Building containerized Spark on a solid foundation with Quobyte and Kubernetes
Sascha Askani
Senior Systems Engineer
sascha.askani@inovex.de

Daniel Bäurer
Head of Operations
daniel.baeurer@inovex.de
inovex is an IT project house with a focus on digital transformation:

- Agile development & management
- Web · UI/UX · replatforming · microservices
- Mobile · apps · smart devices · robotics
- Big data & business intelligence platforms
- Data science · data products · search · deep learning
- Data center automation · DevOps · cloud · hosting
- Trainings & coachings

Using technology to inspire our clients. And ourselves.

inovex is located in Karlsruhe · Pforzheim · München · Köln · Hamburg.
You can also find us at www.inovex.de/en
Scope

› Deploying isolated, ad-hoc Spark clusters on k8s

› Introducing a cloud native storage system

›Demonstrate a possible setup to build upon

› Example code
Spark - Short Overview
Spark - Cluster-Manager
Spark - Storage
Running Spark in containers

› Required to use Spark with Kubernetes

› Separate Spark from Hardware

› Running as Container is lightweight and flexible
Kubernetes (k8s) - Overview
Kubernetes - Intro

› Designed by Google and donated to CNCF

› Managing Containers across a cluster of nodes

› Kubernetes is portable, extensible and self-healing
Kubernetes - Intro

https://en.wikipedia.org/wiki/Kubernetes#/media/File:Kubernetes.png (Khtan66) CC by SA 4.0
Why k8s?

› Unified Cluster-Manager for all container applications

› Share nodes between Spark and Microservices

› Isolate and mix Workloads

› Facilitate autoscaling
Quobyte - Overview
Quobyte
Quobyte - High-Level-Architecture
Quobyte
Why bother?

› Unified filesystem (access via HDFS, S3, NFS, native, ...)
  › with distinct volumes for separation

› Most flexible file placement due to a powerful placement engine
  › Exceeding the possibilities of HDFS

› SPOF-free, linearly scalable, end-to-end checksums
Setup and Demo
Architecture of the Demosystem
Summary & Outlook
Combining Spark with k8s and Quobyte provides a flexible solution.

We are able to almost instantly spawn separate Spark clusters with associated volumes.

Run different Spark clusters with different versions at the same time.
Outlook

› Kubernetes Spark Scheduling in development (SPARK-18278)

› Multiple options for access and ingest

› Authorization / Authentication (RBAC)
Q&A
GitHub Resources
GitHub Resources

› GitHub-Repo with code for assembling k8s on AWS
  › https://github.com/inovex/kubernetes-demo

› GitHub-Repo with code for assembling Spark on k8s
  › https://github.com/inovex/spark-k8s-strata
Vielen Dank

Sascha Askani
Senior Systems Engineer

Daniel Bäurer
Head of Operations

inovex GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe