# Managing and Analyzing Big Data in the Cloud

Klaus M. Miller
Goethe University Frankfurt

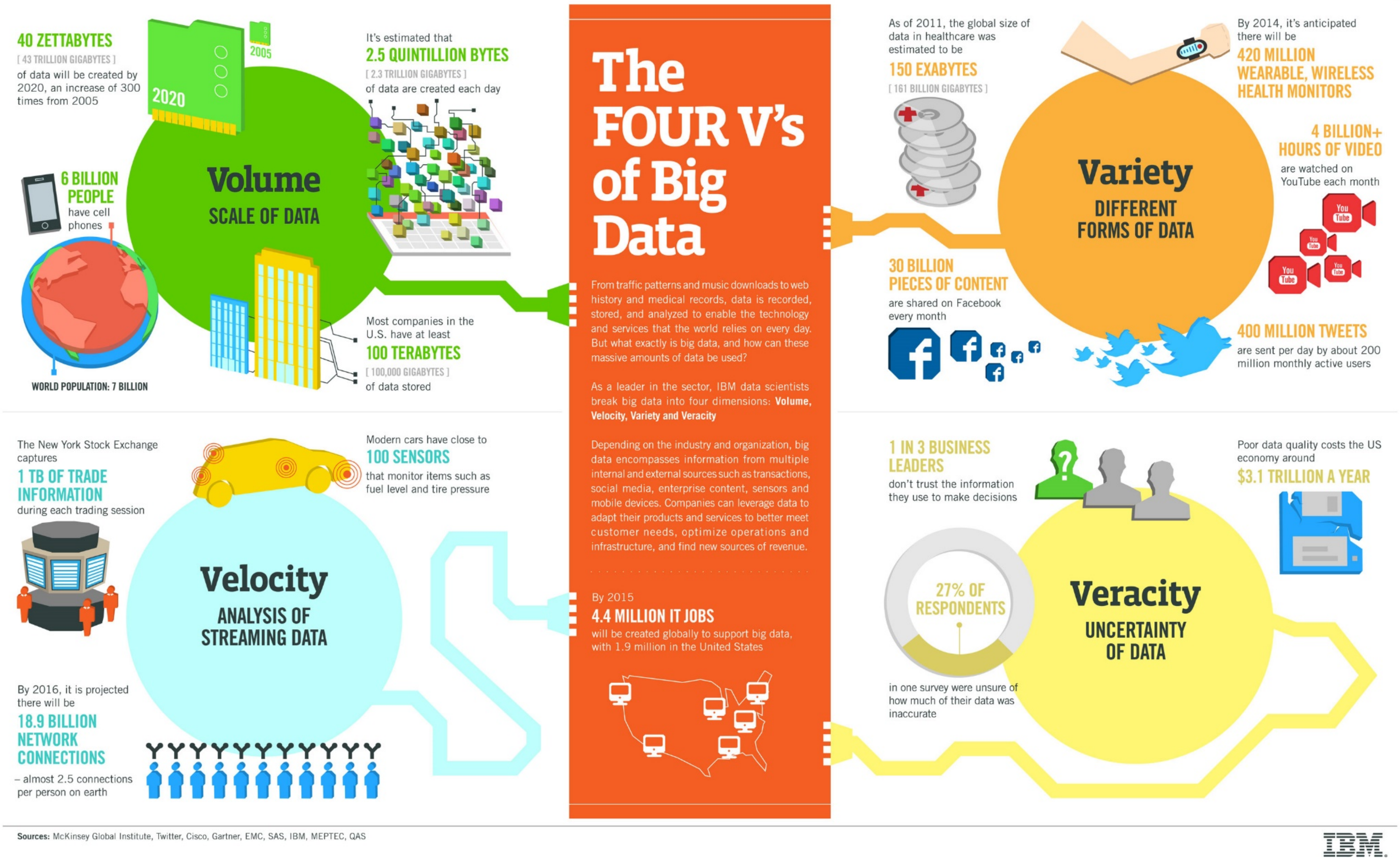
German Academic Association for Business Research
Annual Meeting

Frankfurt, March 18th, 2020

Materials available at: https://github.com/stm/vhb_2020

# Description of Problem

# Definition of Big Data



## The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
**4.4 MILLION IT JOBS**
will be created globally to support big data, with 1.9 million in the United States

### Volume
SCALE OF DATA

**40 ZETTABYTES**
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005

**6 BILLION PEOPLE**
have cell phones

It's estimated that
**2.5 QUINTILLION BYTES**
[ 2.3 TRILLION GIGABYTES ]
of data are created each day

Most companies in the U.S. have at least
**100 TERABYTES**
[ 100,000 GIGABYTES ]
of data stored

### Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
**150 EXABYTES**
[ 161 BILLION GIGABYTES ]

**30 BILLION PIECES OF CONTENT**
are shared on Facebook every month

By 2014, it's anticipated there will be
**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**
are watched on YouTube each month

**400 MILLION TWEETS**
are sent per day by about 200 million monthly active users

### Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
**1 TB OF TRADE INFORMATION**
during each trading session

Modern cars have close to
**100 SENSORS**
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
**18.9 BILLION NETWORK CONNECTIONS**
– almost 2.5 connections per person on earth

### Veracity
UNCERTAINTY OF DATA

**1 IN 3 BUSINESS LEADERS**
don't trust the information they use to make decisions

Poor data quality costs the US economy around
**$3.1 TRILLION A YEAR**

**27% OF RESPONDENTS**
in one survey were unsure of how much of their data was inaccurate

**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

# Definition of Big Data

Data is big anytime it makes you feel it is.

