

www.advancinganalytics.co.uk

Databricks FOR THE MIDDLE ISLE



@ADVANCINGANALYTICS



@ADVALYTICSUK



/ADVANCING ANALYTICS

DATA:Scotland 2022

is proudly supported by



THE DATA LAB
value from data



**ADVANCING
ANALYTICS**



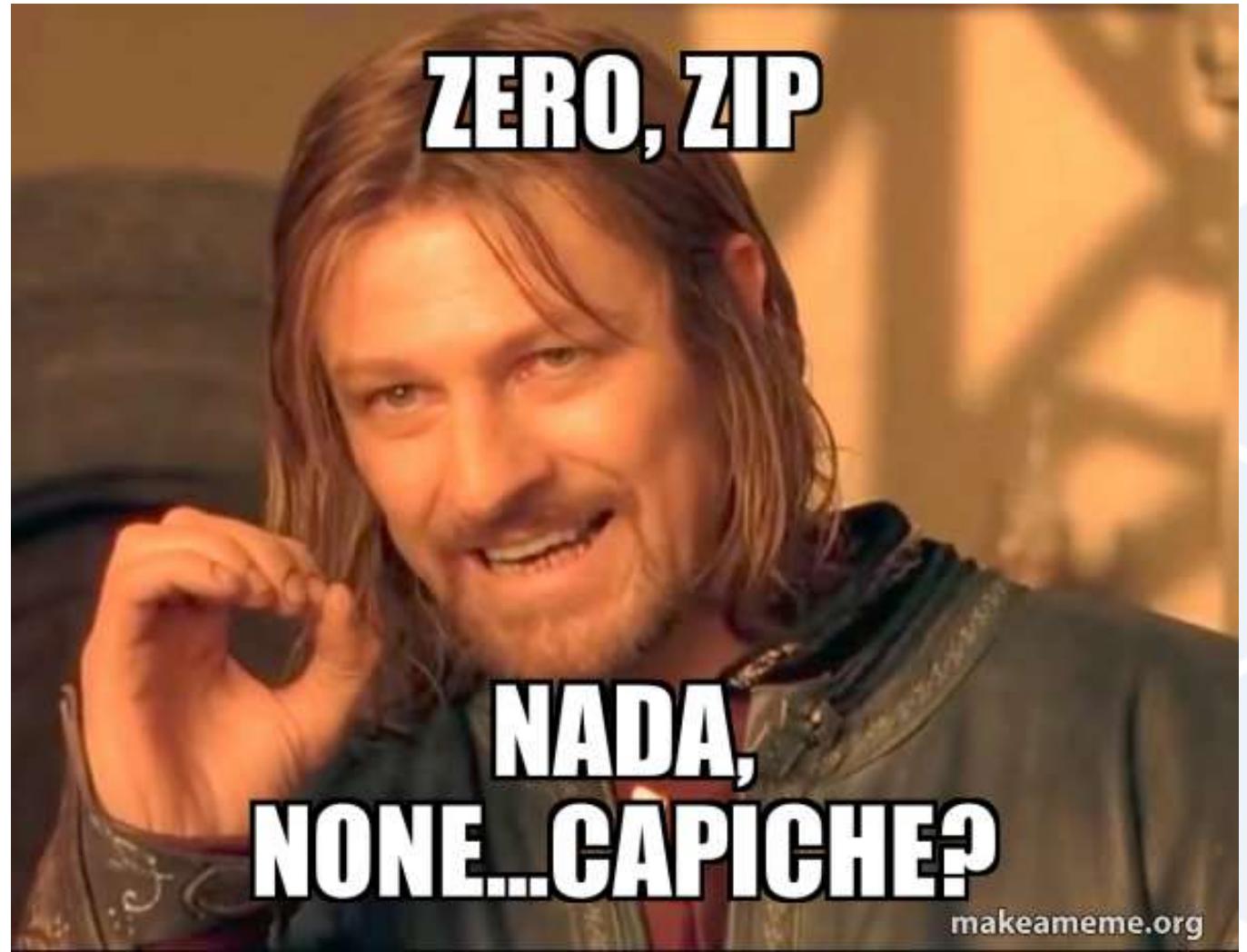
