

Databricks: Fundamentals – the platform and first data products

Databricks Fundamentals is a hands-on introduction to working with the Databricks Data Intelligence Platform. The program focuses on moving from a raw file to an analysis-ready table in a controlled way, using Unity Catalog, Volumes, SQL and Delta tables.

Participants go through the whole basic workflow: organizing data in Unity Catalog, loading a file into a Volume, writing the first transformations in SQL, saving the result as a Delta table and running it as a Lakeflow Job. The exercises are based on a single RetailOps business scenario (a fictional retail company whose data we use for all exercises), so by the end of the day each participant has a complete, repeatable data flow built from start to finish.

The training is run hands-on. The trainer shows every step on screen and participants do it together with the trainer in their own notebooks, so nobody is left alone with a new tool. At the end of the day, in a 45-60-minute workshop, participants combine all the pieces on their own into one working flow.



Training recipients

The training is intended for people who are starting to work with Databricks and want to understand how the platform is built and how to work with data in it.

The course is particularly suitable for:

- new members of Data, BI and IT teams who will be using Databricks,
- data analysts moving from a data warehouse or BI tools to a Lakehouse,
- developers and engineers starting projects on Databricks,

- technical staff who need a shared vocabulary before further courses in the Databricks path.
- The training will be especially valuable for teams that want to follow current platform standards (Unity Catalog, Volumes, serverless compute) from day one.



Benefits

- Understanding the Databricks platform – you will learn the roles of the workspace, compute and Unity Catalog and how these parts work together.
- Organizing data in Unity Catalog – you will learn to navigate the catalog → schema → table/Volume hierarchy and understand who has access to what.
- Working with files in Volumes – you will learn how to load a file into a Volume, read it and check its schema.
- SQL transformations – you will learn to filter, join and aggregate data and to create derived columns.
- Delta tables – you will see how to save a result as a Delta table, document it and use its change history.
- A repeatable data flow – you will build a minimal file → Bronze table → analytical dataset flow and run it as a Lakeflow Job.
- Choosing a data ingestion tool – you will learn when a file is enough, when Lakeflow Connect is worth using and when to use Lakeflow Jobs.



Training program

. Module – Databricks Data Intelligence Platform

Module goal: the participant knows what the platform is made of and where in it they work.

- Platform architecture – what runs on the Databricks side, what runs in the customer's cloud, and where the data physically lives
- Workspace – where the team works: notebooks for code, SQL editor for queries, Catalog Explorer for browsing data, the Genie Code AI assistant (formerly Databricks Assistant) for help with writing and fixing code
- Compute, that is processing power – how serverless (started on demand, with no setup) differs from clusters and SQL warehouses, and how that affects cost
- Notebooks in practice – combining SQL and Python in one document, displaying results, parameters (display(), %sql, widgets)
- Git folders – working with a code repository directly from the workspace (overview)
- Hands-on with the trainer: environment check – running the first notebook and accessing the data

2. Module – Unity Catalog: data organization and permissions

Module goal: the participant understands how data is organized in Databricks and how access to it is

controlled.

- Data hierarchy – catalog → schema → table, view or Volume; compared to a database and folders
- Full object names (catalog.schema.table) – how to refer to another team's data
- Managed and external tables – who manages the table's files and what happens when the table is dropped
- Permissions – granting and revoking access with GRANT / REVOKE, the object owner
- Finding data and its origin – the Discover page, where data is grouped by business area (domains), and lineage: which tables a given table was built from and who uses it
- Hands-on with the trainer: the participant creates their own schema; the trainer shows how to grant permissions to other people

3. Module – Volumes and the first data load

Module goal: the participant can store files in Databricks and load them for analysis.