# Example of Big Data: Cookie Data Set (I)

Data Provider: Large European Ad Exchange

- 84% reach of internet users in relevant market
- Desktop and mobile browsing traffic

Observation Period:

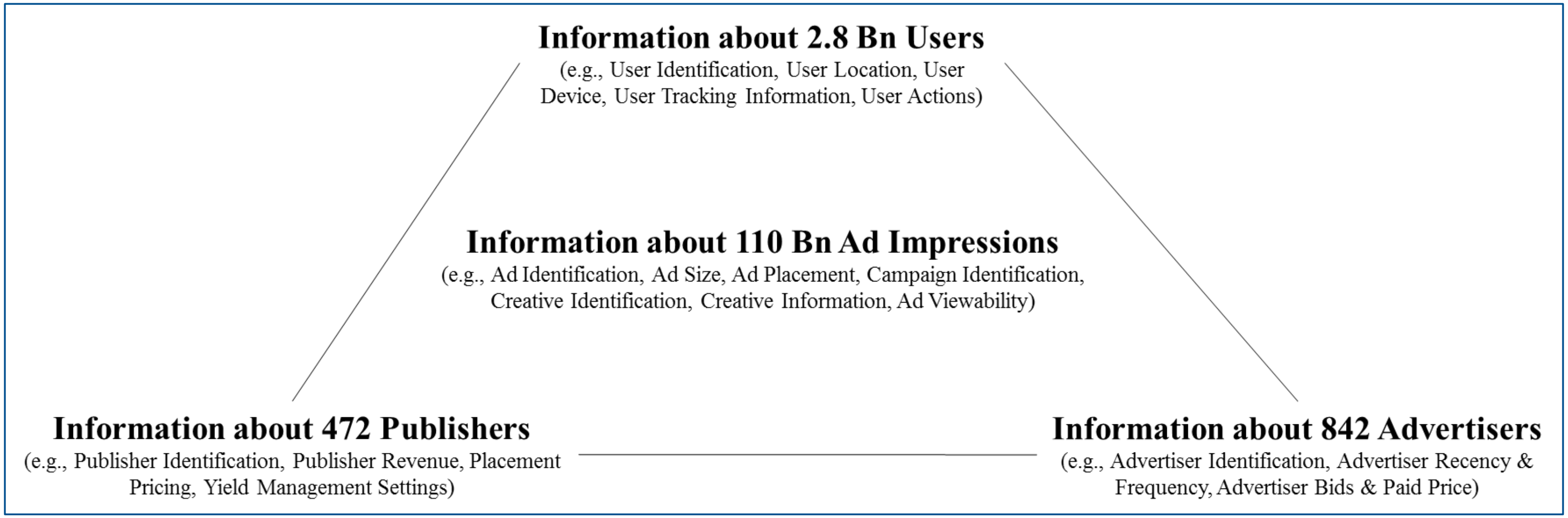
- ~2.5 years

Dimensions

- Log-level data set
- ~ 130 columns
- 550 million auctions (= rows) per day
- Total size: 60-65 TB

# Example of Big Data: Cookie Data Set (II)

## Overview of Type of Information Embedded in the Data



**Information about 2.8 Bn Users**
(e.g., User Identification, User Location, User Device, User Tracking Information, User Actions)

**Information about 110 Bn Ad Impressions**
(e.g., Ad Identification, Ad Size, Ad Placement, Campaign Identification, Creative Identification, Creative Information, Ad Viewability)

**Information about 472 Publishers**
(e.g., Publisher Identification, Publisher Revenue, Placement Pricing, Yield Management Settings)

**Information about 842 Advertisers**
(e.g., Advertiser Identification, Advertiser Recency & Frequency, Advertiser Bids & Paid Price)

