CROZ

Education catalog

Data and AI/ML

Table of contents

Course list

- Architecture, design and data management

- Business and Data Analysis

- Development

Overview

Digital transformation is present around us today. One of the essential elements of digital transformation is data-driven transformation. Data management has never been more challenging, and AI and ML initiatives are emerging and cannot function without supporting data processes and architecture.

Our experience working on modern AI/ML and data projects has shown us the importance of continuous education. It can be a crucial factor in the success of a company. To educate our clients, increase their knowledge level, and keep up with technological changes, we organize a whole series of educational services.

Our data and AI/ML education services include prepared catalogue courses, customized workshops, and consulting on actual projects. We continuously develop new courses. Our data engineers and consultants consistently participate in development projects where they acquire new knowledge. All education services are delivered by the ISO 9001 standard for education, and IT Professional Training. Our instructors receive high grades for providing education services.

We have prepared an array of courses, training programs, and consulting packages that help organizations of all sizes and industries successfully implement their agile transition. This is done thanks to our partnership with the agile42 company, specialized in consultancy in agile methods.

As an accredited provider of ISTQB® training, we offer a preparatory course to ISTQB Foundation as well as ISTQB® certification.

We have more than ten years of experience in educating our clients from the data-driven domain. We improve our education program from year to year, considering best practices and our expertise. What we have realized over time is that interactivity and modularity are essential elements of every education. Besides this, we have transformed some trainings from course to workshops where our participants can work with their data and use cases. It is challenging to create and adapt training that will be acceptable to any client or project, so we created a customizable set of training for you.

For any additional inquiries

Contact us at learn@croz.net

Interactive exercises

We think that course interactivity is crucial - participants can experience the principle of "learning by doing." For this reason, our courses use a minimum of presentation and paper scripts and rely more on interactive notebooks and assignments that contain enough theoretical material but concentrate more on real-world examples. Our classes are designed in a way that allows students to have maximum focus on practical examples - students can immediately complete and verify the correctness while the lecturer explains and teaches. In this way, the student can try out different possibilities and ask specific questions.

Modularity

Most of our education is designed to be modular. Modularity means that one training contains a series of modules, i.e., independent units that can be combined depending on the level of knowledge and the desired preferences. In this way, companies can set an optimal educational plan depending on the company's needs, level of expertise, and future requirements.

Workshops - Customized Data courses

Since current data jobs become more and more complex, with an increasing number of topics, there is no single course that could cover it all. Also, since people come from various backgrounds, there can be a different level of knowledge for topics that are taught. To cover that, apart from courses from the catalogue, Learn@CROZ offers a possibility for customized courses – workshops - tailored to your needs. In cooperation with you, we can modify our courses to provide you with more valuable experience. To give you an idea, these are just some of the options which we can do for you:

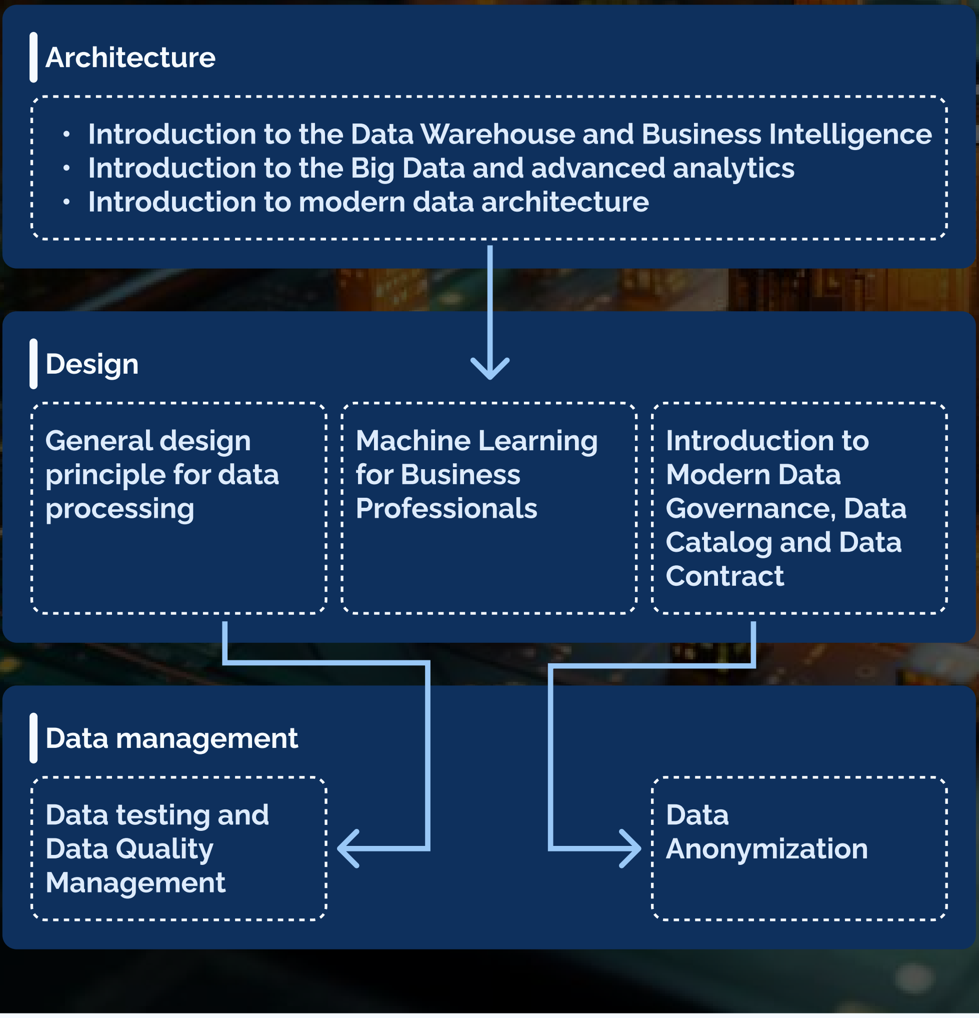
- **Modification of Course Syllabus** – Include topics that are (more) relevant to your current work e.g., if you want some of the course topics covered in more detail.
- **Merging of Course Syllabuses** – We can merge topics and modules from different courses into one course.
- **Bring Your Own Data (BYOD)** – To make exercises more familiar to your students, we can include your data set in the hands-on exercises.
- **Bring Your Own Use Case (BYOUC)** – If you want us to cover your business use case through education, we can do that. Most courses have an initial scenario that drives the examples and exercises, and we can include yours.

Education packages

Educational packages represent set of courses for training for specific data role - Data Architect, Data Engineer and Data Analyst. This is just a suggestion, see the table containing descriptions of all trainings and modules and build the optimal plan that best suits you.

