AWS re:Invent

ANT224-R

Using containers with Amazon EMR and Apache Spark

Matthew Liem

Big Data Architect, Amazon EMR

Amazon Web Services

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Hadoop 3 Docker support

Docker registries

Workshop: Using Amazon EMR Spark with Amazon ECR

- Getting started with Amazon ECR
- Launch and configure Amazon EMR with Amazon ECR registry
- Creating a Docker image
- SparkR and PySpark example

Q&A

Hadoop 3 Docker support

- Available starting in Amazon EMR 6.0.0 with Spark 2.4.3 and Hadoop 3.1.0

- Why Docker on Amazon EMR?
  - Reduce complexity – Bundle libraries and dependencies with application
  - Increased utilization – Run different versions and applications on the same cluster
  - Increased agility – Test and put new versions of software into production quickly
  - Application portability – Run on multiple OS, keeping user execution environment unchanged

Docker registries

A storage and distribution system for Docker images. For Amazon EMR 6.0.0 (beta), the following Docker registries can be configured:

- Docker Hub: A public Docker registry containing over 100,000 popular Docker images
- Amazon ECR: A fully managed Docker container registry that allows you to create your own custom images and host them in a highly available and scalable architecture

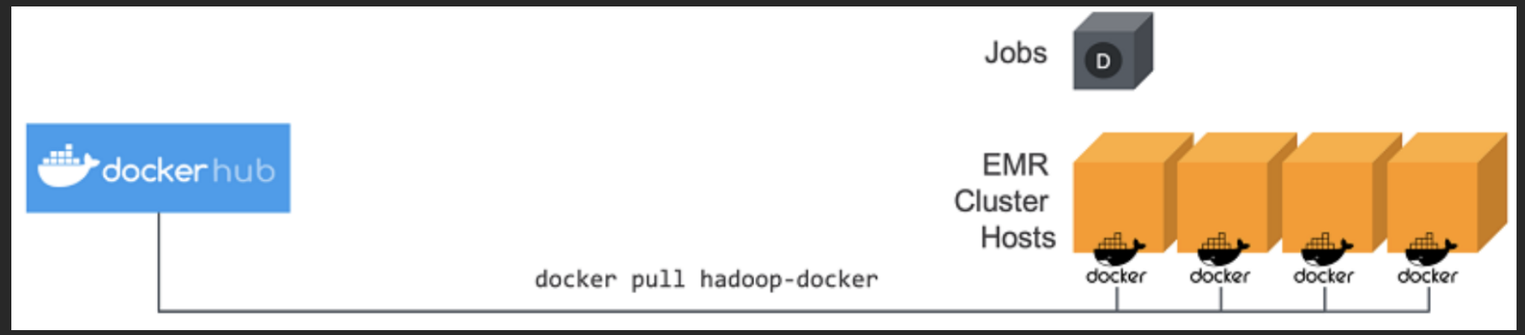
Deployment considerations

- Docker registries require network access from each host in the cluster
- How you choose to deploy your EMR cluster may limit your choice of Docker registry

Docker registries

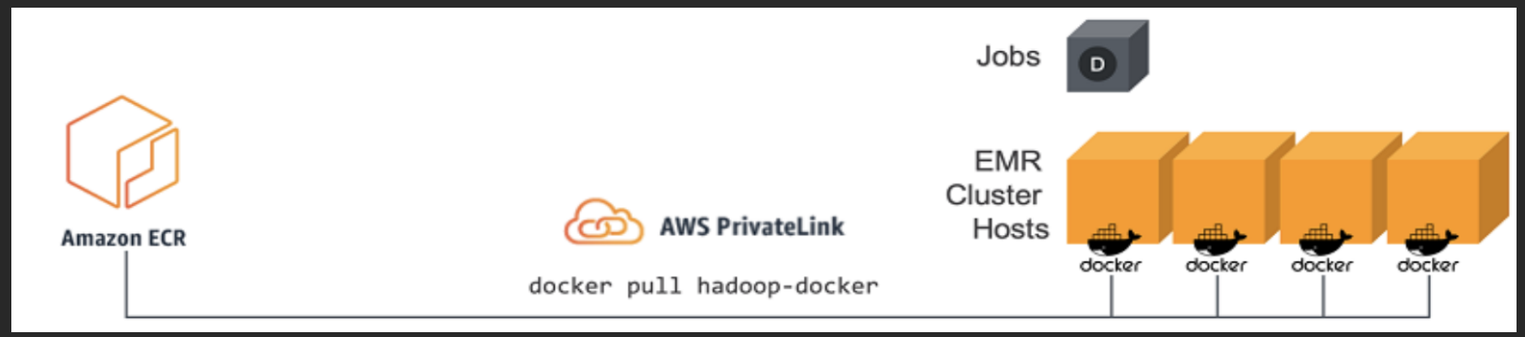
Public subnet

- YARN can directly access any registry available over the internet, such as Docker Hub



Private subnet

- Docker images can be hosted in the Amazon ECR and accessed through AWS PrivateLink



Workshop : Using Amazon EMR Spark with Docker

Getting started with Amazon ECR

Launch and configure Amazon EMR with Amazon ECR registry

Creating a Docker image

SparkR and PySpark example with Docker image

Resources for migrating to Amazon EMR

- Detailed migration guide
- Free on-site two-day workshop
  - Deconstruction of workloads
  - Future-state architecture
  - Migration plan and recommended next steps
  - AWS Professional Services available for implementation

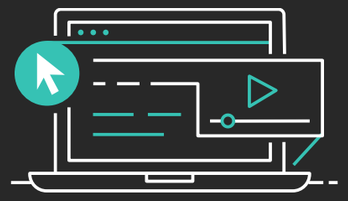


- Visit aws.amazon.com/emr/emr-migration/
- Email emr-migration-help@amazon.com



Learn big data with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills



New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, and Amazon Redshift



Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs



Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws training and certification

Thank you!

Matthew Liem

mattliem@amazon.com

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey in the mobile app.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.