

training code: DBX-DEA / ENG DL 3d / EN

Databricks: Data Engineer Associate – certification preparation

Databricks: Data Engineer Associate is an intensive three-day preparation for the Databricks Certified Data Engineer Associate exam, **delivered only as a dedicated training for a single organization**. The program takes participants through the full skill set of a data engineer on the Databricks platform: from the platform, Unity Catalog and Delta Lake, through ingestion, the Medallion architecture and optimization, to Lakeflow and CI/CD.

Participants cover all areas of the current exam guide, combining theory with hands-on labs and a quiz after each day. The program also covers Spark SQL and PySpark to the extent required by the exam. Day three ends with a practice exam and a review of knowledge gaps, so each participant knows what to work on before taking the certification exam. The SQL and Python scope, the cloud and the examples are adapted to the client's team before the training.



Training recipients

The training is intended for teams that work with Databricks and want to confirm their skills with the Databricks Certified Data Engineer Associate certification.

The course is particularly suitable for:

- data engineers with hands-on Databricks experience (at least 6 months recommended),
- ETL developers and technical analysts moving into a data engineer role,
- teams for which certification is a development goal or a partner requirement,
- organizations that want to align their team's knowledge before a larger Databricks project.

The training will be especially valuable for teams that want to prepare for the exam using examples close to their own projects.



Benefits

- Platform and Unity Catalog – you will learn how to choose compute, the developer tools and access management: permissions, row-level security, column masking and ABAC.
- Spark SQL and PySpark – you will reinforce transformations on complex data, joins, aggregations and tuning basics.
- Delta Lake – you will learn to use MERGE, table history, liquid clustering and predictive optimization.
- Ingestion and Medallion architecture – you will build incremental loading with COPY INTO and Auto Loader and learn Lakeflow Connect.
- Lakeflow – you will learn Lakeflow pipelines (Spark Declarative Pipelines, formerly Delta Live Tables) and Lakeflow Jobs to the extent required by the exam.
- CI/CD – you will learn how to work with Git and Declarative Automation Bundles: dev, test and prod environments, variables, overrides and deployment.
- Exam preparation – you will take a practice exam and receive an individual plan for further study.



Training program

Day 1 – Platform, Unity Catalog, Spark and Delta Lake

Leading theme: “We understand the platform and the data”

1. Module – The Databricks platform and developer tools

Module goal: the participant knows the parts of the platform and can choose the right compute for a task.

- Platform architecture – what runs on the Databricks side and what runs in the customer's cloud
- Types of compute – serverless, interactive clusters, job clusters, SQL warehouses; when to use each and what it costs
- Working tools – notebooks, SQL editor, the Genie Code AI assistant (formerly Databricks Assistant), Git folders
- Debugging code in a notebook and working from a local IDE (Databricks Connect)
- Demo: comparing startup time and the DBU rate for serverless and a cluster

2. Module – Unity Catalog and access management

Module goal: the participant can organize data in Unity Catalog and control access to it.

- Object hierarchy – catalog → schema → table, view, Volume, function
- Managed and external tables, cloud storage locations for data, and Volumes for files
- Changing the table type (managed ↔ external conversion); permissions for users, groups and technical accounts (service principals) in the UI and in SQL
- Permissions – granting and revoking access (GRANT, REVOKE), the object owner; blocking access: in

Unity Catalog a privilege that was not granted means no access, the DENY command works only in the legacy Hive metastore, and selected blocks are handled by ABAC DENY policies (Beta)

- Limiting data visibility – row filters, column masks and tag-based rules (ABAC)
- Lineage, tags and auditing – who uses the data and where it comes from
- Lab: the participant sets up permissions, a row filter and a column mask on their own schema

3. Module – Spark SQL and PySpark

Module goal: the participant transforms data fluently in SQL and Python, to the extent required for the exam.