Data Architect package

This tailored package serves as an ideal learning guide for data architects and team leaders in data-centric organizations. It provides a comprehensive curriculum for those engaged in the orchestration, conceptualization, and leadership of data-based initiatives and teams. While it serves as an excellent starting point for those new to data architecture and design, it also affords seasoned professionals a holistic overview of all contemporary, enterprise-related data topics



Architecture

- Introduction to the Data Warehouse and Business Intelligence
- Introduction to the Big Data and advanced analytics
- Introduction to modern data architecture

Design

General design principle for data processing

Machine Learning for Business Professionals

Introduction to Modern Data Governance, Data Catalog and Data Contract

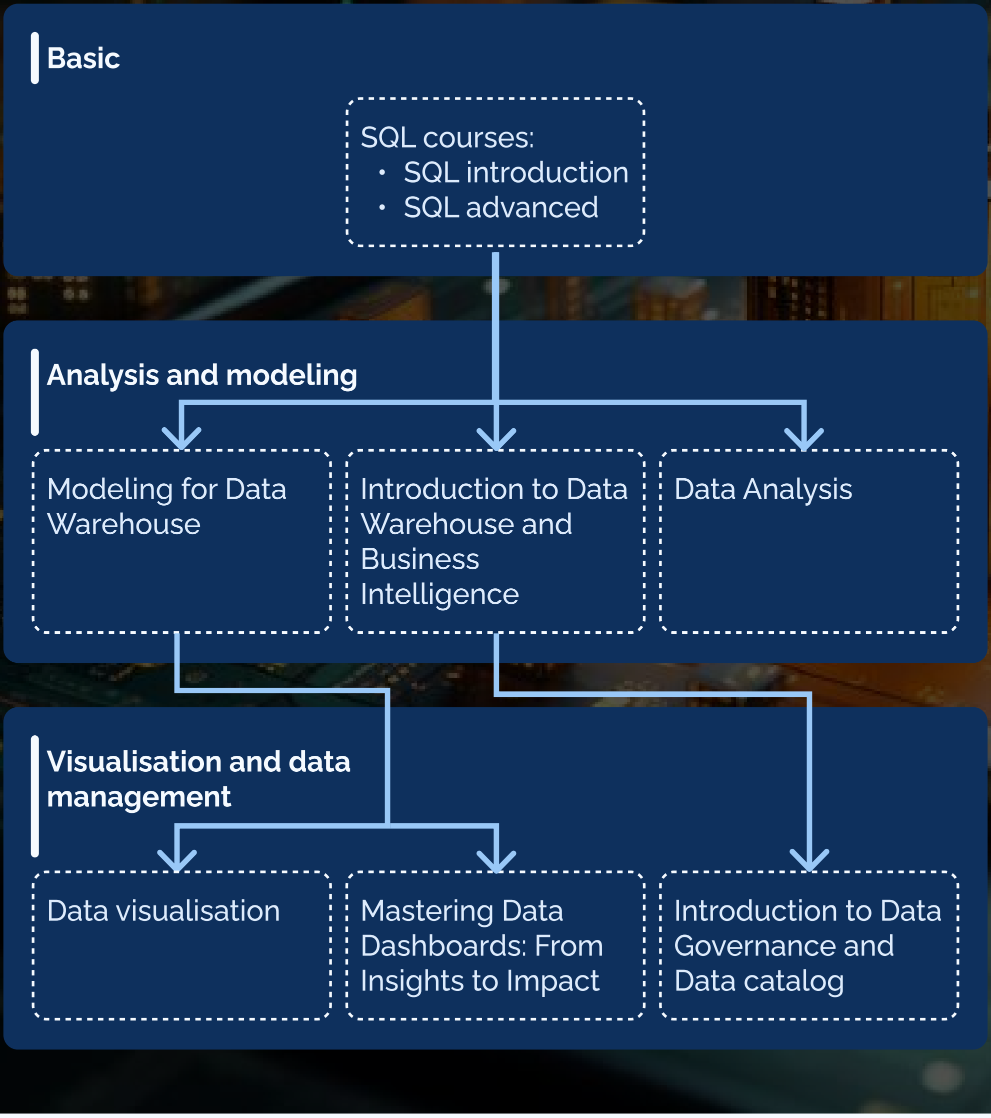
Data management

Data testing and Data Quality Management

Data Anonymization

Data Analyst package

We present this package as an optimal learning path for data and business analysts, along with team leaders in data-focused organizations. It addresses the needs of those engaged in data analysis and the application of data in making decisions. While it offers a solid foundation for novices in data analysis and visualization, it also delves into multiple topics associated with the business aspect of data management.



