SQL

ODS

Operational Data Store

OLAP

Online Analytical Processing

OLTP

Online Transaction Processing

PII

Personally Identifiable Information

RDBMS

Relational Database Management System

ROLAP

Relational Online Analytical Processing

S3

Simple Storage Service (Amazon Web Services)

SCD

Slowly Changing Dimension

SNOWFLAKE

Normalized Reusable Data Model

SQL

Structured Query Language

STAR

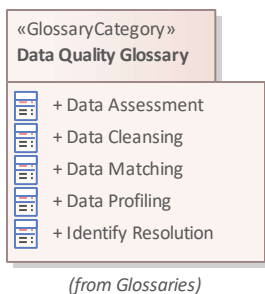
Standardized Architecture for Reusable Data

XML

Extensible Markup Language

Data Quality Glossary

Data Quality Glossary Category



Data Quality Glossary

Data Quality Glossary Entries

<div>«GlossaryEntry» Data Assessment</div> <div>Data Assessment, or data evaluation or analysis, refers to examining and interpreting data to gain insights, identify patterns, detect anomalies, and draw conclusions. It is a critical step in data analysis pipelines essential to decision-making, problem-solving, and understanding a dataset.</div>	<div>«GlossaryEntry» Data Cleansing</div> <div>Also known as data cleaning or data scrubbing is the process of identifying and rectifying errors, inconsistencies, inaccuracies, and incomplete data within a dataset. The principal purpose of data cleansing is to improve the overall quality and reliability of the data, making it suitable for accurate analysis and decision-making.</div>	<div>«GlossaryEntry» Data Matching</div> <div>Also known as record linkage or entity resolution, is the procedure of identifying and linking records from multiple datasets that correspond to the identical real-world entity or entity instance. Data matching aims to consolidate related data from various sources into a harmonious and accurate representation, reducing redundancy and improving data quality.</div>	<div>«GlossaryEntry» Data Profiling</div> <div>Data profiling comprises analyzing, examining, and summarizing the structure and content of a dataset to produce a helpful and comprehensive overview of the organization's data. The summary aids in the discovery of data quality issues, security risks, and other crucial insights.</div>
<div>«GlossaryEntry» Identify Resolution</div> <div>Identity resolution is linking and matching various digital identities or data records corresponding to the same entity or individual across different data sources or datasets. The goal is to accurately identify and consolidate disparate information associated with the same real-world entity, whether a person, a household, a device, or any other entity.</div>			

Data Assessment

Data Assessment, or data evaluation or analysis, refers to examining and interpreting data to gain insights, identify patterns, detect anomalies, and draw conclusions. It is a critical step in data analysis pipelines essential to decision-making, problem-solving, and understanding a dataset.

Data Cleansing

Also known as data cleaning or data scrubbing is the process of identifying and rectifying errors, inconsistencies, inaccuracies, and incomplete data within a dataset. The principal purpose of data cleansing is to improve the overall quality and reliability of the data, making it suitable for accurate analysis and decision-making.

Data Matching

Also known as record linkage or entity resolution, is the procedure of identifying and linking records from multiple datasets that correspond to the identical real-world entity or entity instance. Data matching aims to consolidate related data from various sources into a harmonious and accurate representation, reducing redundancy and improving data quality.

Data Profiling









Data profiling comprises analyzing, examining, and summarizing the structure and content of a dataset to produce a helpful and comprehensive overview of the organization's data. The summary aids in the discovery of data quality issues, security risks, and other crucial insights.

Identify Resolution

Identity resolution is linking and matching various digital identities or data records corresponding to the same entity or individual across different data sources or datasets. The goal is to accurately identify and consolidate disparate information associated with the same real-world entity, whether a person, a household, a device, or any other entity.

Data Modeling Glossary

Data Modeling Glossary Category

«GlossaryCategory» Data Modeling Glossary	
	+ Conceptual Data Model
	+ Logical Data Model
	+ Physical Data Model
	+ Business Key
	+ Primary Key
	+ Foreign Key
	+ Unique Key 

(from Glossaries)

Data Modeling Glossary

Data Modeling Glossary Entries

<p>«GlossaryEntry» Conceptual Data Model</p> <p>A Conceptual Data Model is a high-level representation of a domain's crucial entities and their relationships in a form that both technical and non-technical stakeholders easily understand. Some modelers also choose to include attributes in a conceptual model.</p>	<p>«GlossaryEntry» Logical Data Model</p> <p>A Logical Data Model elaborates the concepts and relationships captured in the conceptual data model and adds more detail to describe attributes and their primitive data types. The logical data model is independent of any specific database management system (DBMS). It typically introduces business and foreign keys and adds some level of normalization.</p>	<p>«GlossaryEntry» Physical Data Model</p> <p>A Physical Data Model details the concepts defined in the logical model, adding information such as primary keys and indexes that will be used in the implementation. The physical data model targets a specific DBMS platform and uses the data type and other structures of the target DBMS.</p>	<p>«GlossaryEntry» Business Key</p> <p>A Business Key is a unique identifier used to uniquely identify a business entity or record within an organization's data. It is a term that the business users will be familiar with to ensure data accuracy, integrity, and consistency of the records. Some modelers use this as a primary key; others use a system-generated key as the primary key. The choice depends on a range of factors.</p>
<p>«GlossaryEntry» Primary Key</p> <p>A Primary Key is a unique identifier for a row in a table. It uniquely identifies individual records, ensures data integrity, and enables efficient data retrieval. Each table in a database can have only one primary key, and it must contain unique values for each record. This constraint prevents duplicate entries and enforces the uniqueness of the values. The primary key can be composite and participates in foreign key relationships.</p>	<p>«GlossaryEntry» Foreign Key</p> <p>A Foreign Key creates a relationship between two tables by referencing the primary key of one table in another table. The purpose of a foreign key is to maintain data integrity and enforce referential integrity, ensuring that the data between related tables remains consistent and accurate. This constraint ensures values entered in the foreign key table exist in the primary key table.</p>	<p>«GlossaryEntry» Unique Key</p> <p>A Unique Key defines a constraint that guarantees the values within a specific column or group of columns are unique across all rows in a table. A table may have any number of unique keys, and values can be NULL in contradistinction to primary keys.</p>	

Business Key

A Business Key is a unique identifier used to uniquely identify a business entity or record within an organization's data. It is a term that the business users will be familiar with to ensure data accuracy, integrity, and consistency of the records. Some modelers use this as a primary key; others use a system-generated key as the primary key. The choice depends on a range of factors.