- Reading data from CSV, JSON and Parquet files and from Delta tables
- Data schema – automatic detection vs. a manually defined schema; data preview and statistics
- Basic operations – selecting columns, filtering, calculated columns, grouping, joining tables
- Combining and summarizing data – broadcast and cross joins, UNION / UNION ALL, approximate distinct counts (approx_count_distinct), summary()
- Complex data – lists, structures and JSON; unpacking them into regular columns
- Cleaning – duplicates, missing values, type conversion, conditions; a short look at window functions
- Basic performance settings – the number of partitions after a shuffle, the broadcast join threshold, driver and executor memory (spark.sql.shuffle.partitions, spark.sql.autoBroadcastJoinThreshold)
- Combining SQL and Python in one notebook (temporary views)
- Lab: the participant cleans and enriches a raw orders file in SQL and PySpark

4. Module – Delta Lake

Module goal: the participant knows the Delta table operations covered by the exam.

- How a Delta table works – the transaction log, consistent writes, schema enforcement and schema changes
- Creating tables and modifying data – inserting, updating, deleting, MERGE INTO
- Change history, reading previous versions and restoring a table
- Performance – liquid clustering, merging small files (OPTIMIZE), removing old versions (VACUUM), statistics
- Automatic table maintenance (predictive optimization)
- Lab: the participant updates the orders table, reverts the change and reviews the history

5. Module – Day 1 end-to-end lab and quiz

Module goal: the participant brings the day's topics together and checks their knowledge.

- Lab: raw files → cleaning in PySpark → Delta table in Unity Catalog with permissions → review of the change history. The outcome is a repeatable Delta table and a quiz score on the Day 1 topics.

Day 2 – Data loading, Medallion, incremental processing and optimization

Leading theme: “We build the Medallion”

6. Module – Medallion architecture and choosing a loading method

Module goal: the participant knows how to split data into layers and how to match the loading method to the source.

- Bronze, Silver and Gold layers – what goes into which layer
- Process repeatability and the business key – how to avoid duplicates when a process is rerun
- Types of sources – files, databases, SaaS applications, streams; which mechanism fits which

7. Module – COPY INTO and Auto Loader

Module goal: the participant loads files incrementally and handles changes in their structure.

- COPY INTO – loading files while skipping those already loaded
- Auto Loader – automatic detection of new files and tracking of progress (checkpoint)
- Auto Loader modes: directory listing vs. file notifications; Databricks recommends notifications based on file events
- Changes in file structure – new columns, a column for values that do not fit (`_rescued_data`), type hints
- Lab: the participant loads successive batches of files into Bronze, including a batch with a new column

8. Module – Lakeflow Connect and other sources

Module goal: the participant knows the ways to pull data from databases, applications and APIs.

- Lakeflow Connect – ready-made connectors to databases and SaaS applications
- Lakeflow Connect connectors: standard, managed (SaaS applications, databases with CDC, files in SharePoint and Google Drive) and partner – selection criteria
- Uploading files from a computer through the UI (local file upload)
- Databases (JDBC, ODBC) and REST APIs as data sources
- Nested data from JSON and APIs – converting it into a flat table
- Lab: the participant reads JSON data from an API (a prepared file) and turns it into a table

9. Module – Structured Streaming: incremental processing

Module goal: the participant processes data incrementally with progress tracking.

- Stream processing – data processed in batches as it arrives, with progress tracked
- Run modes – a one-time pass over all new data (`availableNow`) or scheduled runs
- Lab: the participant builds a Bronze → Silver stream with duplicate removal

10. Module – Monitoring and optimization

Module goal: the participant can find and remove the cause of slow processing.

- Spark UI – how to read a job's execution: stages, data transfer between nodes, out-of-memory errors
- Common problems – uneven data distribution (skew), moving data between nodes (shuffle), writing to disk (spill), out-of-memory errors, cluster start failures, library conflicts
- Query profile and cost in system tables
- Lab: the participant diagnoses a slow query and fixes it

11. Module – Day 2 end-to-end lab

Module goal: the participant builds a complete Medallion architecture.

