kod szkolenia: DBX-EXP / ENG DL 1d

Databricks (Explorer) Data Exploration

The Databricks Explorer training is the second step in the structured training path Fundamental → Explorer → Lakehouse → Transformation. Participants dive deeper into data analysis and exploration using SQL and PySpark. They will learn to combine declarative (SQL) and imperative (PySpark) approaches, perform aggregations, joins, data profiling, and present insights in interactive Databricks notebooks.

Training recipients

The training is designed for data engineers, analysts, and BI specialists who want to explore and analyze data in Databricks using SQL and PySpark. It is the natural next step after completing Databricks Fundamentals.

Benefits

– understand the differences between SQL and the PySpark DataFrame API
– can load and explore data from multiple sources
– are able to perform aggregations, joins, and logical transformations
– can profile data quality and detect missing values, outliers, and inconsistencies
– understand the trade-offs of different file formats (CSV, JSON, Parquet, Delta)
– can create visualizations and simple dashboards in Databricks
– are prepared for the next stage of the training path – Databricks Lakehouse

Training program

1.Data analysis in the Databricks environment
- Environment recap: Unity Catalog, notebooks, SQL Editor
- Loading data from different sources: CSV, JSON, Parquet, Delta
- Creating tables and views in Unity Catalog
- Data exploration: display(), show(), summary(), describe()

2.SQL and PySpark in data analysis
- Differences between SQL and DataFrame API
- Creating SQL queries in Databricks notebooks
- Combining SQL and PySpark in a single analysis (mixed cells)
- Common mistakes and query optimization techniques

3.Data operations and transformations
- Filtering (WHERE, filter) and sorting (ORDER BY, sort)
- Grouping and aggregations (GROUP BY, agg, count, sum, avg)
- Creating and modifying columns (withColumn, alias)
- Joins: INNER, LEFT, RIGHT, FULL, SEMI, ANTI
- Logical transformations (case when, when, otherwise)

4.Data quality analysis and profiling
- Analyzing missing values, uniqueness, and data types (na, distinct, count)
- Data validation and typing (cast, printSchema)
- AI Functions – assisting in data analysis and cleansing
- Generating statistics and column profiling

5.File formats trade-offs (light)
- CSV, JSON, Parquet, Delta – differences and use cases
- Delta vs Parquet: ACID, schema evolution, time travel
- Costs and performance of each format

6.Visualization and presentation of results
- Creating charts and dashboards in the Databricks GUI
- Comparing distributions and key values
- Documenting results and insights in notebooks

7.Final project
- Prepare a data quality analysis and transformation combining SQL and PySpark, supplemented with a simple visualization of results in a Databricks notebook

Expected preparation of the participant

– Completion of Databricks Fundamentals or equivalent knowledge
– Basic SQL knowledge
– Basic experience working with data

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