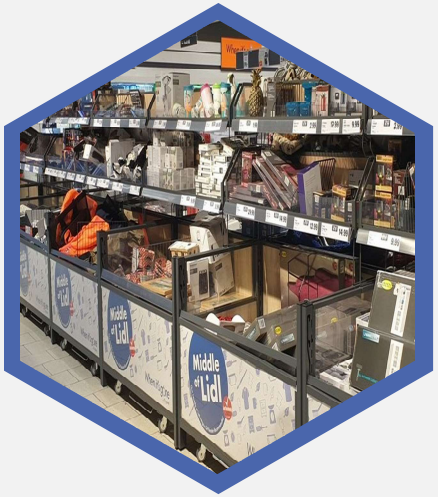
www.advancinganalytics.co.uk

Databricks
FOR THE MIDDLE ISLE



DATA:Scotland 2022

is proudly supported by

THE DATA LAB

value from data

quorum

Octopus Deploy

ADVANCING ANALYTICS

TheDataShed

solarwinds

Microsoft

XTEN

CLOUD | DATA | XOPS

DATAmasterminds

ADVANCING
ANALYTICS

ADVANCING
ANALYTICS

ADVANCING ANALYTICS

WHY

WHY – OPPORTUNITY


I'm Here!
CHAOTIC GOOD

LEARN BY DOING

WHAT IS DATABRICKS

WHAT IS DATABRICKS

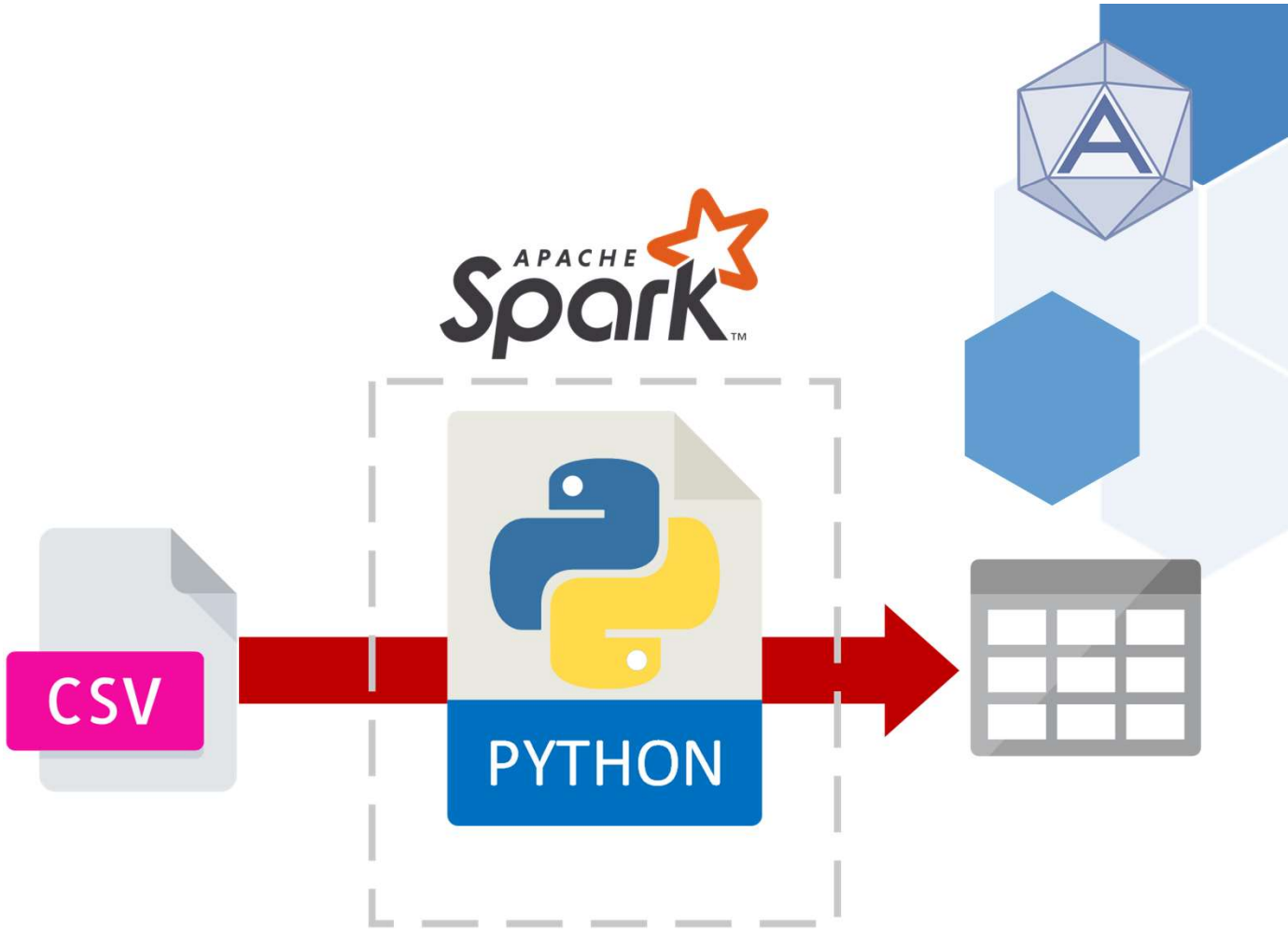
- Spark as a Service
  "Multiclould Lakehouse Platform"
- Built by the team who created Spark at UC Berkley
- Includes Management, Security & Governance