Conceptual Data Model

A Conceptual Data Model is a high-level representation of a domain's crucial entities and their relationships in a form that both technical and non-technical stakeholders easily understand. Some modelers also choose to include attributes in a conceptual model.

Foreign Key

A Foreign Key creates a relationship between two tables by referencing the primary key of one table in another table. The purpose of a foreign key is to maintain data integrity and enforce referential integrity, ensuring that the data between related tables remains consistent and accurate. This constraint ensures values entered in the foreign key table exist in the primary key table.

Logical Data Model

A Logical Data Model elaborates the concepts and relationships captured in the conceptual data model and adds more detail to describe attributes and their primitive data types. The logical data model is independent of any specific database management system (DBMS). It typically introduces business and foreign keys and adds some level of normalization.

Physical Data Model

A Physical Data Model details the concepts defined in the logical model, adding information such as primary keys and indexes that will be used in the implementation. The physical data model targets a specific DBMS platform and uses the data type and other structures of the target DBMS.

Primary Key

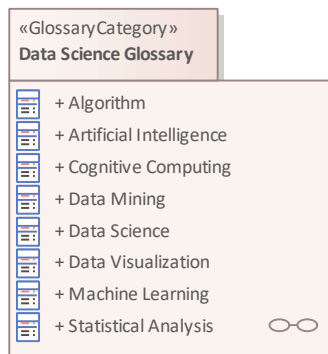
A Primary Key is a unique identifier for a row in a table. It uniquely identifies individual records, ensures data integrity, and enables efficient data retrieval. Each table in a database can have only one primary key, and it must contain unique values for each record. This constraint prevents duplicate entries and enforces the uniqueness of the values. The primary key can be composite and participates in foreign key relationships.

Unique Key

A Unique Key defines a constraint that guarantees the values within a specific column or group of columns are unique across all rows in a table. A table may have any number of unique keys, and values can be NULL in contradistinction to primary keys.

Data Science Glossary

Data Science Glossary Category



(from Glossaries)

Data Science Glossary

Data Science Glossary Entries

<p>«GlossaryEntry» Algorithm</p> <p>A set of computational steps or procedures designed to solve a specific problem or extract meaningful insights from data. It is a systematic approach that utilizes mathematical and statistical techniques to analyze and interpret data, discover patterns, make predictions, and provide valuable information for decision-making.</p>	<p>«GlossaryEntry» Artificial Intelligence</p> <p>The discipline that develops computer systems or machines that can perform tasks that would typically require human intelligence. AI aims to create intelligent systems capable of perceiving, reasoning, learning, and making decisions or taking actions based on available information.</p>	<p>«GlossaryEntry» Cognitive Computing</p> <p>The discipline that develops computer systems that can simulate and augment human cognitive abilities, such as perception, reasoning, learning, and problem-solving. It involves using artificial intelligence (AI) technologies to enable machines to understand and interact with complex, unstructured data more humanistically.</p>	<p>«GlossaryEntry» Data Mining</p> <p>The Data Mining discipline and process discovers patterns, trends, relationships, and insights within large volumes of data. Data mining uses different techniques, algorithms, and tools to extract valuable information from data, which stakeholders can use for decision-making, prediction, and knowledge discovery.</p>
<p>«GlossaryEntry» Data Science</p> <p>The Data Science discipline merges techniques from various domains, including computer science, mathematics, statistics, domain expertise, and presentation skills, to analyze and interpret large volumes of structured, semi-structured, and unstructured data. These techniques help to solve complex problems and make educated decisions.</p>	<p>«GlossaryEntry» Data Visualization</p> <p>Data visualization concerns the graphical representation of data and information to present datasets in a visual format, such as plots, infographics, graphs, charts, maps, and dashboards. These visual displays of data convey complex data relationships and data-driven insights, making the information available and understandable to business and technical stakeholders.</p>	<p>«GlossaryEntry» Machine Learning</p> <p>Machine learning is a subfield of the artificial intelligence (AI) discipline that concerns the creation of models and the development of algorithms that enable computers to learn and make predictions, forecasts, or decisions based on data without the need for explicit programming.</p>	<p>«GlossaryEntry» Statistical Analysis</p> <p>Statistical analysis involves collecting, analyzing, summarizing, and interpreting data to discover patterns, trends, relationships, and meaningful insights. It uses statistical methods, mathematical tools, and other techniques to analyze categorical or numerical data, make inferences, and draw conclusions based on the analysis.</p>

Algorithm

A set of computational steps or procedures designed to solve a specific problem or extract meaningful insights from data. It is a systematic approach that utilizes mathematical and statistical techniques to analyze and interpret data, discover patterns, make predictions, and provide valuable information for decision-making.

Artificial Intelligence

The discipline that develops computer systems or machines that can perform tasks that would typically require human intelligence. AI aims to create intelligent systems capable of perceiving, reasoning, learning, and making decisions or taking actions based on available information.

Cognitive Computing

The discipline that develops computer systems that can simulate and augment human cognitive abilities, such as perception, reasoning, learning, and problem-solving. It involves using artificial intelligence (AI) technologies to enable machines to understand and interact with complex, unstructured data more humanistically.