DATAmasterminds

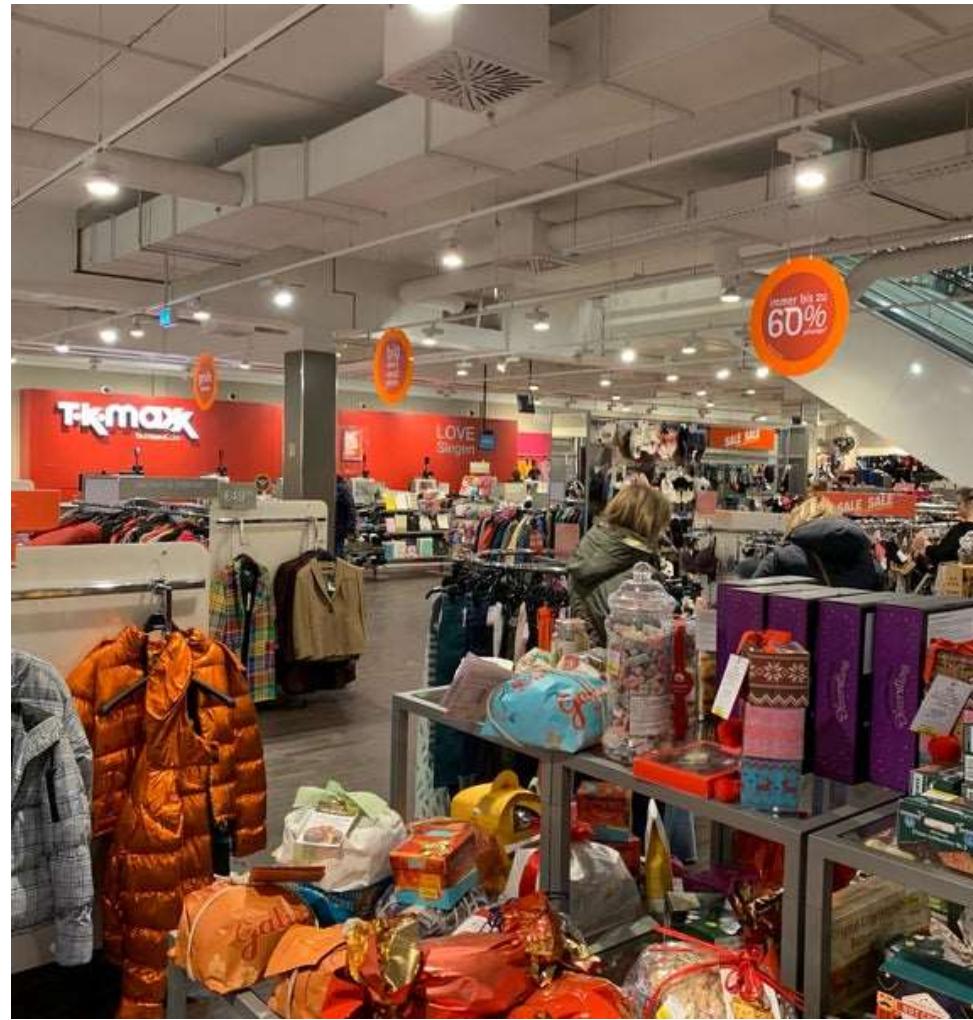




ADVANCING
ANALYTICS

ADVANCING
ANALYTICS





ADVANCING
ANALYTICS







ADVANCING
ANALYTICS



WHY



WHY - OPPORTUNITY

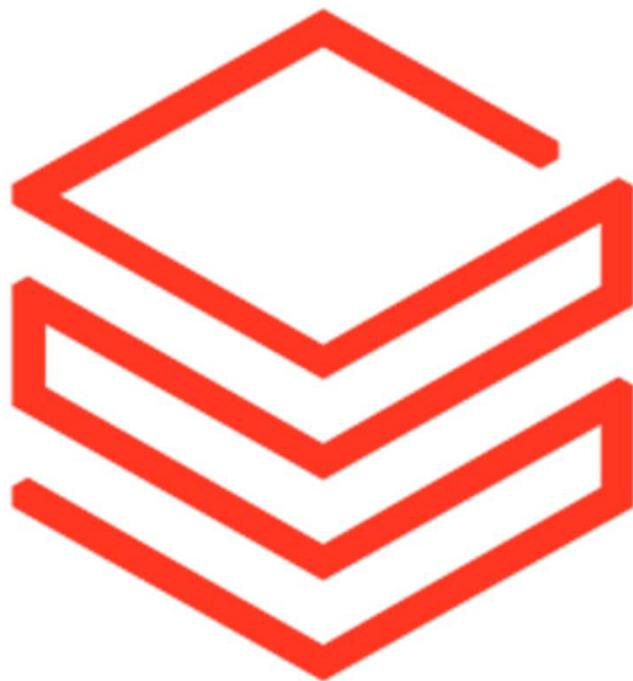




LEARN BY DOING