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

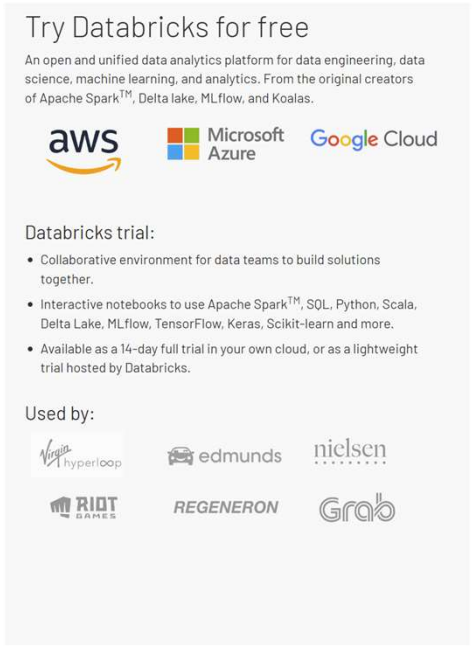
Use it!

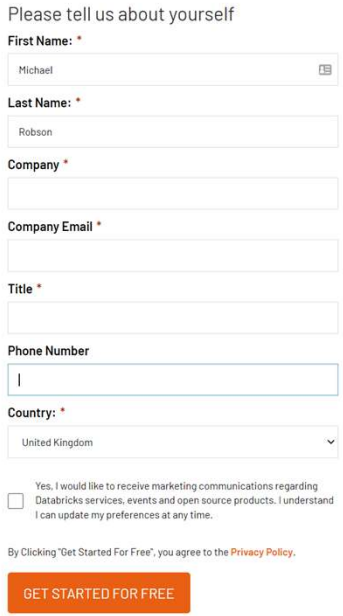
Over to you

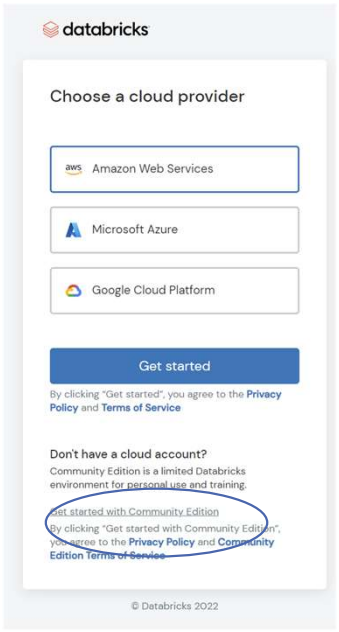
Learning Resources

DATABRICKS COMMUNITY EDITION

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

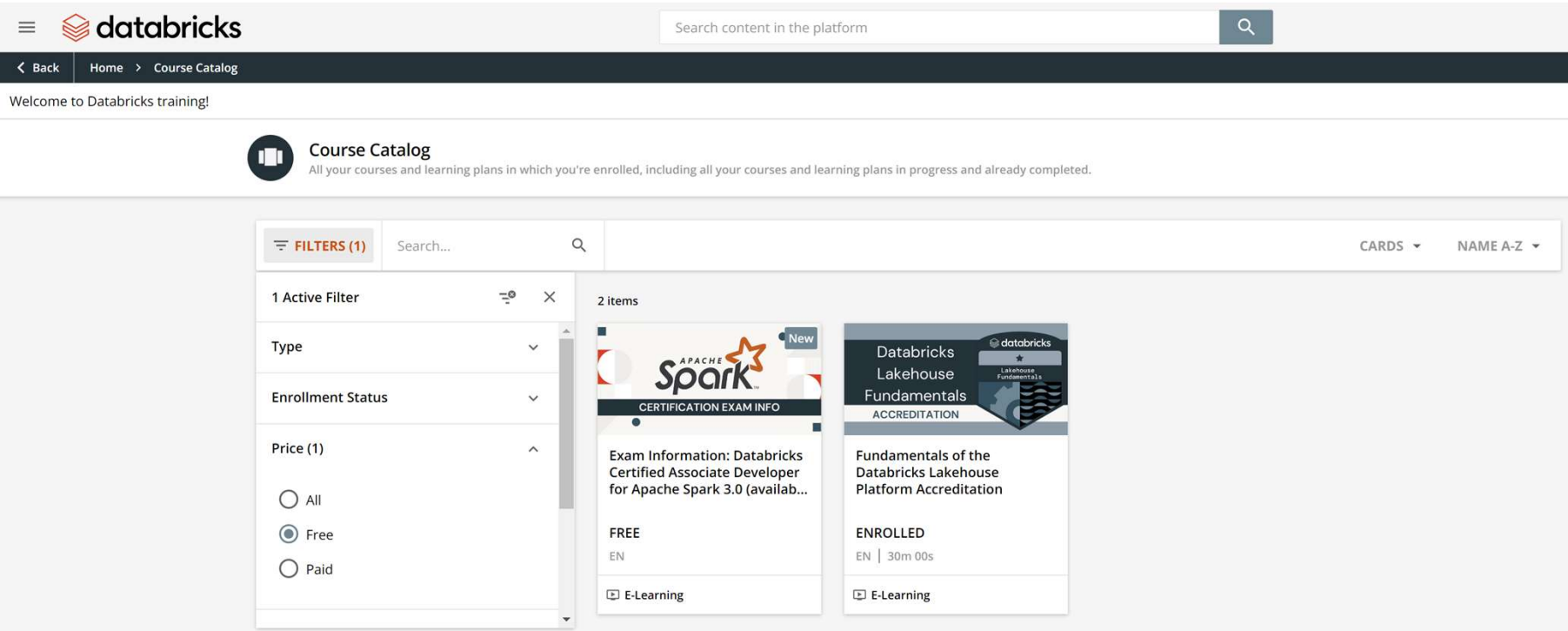