Basic

SQL courses:
- SQL introduction
- SQL advanced

Analysis and modeling

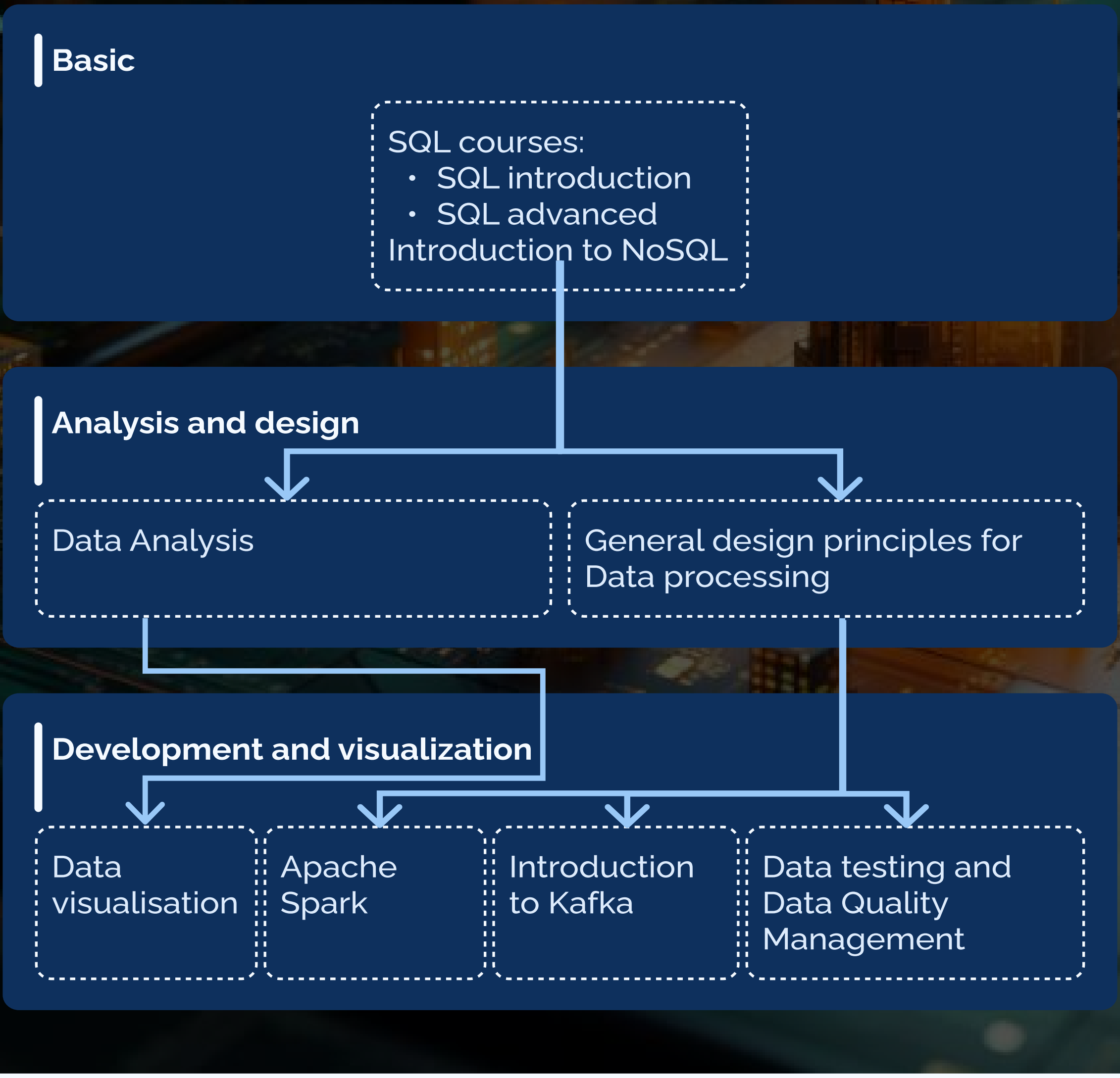
- Modeling for Data Warehouse
- Introduction to Data Warehouse and Business Intelligence
- Data Analysis

Visualisation and data management

- Data visualisation
- Mastering Data Dashboards: From Insights to Impact
- Introduction to Data Governance and Data catalog

Data Engineer package

This is our recommended package and roadmap for data engineers and integration specialists. It focuses on those who handle the infrastructure, development, and integration elements of data. It's an accessible starting point for those venturing into data engineering and integration. Simultaneously, it encompasses a broad range of topics in data engineering, including open-source tools such as Kafka and Spark, which have become standards in the industry. This makes it equally beneficial for seasoned practitioners.



Basic

SQL courses:
- SQL introduction
- SQL advanced

Introduction to NoSQL

Analysis and design

Data Analysis

General design principles for Data processing

Development and visualization

Data visualisation

Apache Spark

Introduction to Kafka

Data testing and Data Quality Management

Course list

Architecture, design, and data management

Introduction to Modern Data Architecture

Duration – **2 days**

Overview

The main objective of this course is to provide a comprehensive understanding of contemporary data architecture principles. Today's modern data architectures must support data-driven transformation, meaning we must be able to implement various concepts that complement each other. This education is designed first to define the needs, added value, and requirements for architecture, and then show the assessment process that will identify which architectures and design patterns are needed to support the desired data-driven transformation. After the introduction, we delve into architectures such as DWH, Data Lake, Lakehouse, Data Fabric, etc. In the end, we will also explore the concepts of Data as a Product and Data Mesh.

Target audience

The course is intended for business and technical teams that need an introduction and a structured overview of Modern Data Architecture. The intended audience for this education includes business stakeholders, architects, data engineers, as well as business and data analysts.

Prerequisites

The prerequisite for this course is familiarity with the concepts and architectures of data management.

Content

Day 1

- What does it mean to be Data-driven?
  - Basic concepts
  - The basis of every data architecture
- Data Warehouse (DWH) and Business Intelligence (BI)
  - General architecture overview and application
  - How to model a DWH platform
  - The process of developing a DWH platform
  - Practical examples and tools
- Data Lake
  - General architecture overview
  - How to design a Data Lake platform (we don't want a data swamp)
  - What is the process of developing a Data Lake platform
  - New derivatives Lakehouse, Data Fabric, tools used

Day 2

- LakeHouse
  - General architecture overview and application
  - How to design a LakeHouse platform
  - The process of development within the LakeHouse platform
  - Practical examples and tools
- Data fabric
  - General architecture overview and application
  - How to design a Data Fabric platform
  - The process of development within the Data Fabric platform
  - Practical examples and tools
- Data Mesh
  - Why Data Mesh?
  - Four basic characteristics of Data Mesh (Domain-driven, Data Product, self-service, and computational federate governance)

General design principles for Data processing

Duration – **2 days**

Overview

Today, successful data architectures and data processing platforms must undertake continuous modernization to maintain data platform sustainability. Various technologies (ETL, Streaming, MLOps) and concepts have emerged over the past decade and sustainable ETL process design that isn't directly dependent on a platform or concept is becoming increasingly challenging. This education is intended for engineers and architects who wish to develop their own Framework or standard for data processing, whether it's for populating a DWH, DataLake, LakeHouse, Data Fabric, or Data Mesh to ensure sustainability on the given platform.

The primary goal of this education is to teach participants how to efficiently use key design patterns depending on the business case and architectural principles. This course covers all the important architectural patterns for data processing, ranging from ETL, ELT, and near real-time, to advanced transformations and support for streaming processing. The education will provide answers on how to approach the analysis of existing processing requirements and define standards for each process.

Target Audience

The course is intended for business and technical teams that need an introduction and a structured overview of Modern Data Architecture. Its intended roles are architects, data engineers, and business and data analysts.

Prerequisites

The prerequisite for this course is that participants are familiar with the concepts and architectures of data management.

Content

Day 1

- Introduction and Overview
  - Review of the agenda
  - Introduction to the topic: Overview of data architectures and data processing
  - Overview of relevant technologies and main requirements for Batch, Streaming, MLOps
- Introduction to Concepts
  - Data Architecture: DWH, DataLake, LakeHouse, Data Fabric, Data Mesh
  - Definition and importance of Data Ingestion processes
  - Definition and importance of traditional ETL processes
  - Definition and importance of self-service data processing
- Workshop: Analysis requirements for building individual Frameworks and Standards
  - A practical workshop where participants analyze examples of their Frameworks and standards
- Discussion and Q&A
  - Discussion of the material covered during the day
  - Answers to questions

Day 2

- Application of Key Design Patterns
  - Learning about key design patterns (CDC, push, pulls, API)
  - Review of business cases and architectural principles
- Data Processing and Architectural Patterns
  - Overview of all important architectural patterns for data processing
  - A detailed discussion of ETL, ELT, near real-time, advanced transformations, and support for streaming processing
  - A detailed discussion of the automation of the whole process
- Workshop: Build and Publish Standards
  - Defining and publishing standards for each process (design, build, control)
- Conclusion and Q&A
  - Review of what was covered during the two days
  - Opportunity for additional questions and discussion

Introduction to Modern Data Governance, Data Catalog, and Data Contract

Duration – **2 days**

Overview

This education consists of two modules. The first module is a full-day education providing an overview of Data Governance core concepts, such as ownership and stewardship processes, how to establish a Data Governance office with virtual and/or physical roles, what are the mandatory parts of a metadata management framework, and how to set up a data quality process. The second module focuses on the operational side of Data Governance, i.e., it explains how and in what way to implement Data governance policies, standards, and processes in the current data integration platform. More specifically, on the second day/module, it will be shown how to decide which Data Catalog to implement and how to integrate Data Contract into a Data Catalog based on knowledge graph concepts.

Target Audience

The course is intended for business and technical teams that need an introduction and a structured overview of Data Governance practices. Business roles like business stakeholders, architects, data engineers, and business and data analysts.

Prerequisites

The prerequisite for this course is familiarity with the concepts and architectures of data management.

Content

Day 1 Introduction to Data Governance

In the beginning, we introduce key Data Governance concepts and initiatives.

- What is and why is data governance important in every data-driven organization
- Introduction to the Data Governance Office and key Chief Data Governance role
- How to define data ownership and stewardship process
- Introduction to Metadata Management and what are the key components of Metadata Management Framework
- Data Quality management and how to define initial data quality measures
- Data Warehouse & Data Lake
- Master Data Management

Day 2 Data Contract and Data Catalog

Building a Data Governance program with a Data Catalog is the most critical part, from a technical implementation standpoint. The Data Catalog is the heart of metadata management where all business and technical aspects are stored (persisted) in one unified view in a single repository. On the second day, all participants will be able to learn how to build a Data Catalog properly with a good assessment.

- What is Data Catalog
- How to effectively implement Data Lineage for business and technical stakeholders
- Why and how to implement a business glossary
- Why and how to properly classify data using classification or tags
- How to implement Data Privacy policy and standards
- What is a Data Contract and how is it effectively stored in a Data Catalog
- How to correctly assess and choose the best tool for a Data Catalog

Data testing and Data Quality Management

Duration – **2 days**

Overview

This course is designed to provide participants understanding of Data Quality concepts and principles. During the course participants will learn the importance of high-quality data, the key differences between testing data and Data Quality, are dimensions of Data Quality, how to assess Data Quality, and what steps are to be taken to improve it. It also shows how Data Quality affects business and is important on an organizational level.

Target Audience

- Data Engineers
- Data Analysts
- Data Product Owners
- Business Managers
- Project Managers

Prerequisites

- Basic understanding of databases and SQL

Content

Day 1

- Introduction to Data Testing and Data Quality Management
  - Overview of the course objectives and structure
  - Importance of data quality in decision-making and business processes
  - Difference between testing data and data quality
- Understanding Data Quality
  - Definition and dimensions of data quality
  - Common data quality issues and challenges
  - Impact of poor data quality on organizations and decision-making
- Organizational Processes and Data Governance
  - Importance of data governance in ensuring data quality
  - Roles and responsibilities of data stewards and data owners
  - Establishing data governance frameworks and processes
- Data Quality Assessment
  - Techniques for assessing data quality
  - Data quality assessment tools and methodologies
  - Designing and implementing data quality assessment processes
- Data Quality Metrics and Monitoring
  - Defining key data quality metrics and indicators
  - Establishing data quality thresholds and benchmarks
  - Implementing data quality monitoring and reporting processes
- Data Cleansing and Remediation
  - Techniques and approaches for data cleansing
  - Addressing data quality issues and errors
  - Implementing data remediation processes

Day 2

- Data Quality Improvement Strategies
  - Data quality improvement methodologies
  - Best practices for maintaining data quality over time
- Data Quality Management Tools and Technologies
  - Overview of data quality management tools and technologies
  - Selection criteria for data quality tools
  - Integration of data quality tools into existing data management systems
- Data Quality Governance and Regulatory Compliance
  - Relationship between data quality governance and regulatory compliance
  - Ensuring data privacy and security in data quality management processes
  - Compliance considerations in data quality management
- Case Studies and Practical Applications
  - Real-world examples of data quality management implementations
  - Case studies illustrating the impact of data quality on organizations
  - Practical exercises and assignments to reinforce learning
- Course Review and Conclusion
  - Recap of key concepts
  - Q&A session

Data Anonymization

Duration – **2 days**

Overview

A Data Anonymization workshop is designed to provide participants with a comprehensive understanding of data anonymization techniques, best practices, and regulatory considerations. The workshop aims to equip attendees with the knowledge and skills necessary to effectively anonymize sensitive data while preserving its utility for analysis and research purposes. Throughout the workshop, there will be interactive sessions, hands-on exercises, and discussions to reinforce the concepts learned. Participants will have the opportunity to apply anonymization techniques to sample datasets and discuss practical implementation challenges. Participants gain a deeper understanding of privacy-preserving data practices, learn practical techniques for anonymization and become equipped to navigate the legal and ethical challenges associated with handling sensitive data.

By the end of the workshop, participants will have a solid understanding of data anonymization concepts and different anonymization techniques.

Target audience

- Data Privacy and Security Professionals
- Data Engineers and Data Governance Professionals
- Data Scientists and Analysts
- Legal and Compliance Professionals

Prerequisites

- Basic knowledge of data management and regulations related to data (e.g., GDPR)

Content

- Introduction to Data Anonymization:
  - Overview of data anonymization and its importance
  - Legal and ethical considerations related to data privacy and protection
- Anonymization Techniques:
  - Different methods and algorithms for data anonymization
  - Approaches for masking personally identifiable information (PII)
- Evaluating Anonymization Effectiveness
  - Metrics for assessing the privacy risk and utility of anonymized data
  - Methods for measuring re-identification risk
  - Balancing privacy and data utility trade-offs
- Practical Anonymization Strategies:
  - Anonymization techniques for structured and unstructured data
  - De-identification methods for various data types (e.g., numerical, categorical, text)
  - Data generalization, suppression, perturbation, and other anonymization approaches
- Regulatory Compliance and Privacy Laws:
  - Understanding relevant data protection regulations (e.g., GDPR, CCPA)
  - Compliance requirements for anonymized data
  - Legal considerations for sharing and using anonymized data
- Challenges and Limitations:
  - Limitations and vulnerabilities of anonymization techniques
  - Risks associated with re-identification attacks
  - Mitigation strategies and ongoing research in data anonymization
- Anonymization Tools and Technologies:
  - Overview of software tools and libraries for data anonymization
  - **Hands-on demonstrations** of popular anonymization platforms
  - Integration of anonymization techniques into data processing workflows
- Case Studies and Practical Applications
  - Real-world case studies highlighting anonymization challenges and solutions
  - Group discussions on anonymization in specific industries and use cases
  - Best practices for implementing anonymization in different contexts

