kod szkolenia: DBX-PFE / ENG DL 1d

Databricks Data Transformation

The Databricks Transformation training is the final step in the structured training path Fundamental → Explorer → Lakehouse → Transformation. Participants will learn how to design modular pipelines, combine batch and streaming, apply advanced transformations, use Delta Live Tables, and integrate workflows with Git and CI/CD.

Training recipients

The training is designed for data engineers and DataOps teams responsible for implementing and maintaining production data processing workflows in the Lakehouse architecture.

Benefits

– can design modular Silver → Gold pipelines
– understand how to combine batch and stream in a single workflow
– can apply PySpark window functions for data transformations
– can use Delta Live Tables to automate pipelines
– know best practices for orchestration and CI/CD in Databricks
– can ensure data quality with expectations and monitor lineage
– are prepared to maintain production workflows in Databricks

Training program

1.Data processing architecture

- Recap of Bronze–Silver–Gold in the context of transformation pipelines

- Designing data flows in Silver and Gold layers
- Modularity and separation of processing logic (load → transform → save)

2.Batch and stream load in practice

- Differences between batch and streaming processing
- Batch ingest using COPY INTO and writing to Delta tables
- Streaming ingest with Auto Loader (cloudFiles)
- Structured Streaming: readStream, writeStream, checkpointing, and fault tolerance
- Integrating batch and stream

3.Advanced data transformations

- Creating numerical, text, and binary features
- Logical transformations (case when, when, otherwise)
- Window functions (lag, lead, row_number, rolling average)
- Creating time-based and session features

4.Delta Live Tables – pipeline automation

- Declarative processing approach: CREATE LIVE TABLE
- Creating DAGs and scheduling in DLT
- Integrating DLT with Auto Loader and Structured Streaming
- Expectations – real-time data quality control
- Monitoring and lineage in the DLT interface

5.Orchestration and automation

- Databricks Workflows – multi-task jobs, dependencies, retries
- Pipeline parameterization (dbutils.widgets, dbutils.notebook.run)
- Best practices for CI/CD and code maintenance (Repos, versioning notebooks)

6.CI/CD – practical Git (Repos) demo

- Cloning a repository in Databricks Repos
- Commit and push notebooks to Git
- Running a pipeline from Workflows based on a repo
- DevOps best practices for Databricks

7.Final project

- Design and run a Silver → Gold pipeline using batch and stream load, Delta Live Tables, quality control rules, and Git integration

Expected preparation of the participant

– Completion of Databricks Lakehouse or equivalent knowledge
– Knowledge of SQL and PySpark
– Basic experience implementing data pipelines

Training Includes

- access to Altkom Akademia student

Training method:

The training is conducted in the Databricks cloud environment. Each participant receives their own workspace with access to Unity Catalog, SQL Editor, Notebooks, and a catalog with test data.

Czas trwania

1 dni / 7 godzin

Language

Training: English

- Materials: English