- Volumes – a controlled place for files (CSV, JSON, PDF) covered by Unity Catalog permissions
- DBFS and mounts – the older way of working with files; how to recognize it in existing code and why it is being phased out
- Uploading a file through the UI and browsing the contents of a Volume
- Reading CSV and JSON files in SQL (read_files()) and in Python (spark.read)
- Data schema – automatic detection of column types vs. a manually defined schema; common problems: delimiter, headers, empty values
- Hands-on with the trainer: the participant uploads a RetailOps orders file and checks whether the column types were detected correctly

4. Module – SQL and basic transformations

Module goal: the participant can prepare data for analysis using SQL.

- Exploring data – selecting columns, filtering, sorting
- Joining tables (JOIN) – for example, orders with customer and product data
- Aggregations – sums, averages and counts broken down by group (GROUP BY)
- Calculated columns – conditions (CASE WHEN), type conversion, operations on dates and text
- Missing values and duplicates – how to find and handle them
- Views – saving a query under a name so others can use it
- Hands-on with the trainer: the participant joins orders with customers and products and calculates sales by category and month

5. Module – Delta Lake: the table, its history and quality

Module goal: the participant saves the result of their work as a Delta table and knows how to use it safely.

- What a Delta table is – the table format in Databricks that guarantees consistent writes and keeps a history of changes
- Saving a query result as a new table and appending data
- Table and column descriptions – documentation for other data users
- Change history and going back to a previous version of the table (time travel, RESTORE) – for example, after a faulty update
- Schema protection – what happens when someone tries to write data in the wrong format

- Simple quality rules – required columns and conditions that values must meet
- Hands-on with the trainer: the participant saves a table, deliberately introduces an error and restores the previous version

6. Module – Lakeflow Jobs and Lakeflow Connect

Module goal: the participant can run a prepared process automatically and knows the options for bringing data in.

- Lakeflow Jobs – running a notebook on a schedule, with a notification when it fails
- Run history – how to check whether a process succeeded and what went wrong
- Lakeflow Connect – ready-made connectors to source systems (SaaS applications, databases, files in SharePoint and Google Drive) and their requirements (overview)
- Choosing how to bring data in – when a file is enough, when to use a connector, and when a scheduled job
- Hands-on with the trainer: creating a job from a notebook, running it manually and reviewing the run history (participants set the schedule on their own in the workshop)

7. Module – End-to-end workshop and good practices

Module goal: the participant puts all the pieces of the day together into one working flow on their own.

- Independent workshop (45–60 min): without step-by-step guidance, participants combine what they did with the trainer in modules 1–6: file in a Volume → raw data table → cleaned dataset for analysis → scheduled job. The trainer is available for questions and reviews the solutions at the end. The outcome is a complete, repeatable data flow with documented tables.
- Cost control – automatic shutdown of compute, choosing between serverless and a cluster, viewing usage
- Everyday security – least privilege, no passwords in code
- Recommended next steps depending on the participant's role



Expected preparation of the participant

- No formal prerequisites
- A basic understanding of tables and SQL fundamentals (SELECT, WHERE) is helpful
- Knowledge of Python or PySpark is not required

To increase the comfort of work and training's effectiveness we suggest using an additional monitor. The lack of additional monitor does not exclude participation in the training, however, it significantly influences the comfort of work during classes.

You can find information and requirements of participation in Distance Learning trainings at:

<https://www.altkomakademia.pl/en/how-it-works/about-courses/>



Training Includes

Training includes

- access to the Altkom Akademia student portal
- a prepared Databricks environment with Unity Catalog and RetailOps scenario data for the duration of the training
- notebooks covering all exercises done with the trainer, plus the workshop task, in electronic form

Note: During selected training sessions, the Academy provides Participants with access to paid tools and services necessary to complete the training program (in particular Azure, Fabric, and other cloud resources). The Participant is required to use the provided resources solely within the scope of the training agenda and in accordance with the trainer's instructions. Use of these resources beyond the training scope—particularly for private projects, testing additional services, modifying the environment configuration, or undertaking activities that generate costs unrelated to the exercises—will result in the Participant being charged for costs exceeding the allocated limit, based on expenses incurred by the Academy.



Duration

1 days / 7 hours

Language

- Training: English
- Materials: English