Business and Data Analysis

Machine Learning for Business Professionals

Duration – **2 days**

Overview

This course is designed to provide business professionals with an understanding of machine learning concepts, practical applications, lifecycle, and the ability to frame and evaluate machine learning problems. The course will focus on real-world machine learning solutions in business, covering various industries and use cases. Through hands-on workshops, participants will learn how to identify opportunities, and examine and evaluate prerequisites for machine learning solution implementation.

Target audience

- Business analysts
- Data analysts
- Business Professionals
- Management

Prerequisites

- Basic understanding of business processes and challenges
- No prior knowledge of machine learning or programming is required.

Content

- Introduction to Machine Learning
  - What is machine learning and why it matters for business?
  - Key terminology and concepts
  - Types of machine learning: supervised, unsupervised, and reinforcement learning
  - Intuitive explanation of main machine learning algorithms
  - Overview of machine learning tools and platforms
- Machine Learning Applications in Business
  - Real-world case studies from various industries
- Machine Learning Lifecycle
  - Machine learning projects and solutions' lifecycle
  - MLOps
- Machine Learning Problem Framing
  - Detailed hands-on exercise for framing and evaluating machine learning solutions for real world problem

Data Analysis

Duration – **2 days**

Overview

This two-day in-person data analytics course is designed to provide business and data analysts with a comprehensive understanding of the practical and theoretical aspects of data analysis. The course will focus on exploratory data analysis using Python, SQL, and MS Excel. Participants will learn how to clean, analyze, and visualize data to extract insights and make data-driven decisions.

Target audience

- Business analysts
- Data analysts
- Data Scientists
- Data Engineers

Prerequisites

- Basic programming knowledge (preferably Python)
- Basic understanding of databases and SQL
- Familiarity with MS Excel

Content

Introduction to Data Analytics

- The importance of data analytics in business
- Types of data: structured, semi-structured, and unstructured
- Overview of Python, SQL, and MS Excel as data analytics tools

Data Collection

- APIs, web scraping, and databases
- Importing data into Python, SQL, and Excel
- **Hands-on exercise**: Quick introduction to tools/notebooks

Data Cleaning and Pre-processing

- Handling missing and erroneous values
- Data transformation and normalization
- Feature selection and engineering
- Hands-on exercise: Data cleaning and pre-processing using Python, SQL, and MS Excel

Introduction to Probability and Statistics for Data Analysis

- Basic probability concepts and distributions
- Types of variables
- Descriptive statistics: measures of central tendency and dispersion
- Outliers
- **Hands-on exercise**: Descriptive statistics on the dataset using Python/SQL

Exploratory Data Analysis

- Business rules valuation
- Visualizing data using Python
    - Histograms, bar plots, box plots, and scatter plots
    - Heatmaps and correlation analysis
- **Hands-on exercise:** EDA using Python
- SQL for data analysis: aggregation and summary statistics
    - GROUP BY, HAVING, and JOIN clauses
    - Window functions and analytical functions
- Excel Pivots and Pivot Charts
- **Hands-on exercise**: EDA using SQL and Excel

Introduction to Machine Learning and Data Preparation for Models

- Supervised learning (regression and classification)
- Unsupervised learning (clustering and anomaly detection)
- Model evaluation and validation
- Data Preparation
    - Handling outliers and missing values
    - Standardization
    - Normalization
    - Handling of categorical variables (one-hot, ordinal encoding)
    - Feature Engineering (e.g. extracting time features)
- **Hands-on exercise**: Performing dataset preparation and building a simple machine learning model using Python and scikit-learn.

Data Analysis Best Practices

- Reproducible research and version control (Git and GitHub)
- Ensuring data quality and integrity
- Communicating results and insights effectively

Data Visualization

Duration – **1 day**

Overview

This course will focus on the application of data visualization, methods, and best practices of data visualization. The course consists of presentation, demonstration, and practical tasks for participants.

The goal of this course is to educate participants about basic steps in data visualization. Because of the way how the human brain processes information, the best impact and memorability of the information is achieved by using visual objects, instead of using exhausting reports, or extremely large tables.

Data visualization means the usage of visual objects, like tables and charts, so that they can carry out the messages to the final users in a simple, clear, and correct shape. The course will cover application, methods and the best practices for data visualization.

Goals of the course

- To introduce the participants to the advantages of data visualization and how the audience visually perceives the information
- To introduce the participants to available tools/techniques for data visualization
- To educate the participants on how to emphasize the message by choosing the correct type of visualization for the specific type of data
- To introduce the participants to the most common mistakes in the design of the visualizations and give participants the guidelines for the design of more effective visualizations.

Target audience

- IT managers
- Directors
- Analysts (both business and technical analysts)
- Everyone whose job is to present data and is important to tell a story by using appropriate visualizations
- Everyone who would like to introduce creativity in their work and at the same time improve the effectiveness of their reports, conclusions, and messages

Prerequisites

- None

Content

- **Introduction to the data visualization** – what is data visualization; current status and challenges; visual perception of the information; business value of visualizations
- **Ways for data visualization** – techniques and methods for data visualization; table and chart design
- **Application of data visualization** – when is the best time to use the data visualization; steps in the process of the virtual design – how to find a story in the data
- **The best practices**– how to use the visual language of the organization; most common mistakes and the best practices in visualization design
- **Tools for data visualization**– the overview of the available tools for data visualization; workshop - making tables and charts by applying what was learned

Mastering Data Dashboards: From Insights to Impact

Duration – **2 days**

Overview

The Dashboard Design course provides participants with the essential knowledge and skills to create effective dashboards. Covering the fundamentals of data visualization, participants learn about design principles, chart selection and the psychology of information retrieval. The course explores the differences between reports and dashboards, highlights dashboards as communication tools, and delves into the process of designing and preparing data for dashboards. Participants also gain insights into various types of dashboards and study best practices through examples of both poorly and well-designed dashboards. Throughout the course, there will be hands-on exercises for active participation and practical learning. By the end of the course, participants will be equipped to create impactful dashboards that effectively communicate data-driven insights.

