


Spark → advance optimization ✓
→ memory mgmt. ✓



The diagram shows a horizontal line with three vertical lines intersecting it. The leftmost vertical line has a downward-pointing arrow. The middle vertical line has a downward-pointing arrow. The rightmost vertical line has a downward-pointing arrow.

Databricks ⇒ mini project

Hive ⇒

Datavricks

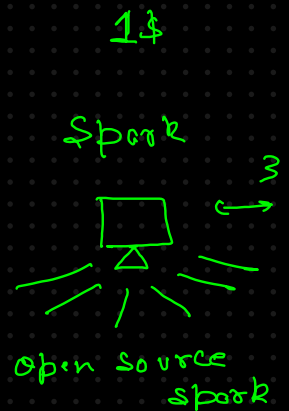
Datatrik is a cloud-based, managed data analytics platform built on Apache Spark

interactive workshop ← data eng.
data analyst



data eng.
data analyst

Spark



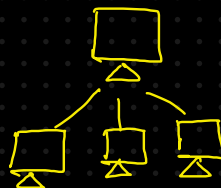
Datatricks is created by inventor of spark.

Spark \rightarrow Apache Foundation

↓
Dokumente (papier)



Azure
AWS
GCP



Apodize
spark

performance optimization lecture

Open Source Spark

Complex Infrastructure setup

Manual software install & updates

Lack of user Interface

Difficult Security management

Version compatibility issue

Databricks

Fully managed cluster

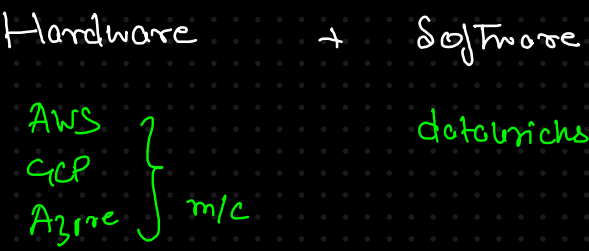
Auto configured environment

Web-based notebooks

Built in security & governance

Optimized spark runtime

Databricks architecture



Workspace

Notebooks

Feature	Description
Workspace	Organizes notebooks, libraries, and files.
Clusters	Manages Spark clusters for processing.
Jobs	Runs scheduled workloads and ETL tasks.
Data	Stores datasets and connects external storage.
Notebooks	Interactive coding interface (supports Python, Scala, SQL, R).

Cluster Type	Purpose
All-Purpose Cluster	Interactive work (for development, shared by multiple users).
Job Cluster	Runs specific jobs and shuts down after completion (cost-effective).
Cluster Pool	Manages pre-warmed clusters for faster job execution.