ADVANCING ANALYTICS

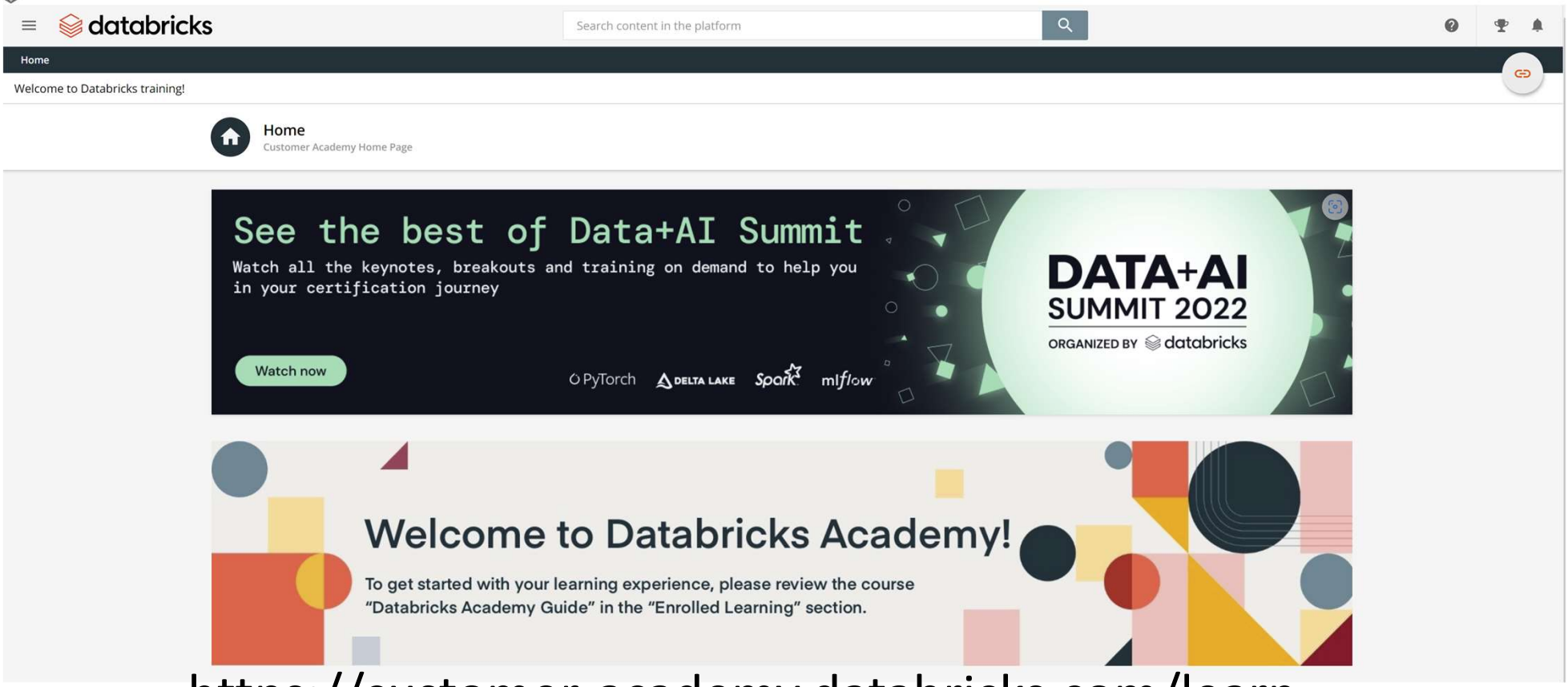
DATABRICKS ACADEMY



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

ADVANCING ANALYTICS

DATABRICKS ACADEMY



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

HOW - ACCESS DATABRICKS ACADEMY ON GITHUB

Databricks Academy (github.com)