# Possibilities to Deal with Large Data Sets in R

1. Allocate more memory („memory.limit“)

2. Vectorize (use „apply“ family instead of „for loops“)

3. Collect garbage („gc“)

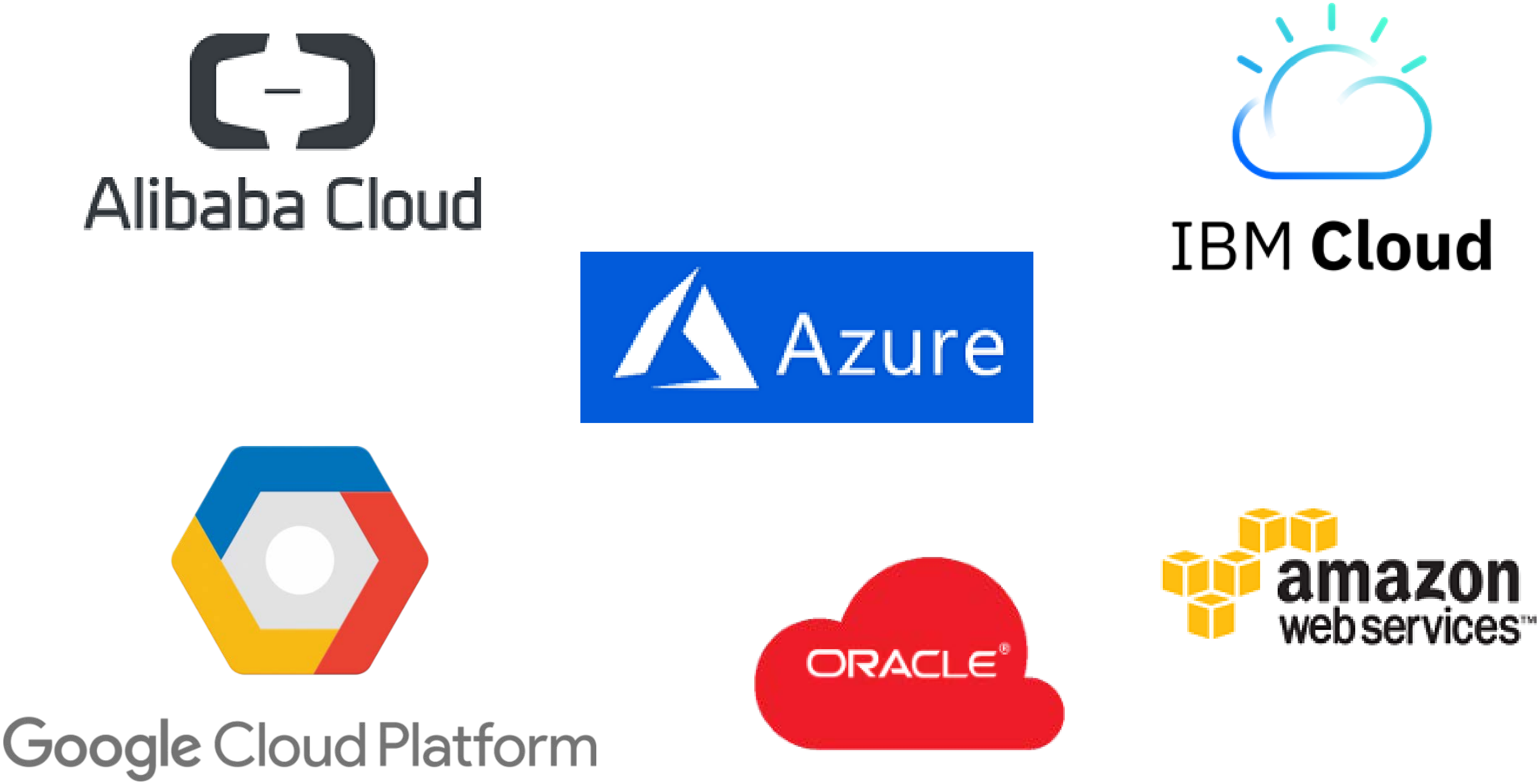
4. Parallelize („parApply“, „doParallel“)

5. Use Command Line Interface (e.g., GIT Bash)

6. Scale to the Cloud

# Scaling to the Cloud

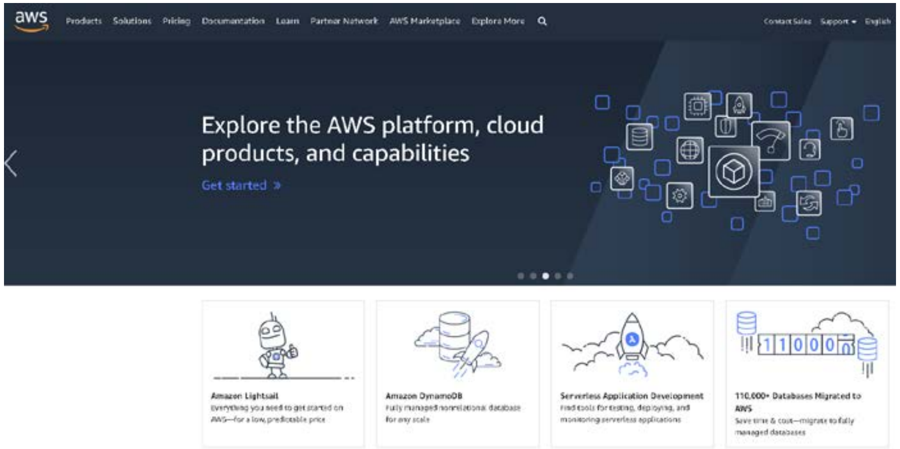
# Cloud Services



# Creating an AWS Account

- Free Basic Account and Credits
- Choose Region (e.g., EU (Frankfurt))
- Access AWS Services:
  - Data Storage (S3)
  - Elastic Map Reduce (EMR)

