INTRODUCTION TO PYSPARK

SAM BAIL // HI@SAMBAIL.COM

JULY 2026

HI, I'M SAM

Data engineer, tech educator, comp sci PhD, event producer, runner, serial concert goer

German native, spent 18 years in the UK / US, just moved to Berlin

Worked for tech startups in healthcare, data infra, marketplace (Flatiron Health, Great Expectations, Collectors)

CONTENT

GOAL: AN INTRODUCTION TO PYSPARK SYNTAX AND AN OVERVIEW OF HOW IT CAN BE USED IN PRODUCTION

PART 1

INTRO & BACKGROUND

PART 2

LIVE DEMO

PART 3

PYSPARK IN PRODUCTION

CONTENT

GOAL: AN INTRODUCTION TO PYSPARK SYNTAX AND AN OVERVIEW OF HOW IT CAN BE USED IN PRODUCTION

PART 1

INTRO & BACKGROUND

PART 2

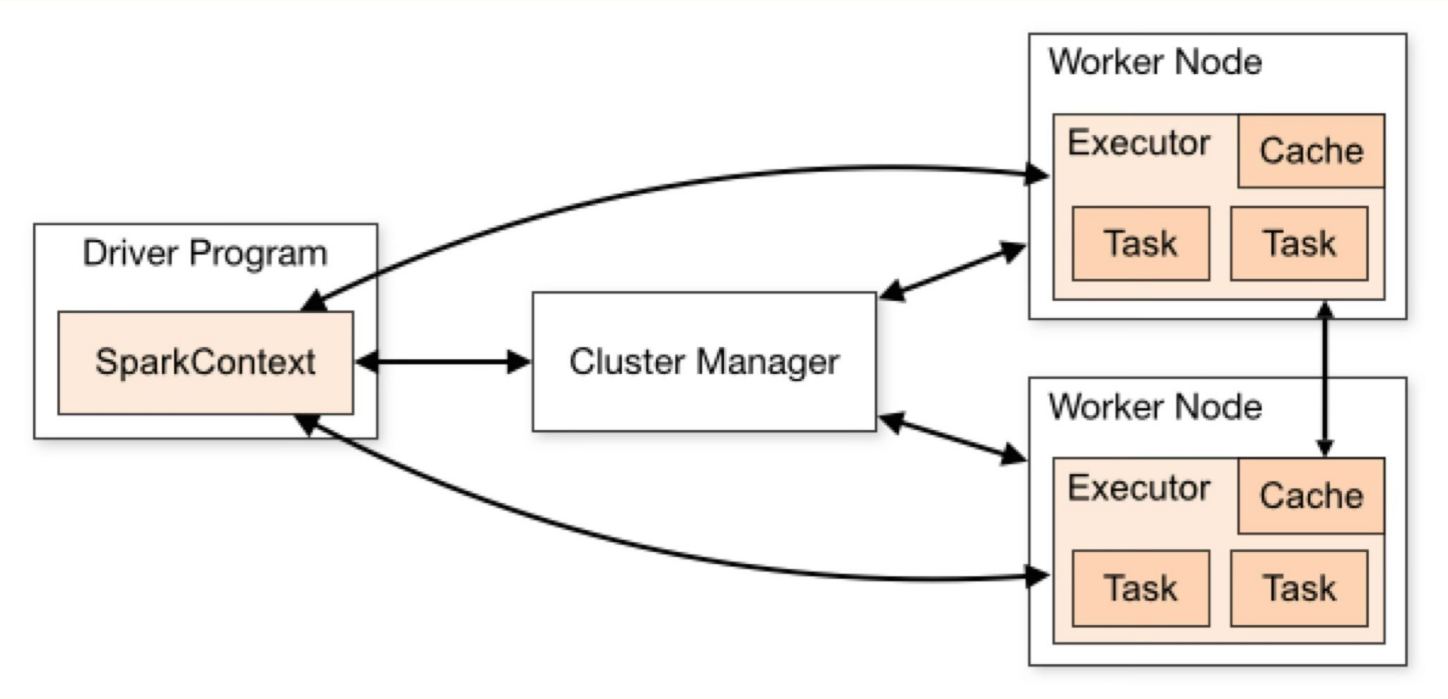
LIVE DEMO

PART 3

PYSPARK IN PRODUCTION

WHAT IS PYSPARK?