Data Mining

The Data Mining discipline and process discovers patterns, trends, relationships, and insights within large volumes of data. Data mining uses different techniques, algorithms, and tools to extract valuable information from data, which stakeholders can use for decision-making, prediction, and knowledge discovery.

Data Science

The Data Science discipline merges techniques from various domains, including computer science, mathematics, statistics, domain expertise, and presentation skills, to analyze and interpret large volumes of structured, semi-structured, and unstructured data. These techniques help to solve complex problems and make educated decisions.

Data Visualization

Data visualization concerns the graphical representation of data and information to present datasets in a visual format, such as plots, infographics, graphs, charts, maps, and dashboards. These visual displays of data convey complex data relationships and data-driven insights, making the information available and understandable to business and technical stakeholders.

Machine Learning

Machine learning is a subfield of the artificial intelligence (AI) discipline that concerns the creation of models and the development of algorithms that enable computers to learn and make predictions, forecasts, or decisions based on data without the need for explicit programming.

Statistical Analysis

Statistical analysis involves collecting, analyzing, summarizing, and interpreting data to discover patterns, trends, relationships, and meaningful insights. It uses statistical methods, mathematical tools, and other techniques to analyze categorical or numerical data, make inferences, and draw conclusions based on the analysis.

Machine Learning Glossary

Machine Learning Glossary Category

«GlossaryCategory»	
Machine Learning Glossary	
+ Algorithm	
+ Artificial Neural Network (ANN)	
+ Convolutional Neural Network (CNN)	
+ Deep Learning	
+ Feature Engineering	
+ Hyperparameter	
+ Machine Learning (ML)	
+ Neural Networks	
+ Overfitting	
+ Recurrent Neural Network (RNN)	
+ Reinforcement Learning	
+ Supervised Learning	
+ Test Set	
+ Training Set	
+ Underfitting	
+ Unsupervised Learning	
+ Validation Set	

(from Glossaries)

Machine Learning Glossary

Machine Learning Glossary Entries

<p>«GlossaryEntry» Algorithm</p> <p>A set of computational steps or procedures designed to solve a specific problem or extract meaningful insights from data. It is a systematic approach that utilizes mathematical and statistical techniques to analyze and interpret data, discover patterns, make predictions, and provide valuable information for decision-making.</p>	<p>«GlossaryEntry» Artificial Neural Network (ANN)</p> <p>Artificial Neural Network (ANN) comprises a series of layers of interconnected nodes that process information. Connections behave like synapses in an animal brain; the neuron receives input and, in response, computes an output that it transmits to connected neurons, typically if a threshold is reached.</p>	<p>«GlossaryEntry» Convolutional Neural Network (CNN)</p> <p>A Convolutional Neural Network (CNN) is a specialized feed-forward neural network for deep learning algorithms that learns directly from data. CNNs help discover image patterns to recognize objects, classes, and categories. They are also commonly used for processing grid-structured data in image and video recognition tasks and for time-series and signal data.</p>	<p>«GlossaryEntry» Deep Learning</p> <p>Deep Learning is a machine learning paradigm that utilizes neural networks with multiple layers (deep neural networks) to learn intricate patterns and representations from data. These networks attempt to approximate how animal brains function and are typically used in automation and tasks where a machine needs to replace human input.</p>	<p>«GlossaryEntry» Validation Set</p> <p>A Validation Set is a subset of the available data used to fine-tune model hyperparameters and evaluate model performance during training. The validation set provides an unbiased evaluation of the models and allows the most reliable model to be selected and optimized to avoid problems like overfitting.</p>
<p>«GlossaryEntry» Feature Engineering</p> <p>Feature Engineering concerns extracting and transforming variables such as characteristics, properties, and attributes from raw data into a format that improves the performance of machine learning models.</p>	<p>«GlossaryEntry» Hyperparameter</p> <p>A Hyperparameter is a model-level parameter applied before the learning process begins that influences the learning algorithm's behavior. It does not affect the model performance but impacts the quality and speed of the learning process.</p>	<p>«GlossaryEntry» Machine Learning (ML)</p> <p>Machine learning is a subfield of the artificial intelligence (AI) discipline that concerns the creation of models and the development of algorithms that enable computers to learn and make predictions, forecasts, or decisions based on data without the need for explicit programming.</p>	<p>«GlossaryEntry» Neural Networks</p> <p>A Neural Network is a series of interconnected nodes structured in layers that attempt to identify underlying relationships in a dataset by mimicking how an animal brain functions. They process information hierarchically and at various levels of abstraction.</p>	
<p>«GlossaryEntry» Overfitting</p> <p>Overfitting arises in machine learning algorithms when a model provides accurate predictions for training data but fails to generalize to new, unseen data, resulting in inaccurate predictions or classifications. An overly complex model or a lengthy training period resulting in noise or irrelevant patterns can lead to the algorithm's failure or the attenuation of its value.</p>	<p>«GlossaryEntry» Recurrent Neural Network (RNN)</p> <p>A Recurrent Neural Network (RNN) is an artificial neural network suitable for sequential or time series data, where node connections form directed cycles. They can use prior inputs to influence both the current inputs and output, making them suitable for problems such as handwriting and speech recognition, image captioning, and natural language processing and translation.</p>	<p>«GlossaryEntry» Reinforcement Learning</p> <p>Reinforcement Learning is a machine learning training paradigm that rewards desired behaviors and punishes undesired ones. An agent learns to make decisions by interacting with a dynamic environment and learns from the consequences of its actions.</p>	<p>«GlossaryEntry» Supervised Learning</p> <p>Supervised Learning is a machine learning paradigm where the algorithm is trained on a labeled dataset. Data scientists and engineers provide input features and output labels that allow the algorithms to learn patterns and relationships.</p>	
<p>«GlossaryEntry» Test Set</p> <p>A test set refers to a part of the available data kept aside and not used during the training and validation of the model. Once the model is trained, the test set is critical for evaluating its performance and generalization capabilities before using it in production.</p>	<p>«GlossaryEntry» Training Set</p> <p>A training set refers to a part of the available dataset used to train a machine-learning model. For supervised learning, the dataset comprises examples with both input features (variables) and known output labels. For unsupervised learning, it only includes the input data.</p>	<p>«GlossaryEntry» Underfitting</p> <p>Underfitting arises in machine learning algorithms when a model is too simple to capture the underlying patterns in the training data and thus cannot generalize well when presented with unseen or new data. When used, the model results in poor performance and potentially unreliable predictions.</p>	<p>«GlossaryEntry» Unsupervised Learning</p> <p>Unsupervised Learning is a machine learning paradigm or technique where an algorithm learns from an unlabeled dataset, uncovering patterns and structures in the data without requiring direct supervision.</p>	