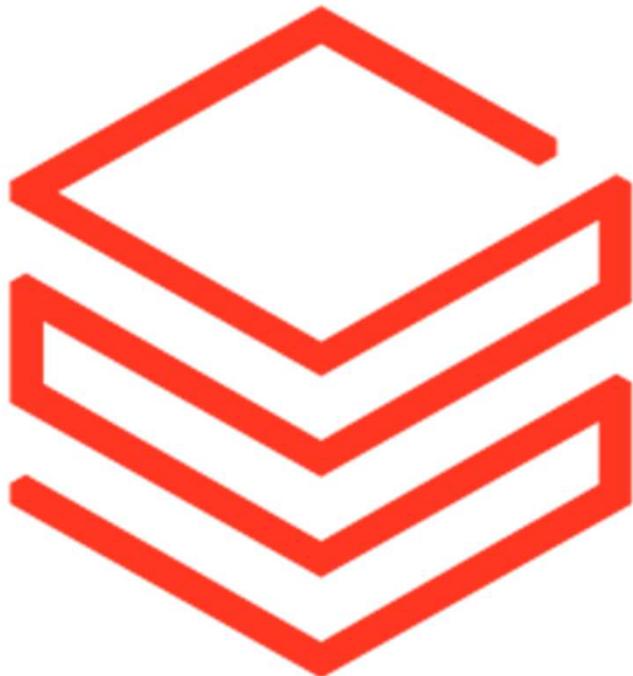


WHAT IS DATABRICKS



ADVANCING
ANALYTICS

WHAT IS DATABRICKS



- Spark as a Service
“Multicloud Lakehouse Platform”
- Built by the team who created
Spark at UC Berkley
- Includes Management, Security
& Governance