https://github.com/databricks-academy

START WITH…

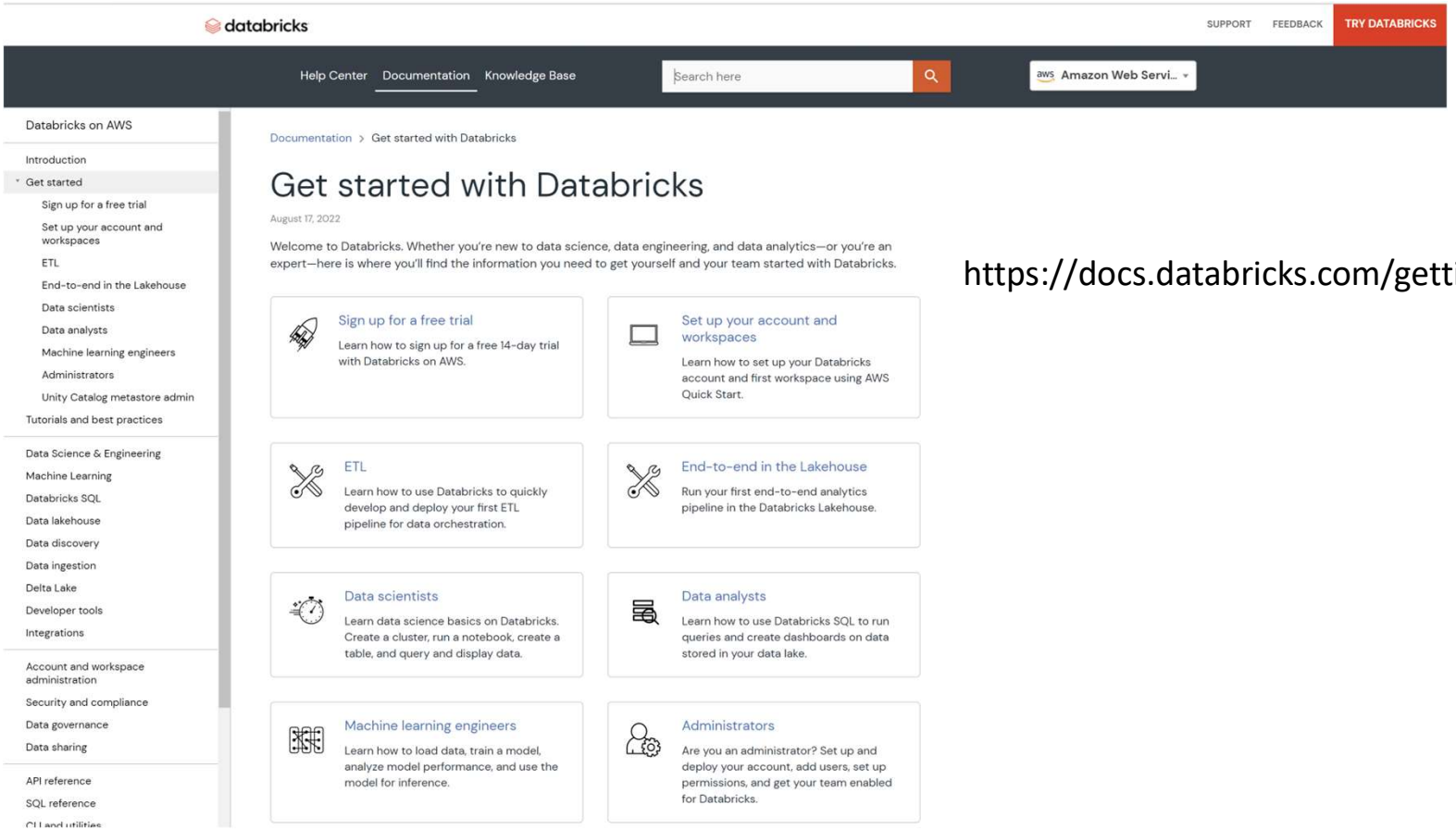
Data Engineering with databricks (github.com)

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

DATABRICKS DOCUMENTATION



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

DATABRICKS DOCUMENTATION

databricks

Help Center | Documentation | Knowledge Base

- Databricks on AWS
- Introduction
- Get started
  - Sign up for a free trial
  - Set up your account and workspaces
  - ETL
  - End-to-end in the Lakehouse
  - Data scientists
  - Data analysts
  - Machine learning engineers
  - Administrators
  - Unity Catalog metastore a
  - Tutorials and best practices
- Data Science & Engineering
- Machine Learning
- Databricks SQL
- Data lakehouse
- Data discovery
- Data ingestion
- Delta Lake
- Developer tools
- Integrations
- Account and workspace administration
- Security and compliance
- Data governance
- Data sharing
- API reference
- SQL reference
- CLI and utilities

