Big Data on OSS @ SBB

Adrian Burri

Zug, 14.05.2024

Agenda.

- OSS @ SBB
- Train Delay Analysis
- OSS for Big Data
- Discussion

Usage:
Transition from Closed Source to Open Source

Dimension	until 2015	today
Application Runtime	IBM Mainframe System, IBM WebSphere, Spring	OpenShift, Node.js, Spring Boot, Angular
Messaging	IBM MQ, IBM Integration Bus	Kafka, NATS
Databases/Caches	Oracle, Microsoft SQL Server	PostgreSQL, MongoDB, hazelcast
Monitoring	splunk>, ++ ganz viele mehr	splunk>, New Relic, Instana, Grafana, Prometheus, Kibana, Elastic
DevOps Tools	IBM Rational, Nexus, Jenkins	Jenkins, Gogs, Ansible, Terraform, Vault
Infrastructure	vmware, Hardware	aws, Azure, OpenStack, OpenShift

Open Source

Closed Source

OpenRail Association to Foster Collaboration of European Rail Companies

OPENRAIL ASSOCIATION

AUTOMOTIVE GRADE LINUX

LFENERGY

open logistics foundation

OpenRail Association Governance

- **Neutral basis**
  - Open source licenses
  - Open source collaboration model
  - Stimulate active contribution
- **Project governance** on project level

SBB Open Source Guide

- Strategic goals
- Principles for Use/Create
- Structured process

Cooperation in Switzerland: Driver Advisory System

- **Cost Sharing: SBB** develops for itself and different other railway companies (BLS, SOB)
- Open Source **License**
  - Use
  - Enhancements
- **Transparency** for partners
- Built on open **standards**
- **International** opportunities

Agenda.

OSS @ SBB

Train Delay Analysis

OSS for Big Data

Discussion

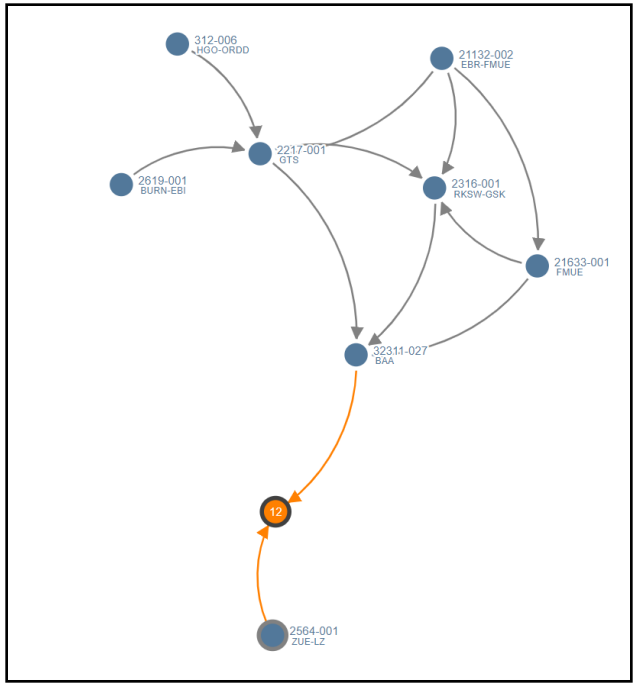
Punctuality: Our DNA

Definition:

Arrival time < 3 min.

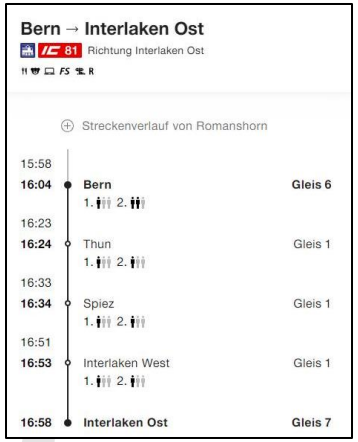
after schedule

Train Delay Analysis

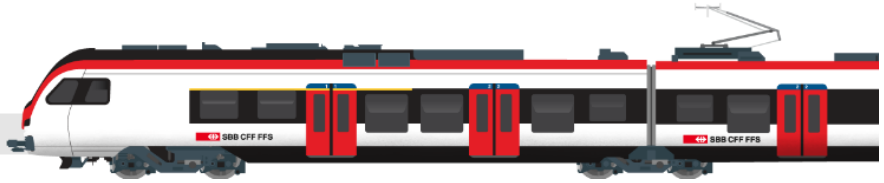


Delay Analysis:
Root Cause Statistics,
Deviation Graphs

Planning:
Timetables, Connections,
Rolling Stock, etc.



Operations:
Actual Times, Deviations



Dispatching:
Incidents, Dispatcher
Actions

Example Code Assignment
91: Special Environmental Conditions

Agenda.

OSS @ SBB

Train Delay Analysis

OSS for Big Data

Discussion

Overview Tech Stack

Workflow

Apache Airflow

Analysis

jupyter

Monitoring

Prometheus

Grafana

CI/CD

argo

TEKTON

JFrog ARTIFACTORY

Renovate

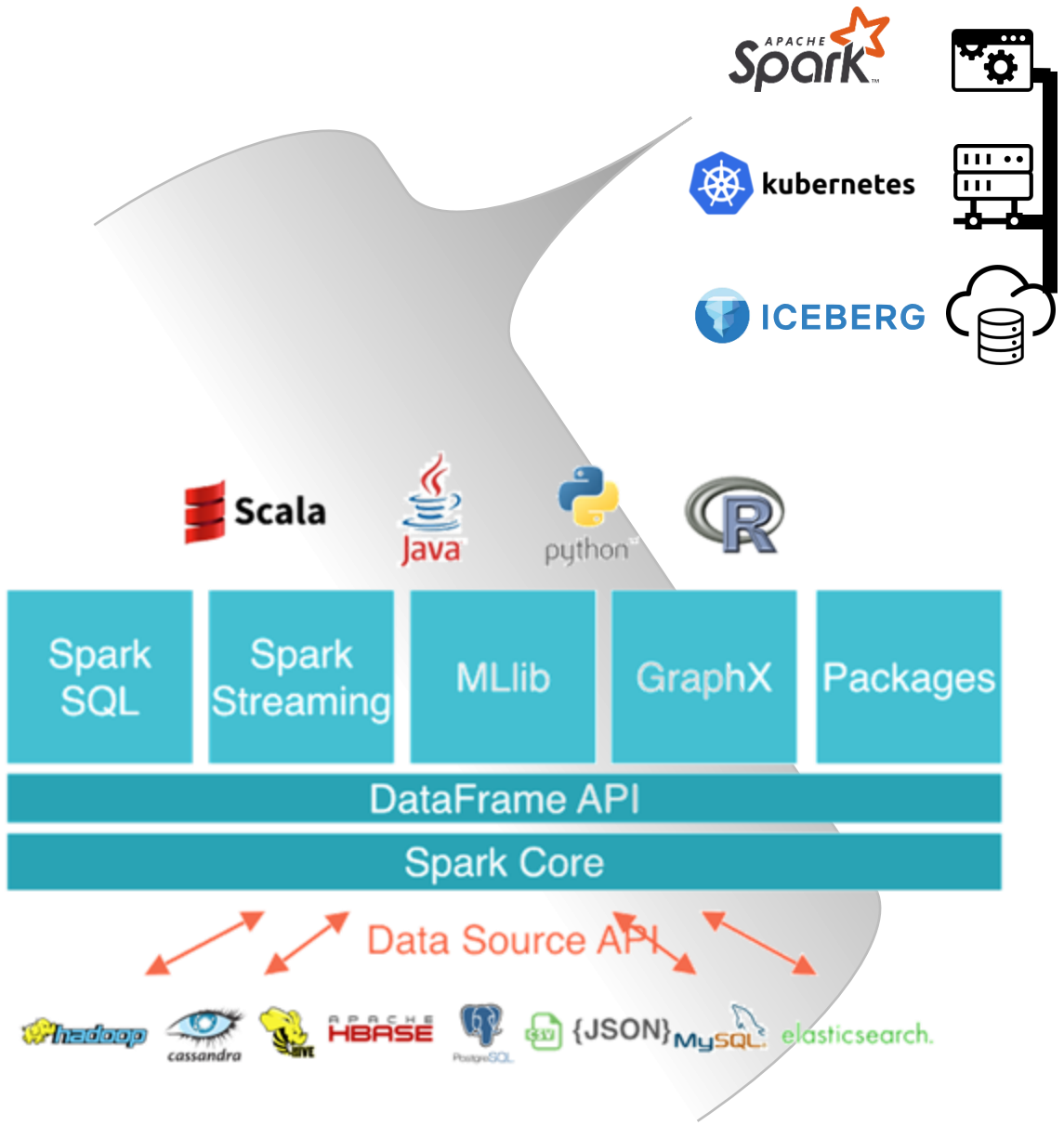
Apache Spark

Type: Unified engine for data analytics
OSS since: 2010

Spark is based on the concept of an RDD (Resilient Distributed Dataset)