- AWS Educate for Usage in Class



Explore the AWS platform, cloud products, and capabilities

**Amazon Lightsail**
Everything you need to get started on AWS—for a low, predictable price

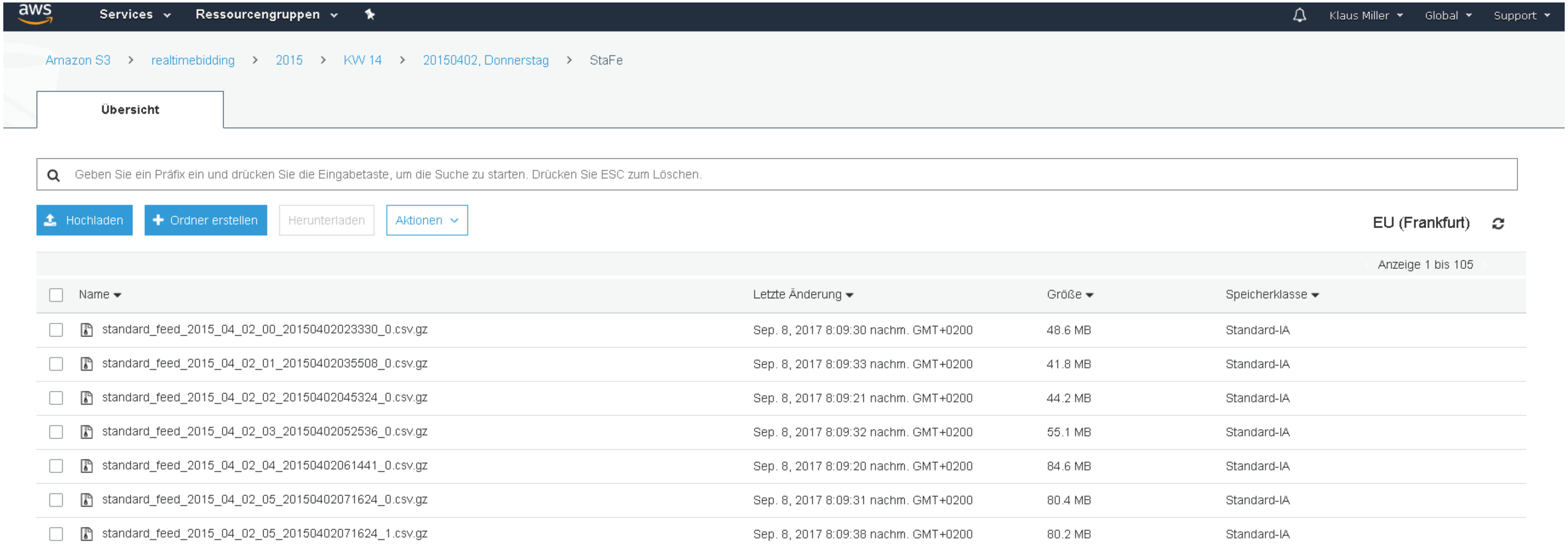
**Amazon DynamoDB**
Fully managed nonrelational database for any scale

**Serverless Application Development**
Find tools for testing, deploying, and monitoring serverless applications

**110,000+ Databases Migrated to AWS**
Save time & cost—migrate to fully managed databases

https://aws.amazon.com/

# Data Storage: AWS S3



Amazon S3 > realtimebidding > 2015 > KW 14 > 20150402, Donnerstag > StaFe

Geben Sie ein Präfix ein und drücken Sie die Eingabetaste, um die Suche zu starten. Drücken Sie ESC zum Löschen.

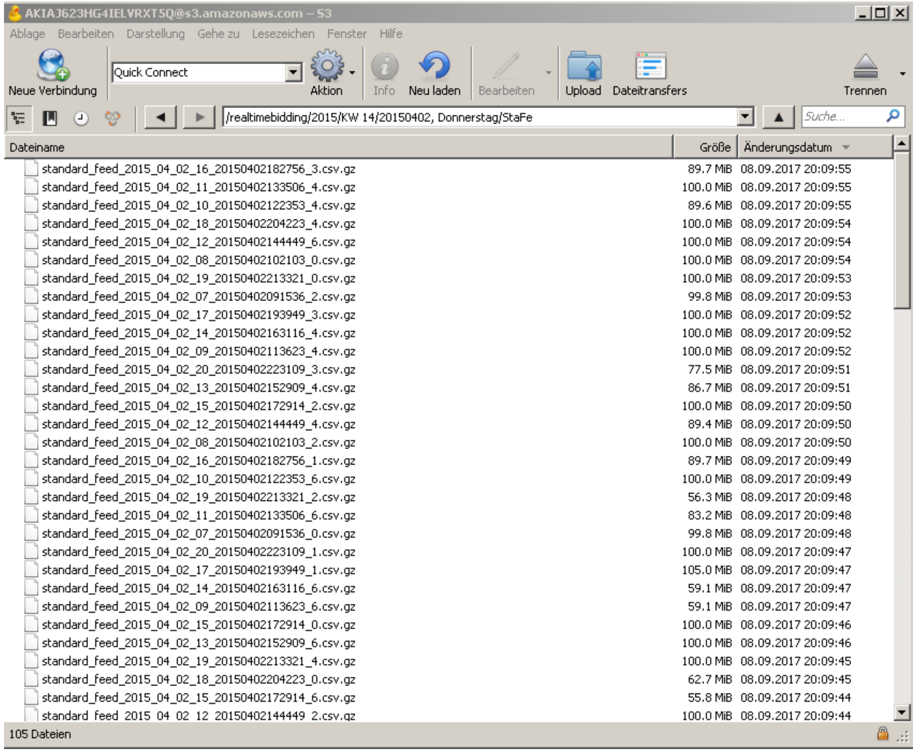
EU (Frankfurt)