- Apache Spark = analytics engine for large-scale data processing, written in Scala
- Developed in 2012 to address shortcomings of MapReduce (great for handling big data, but can be slow)
- Uses resilient distributed dataset (RDD) - immutable, fault-tolerant
- PySpark = Python API for Spark



SPARK CLUSTER

SPARK

- Does not have its own native file system
- Lower latency due to in-memory data storage during processing
- More suitable for real-time processing
- Built-in ML library

HADOOP

- Has native file system - Hadoop Distributed File System (HDFS)
- Higher latency due to data writes at each processing step
- Better suited for batch jobs
- No native ML capabilities

THE APACHE SPARK ECOSYSTEM (1)

SPARK CORE

Distributed task dispatching, scheduling, and basic input/output functionalities

RDD

Resilient Distributed Dataset: Read-only multiset of data distributed over a cluster of machines.

DATAFRAME

Higher-level abstraction on top of RDDs that's optimized for structured data and tabular data processing.

THE APACHE SPARK ECOSYSTEM (2)

SPARK SQL

Allows querying data both in RDDs and in external sources, such as relational databases

SPARK DATASET

Interface that combines the benefits of RDDs with the benefits of Spark SQL's optimized execution engine

SPARK WEB UI

Web app that shows helpful information about Spark jobs using different visualizations

PYSPARK CAPABILITIES

- Similar to Pandas, but for larger datasets
- Read, write, transform, analyze large datasets using Python & SQL
- Processing real-time streaming data
- Machine learning
- Interactive shell for ad-hoc analysis

PYSPARK DF

- Designed for large, distributed datasets
- Immutable
- Lazy evaluation (operations are queued and optimized)

PANDAS DF

- Designed for small to medium size datasets
- Mutable
- Eager evaluation (executed as soon as they occur in the code)

CONTENT

GOAL: AN INTRODUCTION TO PYSPARK SYNTAX AND AN OVERVIEW OF HOW IT CAN BE USED IN PRODUCTION

PART 1

INTRO & BACKGROUND

PART 2

LIVE DEMO

PART 3

PYSPARK IN PRODUCTION

https://colab.research.google.com/drive/1LpLePrxXP5wVZt0yqJLLl63ldW2o-qj9?usp=sharing

CONTENT

GOAL: AN INTRODUCTION TO PYSPARK SYNTAX AND AN OVERVIEW OF HOW IT CAN BE USED IN PRODUCTION

PART 1

INTRO & BACKGROUND

PART 2

LIVE DEMO

PART 3

PYSPARK IN PRODUCTION

PRODUCTION REQUIREMENTS

- Data sources ingested through extraction tools
- Distributed storage
- Cluster management
- Job scheduling
- Monitoring & logging
- Security & access control