Goals of the course

- To introduce the participants to the advantages of data visualization and how the audience visually perceives the information
- To introduce the participants to available tools/techniques for data visualization
- To educate the participants on how to emphasize the message by choosing the correct type of visualization for the specific type of data
- To introduce the participants to the most common mistakes in the design of the visualizations and give participants the guidelines for the design of more effective visualizations.

Target audience

- Data Analysts
- Business Intelligence Professionals
- Data Scientists
- Business Managers
- Marketers and Sales Professionals
- Project Managers
- Researchers and Academics
- Anyone interested in improving their dashboard design skills for better data communication and decision-making.

Prerequisites

- Fundamental Data Analysis Concepts
- Basic Excel Skills
- Basic Knowledge of Data Visualization Tools (Tableau, Microsoft Power BI...)
- Completed the data visualization course (preferable, but not mandatory)

Content

- Module 1: Fundamentals of Data Visualization
  - Introduction to data visualization principles
  - Understanding the psychology of information retrieval
  - Design principles for effective visualization elements
  - Selecting the right charts for data representation
- Module 2: Introduction to Dashboards
  - Differentiating between reports and dashboards
  - Identifying the ideal scenarios for dashboard creation
  - Exploring dashboards as powerful communication tools
- Module 3: Dashboard Design Guidelines
  - Defining the target audience and their goals
  - Designing an intuitive and user-friendly layout
  - Data preparation for dashboard implementation
- Module 4: Types of Dashboards by Use Cases
  - Overview of various types of dashboards (e.g., operational, strategic, analytical)
  - Understanding the specific use cases for each dashboard type
  - Tailoring dashboards to meet different business objectives
- Module 5: Best Practices in Dashboard Design
  - Analysing examples of poorly designed dashboards and common mistakes to avoid
  - Studying well-designed dashboards and effective visualization techniques
  - Incorporating best practices for creating impactful and engaging dashboards

NoSQL introduction

Duration – **2 days**

Overview

Today modern architecture systems use different databases to upgrade functionalities that cannot be efficiently resolved using traditional relational databases. Here come the Polyglot Persistence databases designed for special customer cases.

The goal of this education is to become familiar with the architectures and concepts of Polyglot Persistence databases. Besides the theoretical part participants will get practical experience working with customer cases using Neo4J, Cassandra, and MongoDB.

Participants will become familiar with:

- Graf, key-value, column, document db
- Key customer cases
- Best practices

Target audience

- Software and Data Engineers and Developers
- Software and Data Architects

Prerequisites

- Basic knowledge of database concepts
- Basic knowledge of Java programming language.

Content

Day 1

- NoSQL Introduction
    - Differences between relational and NoSQL databases
    - Use Cases
- NoSQL database types
    - key-value
    - document
    - column
    - graph
    - multi-model
- Introduction to MongoDB (high-level)

Day 2

- Introduction to MongoDB, Cassandra and Neo4J
    - Distributed database on commodity hardware
- Topics (theory and exercises)
    - Json
    - Schema design
    - Standalone
    - Config file
    - Crud
    - Index
    - Replica set
    - Majority read and write
    - Sharding
    - Security

Big Data and Advanced Analytics

Duration – **1-3 days**

Overview

The purpose of this course is to explain all the key elements when setting up a big data environment for advanced analytics.

Through the course, the participants will learn about the basic concepts of big data methodology and technology. The course includes some basic terms, such as predictive analytics, text processing, and sentiment analysis. The participants will get a detailed insight into data-at-rest technologies, such as Hadoop and NoSQL bases for batch processing a large amount of data. Other than that, data-in-motion technologies for processing data in real-time (real-time, Streaming, IoT,...) will also be explained.

Target audience

- Architects
- Business Analysts
- Data Scientists
- Data Engineers
- Integration development process engineers

Prerequisites

- The prerequisite for this course is that the participants are familiar with the concepts and architectures of data management

Duration

The course is divided into three days and there is a possibility of arranging the course by certain units:

- Day 1: The introduction to big data technologies
- Day 2: Architecture, technology, and development of big data managing
- Day 3: Architecture, technology, and development of data management in real time.

It is possible to participate in only one or two days or the whole three-day course.

Content

Day 1 The introduction to the big data environment

- What is advanced analytics?
- The concepts of advanced analytics (predictive, sentiment, …)
- How to establish a Data Science environment?
- Data quality in a big data environment
- Data visualization

Day 2 Data-at-rest – managing all the available data

- General architecture for big data solutions and „Data Lake“ (data-at-rest overview)
- The introduction to basic technologies for big data environment (Hadoop and Spark)
- What is the „Polyglot persistence“ architecture and how to use NoSQL databases?
- How to modernize the existing architecture for the needs of advanced analytics?
- Program storage devices (DWH appliance machine)
- Exercise: Building an application for big data processing

Day 3 Data-in-motion – managing data in real-time

- What is real-time, streaming, and sensor data?
- Lambda architecture – how to adapt batch and real-time analytics?
- General architecture for big data solutions (data-in-motion overview)
- Exercise: Building an application for real-time processing

Introduction to Data Warehousing and Business Intelligence

Duration – **2 days**

Overview

The Introduction to Data Warehousing and Business Intelligence course offers a comprehensive exploration of the fundamental terms and concepts in data warehousing and business intelligence. Participants will gain valuable insights into the architecture, processes, infrastructure, and technologies that constitute a robust DWH/BI system. Through a well-rounded curriculum comprising presentations, interactive demonstrations, and hands-on exercises, attendees will develop a deep understanding of these key topics.

Whether you are a business analyst seeking a deeper understanding of data-driven insights, a data architect looking to optimize your DWH/BI infrastructure, a database administrator aiming to enhance data management, or a software development engineer engaged in DW/BI implementation, this course offers an invaluable opportunity to expand your expertise and play a pivotal role in developing impactful business intelligence systems.

By the end of the course, participants will have a solid understanding of data warehouse concepts and different usages of BI applications.