Anzeige 1 bis 105

| | Name | Letzte Änderung | Größe | Speicherklasse |
|---|---|---|---|---|
| □ | standard_feed_2015_04_02_00_20150402023330_0.csv.gz | Sep. 8, 2017 8:09:30 nachm. GMT+0200 | 48.6 MB | Standard-IA |
| □ | standard_feed_2015_04_02_01_20150402035508_0.csv.gz | Sep. 8, 2017 8:09:33 nachm. GMT+0200 | 41.8 MB | Standard-IA |
| □ | standard_feed_2015_04_02_02_20150402045324_0.csv.gz | Sep. 8, 2017 8:09:21 nachm. GMT+0200 | 44.2 MB | Standard-IA |
| □ | standard_feed_2015_04_02_03_20150402052536_0.csv.gz | Sep. 8, 2017 8:09:32 nachm. GMT+0200 | 55.1 MB | Standard-IA |
| □ | standard_feed_2015_04_02_04_20150402061441_0.csv.gz | Sep. 8, 2017 8:09:20 nachm. GMT+0200 | 84.6 MB | Standard-IA |
| □ | standard_feed_2015_04_02_05_20150402071624_0.csv.gz | Sep. 8, 2017 8:09:31 nachm. GMT+0200 | 80.4 MB | Standard-IA |
| □ | standard_feed_2015_04_02_05_20150402071624_1.csv.gz | Sep. 8, 2017 8:09:38 nachm. GMT+0200 | 80.2 MB | Standard-IA |

# Data Access

- AWS Console Access (Browser)
- API Access
  - Data sharing with collaborators worldwide
  - Obtaining data from data providers
- Cloud Storage Browsers (e.g., S3 Browser, Cyberduck)