Algorithm

A set of computational steps or procedures designed to solve a specific problem or extract meaningful insights from data. It is a systematic approach that utilizes mathematical and statistical techniques to analyze and interpret data, discover patterns, make predictions, and provide valuable information for decision-making.

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) comprises a series of layers of interconnected nodes that process information. Connections behave like synapses in an animal brain; the neuron receives input and, in response, computes an output that it transmits to connected neurons, typically if a threshold is reached.

Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a specialized feed-forward neural network for deep learning algorithms that learns directly from data. CNNs help discover image patterns to recognize objects, classes, and categories. They are also commonly used for processing grid-structured data in image and video recognition tasks and for time-series and signal data.

Deep Learning

Deep Learning is a machine learning paradigm that utilizes neural networks with multiple layers (deep neural networks) to learn intricate patterns and representations from data. These networks attempt to approximate how animal brains function and are typically used in automation and tasks where a machine needs to replace human input.

Feature Engineering

Feature Engineering concerns extracting and transforming variables such as characteristics, properties, and attributes from raw data into a format that improves the performance of machine learning models.

Hyperparameter

A Hyperparameter is a model-level parameter applied before the learning process begins that influences the learning algorithm's behavior. It does not affect the model performance but impacts the quality and speed of the learning process.

Machine Learning (ML)

Machine learning is a subfield of the artificial intelligence (AI) discipline that concerns the creation of models and the development of algorithms that enable computers to learn and make predictions, forecasts, or decisions based on data without the need for explicit programming.

Neural Networks

A Neural Network is a series of interconnected nodes structured in layers that attempt to identify underlying relationships in a dataset by mimicking how an animal brain functions. They process information hierarchically and at various levels of abstraction.

Overfitting

Overfitting arises in machine learning algorithms when a model provides accurate predictions for training data but fails to generalize to new, unseen data, resulting in inaccurate predictions or classifications. An overly complex model or a lengthy training period resulting in noise or irrelevant patterns can lead to the algorithm's failure or the attenuation of its value.

Recurrent Neural Network (RNN)

A Recurrent Neural Network (RNN) is an artificial neural network suitable for sequential or time series data, where node connections form directed cycles. They can use prior inputs to influence both the current inputs and output, making them suitable for problems such as handwriting and speech recognition, image captioning, and natural language processing and translation.

Reinforcement Learning

Reinforcement Learning is a machine learning training paradigm that rewards desired behaviors and punishes undesired ones. An agent learns to make decisions by interacting with a dynamic environment and learns from the consequences of its actions.

Supervised Learning

Supervised Learning is a machine learning paradigm where the algorithm is trained on a labeled dataset. Data scientists and engineers provide input features and output labels that allow the algorithms to learn patterns and relationships.

Test Set

A test set refers to a part of the available data kept aside and not used during the training and validation of the model. Once the model is trained, the test set is critical for evaluating its performance and generalization capabilities before using it in production.

Training Set

A training set refers to a part of the available dataset used to train a machine-learning model. For supervised learning, the dataset comprises examples with both input features (variables) and known output labels. For unsupervised learning, it only includes the input data.

Underfitting

Underfitting arises in machine learning algorithms when a model is too simple to capture the underlying patterns in the training data and thus cannot generalize well when presented with unseen or new data. When used, the model results in poor performance and potentially unreliable predictions.

Unsupervised Learning

Unsupervised Learning is a machine learning paradigm or technique where an algorithm learns from an unlabeled dataset, uncovering patterns and structures in the data without requiring direct supervision.

Validation Set

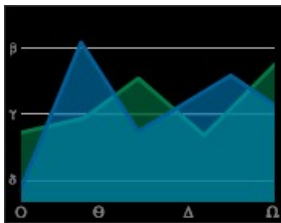
A Validation Set is a subset of the available data used to fine-tune model hyperparameters and evaluate model performance during training. The validation set provides an unbiased evaluation of the models and allows the most reliable model to be selected and optimized to avoid problems like overfitting.

Image Libraries

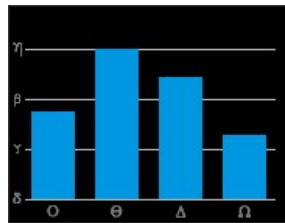
Een library met images en icons rond data gerelateerde zaken. Kunnen gebruikt worden om de opmaak van diagrammen wat aantrekkelijker te maken voor niet architecten en metadata specialisten