SPARK PRODUCTION EXAMPLE - DIY

Spark engine

Run Spark on cluster of EC2 instances & YARN as cluster manager

Distributed storage

Amazon S3

Job scheduling

Apache Airflow on EC2 instance

Monitoring & logging

AWS CloudWatch or Spark Web UI

SPARK PRODUCTION EXAMPLE - MANAGED

Spark engine

AWS Glue or EMR (Elastic MapReduce) Serverless, GCP Dataproc, Azure Synapse

Distributed storage

Amazon S3

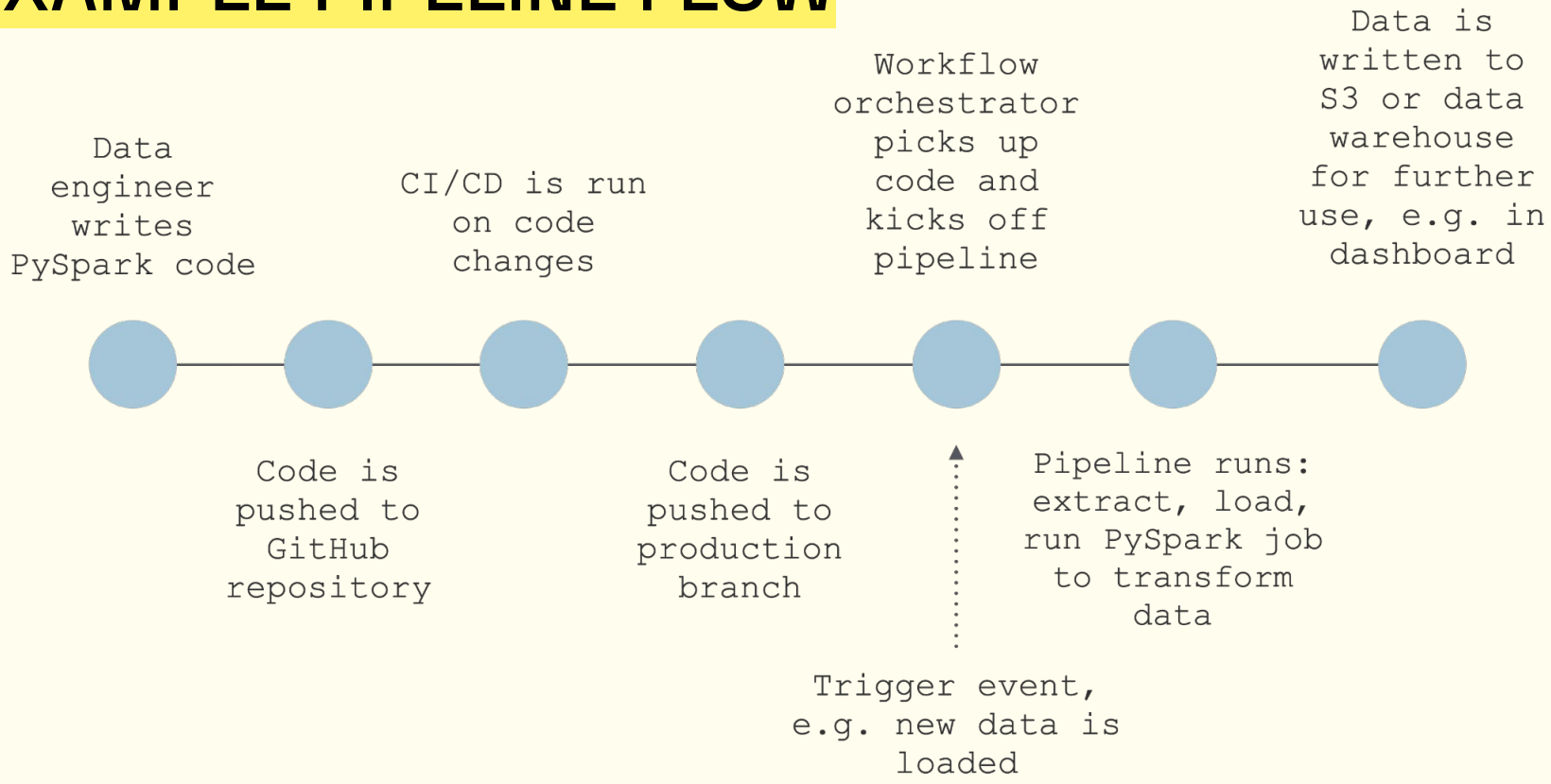
Job scheduling

MWAA (Managed Workflows for Airflow)

Monitoring & logging

AWS CloudWatch

EXAMPLE PIPELINE FLOW



OR... DATABRICKS

- One of the most popular platforms for running PySpark
- Developed by creators of Apache Spark
- Collaborative notebook environment
- Auto-scaling Spark clusters
- Works on AWS, Azure, Google Cloud

MORE CONCEPTS TO EXPLORE

- Structured Streaming - real-time data processing
- Delta Lake - versioning and data consistency
- MLlib - Machine learning capabilities

THANK YOU

SAM BAIL

hi@sambail.com
www.sambail.com for slides

Watch my PySpark LinkedIn Learning course here!