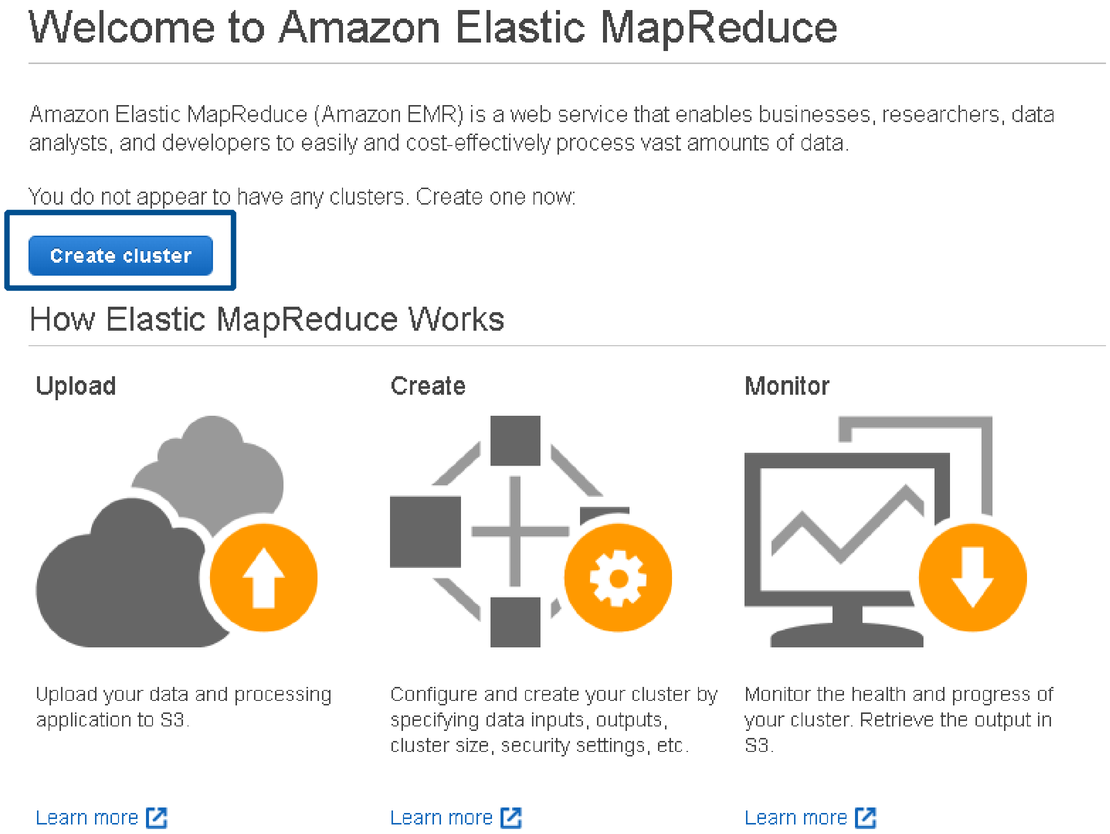


/realtimebidding/2015/KW 14/20150402, Donnerstag/StaFe

| Dateiname | Größe | Änderungsdatum |
|---|---|---|
| standard_feed_2015_04_02_16_20150402182756_3.csv.gz | 89.7 MiB | 08.09.2017 20:09:55 |
| standard_feed_2015_04_02_11_20150402133506_4.csv.gz | 100.0 MiB | 08.09.2017 20:09:55 |
| standard_feed_2015_04_02_10_20150402122353_4.csv.gz | 89.6 MiB | 08.09.2017 20:09:55 |
| standard_feed_2015_04_02_18_20150402204223_4.csv.gz | 100.0 MiB | 08.09.2017 20:09:54 |
| standard_feed_2015_04_02_12_20150402144449_6.csv.gz | 100.0 MiB | 08.09.2017 20:09:54 |
| standard_feed_2015_04_02_08_20150402102103_0.csv.gz | 100.0 MiB | 08.09.2017 20:09:54 |
| standard_feed_2015_04_02_19_20150402213321_0.csv.gz | 100.0 MiB | 08.09.2017 20:09:53 |
| standard_feed_2015_04_02_07_20150402091536_2.csv.gz | 99.8 MiB | 08.09.2017 20:09:53 |
| standard_feed_2015_04_02_17_20150402193949_3.csv.gz | 100.0 MiB | 08.09.2017 20:09:52 |
| standard_feed_2015_04_02_14_20150402163116_4.csv.gz | 100.0 MiB | 08.09.2017 20:09:52 |
| standard_feed_2015_04_02_09_20150402113623_4.csv.gz | 100.0 MiB | 08.09.2017 20:09:52 |
| standard_feed_2015_04_02_20_20150402223109_3.csv.gz | 77.5 MiB | 08.09.2017 20:09:51 |
| standard_feed_2015_04_02_13_20150402152909_4.csv.gz | 86.7 MiB | 08.09.2017 20:09:51 |
| standard_feed_2015_04_02_15_20150402172914_2.csv.gz | 100.0 MiB | 08.09.2017 20:09:50 |
| standard_feed_2015_04_02_12_20150402144449_4.csv.gz | 89.4 MiB | 08.09.2017 20:09:50 |
| standard_feed_2015_04_02_08_20150402102103_2.csv.gz | 100.0 MiB | 08.09.2017 20:09:50 |
| standard_feed_2015_04_02_16_20150402182756_1.csv.gz | 89.7 MiB | 08.09.2017 20:09:49 |
| standard_feed_2015_04_02_10_20150402122353_6.csv.gz | 100.0 MiB | 08.09.2017 20:09:49 |
| standard_feed_2015_04_02_19_20150402213321_2.csv.gz | 56.3 MiB | 08.09.2017 20:09:48 |
| standard_feed_2015_04_02_11_20150402133506_6.csv.gz | 83.2 MiB | 08.09.2017 20:09:48 |
| standard_feed_2015_04_02_07_20150402091536_0.csv.gz | 99.8 MiB | 08.09.2017 20:09:48 |
| standard_feed_2015_04_02_20_20150402223109_1.csv.gz | 100.0 MiB | 08.09.2017 20:09:47 |
| standard_feed_2015_04_02_17_20150402193949_1.csv.gz | 105.0 MiB | 08.09.2017 20:09:47 |
| standard_feed_2015_04_02_14_20150402163116_6.csv.gz | 59.1 MiB | 08.09.2017 20:09:47 |
| standard_feed_2015_04_02_09_20150402113623_6.csv.gz | 59.1 MiB | 08.09.2017 20:09:47 |
| standard_feed_2015_04_02_15_20150402172914_0.csv.gz | 100.0 MiB | 08.09.2017 20:09:46 |
| standard_feed_2015_04_02_13_20150402152909_6.csv.gz | 100.0 MiB | 08.09.2017 20:09:46 |
| standard_feed_2015_04_02_19_20150402213321_4.csv.gz | 100.0 MiB | 08.09.2017 20:09:45 |
| standard_feed_2015_04_02_18_20150402204223_0.csv.gz | 62.7 MiB | 08.09.2017 20:09:45 |
| standard_feed_2015_04_02_15_20150402172914_6.csv.gz | 55.8 MiB | 08.09.2017 20:09:44 |
| standard_feed_2015_04_02_12_20150402144449_2.csv.gz | 100.0 MiB | 08.09.2017 20:09:44 |

105 Dateien

# Create EMR Cluster



## Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

## How Elastic MapReduce Works

### Upload

Upload your data and processing application to S3.