Target Audience

- Data Scientists and Business Analysts
- SW and data architects
- Database administrators
- Data Engineers and Software development engineers

Prerequisites

- Basic understanding of databases and SQL

Content

- Introduction to BI
  - Definitions and concepts (BI, DWH, storage methods, all-in-one devices)
- Business application of BI systems
  - Specifics of business requirements
  - Impact on the organization,
  - Business analytics (measures, metrics, KPIs, dashboards)
- Architecture and processes
  - Data warehouse components
  - DW architecture
  - DW processes
- BI infrastructure
  - Processes
  - Technologies and people
- BI processes
  - Project management
  - Change management
  - Data quality
  - Data governance
  - Data warehouse administration
  - Metadata management
- BI system development methodology
  - Project phases and activities
  - Roles and responsibilities of team members
- The most common mistakes and best practices
  - Valuable tips to apply to your own BI system implementation projects

Modeling for Data Warehouse

Duration – **1 day**

Overview

Data modeling for a data warehouse is the process of designing the structure and relationships of data entities within a data warehouse environment. It involves creating a conceptual, logical, and physical representation of the data to support efficient data storage, retrieval, and analysis. The data modeling process begins with a thorough understanding of the business requirements and objectives of the data warehouse. This includes identifying the key data entities, their relationships, and the desired analytical outcomes. The data modelling process goes through Conceptual, Logical, and Physical phases understanding business requirements and how to identify dimensions and facts, as One of the key steps in data modeling for a data warehouse where dimensions represent the descriptive attributes used for analysis, while facts are the numeric measures or metrics that provide the basis for analysis.

By the end of the workshop, participants will have a solid understanding of data modeling concepts and different techniques such as Dimensional modeling or Data Vault and will have created a data model based on a sample business case.

Target Audience

- Data architects
- Data mart developers
- Business analysts

Prerequisites

- Basic understanding of databases and SQL

Content

- Overview of data modeling for data warehouse
  - Introduction to Entity-Relationship (ER) modeling and dimensional modeling
  - Key concepts and terminology
  - Business requirements and objectives
  - Techniques for capturing and documenting business requirements
- Creating high-level data entities and relationships
  - Techniques for creating a logical data model
  - Normalization and denormalization techniques
  - Mapping business requirements to data entities and relationships
  - Creating a logical data model based on the conceptual data model
- Introduction to dimensional data modeling
  - Star schema and snowflake schema design techniques
  - Identifying dimensions and facts
  - Creating a dimensional data model based on the logical data model
- Introduction to physical data modeling
  - Techniques for creating a physical data model
  - Mapping the dimensional data model to a specific database technology
  - Indexing, partitioning, and other physical implementation details
- Iterative refinement of the data model based on performance considerations and evolving business requirements
  - Best practices for data modeling for data warehouse

Development

SQL course

Duration – **2 days**

Overview

The two–day course provides a theoretical and practical overview of database introduction and simple SQL queries.

The SQL course introduces students to relational database usage. It consists of 8 units and begins with an introduction to databases and defining basic terms and types of data. The course includes the basics of Data Manipulation Language (DML) and Data Definition Language (DDL). Participants are introduced to the basic scalar and aggregate functions and how to apply them in everyday work. The second day of the seminar discusses the ways of grouping data and mandatory rules to create a union or difference of data sets.

Through the course, participants will learn to use subqueries, the last lesson contains basic DDL commands.

Through the exercises, the understanding of the presented material is checked, and the presence of an instructor facilitates the solution and correction of mistakes.

All materials and exercises are available online. The course could be organized online or onsite. All materials for repeating and practice are available a week after the course is finished. After each section, students have a quiz to check their knowledge. After completing the course there is a link for outside checking knowledge.

Target audience

- Data Analysts
- Business Analysts
- Data Engineers

Prerequisites

No prerequisites are needed, elementary knowledge of databases and programming concepts is sufficient.

Content

Day 1

- Introduction
- Simple SQL queries
- Retrieving data from multiple tables
- Scalar and arithmetic functions

Day 2

- Aggregate functions and grouping
- Retrieving data from multiple data sets
- Subqueries
- Data Editing

Day 3 Data-in-motion – managing data in real-time

- What is real-time, streaming, and sensor data?
- Lambda architecture – how to adapt batch and real-time analytics?
- General architecture for big data solutions (data-in-motion overview)
- Exercise: Building an application for real-time processing

Advanced SQL course

Duration – **2 days**

Overview

Advanced SQL course teaches how to use advanced SQL techniques to access a database. This seminar is intended for students who have a basic knowledge of SQL or have completed a basic SQL course.

Target Audience

- Data Analysts
- Business Analysts
- Data Engineers
- SW Engineers

Prerequisites

- Basic understanding of databases and SQL (e.g. our Basic SQL course)

Content

Day 1

- Overview of the basics of SQL and OLAP functions
- Creating objects
- Join tables
- CASE, CAST, summary (MQT) and provisional tables

Day 2

- Subqueries
- Scalar functions
- Tabular expressions and recursive SQL
- UDT, UDF and performance

PL/SQL Workshop

Duration – **5 days**

Overview

The goal of this course is to become familiar with widely used PL/SQL concepts, such as anonymous blocks, conditional execution, and loops, usage of procedures and functions, cursors, collections, defining exceptions, understanding of transactions, creation of triggers, usage of packages, writing dynamic SQL, performance tuning and code debugging through SQL developer.

Target Audience

- Data Analysts
- Business Analysts
- Data Engineers
- SW Engineers

Prerequisites

- Basic programming concepts
- Basic database concepts
- Basic SQL queries

Content

Day 1

- SQL recapitulation
- Introduction to PL/SQL
- Anonymous blocks
- PL/SQL data types

Day 2

- SQL in PL/SQL
- Loops
- Conditional execution
- Sequences
- Triggers

Day 3

- Cursors
- Exceptions
- Procedures and functions

Day 4

- Local subprograms
- Packages
- Transactions

Day 5

- Dynamic SQL
- Debugging SQL Developer
- Performance Tuning
- Hints

Apache Kafka

Duration – **1-5 days**

Overview

Apache Kafka is one of the **most popular distributed streaming platforms** today. Some of the biggest companies in the world use Kafka for their streaming data pipelines, streaming analytics, data integration, or communication between microservices. Kafka is a distributed system that scales easily to hundreds or even thousands of servers. It's fast and reliable.

Integrating other systems with Kafka is not a problem. Applications easily integrate with Kafka through its Producer and Consumer API. If you want to connect to other system like RDBMS, Elasticsearch, or Hadoop, Kafka Connect comes to the rescue. You can even develop streaming applications on Kafka using Kafka Streams. **All of that and more can be learned in our new Apache Kafka course.**

One module = one day. The course is divided into 5 modules, allowing you to choose modules depending on your previous knowledge and preferences – whether it's application development with Kafka or maybe installation, configuration, and security. Every module is followed by examples and exercises based on **real-life scenarios**. Programming exercises will be in Java, so a basic knowledge of Java programming language is required.

Target Audience

- Data Engineers
- SW Engineers

Prerequisites

- Programming exercises will be in Java, so a basic knowledge of Java programming language is required.