Documentation › Quickstarts, tutorials, and best practices

Quickstarts, tutorials, and best practices

August 11, 2022

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

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

SUPPORT | FEEDBACK | TRY DATABRICKS

ADVANCING ANALYTICS

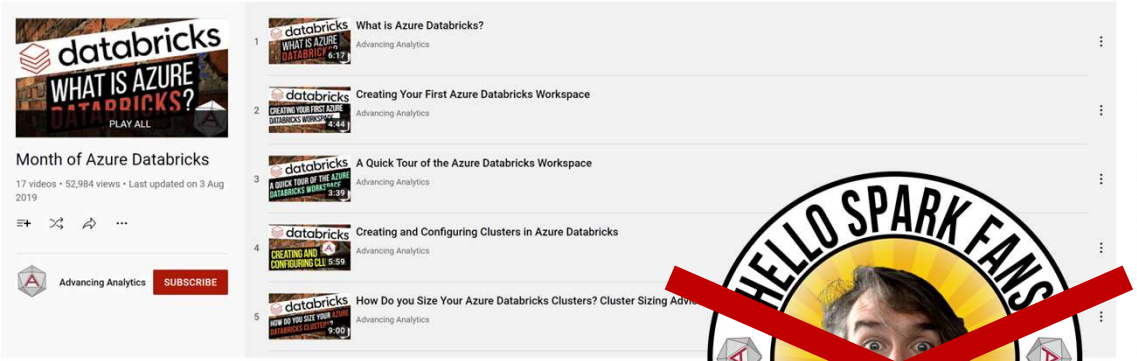
YOUTUBE RESOURCES

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

YOUTUBE RESOURCES

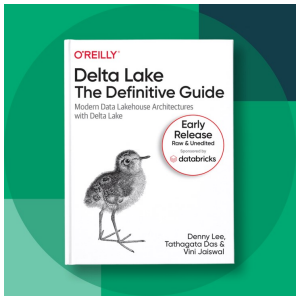
https://www.youtube.com/c/Databricks

Month of Azure Databricks – YouTube



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

ADVANCING ANALYTICS

Learning Resources

Use it!

Over to you

Use it!

www.advancinganalytics.co.uk

COMMUNITY EDITION

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

COMMUNITY EDITION DEMO

Learning Resources

Use it!

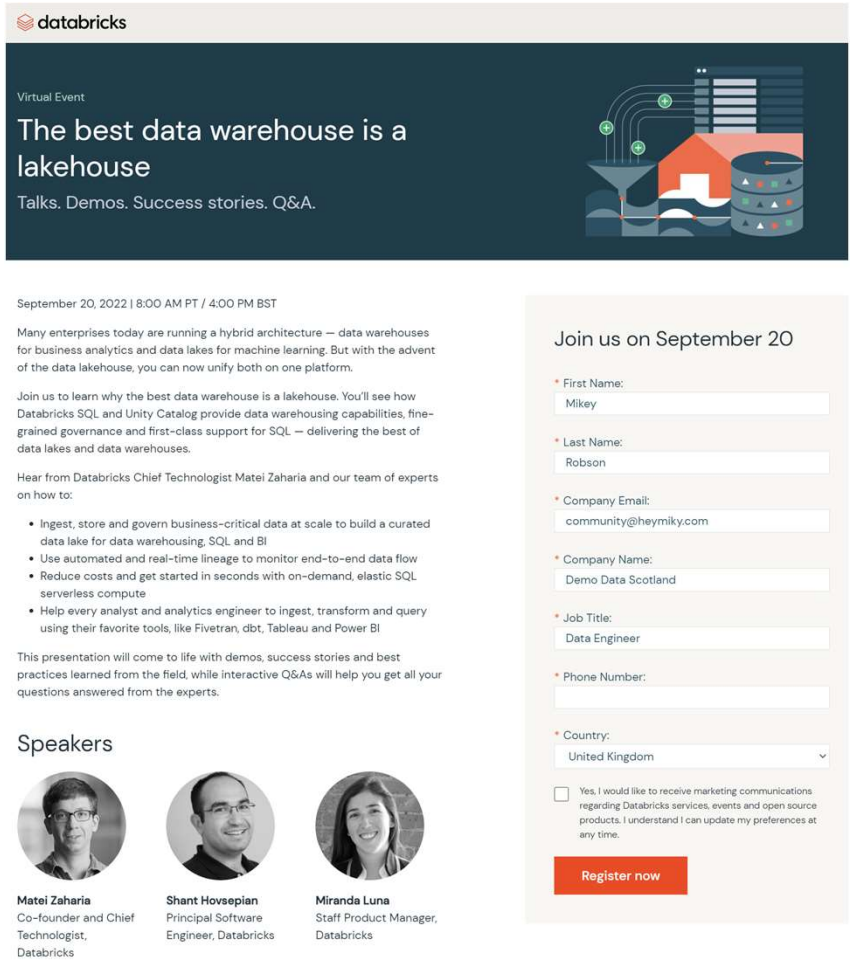
Over to you

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

LOOK OUT FOR VIRTUAL EVENTS (THANKS ZACH)



databricks

Virtual Event

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: Mikey
* 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

ADVANCING ANALYTICS

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

Twitter: @heymiky

https://github.com/michaelrobson/presentations

AdvancingAnalytics.co.uk

Session Feedback

Event Feedback