Learn more

### Create

Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

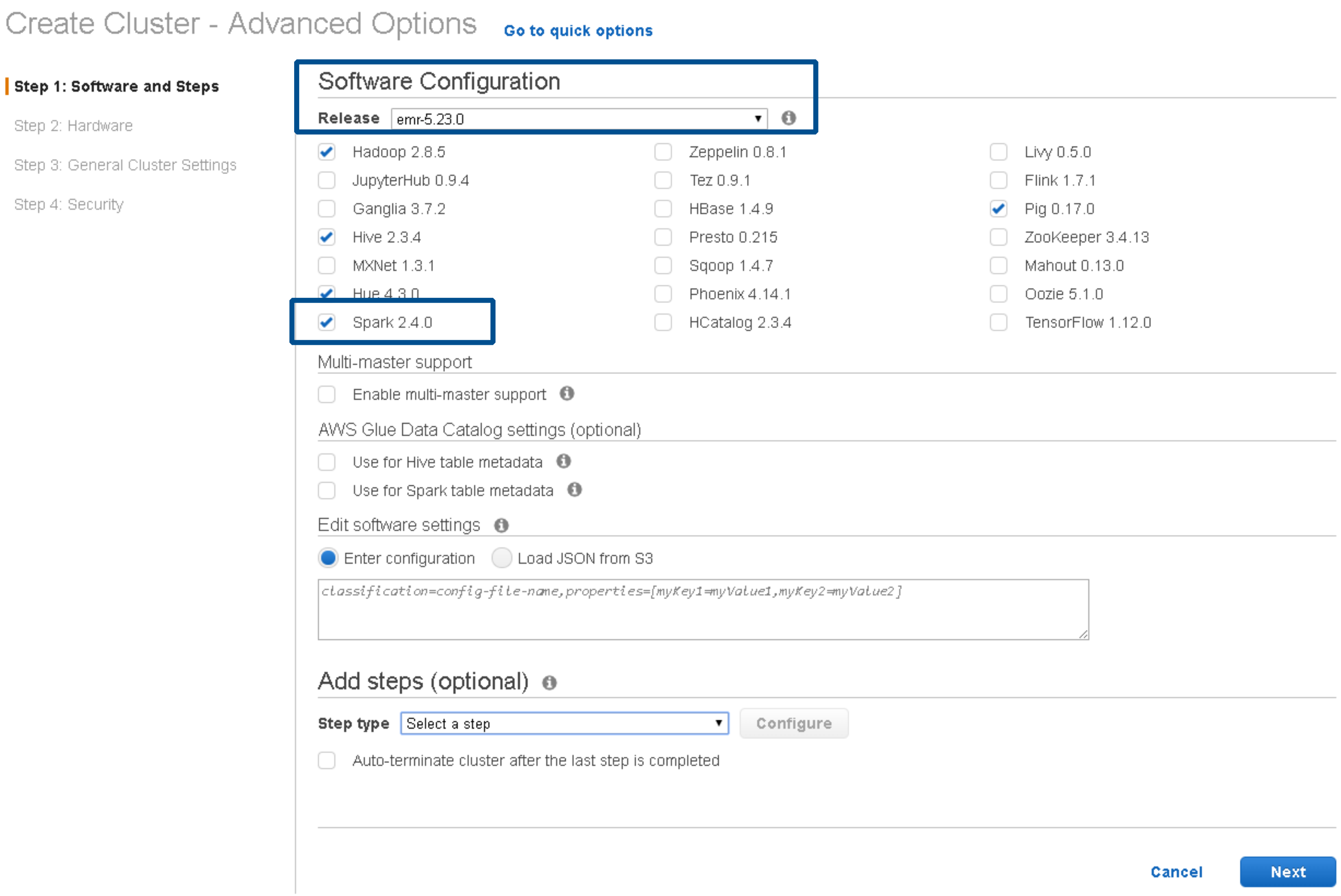
Learn more

### Monitor

Monitor the health and progress of your cluster. Retrieve the output in S3.

Learn more

# Step 1: Software and Steps



## Create Cluster - Advanced Options

**Step 1: Software and Steps**
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

### Software Configuration

**Release** emr-5.23.0

- ☑ Hadoop 2.8.5
- ☐ JupyterHub 0.9.4
- ☐ Ganglia 3.7.2
- ☑ Hive 2.3.4
- ☐ MXNet 1.3.1
- ☑ Hue 4.3.0
- ☑ Spark 2.4.0
- ☐ Zeppelin 0.8.1
- ☐ Tez 0.9.1
- ☐ HBase 1.4.9
- ☐ Presto 0.215
- ☐ Sqoop 1.4.7
- ☐ Phoenix 4.14.1
- ☐ HCatalog 2.3.4
- ☐ Livy 0.5.0
- ☐ Flink 1.7.1
- ☑ Pig 0.17.0
- ☐ ZooKeeper 3.4.13
- ☐ Mahout 0.13.0
- ☐ Oozie 5.1.0
- ☐ TensorFlow 1.12.0