Content

Day 1

- Introduction to Kafka
- Introduction to Confluent Platform
- Consumer and Producer API
- Using Schema Registry with Kafka and Avro

Day 2

- Introduction to Kafka Connect
- Kafka Connect examples of sink and source connectors
- Integrating Kafka Connect with Schema Registry

Day 3

- Introduction to Kafka Streams
- Types of transformations in Kafka Streams
- Developing and deploying Kafka Streams applications
- Introduction to KsqlDB

Day 4

- Kafka security
- Encryption
- Authentication
- Authorization
- Details around Kafka Connect

Day 5

- Administration and monitoring
- Advanced Kafka topics and best practices
- Multicluster deployments
- Cluster planning and sizing

Today, Apache Kafka is everywhere and the de facto standard when you need to exchange messages in real time. We are using Apache Kafka on our various integration projects during application development, i.e., Fraud detection systems or modern data architectures such as Data Lake. After this course, you will be able to successfully use Apache Kafka in your environment.

Introduction to Apache Spark

Duration – **2 days**

Overview

Apache Spark is a framework for fast processing a large amount of data. It is extremely useful for system architects, development engineers, and business analysts, due to the ability to use it for any kind of data processing in any kind of environment. Apache Spark is a very popular system, often used for advanced analytics, data science, and modern BigData architecture, as well as for complex batch (ETL) processing and for processing in real-time.

Spark contains a few key components such as Spark SQL for data structuring, Spark Streaming for processing a large amount of data in real-time, Spark MLib for machine learning, Spark GraphX for graph processing, and SparkR for statistical data processing using R language.

Spark can be started by itself, on a YARN (Hadoop) cluster, or in a Mesos environment, so it can start in any environment. Spark is a polyglot framework, which means that it abstracts its usage to the maximum, and it imposes using a programming language (Python, Java, Scala, R), to the development environment, which fits the organization or the business type the best. All the examples in this education will be primarily processed in Python, but other program languages, e.g., Scala, will also be used. The exercise will be done in an independent and cluster environment, depending on the assignment the participants will be working on.

Target Audience

- System architects
- Development engineers
- Business analysts.

Prerequisites

- Basic knowledge of Python
- Knowledge of OO programming
- Advanced knowledge of the SQL language

Content

The participants will get a brief introduction to Spark at the course, as well as a basic explanation of how Spark functions, and will, through interactive examples, go through an advanced analytics assignment and work on the target DataSet from the big data set download to the final visualization.

Apache Spark - Advanced usage

Duration – **2 days**

Overview

The course is aimed at the participants who want to advance their knowledge in the Spark environment, such as Spark Streaming. All the examples in this education will be primarily processed in Python, but other programming languages, e.g. Scala, will also be used. The exercise will be done in an independent and cluster environment, depending on the assignment the participants will be working on.

Target Audience

- System architects
- Development engineers
- Business analysts.

Prerequisites

- Basic knowledge of Python
- Knowledge of OO programming
- Advanced knowledge of the SQL language

Content

The participants will get all the necessary info about how to establish a streaming process for data processing in real-time. They will learn about the MLib library for machine learning, where they will build a model for machine learning and a process of model training will be shown to them as well. By using the GraphX library for processing graph databases through a few examples, we will show how to use it efficiently in practice.

Machine Learning Operations (MLOps) - ML Lifecycle

Duration – **2 days**

Overview

This course focuses on the implementation and management of machine learning models throughout their lifecycle. It provides ML engineers, data scientists, and data engineers with a comprehensive understanding of MLOps concepts, tools, and best practices. The course covers key stages of the ML lifecycle, including data versioning, feature storing, model governance, deployment, scaling, optimization, and monitoring. Participants will learn how to automate processes, establish feedback loops, and ensure continuous integration, continuous delivery (CI/CD), and retraining of models.

Target Audience

- ML engineers
- Data scientists
- Data engineers

Prerequisites

- Basic knowledge of Python
- Familiarity with machine learning concepts and ML in Python

Content

Day 1

- Introduction to MLOps
  - Overview of MLOps and its importance
  - Key components and challenges in the ML lifecycle
  - Roles and responsibilities of stakeholders
- ML Lifecycle and Steps
  - Overview of the ML lifecycle stages
  - Data collection, preprocessing, and labeling
  - Model training, validation, and evaluation
  - Model deployment, monitoring, and maintenance
- Data Versioning
  - Importance of data versioning in ML projects
  - Techniques for managing data versioning
  - Tools and platforms for data versioning
- Feature Store
  - Introduction to feature stores and their benefits
  - Building and managing feature stores
  - Integration of feature stores with ML pipelines
- Model Governance and Experiment Management
  - Best practices for managing ML experiments
  - Version control and tracking of models
  - Model governance

Day 2

- Model Deployment
    - Strategies for deploying ML models
    - Containerization and orchestration of ML models
- Model Scaling and Optimization
    - Techniques for scaling ML models
    - Optimization methods for improving model performance
- Monitoring ML Models
    - Importance of model monitoring in production
    - Health checks and performance metrics
    - Detection of data and model drift
- Feedback Loop and Continuous Improvement
    - Establishing a feedback loop for model updates
    - Incorporating user feedback into model retraining
- Automation and Retraining
    - Automating ML workflows and pipelines
    - Continuous retraining of models
- Integration of MLOps with CI/CD processes

Introduction to R programming language

Duration – **2 days**

Overview

In the course, participants learn about objects in R, defining and using variables and functions, basic data types, vectors, and operations on vectors. R is a specific programming language designed for data processing, statistical computing, and data visualization. Besides various statistical and visualization packages it includes machine learning and advanced data processing packages. The goal of this course is to give an introduction to the basic concepts of R programming language. The course is held entirely through interactive R notebooks that give course attendees a unique hands-on experience. In the course, attendees learn about objects in R, defining and using variables and functions, basic data types, vectors, and operations on vectors. R is famous for its visualization capabilities, especially using the ggplot package which is learned in the course. R is primarily designed as a statistical language; therefore it has a wide range of statistical and machine-learning packages that are taught in the course. Entire course lectures and exercises are done in RStudio, one of the most popular development environments for R.

Target Audience

- Business Analysts
- Data Scientists
- Data analysts
- Software Engineers

Prerequisites

- Knowledge of basic programming concepts such as variables, functions, and loops is desirable

Content

Day 1

- R basics
  - Objects, variables
  - Functions
  - Data types
  - Vectors and arithmetic with vectors
  - Basic data wrangling
  - Programming in R basics
- Data visualization
  - Introduction to data visualization
  - Data distribution, quantiles, boxplots
  - Plotting data with ggplot
  - Summarizing data with dplyr
  - Data visualization principles

Day 2

- Data wrangling
  - Tidy data – cleaning and preparing data
  - String processing
  - Working with dates and times
- Introduction to statistics
  - Basics of statistics, probability
  - Inference and Modeling
- Machine learning
  - Classification and regression
  - Model evaluation