Key benefits:

- Serial programming model with parallel execution (incl. fault tolerance)
- In-memory processing
- Versatility
  - Many connectors (file, DB, messaging,…)
  - Many use cases (SQL, streaming, ML,…)



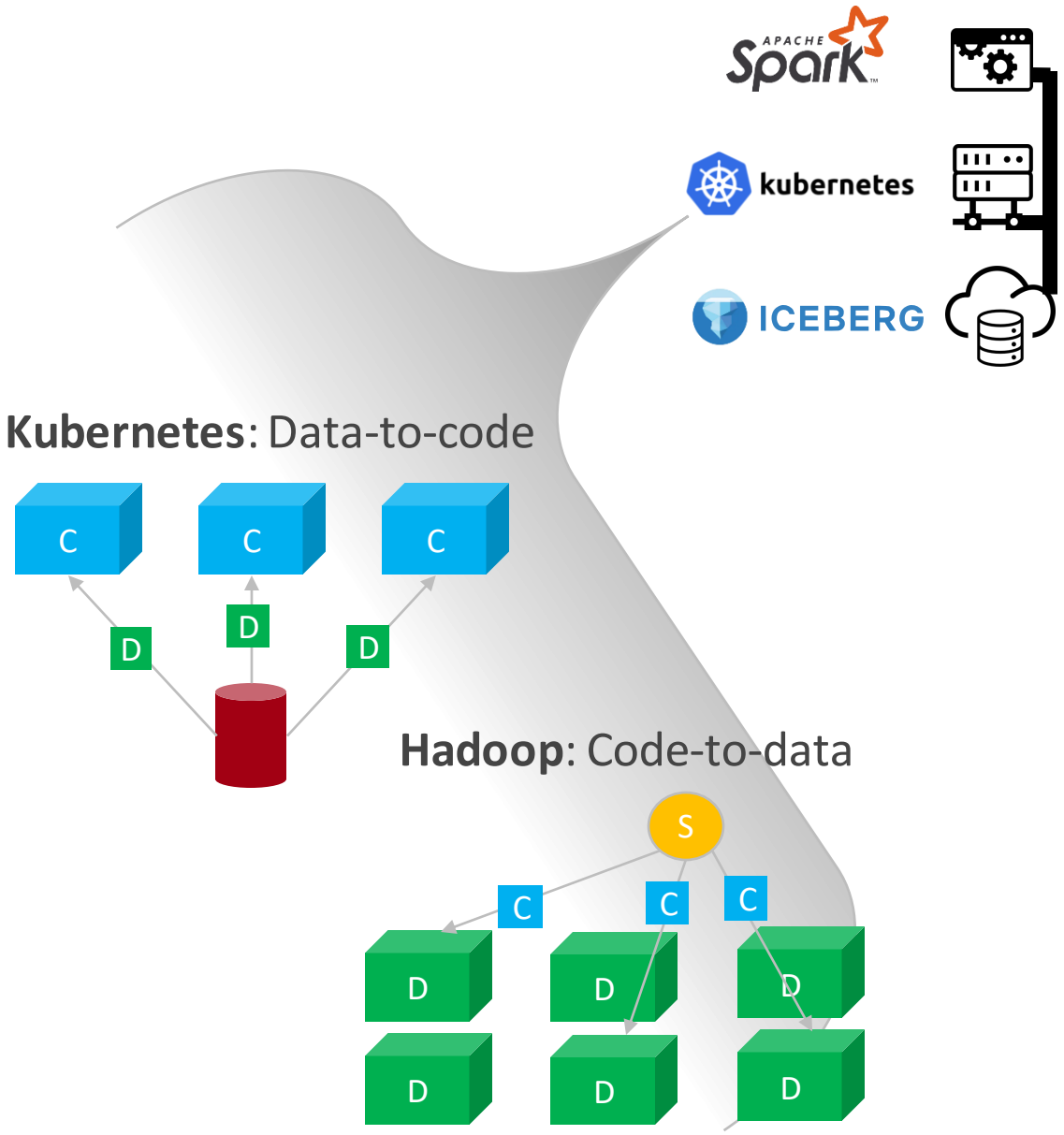
Spark on Kubernetes

Type: Container orchestration
OSS since: 2015

Runtime for transactional and analytical workloads in cloud environments.