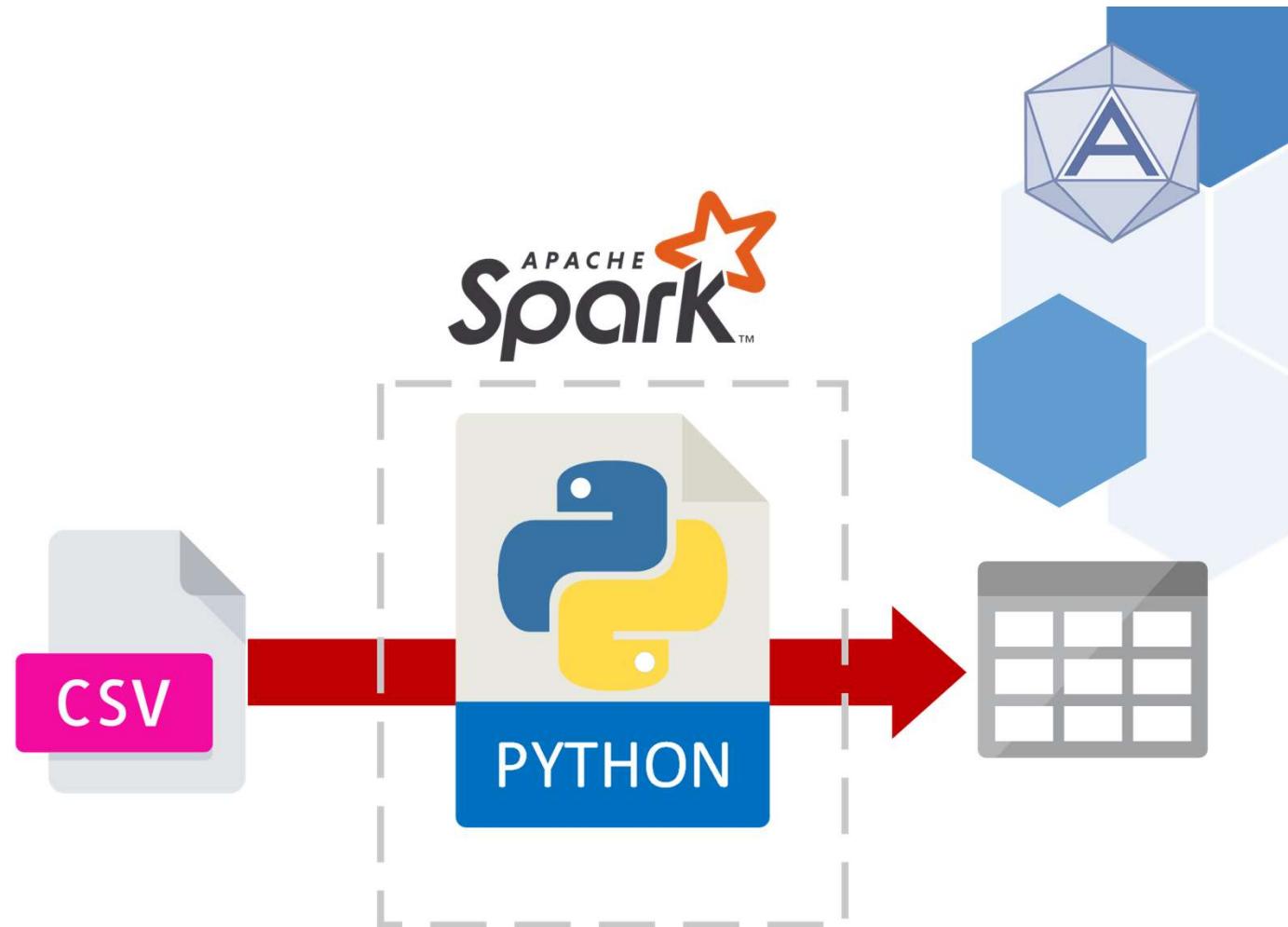


QUICK SPARK OVERVIEW

Spark is a distributed, scalable data processing engine.

It can query **structured** and **non-structured** data

You can use **Python**, **Scala**, **R**, **C#** or **SQL** to interact with it



Learning
Resources



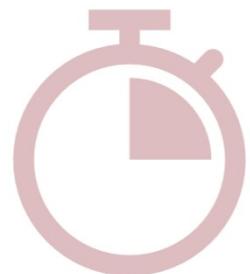
Use it!



Over to you



Learning
Resources



DATABRICKS COMMUNITY EDITION

- <https://www.databricks.com/try-databricks>

Try Databricks for free

An open and unified data analytics platform for data engineering, data science, machine learning, and analytics. From the original creators of Apache Spark™, Delta lake, MLflow, and Koalas.

Databricks trial:

- Collaborative environment for data teams to build solutions together.
- Interactive notebooks to use Apache Spark™, SQL, Python, Scala, Delta Lake, MLflow, TensorFlow, Keras, Scikit-learn and more.
- Available as a 14-day full trial in your own cloud, or as a lightweight trial hosted by Databricks.

Used by:

Please tell us about yourself

First Name: *
Michael

Last Name: *
Robson

Company *

Company Email *

Title *

Phone Number

Country: *
United Kingdom

Yes, I would like to receive marketing communications regarding Databricks services, events and open source products. I understand I can update my preferences at any time.

By Clicking "Get Started For Free", you agree to the [Privacy Policy](#).

GET STARTED FOR FREE

 databricks

Choose a cloud provider

aws Amazon Web Services

Microsoft Azure

Google Cloud Platform

Get started

By clicking "Get started", you agree to the [Privacy Policy](#) and [Terms of Service](#).

Don't have a cloud account?
Community Edition is a limited Databricks environment for personal use and training.

[Get started with Community Edition](#)
By clicking "Get started with Community Edition", you agree to the [Privacy Policy](#) and [Community Edition Terms of Service](#).

