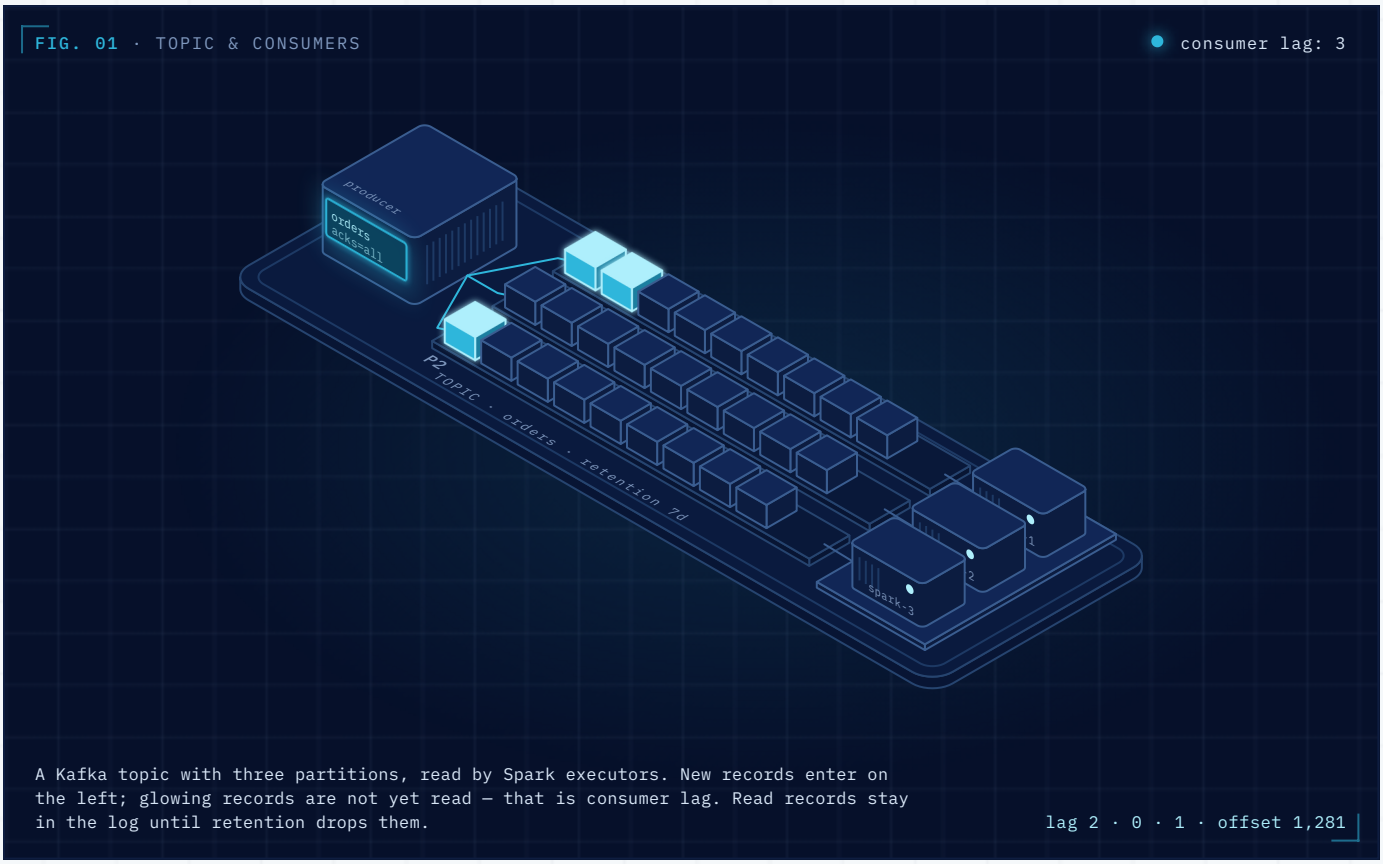
TRAINING & CONSULTING · DATA PLATFORMS

Data pipelines that *keep up* with the stream.

Consulting and training for Apache Kafka, Apache Spark and the wider big data stack — Hadoop, Hive and Trino. Since 2014 we have designed streaming and batch pipelines, tuned the jobs that are slow and expensive, and trained the engineers who run them.



2014
Building data platforms for data-heavy teams

4
Programmes: Kafka, Spark, the Hadoop ecosystem, PySpark

≤15
People per class, with a cloud lab for each of them

EN · TR
Delivered in English or Turkish, remote or on-site

CONSULTING

Make the pipeline boring. Topic design, job tuning and table layouts that keep latency low and costs predictable.

TRAINING

From first topic to platform. Kafka, Spark, Hive, Trino and PySpark, taught with labs on realistic data volumes.

SUPPORT

Stay on top of the load. Capacity reviews, upgrades such as ZooKeeper to KRaft, and help when lag will not drain.