Business Visualization Images

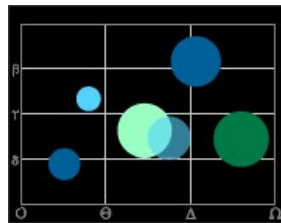
Visualization Images



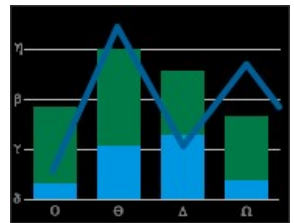
Area Chart



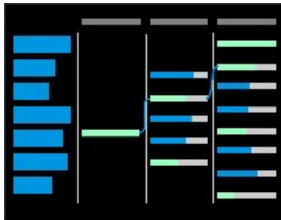
Bar Chart



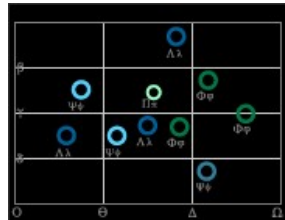
Bubble Chart



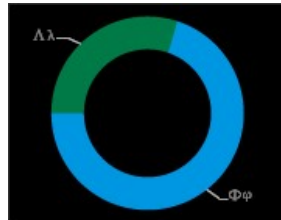
Combo Chart



Decomposition Chart



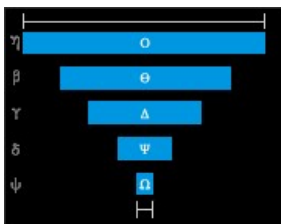
Dot Plot Chart



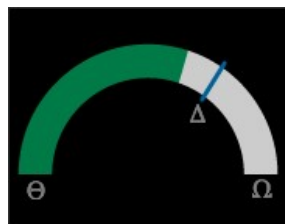
Doughnut Chart



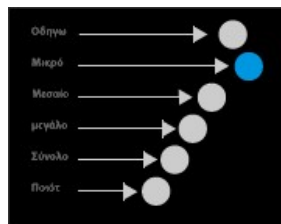
Filled Map Chart



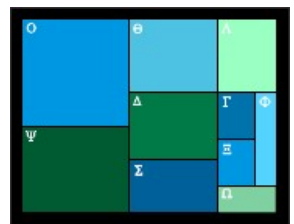
Funnel Chart



Gauge Chart



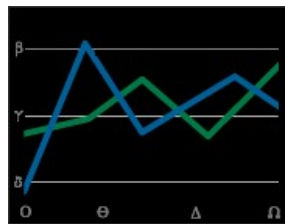
Graph Chart



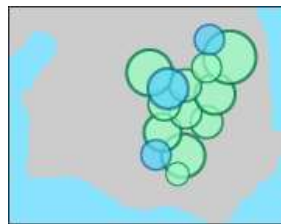
HeatMap



Video



Line Chart



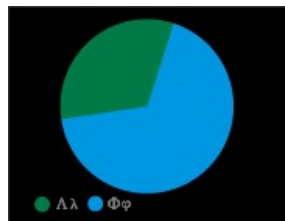
Map

	Περιφέρ.	Δυστυχών	αριθμός
Οδηγώ	11	11,000	1,300
Μικρό	18	18,000	1,700
Μεσαίο	18	18,000	1,700
μεγάλο	47	47,000	4,500
Σύνολο	5	500	20
Ποιότητα	8	800	90
Μικρό	2	200	50
Μεσαίο	15	1,500	210

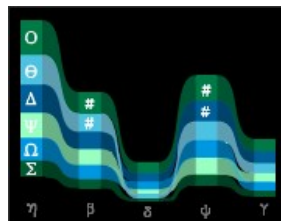
Multi Dimension Matrix



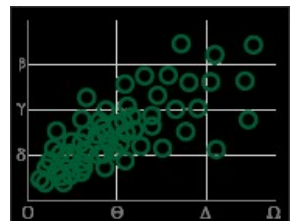
Narrative Text



Pie Chart



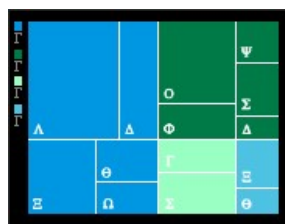
Ribbon Chart



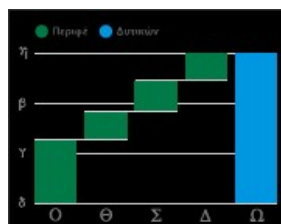
Scatter Chart

Κατηγορία	Περιφέρ.	Δυστυχών	αριθμός
Οδηγώ	11	11,000	1,300
Μικρό	18	18,000	1,700
Μεσαίο	18	18,000	1,700
μεγάλο	47	47,000	4,500
Σύνολο	5	500	20
Ποιότητα	8	800	90
Μικρό	2	200	50
Μεσαίο	15	1,500	210

Schematic



Tree Map



Waterfall Chart

Bar Chart

Bubble Chart

Combo Chart

Decomposition Chart

Dot Plot Chart

Doughnut Chart

Filled Map Chart

Funnel Chart

Gauge Chart

Graph Chart

HeatMap

Line Chart

Map

Multi Dimension Matrix

Narrative Text

Pie Chart

Ribbon Chart

Scatter Chart

Schematic

Tree Map

Video

Waterfall Chart

Personally Identifiable Information (PII) Images

Contains image elements that represent Personally Identifiable Information (PII). The data refers to any information that can be used to identify an individual. It includes any data that can distinguish or trace back to a specific person, either on its own or in combination with other information.

Personally Identifiable Information



Address



Bank Account Number



Biometric Identifiers



Credit Card Number



Date of Birth



Drivers License Number



Education Level



Email Address



Employee Identification Number



Facial Recognition Data



Fingerprints



Full Face Photo



Full Name



Gender



Genetic Identifier



GPS Coordinates



Health Insurance Identifier



IP Address



Iris Retinal Scan



Marital Status



Medical Record



National Identification Number



Online User Name



Passport Number



Phone Number



Social Media Profile



Social Security Number



Tax Identification Number



Vehicle License Plate



Voice Print

The diagram shows the Personally Identifiable Information of a person. The elements are associated with the person and are grouped in the diagram by their type (classification). The two legends color the elements - the fill color signifies the type, and the border color is the sensitivity level. A modeler can use the elements in conjunction with the Personally Identifiable Information images to create a compelling representation of the PII data.

Internet of Things (IoT) Images

Internet of Things (IoT) Images



Generic Sensor



Surveillance Camera



Traffic Flow Camera



Traffic Light



Rail Crossing



Web Site Clicks



Mobile Phone



Social Media



Biometrics Scanner



Alarm



Building Monitor



Infrastructure
Monitor



and



Weather Station



Weather Balloon



Farm Animal



Harvester



Farm Vehicle



Package Location



Warehouse Stock



Retail Product
Placement



Package Delivery



Warehouse
Automation



Warehouse Robotic
Picking



Package Tracking



Robot



Factory



Digger



Drill Press



Soil Sensor



Fraud Detection



Heart Monitor



Smart Watch



Train



Ferry










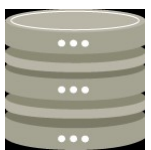



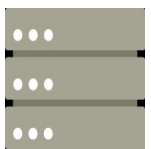
Tram



Bus

Data Storage Images

Data Storage Images

			
Database	Data Store	Data File System	Data Lake
			
Data Mesh	Data Fabric	Data Hub	Data Mart
			
Delta Lake	Data Warehouse	Data Lakehouse	Database

Principles

Lijst van uitgewerkte architectuur principes voor data gerelateerde activiteiten in de breedste zin van het woord van data gedreven werken

Universal Data Principles

Universal data principles are foundational guidelines and best practices that apply broadly to data management, regardless of the specific context, organization, or industry. These principles help ensure the responsible and effective use of data assets.

Universal Data Principles

<p>«data principle» Data is an Asset</p> <p>Data is an important and valuable organizational asset and must be managed through its lifecycle from creation, enrichment, storage, and visualization to destruction.</p>	<p>«data principle» Data has a Custodian</p> <p>The operational aspects of data are managed and safeguarded by the data custodian on behalf of the data owner.</p>	<p>«data principle» Data is Defined</p> <p>Data is defined consistently throughout the organization, and the definitions are comprehensible and accessible to all users.</p>
<p>«data principle» Data is Shared</p> <p>Data is used and consumed by various organizational functions and roles that need access to specific data to perform their roles.</p>	<p>«data principle» Data has an Owner</p> <p>The data owner is typically a business stakeholder or department within an organization with ultimate accountability and responsibility for a specific dataset.</p>	<p>«data principle» Data is Secure</p> <p>Data classified as confidential, sensitive, or personal will be secured and protected in transit, at rest, and in use from unauthorized access or dissemination.</p>
<p>«data principle» Data is Available</p> <p>Data classified for users will be available and accessible to the roles that need the data to perform their business or technical functions.</p>	<p>«data principle» Data is Curated</p> <p>Data is collected, organized, validated, preserved, and managed through its lifecycle to ensure its quality, usability, and long-term value, from acquisition or creation to archiving or disposal.</p>	<p>«data principle» Data Flows are Optimized</p> <p>The movement of data between environments and parts of a pipeline should be optimized to reduce costs and processing time and increase data currency.</p>

Universal data principles are foundational guidelines and best practices that apply broadly to data management, regardless of the specific context, organization, or industry. These principles help ensure the responsible and effective use of data assets.

Data Flows are Optimized

The movement of data between environments and parts of a pipeline should be optimized to reduce costs and processing time and increase data currency.

Data has a Custodian

The operational aspects of data are managed and safeguarded by the data custodian on behalf of the data owner.

Data has an Owner

The data owner is typically a business stakeholder or department within an organization with ultimate accountability and responsibility for a specific dataset.

Data is an Asset

Data is an important and valuable organizational asset and must be managed through its lifecycle from creation, enrichment, storage, and visualization to destruction.

Data is Available

Data classified for users will be available and accessible to the roles that need the data to perform their business or technical functions.

Data is Curated

Data is collected, organized, validated, preserved, and managed through its lifecycle to ensure its quality, usability, and long-term value, from acquisition or creation to archiving or disposal.

Data is Defined

Data is defined consistently throughout the organization, and the definitions are comprehensible and accessible to all users.

Data is Secure

Data classified as confidential, sensitive, or personal will be secured and protected in transit, at rest, and in use from unauthorized access or dissemination.

Data is Shared

Data is used and consumed by various organizational functions and roles that need access to specific data to perform their roles.

Data Architecture Principles

Data architecture principles serve as guiding rules for designing and developing effective data architectures within organizations.

Data Architecture Principles

<p>«data principle» Governed and Managed</p> <p>Refers to the structured and regulated approach to overseeing, organizing, and maintaining an organization's data assets to ensure that all personnel and systems use data securely, effectively, and in compliance with regulatory requirements.</p>	<p>«data principle» Scalable</p> <p>Refers to a system's ability or architecture to manage increasing amounts of data or expanding workloads without sacrificing performance, reliability, or maintainability.</p>
<p>«data principle» Robust and Resilient</p> <p>Refers to designing data systems capable of withstanding disruptions, maintaining data integrity, ensuring high availability, and guaranteeing data operations function effectively despite unanticipated challenges.</p>	<p>«data principle» Driven by Artificial Intelligence</p> <p>Refers to incorporating and operating artificial intelligence (AI) and machine learning (ML) technologies within the design and management of a data architecture or system. Uses include predictive analysis, anomaly detection, cost optimization, and personalization.</p>
<p>«data principle» Adaptable and Flexible</p> <p>Refers to the capacity of a data system or architecture to respond effectively to changes in business processes, requirements, data sources, and technology environments. Flexibility allows a system to accommodate variations in data formats, data models, and data processing workflows. An adaptable data architecture can evolve and change to meet new opportunities and challenges without significant refurbishment or disruptions.</p>	<p>«data principle» Automated Pipelines</p> <p>Automate pipelines to streamline the process of data ingestion, data integration, data transformation, and data analysis. Efficiently manage and process large volumes of data, derive meaningful insights, make informed decisions, and automate business processes.</p>

Data architecture principles serve as guiding rules for designing and developing effective data architectures within organizations.