© Databricks 2022





DATABRICKS ACADEMY

databricks

Search content in the platform



< Back | Home > Course Catalog

Welcome to Databricks training!



Course Catalog

All your courses and learning plans in which you're enrolled, including all your courses and learning plans in progress and already completed.

FILTERS (1) Search... CARDS NAME A-Z

1 Active Filter

Type

Enrollment Status

Price (1)

All

Free

Paid

2 items

CERTIFICATION EXAM INFO

Exam Information: Databricks Certified Associate Developer for Apache Spark 3.0 (availab...

FREE

EN

E-Learning

Databricks Lakehouse Fundamentals ACCREDITATION

Fundamentals of the Databricks Lakehouse Platform Accreditation

ENROLLED

EN | 30m 00s

E-Learning

<https://customer-academy.databricks.com/learn>

ADVANCING
ANALYTICS



DATABRICKS ACADEMY

The screenshot shows the Databricks Academy platform interface. At the top, there's a navigation bar with the Databricks logo, a search bar, and user icons. Below the bar, a banner for the "DATA+AI SUMMIT 2022" is displayed, featuring logos for PyTorch, Delta Lake, Spark, and mlflow. The main content area has a dark background with abstract geometric shapes and a large green circle containing the summit logo. Text on the page includes "See the best of Data+AI Summit", "Watch now", and "Welcome to Databricks Academy!". A call-to-action at the bottom encourages users to review the "Databricks Academy Guide".

Welcome to Databricks training!

Home Customer Academy Home Page

See the best of Data+AI Summit

Watch all the keynotes, breakouts and training on demand to help you in your certification journey

Watch now

PyTorch DELTA LAKE Spark mlflow

DATA+AI SUMMIT 2022
ORGANIZED BY databricks

Welcome to Databricks Academy!

To get started with your learning experience, please review the course "Databricks Academy Guide" in the "Enrolled Learning" section.

<https://customer-academy.databricks.com/learn>

ADVANCING
ANALYTICS

HOW - ACCESS DATABRICKS ACADEMY ON GITHUB



[Databricks Academy \(github.com\)](https://github.com/databricks-academy)

<https://github.com/databricks-academy>

START WITH...

[Data Engineering with databricks \(github.com\)](https://github.com/databricks-academy/data-engineering-with-databricks-english)

<https://github.com/databricks-academy/data-engineering-with-databricks-english>

[Apache Spark Programming with Databricks \(github.com\)](https://github.com/databricks-academy/apache-spark-programming-with-databricks)

<https://github.com/databricks-academy/apache-spark-programming-with-databricks>



DATABRICKS DOCUMENTATION

The screenshot shows the Databricks Documentation website. The top navigation bar includes the Databricks logo, a search bar, and links for SUPPORT, FEEDBACK, and TRY DATABRICKS. The main content area is titled "Get started with Databricks" and features a "Sign up for a free trial" section with a link to learn how to sign up for a free 14-day trial with Databricks on AWS. Below this are sections for ETL, Data scientists, Data analysts, Machine learning engineers, and Administrators, each with a brief description and a corresponding icon.

<https://docs.databricks.com/getting-started/index.html>

ADVANCING
ANALYTICS



DATABRICKS DOCUMENTATION

The screenshot shows the Databricks Documentation homepage. The top navigation bar includes links for Help Center, Documentation (which is currently selected), and Knowledge Base. The main content area is titled "Data on AWS" and contains sections for Introduction, Get started, ETL, Data湖 (Lakehouse), Data scientists, Data analysts, Machine learning engineers, Administrators, Unity Catalog (marked with a red circle), and Tutorials and best practices (also marked with a red circle). A large blue arrow points from the circled "Tutorials and best practices" link in the sidebar to the "Quickstarts and tutorials" section in the main content area.

Documentation > Quickstarts, tutorials, and best practices

Quickstarts, tutorials, and best practices

August 11, 2022

Databricks documentation includes many tutorials, quickstarts, and best practices guides.

SUPPORT FEEDBACK TRY DATABRICKS

Quickstarts and tutorials

Quickstarts provide a shortcut to understanding Databricks features or typical tasks you can perform in Databricks. Most of our quickstarts are intended for new users.

Tutorials provide more complete walkthroughs of typical workflows in Databricks.