Key benefits:

- Scalability of workloads
- Portability between different cloud providers
- Rich ecosystem for DevOps



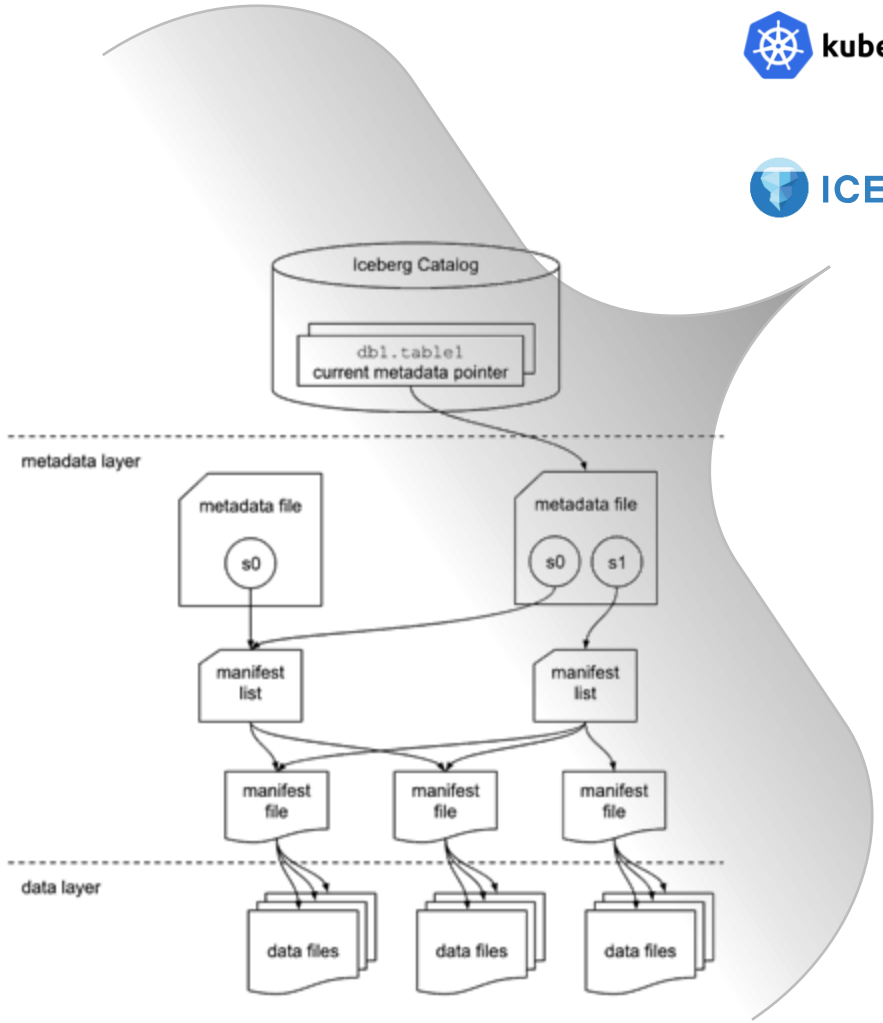
Apache Iceberg

Type: Open table format for analytics
OSS since: 2018

Iceberg provides a table abstraction on unstructured blob storage through hierarchical metadata.

Key benefits:

- Correctness through ACID transactions
- Efficient query planning
- Easy schema evolution (incl. partition change)
- Timetravel



Agenda.