Adaptable and Flexible

Refers to the capacity of a data system or architecture to respond effectively to changes in business processes, requirements, data sources, and technology environments. Flexibility allows a system to accommodate variations in data formats, data models, and data processing workflows. An adaptable data architecture can evolve and change to meet new opportunities and challenges without significant refurbishment or disruptions.

Automated Pipelines

Automate pipelines to streamline the process of data ingestion, data integration, data transformation, and data analysis. Efficiently manage and process large volumes of data, derive meaningful insights, make informed decisions, and automate business processes.

Driven by Artificial Intelligence

Refers to incorporating and operating artificial intelligence (AI) and machine learning (ML) technologies within the design and management of a data architecture or system. Uses include predictive analysis, anomaly detection, cost optimization, and personalization.

Governed and Managed

Refers to the structured and regulated approach to overseeing, organizing, and maintaining an organization's data assets to ensure that all personnel and systems use data securely, effectively, and in compliance with regulatory requirements.

Robust and Resilient

Refers to designing data systems capable of withstanding disruptions, maintaining data integrity, ensuring high availability, and guaranteeing data operations function effectively despite unanticipated challenges.

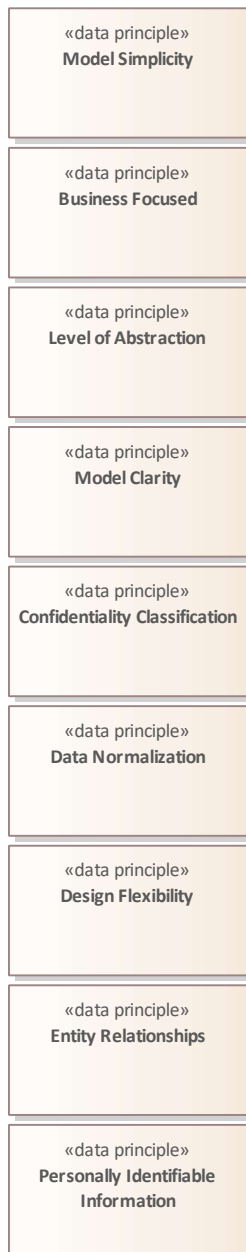
Scalable

Refers to a system's ability or architecture to manage increasing amounts of data or expanding workloads without sacrificing performance, reliability, or maintainability.

Data Modeling Principles

Data modeling principles refer to the foundational guidelines and best practices that data modelers use to create data models. Data models describe data structures, entities, tables, relationships, and constraints that help define data-driven applications, including conceptual, logical, and physical data models. The principles include ensuring that conceptual data models are business-focused and handle levels of abstraction.

Data Modeling Principles



Data modeling principles refer to the foundational guidelines and best practices that data modelers use to create data models. Data models describe data structures, entities, tables, relationships, and constraints that help define data-driven applications, including conceptual, logical, and physical data models. The principles include ensuring that conceptual data models are business-focused and handle levels of abstraction.

Business Focused

Data models should focus on essential aspects of the business domain and abstract away unnecessary details. Regularly validate and verify the data model with business stakeholders to ensure it accurately represents the real-world domain and is fit for purpose.

Confidentiality Classification

Include a field to model data with a confidentiality level for documents, communications, locations, and facilities to ensure this data is distinguishable from other information. Include a reference entity that lists the confidentiality level.

Data Normalization

Apply normalization methods to reduce data redundancy and improve data integrity. Normalization increases efficiency and consistency.

Design Flexibility

Design the data model to be flexible enough to accommodate future business and technical changes and requirements. Avoid inflexible and brittle structures that impede scalability and adaptability.

Entity Relationships

Define relationships between entities accurately. Depending on the business rules and requirements, these relationships can be one-to-one, one-to-many, or many-to-many.

Level of Abstraction

Represent entities, relationships, and attributes at an appropriate level of granularity. Conceptual Data Models typically do not model attributes and naming is business focused. Logical Data Models add attributes with generic data types. Physical Data Models often change names to suit an implementation and add platform-specific data types.

Model Clarity

Ensure that the data model is straightforward and easy to understand by all stakeholders, including business users, developers, and data professionals. Use clear and consistent naming conventions for entities, attributes, and relationships.

Model Simplicity

Ensure the data model is as simple as possible without sacrificing its ability to represent the real-world domain accurately. Avoid unnecessary complexity that attenuates model comprehensibility.

Personally Identifiable Information

Include a field to model PII data to ensure this data is distinguishable from other information. Include a reference entity that lists the privacy type and another that lists the sensitivity level.

General Data Protection Regulation (GDPR) Principles

The General Data Protection Regulation (GDPR) is a comprehensive data protection law that sets out several fundamental principles for processing personal data within the European Union (EU). These principles provide a framework for organizations to ensure the lawful, fair, and transparent handling of personal data.

General Data Protection Regulation (GDPR)

<div>«data principle» Lawfulness, fairness and transparency</div> <div>Processing must be lawful, fair, and transparent to the data subject.</div>	<div>«data principle» Purpose limitation</div> <div>You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.</div>	<div>«data principle» Data minimisation</div> <div>You should collect and process only as much data as absolutely necessary for the purposes specified.</div>
<div>«data principle» Accuracy</div> <div>You must keep personal data accurate and up to date.</div>	<div>«data principle» Storage limitation</div> <div>You may only store personally identifying data for as long as necessary for the specified purpose.</div>	<div>«data principle» Integrity and confidentiality (security)</div> <div>Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).</div>
<div>«data principle» Accountability</div> <div>The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.</div>		

The General Data Protection Regulation (GDPR) is a comprehensive data protection law that sets out several fundamental principles for processing personal data within the European Union (EU). These principles provide a framework for organizations to ensure the lawful, fair, and transparent handling of personal data.

Accountability

The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles.

Accuracy

You must keep personal data accurate and up to date.

Data minimisation

You should collect and process only as much data as absolutely necessary for the purposes specified.