These quickstarts and tutorials are listed according to the Databricks persona-based environment they apply to. However some apply more broadly. For example, the Data Science & Engineering quickstarts are useful for machine learning engineers first encountering Databricks, and both Run your first ETL workload on Databricks and Get started as a Databricks administrator are useful regardless of which environment you are working in.

Databricks Data Science & Engineering

- Quickstart: Get started with Databricks as a data scientist
- Quickstart: Get started with Databricks as a data engineer
- Tutorial: Get started as a Databricks administrator
- Quickstart: Create data pipelines with Delta Live Tables
- Tutorial: Create a workspace with the Databricks Terraform provider
- Bulk load data into a table with COPY INTO with Spark SQL
- Tutorial: Continuously ingest data into Delta Lake with Auto Loader

Databricks Machine Learning

- Quickstart: Get started with Databricks as a machine learning engineer
- Quickstart: Model training
- MLflow quickstarts
- 10-minute tutorials: Get started with machine learning on Databricks

Databricks SQL

- Databricks SQL user quickstart: Import and explore sample dashboards
- Databricks SQL user quickstart: Run and visualize a query
- Databricks SQL admin quickstart: Onboarding
- Databricks SQL admin quickstart: Set up a user to query a table

Best practices

The Databricks documentation includes a number of best practices articles to help you get the best performance at the lowest cost when using and administering Databricks.

For users:

- Delta Lake
- Hyperparameter tuning with Hyperopt
- Deep learning in Databricks
- Delta Lake Structured Streaming with Amazon Kinesis
- CI/CD

For administrators:

- Cluster configuration
- Pools
- Cluster policies
- Data governance
- GDPR and CCPA compliance using Delta Lake

ADVANCING
ANALYTICS

YOUTUBE RESOURCES

<https://www.youtube.com/c/Databricks>



YOUTUBE RESOURCES

<https://www.youtube.com/c/Databricks>

Month of Azure Databricks – YouTube

Month of Azure Databricks

17 videos • 52,984 views • Last updated on 3 Aug 2019

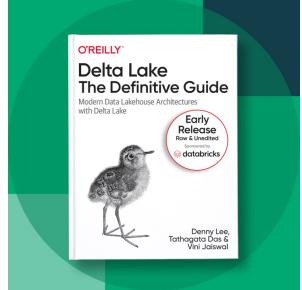
PLAY ALL

Advancing Analytics SUBSCRIBE

- 1 databricks What is Azure Databricks? Advancing Analytics
- 2 databricks Creating Your First Azure Databricks Workspace Advancing Analytics
- 3 databricks A Quick Tour of the Azure Databricks Workspace Advancing Analytics
- 4 databricks Creating and Configuring Clusters in Azure Databricks Advancing Analytics
- 5 databricks How Do you Size Your Azure Databricks Clusters? Cluster Sizing Advice Advancing Analytics



FREE DATABRICKS E-BOOKS



https://www.databricks.com/p/ebook/delta-lake-the-definitive-guide-by-oreilly?itm_data=blog-promo-deltalakeOreilly

<https://www.databricks.com/p/thank-you/ebook-the-data-lakehouse-platform-for-dummies-172327>



<https://www.databricks.com/wp-content/uploads/2021/10/Big-Book-of-Data-Engineering-Final.pdf>



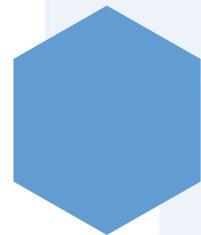
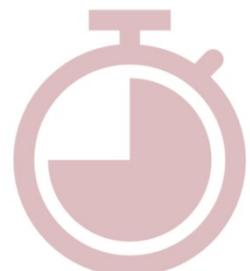
Learning
Resources



Use it!