OSS @ SBB

Train Delay Analysis

OSS for Big Data

Discussion

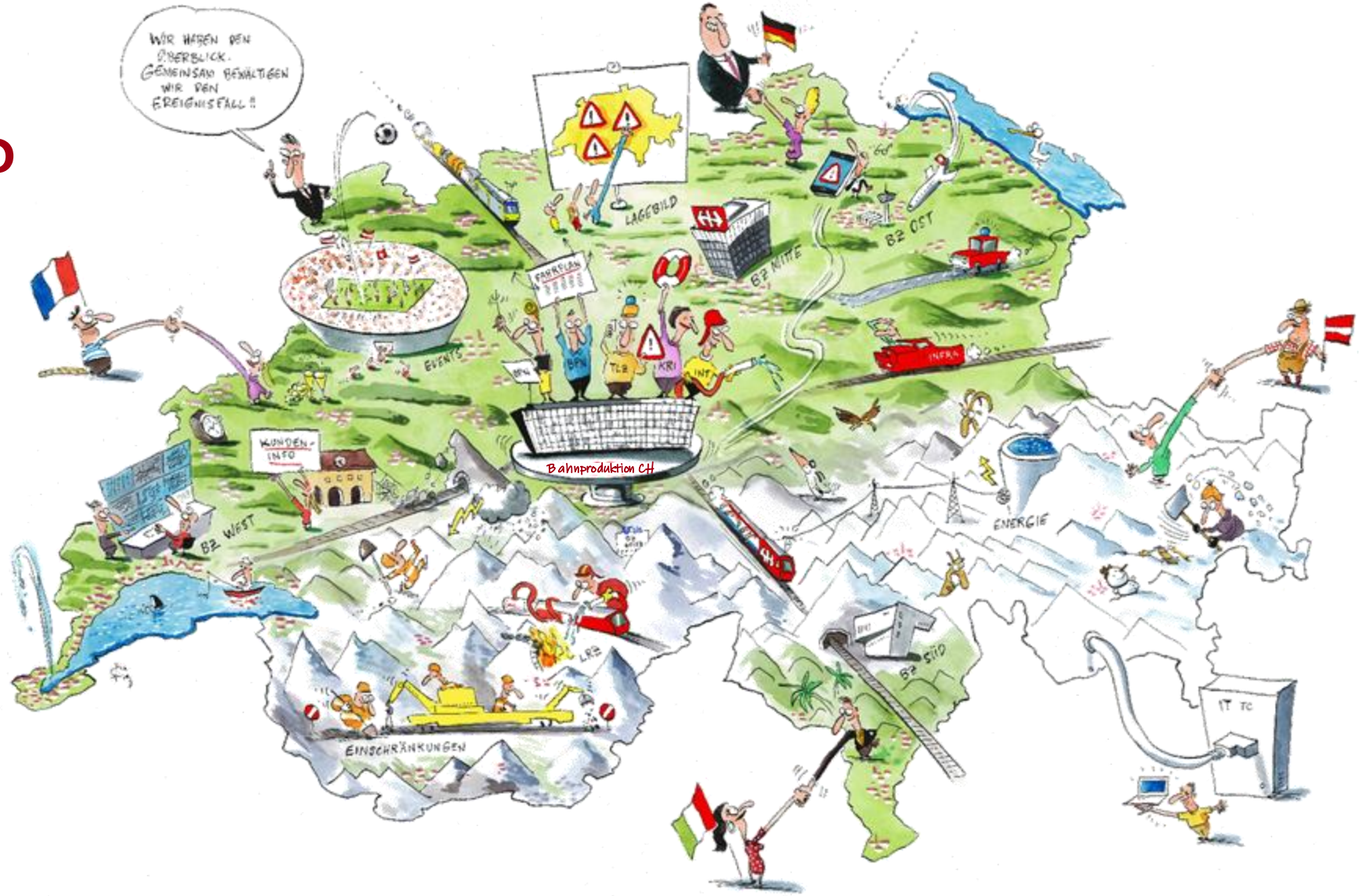
Key Takeaways

Analytics workloads can be processed in a cost-efficient on a purely OSS tech stack.

Key incredients:

- Spark as analytics engine offers a simple and efficient computing model for various Big Data use cases
- In a cloud environment, the same runtime environment (Kubernetes) as for transactional workloads can be re-used
- Iceberg (or other table formats) allow for cheap table-like storage in unstructured cloud storage

Questions?


WIR HABEN DEN ÜBERBLICK. GEMEINSAM BEWÄLTIGEN WIR DEN EREIGNISFALL !
LAGEBILD
FAHRPLAN
EVENTS
KUNDEN-INFO
B2 WEST
B2 MITTE
B2 OST
INFRA
Bahnproduktion CH
ENERGIE
B2 SÜD
IT TC
EINSCHRÄNKUNGEN


SBB CFF FFS
Danke, merci, grazie
and thank you.