- Lab: a Bronze → Silver → Gold flow with Auto Loader, stream processing, transformations and liquid clustering for the client's scenario. The outcome is a working incremental flow reviewed with the trainer.

Day 3 – Lakeflow, CI/CD and practice exam

Leading theme: “Production, CI/CD and certification”

12. Module – Lakeflow pipelines

Module goal: the participant builds a declarative pipeline with data quality checks.

- The declarative approach – we describe the tables, and the platform works out the order and

incremental updates

- Streaming tables and materialized views – when to use which
- Gold layer objects for BI: views, materialized views, streaming tables and tables – when to use which
- Quality rules (expectations) – warn, drop the record or stop the pipeline
- Auto CDC – applying changes from the source with or without history (SCD type 1 and 2)
- Pipeline event log, dependency graph, refresh modes and pipeline parameters (the same code for different catalogs and schemas)
- Lab: the participant builds a declarative pipeline with quality rules

13. Module – Lakeflow Jobs

Module goal: the participant schedules and monitors processes made up of many tasks.

- Task types: notebook, SQL query, dashboard, pipeline
- Tasks and the dependencies between them, conditions and loops
- Parameters, retries, timeouts and notifications
- Triggers – a schedule, a new file, a table change, continuous mode
- Choosing a trigger: time-based or data-driven
- Run history – comparing run times with baselines, spotting blocking tasks; rerunning only the failed tasks and diagnosing errors
- Lab: the participant builds a job that runs a pipeline and a validation task with an alert

14. Module – Git, CI/CD and Declarative Automation Bundles

Module goal: the participant deploys a solution from the repository to different environments.

- Working with Git in Databricks (Git folders) – branches, committing and pushing changes (commit, push), pull requests
- Declarative Automation Bundles – jobs and pipelines described as code, dev, test and prod environments, variables and overrides, Databricks CLI
- Validating, deploying and running a bundle (databricks bundle validate, deploy, run); deployment also from the workspace, without installing the CLI
- Automated deployments in GitHub Actions or Azure DevOps (overview)
- Lab: the participant creates a bundle with dev and prod environments and deploys it to dev

15. Module – Practice exam and preparation plan

Module goal: the participant checks their readiness for the exam and knows what still needs work.

- Exam areas according to the guide in effect from May 4, 2026, and their weights: platform 6%, data ingestion 21%, data transformation 22%, Lakeflow Jobs 16%, CI/CD 10%, troubleshooting 10%, governance 15%
- Exam strategy – eliminating answers, flagging questions to come back to, managing time (10 min)
- Workshop: practice exam – 45 questions, 90 minutes, in English, in a format close to the certification exam
- Review of the answers and the most common pitfalls (40 min)
- An individual exam preparation plan (10 min)



Expected preparation of the participant

- Basic knowledge of SQL and Python or PySpark
- At least 6 months of hands-on Databricks experience recommended (in line with the exam guide); people with less experience will need additional study after the course
- The training is not intended for people without technical experience

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

- access to the Altkom Akademia student portal
- notebooks and exercise materials in electronic form
- quizzes after each day and a practice exam with a review
- a prepared Databricks environment for the duration of the training
- agreement on the group level, cloud and scope of examples before the training

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.



Language

- Training: English

- Materials: English

Duration

3 days / 21 hours

Examination description

The training does not include a formal certification exam; the practice exam is part of the program. The training prepares participants for the Databricks Certified Data Engineer Associate exam; the program follows the exam guide in effect from May 4, 2026 and is checked again before each delivery. Participants take the certification exam on their own after the course.

Exam: 45 single-choice questions, 90 minutes, fee of USD 200 plus applicable taxes; available in English, Japanese, Portuguese and Korean (not in Polish).

DBX-DEA prepares for a cloud-agnostic Databricks certification; for teams working in Azure that need a Microsoft certification, we recommend the DP-750 training.