Integrity and confidentiality (security)

Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. by using encryption).

Lawfulness, fairness and transparency

Processing must be lawful, fair, and transparent to the data subject.

Purpose limitation

You must process data for the legitimate purposes specified explicitly to the data subject when you collected it.

Storage limitation

You may only store personally identifying data for as long as necessary for the specified purpose.

Data Role Library

Lijst met bedrijfsrollen die voorkomen binnen het werkveld data (gedreven werken)

Data Role Library



Business Intelligence Specialist

Business Intelligence Specialists are responsible for designing, developing, and maintaining business intelligence solutions, such as data warehouses, data marts, and reporting systems. They enable users to access, visualize and analyze data easily. The visualizations and analysis support meaningful and actionable insights that stakeholders can use to support decision-making and drive business strategies.

Chief Data Officer

Chief Data Officers (CDOs) are senior executives responsible for managing and leveraging an organization's data assets to drive strategic decision-making, improve operational efficiency, ensure regulatory compliance, and foster data-driven innovation. The CDO plays a critical role in developing and implementing data strategies, ensuring data quality and governance, and promoting a data-centric culture within the organization.

Chief Information Officer

Chief Information Officers (CIOs) are senior-level executives responsible for overseeing and managing an organization's information technology (IT) systems and strategies. The CIO typically reports directly to the CEO or another top-level executive. The role aligns the organization's IT initiatives and resources with its high-level business goals and objectives. This alignment involves developing and implementing IT strategies that support the organization's vision, mission, and long-term plans.

Data Analyst

Data Analysts have the knowledge and skills to change raw data into information and insights that benefit data consumers. They are typically given a specific requirement or problem that needs to be solved and use their knowledge of the structure and content of available data to provide answers to the business.

Data Custodian

Data custodians are responsible for operating and managing technology, including systems that collect, store, process, manage, and provide access to the organization's data. They are commonly associated with the technology services and functions of the organization but may also include systems administrators working within one or more functional areas.

Data Engineer

Data Engineers are responsible for designing, developing, and maintaining the infrastructure, systems, and pipelines required for the efficient and reliable processing, storage, and retrieval of large volumes of data. They work closely with data scientists, analysts, and other stakeholders to ensure data is collected, transformed, and accessible for analysis and decision-making.

Data Governance Manager

A Data Governance Manager establishes and enforces the organization's data governance policies and procedures. The role defines data ownership, establishes data quality standards, and ensures compliance with data protection regulations. They collaborate with other stakeholders to develop data governance frameworks and oversee data stewardship activities, including developing, implementing, and managing data governance initiatives.

Data Modeler

Data modelers create graphical representations of business or information systems. The models are used both to communicate and to define the data entities and their relationships, attributes, and data types. There are conceptual, logical, and physical data modelers. Business requirements and reporting needs drive the creation of conceptual models, and platform and system constraints drive the creation of physical data models.

Data Owner

Data Owners are typically directors or managers with the authority to determine business definitions of data, grant access to data, and approve secure data usage for the functional areas within their jurisdiction of authority. By understanding the organization's information needs, data owners can anticipate how data can be used to meet the organization's strategic goals.

Data Protection Officer

A Data Protection Officer is an individual designated by an organization to oversee data protection and privacy matters within the organization and information shared with external entities. The role is typically associated with compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union, and analogous authorities in other jurisdictions concerned with privacy and human rights law.

Data Quality Analyst

Data Quality Analysts focus on assessing and improving the quality of data within an organization. They develop and implement data quality metrics, conduct data quality assessments, and collaborate with data owners to address data quality issues.

Data Scientist

Data scientists are advanced analysts responsible for collecting, analyzing, and interpreting data to help drive organizational decision-making. They use sophisticated analytics techniques, such as machine learning and predictive modeling, to gain hidden insights and create predictive models. They can work with structured, semi-structured or unstructured data sets.

Data Stakeholder

Data Stakeholders are any individual or group interested in or involved in the management, access, quality, and use of data within an organization or a specific project. Data stakeholders can include various roles and departments across an organization, each with different perspectives and responsibilities related to data.

Data Steward

Data stewards are responsible for implementing data policies and managing one or more types of organizational data. They authorize and monitor the secure use of data within their assigned business areas. They ensure appropriate access, accuracy, classification, privacy, and data security is maintained at all times.

Data User

Data users are authorized individuals with access to organization data to perform their assigned duties or functions. When users are given access to data, they assume responsibility for the appropriate use, management, and application of privacy and security standards for the data they are authorized to use.

Database Administrator

Database Administrators (DBAs) are responsible for managing and maintaining an organization's databases. They ensure data integrity, security, and performance by monitoring database performance, optimizing queries, and managing backups.

Executive Sponsor

An Executive Sponsor is a senior leadership member with planning and policy responsibility and accountability for major administrative data systems within their functional areas. By understanding the organization's functions, they can anticipate how it will use data to meet internal and external organizational needs.

Information Architect

Information Architects focus on organizing, structuring, and designing information systems to facilitate efficient access, retrieval, and use of information. They are responsible for analyzing an organization's information needs, understanding user requirements, and creating secure and compliant information architectures that meet those needs. They work closely with stakeholders, such as business leaders, content creators, and technology teams, to develop solutions that enhance information management and usability.

Tot slot

Dit model is gegenereerd met Sparx Enterprise Architect en een nieuwe functionaliteit Custom Documents. Heb je interesse in een filebased repository of XML file met de inhoud van dit document schroom dan niet om me te contacteren.

Over de auteur



Bert Dingemans is trainer op het vlak van data architectuur, data management en Big Data. Hij heeft een passie voor modelleren, modelleertools en het effectief inzetten van geautomatiseerde hulpmiddelen om modellen effectief in te zetten in de praktijk. Meer whitepapers zijn te vinden op <https://data-docent.nl>. Bert is per mail te bereiken via bert@data-docent.nl