Over to
you



Use it!





www.advancinganalytics.co.uk

COMMUNITY EDITION



@ADVANCINGANALYTICS



@ADVANALYTICSUK



/ADVANCING ANALYTICS

WHAT YOU CAN'T DO

- Access to Single Cluster
 - Cluster has a single driver, no workers
- Can't administer the Cluster
- Limited Security Options
- No REST API



[Databricks Community Edition FAQ – Databricks](#)

WHAT YOU CAN'T DO

- Persist the Hive Metastore*
- No Workflows
- No Repo Integration

*Believe you can add an External Metastore in Community Edition, but I haven't tested

<https://docs.microsoft.com/en-us/azure/databricks/data/metastores/external-hive-metastore>



[Databricks Community Edition FAQ – Databricks](#)





www.advancinganalytics.co.uk

COMMUNITY EDITION DEMO



@ADVANCINGANALYTICS



@ADVANALYTICSUK



/ADVANCING ANALYTICS



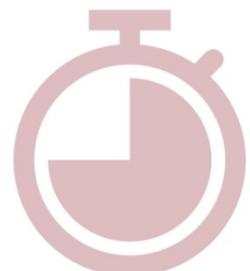
Learning
Resources



Use it!



Over to
you





GET A FREE DATABRICKS ACCREDITATION!

1. Sign up to Databricks Academy

- <https://customer-academy.databricks.com/learn/register>

2. Free Lakehouse Fundamentals

- <https://www.databricks.com/learn/training/lakehouse-fundamentals-accreditation#videocomp>
- <https://www.youtube.com/c/Databricks>



ADVANCING
ANALYTICS

LOOK OUT FOR VIRTUAL EVENTS (THANKS ZACH)



ADVANCING
ANALYTICS

The best data warehouse is a lakehouse

Talks. Demos. Success stories. Q&A.

September 20, 2022 | 8:00 AM PT / 4:00 PM BST

Many enterprises today are running a hybrid architecture — data warehouses for business analytics and data lakes for machine learning. But with the advent of the data lakehouse, you can now unify both on one platform.

Join us to learn why the best data warehouse is a lakehouse. You'll see how Databricks SQL and Unity Catalog provide data warehousing capabilities, fine-grained governance and first-class support for SQL — delivering the best of data lakes and data warehouses.

Hear from Databricks Chief Technologist Matei Zaharia and our team of experts on how to:

- Ingest, store and govern business-critical data at scale to build a curated data lake for data warehousing, SQL and BI
- Use automated and real-time lineage to monitor end-to-end data flow
- Reduce costs and get started in seconds with on-demand, elastic SQL serverless compute
- Help every analyst and analytics engineer to ingest, transform and query using their favorite tools, like Fivetran, dbt, Tableau and Power BI

This presentation will come to life with demos, success stories and best practices learned from the field, while interactive Q&As will help you get all your questions answered from the experts.

Speakers



Matei Zaharia
Co-founder and Chief Technologist, Databricks



Shant Hovsepian
Principal Software Engineer, Databricks



Miranda Luna
Staff Product Manager, Databricks



Join us on September 20

* First Name:
Mikay

* Last Name:
Robson

* Company Email:
community@heymiky.com

* Company Name:
Demo Data Scotland

* Job Title:
Data Engineer

* Phone Number:

* Country:
United Kingdom

Yes, I would like to receive marketing communications regarding Databricks services, events and open source products. I understand I can update my preferences at any time.

Register now

ALTERNATIVES TO COMMUNITY EDITION

- Install Spark Locally
 - Great article on Spark by Example

<https://sparkbyexamples.com/spark/apache-spark-installation-on-windows/>

<https://sparkbyexamples.com/spark/install-apache-spark-on-mac/>

[Spark Installation on Linux Ubuntu - Spark by {Examples} \(sparkbyexamples.com\)](#)



ALTERNATIVES TO COMMUNITY EDITION

- Get a Free Trail of Databricks
 - use a full version of Databricks on AWS, Azure, or GCP

- Set spending limits on your subscription

<https://aws.amazon.com/getting-started/hands-on/control-your-costs-free-tier-budgets/>

<https://docs.microsoft.com/en-us/azure/cost-management-billing/manage/spending-limit>

[Protect your Google Cloud spending with budgets | Google Cloud Blog](#)



THANKS FOR LISTENING

www.advancinganalytics.co.uk



Twitter: @heymiky

<https://github.com/michaelrobson/presentations>
AdvancingAnalytics.co.uk



@ADVANCINGANALYTICS



@ADANALYTICSUK



/ADVANCING ANALYTICS

Session Feedback



Event Feedback