02 · CONSULTING

Streams that drain, jobs that *finish*.

Data platforms slow down in predictable places: partitions chosen years ago, skewed joins, small files, lag nobody watches. We find the bottleneck with your metrics, fix it with your team and leave the monitoring that shows it stays fixed.

01

Streaming architecture

Topic and partition design, keys and ordering, retention and compaction, schema registry, exactly-once.

YOU GET

A reviewed event model and a capacity plan.

02

Kafka operations

KRaft migration, broker sizing, rack awareness, replication and min ISR, rebalancing storms, MirrorMaker.

YOU GET

Tuned clusters and lag alerts that mean something.

03

Spark performance

Skew, shuffles, broadcast joins and salting, partitioning, AQE, executor memory, small files.

YOU GET

Faster, cheaper jobs with before/after run times.

04

Lakehouse & warehouse

Parquet, Iceberg and Delta tables, partitioning, the Hive metastore and Trino for interactive SQL.

YOU GET

A table layout that queries fast and costs less.

05

Integration & CDC

Kafka Connect, Debezium change data capture, sinks into search, vector and warehouse stores; RabbitMQ bridges.

YOU GET

Pipelines with replay, dead-letter queues and monitoring.

06

Platform on Kubernetes

Strimzi for Kafka, the Spark operator and Trino on Kubernetes; autoscaling and cost control.

YOU GET

A reproducible platform defined as code.

SIGNALS IT IS TIME TO CALL

- Consumer lag grows every evening and drains by morning
- One Spark task runs for an hour while the rest take minutes
- The nightly batch no longer finishes before the workday
- Partition counts were picked years ago and never revisited
- Millions of small files slow every query
- ZooKeeper to KRaft is on the roadmap but nobody owns it

HOW AN ENGAGEMENT RUNS

01 **Scoping call** — 45 min · free
We hear the problem, agree goals and the numbers success is judged by.

02 **Assessment** — 1–2 weeks
Read-only access, diagnostics, logs and short interviews. Findings in writing.

03 **Fix & tune** — 2–6 weeks
Changes made with your team in small steps, each measured before and after.

04 **Hand-over** — ongoing
Runbooks, a focused workshop for your engineers and optional retained support.

FIELD EXPERIENCE: Banking & finance · Telecom · Public sector · Defence · Retail & e-commerce · Logistics

03 · TRAINING

Four programmes, one *platform*.

Take them on their own or as a path. Each programme runs on a cloud lab with realistic data volumes and ends with a project built by the participants.

PROGRAMME 01 · STREAMING

Apache Kafka

5 days · 30 hours

LEVEL Beginner – intermediate
PREREQ. Basic Linux

- Architecture and installation: brokers, ZooKeeper and KRaft
- Partitions, replication, topic design and retention
- Kafka Connect: sources, sinks and connector operations
- Streams API, REST proxy and custom serializers
- Kafka and RabbitMQ: when to use which, and bridging them

PROGRAMME 02 · PROCESSING

Apache Spark

4 days · 24 hours

LEVEL Intermediate – advanced
PREREQ. Linux and containers

- RDDs, DataFrames and Spark SQL; ETL jobs
- Performance: joins, broadcasting, salting, cluster management
- Structured Streaming from Kafka and files, windowing
- Reporting with Grafana and Kibana
- Project day: an analysis on real data, built in teams

PROGRAMME 03 · PLATFORM

Big Data Ecosystem on Kubernetes

5 days · 30 hours

LEVEL Intermediate – advanced
PREREQ. Linux and containers

- Big data concepts, batch and stream; HDFS architecture
- Hadoop, HDFS file operations and MapReduce basics
- Hive: architecture, DDL and DML, performance
- Real-time ingestion with Kafka
- Spark, Kafka and Trino on Kubernetes, managed with operators

PROGRAMME 04 · ANALYSIS

Python for Data Analysis & PySpark

5 days · 30 hours

LEVEL Beginner – intermediate
PREREQ. Basic programming

- Python fundamentals, environments and Jupyter
- SQL, SQLite and SQLAlchemy; Pandas with SQL
- NumPy arrays, broadcasting and linear algebra
- Pandas: selection, cleaning, groupby, joins, pivots, time series
- PySpark: DataFrames, transformations and actions, Spark SQL

DELIVERY
Remote or on-site; daytime, evening or weekend sessions

LABS
A cloud lab for every participant, on realistic data

CLASS SIZE
Up to 15 people, so everyone gets hands-on time

LANGUAGE
English or Turkish, with all materials provided

TAILORING
Content, length and examples fitted to your stack

NEXT STEP

Start with a 45-minute scoping call.

Tell us what is slow, wrong or about to grow. We reply within one working day with a suggested next step — consulting, training or both.

info@scaleyazilim.com

scaleyazilim.com/services/big-data

Ankara · Türkiye — remote or on-site, worldwide