Multi-master support

- ☐ Enable multi-master support

AWS Glue Data Catalog settings (optional)

- ☐ Use for Hive table metadata
- ☐ Use for Spark table metadata

Edit software settings

● Enter configuration ○ Load JSON from S3

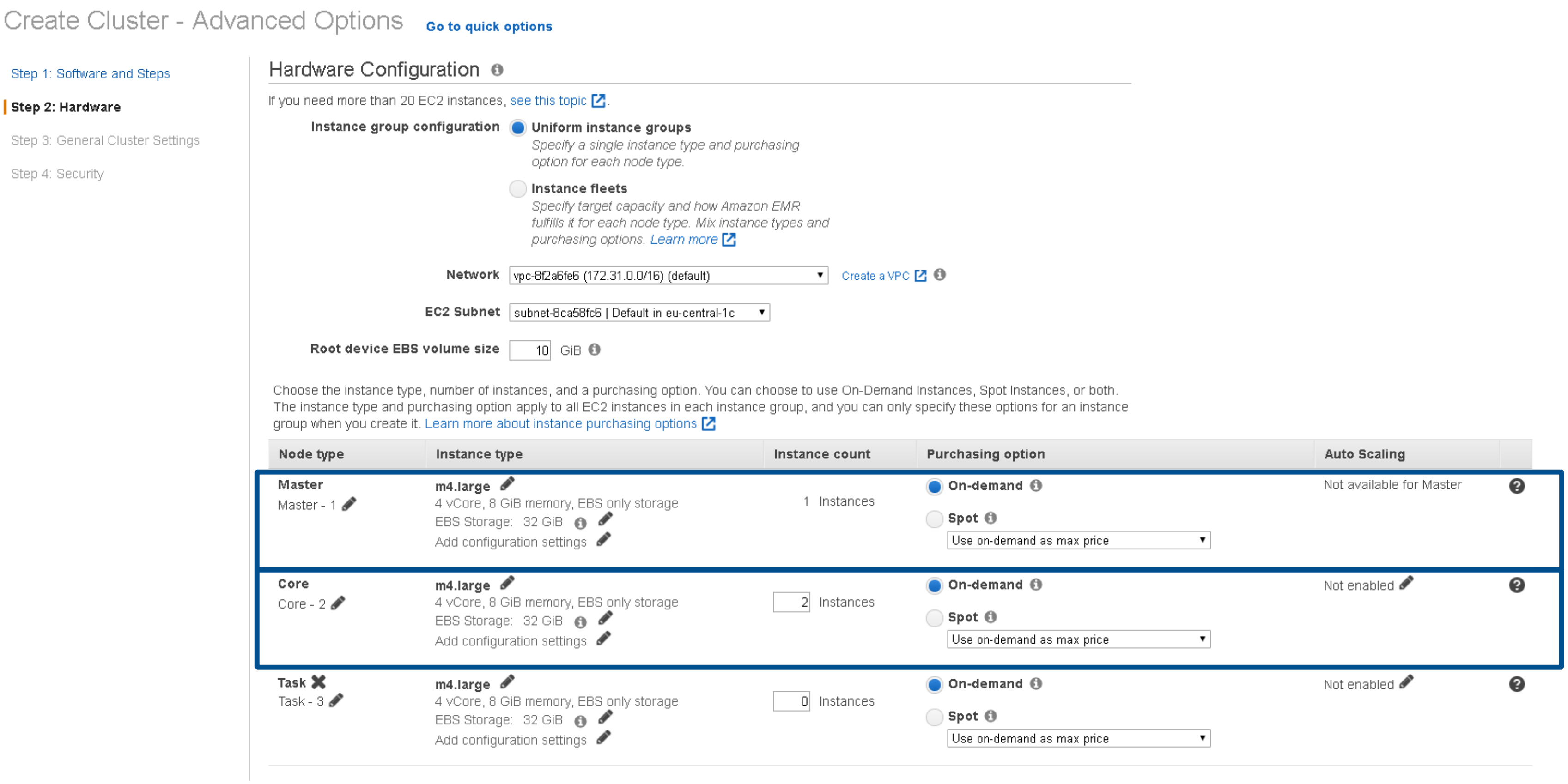
```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

### Add steps (optional)

**Step type** Select a step

- ☐ Auto-terminate cluster after the last step is completed

# Step 2: Hardware



Create Cluster - Advanced Options   Go to quick options

Step 1: Software and Steps

**Step 2: Hardware**

Step 3: General Cluster Settings

Step 4: Security

## Hardware Configuration

If you need more than 20 EC2 instances, see this topic.

**Instance group configuration**

● **Uniform instance groups**
*Specify a single instance type and purchasing option for each node type.*

○ **Instance fleets**
*Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. Learn more*

**Network** vpc-8f2a6fe6 (172.31.0.0/16) (default)   Create a VPC

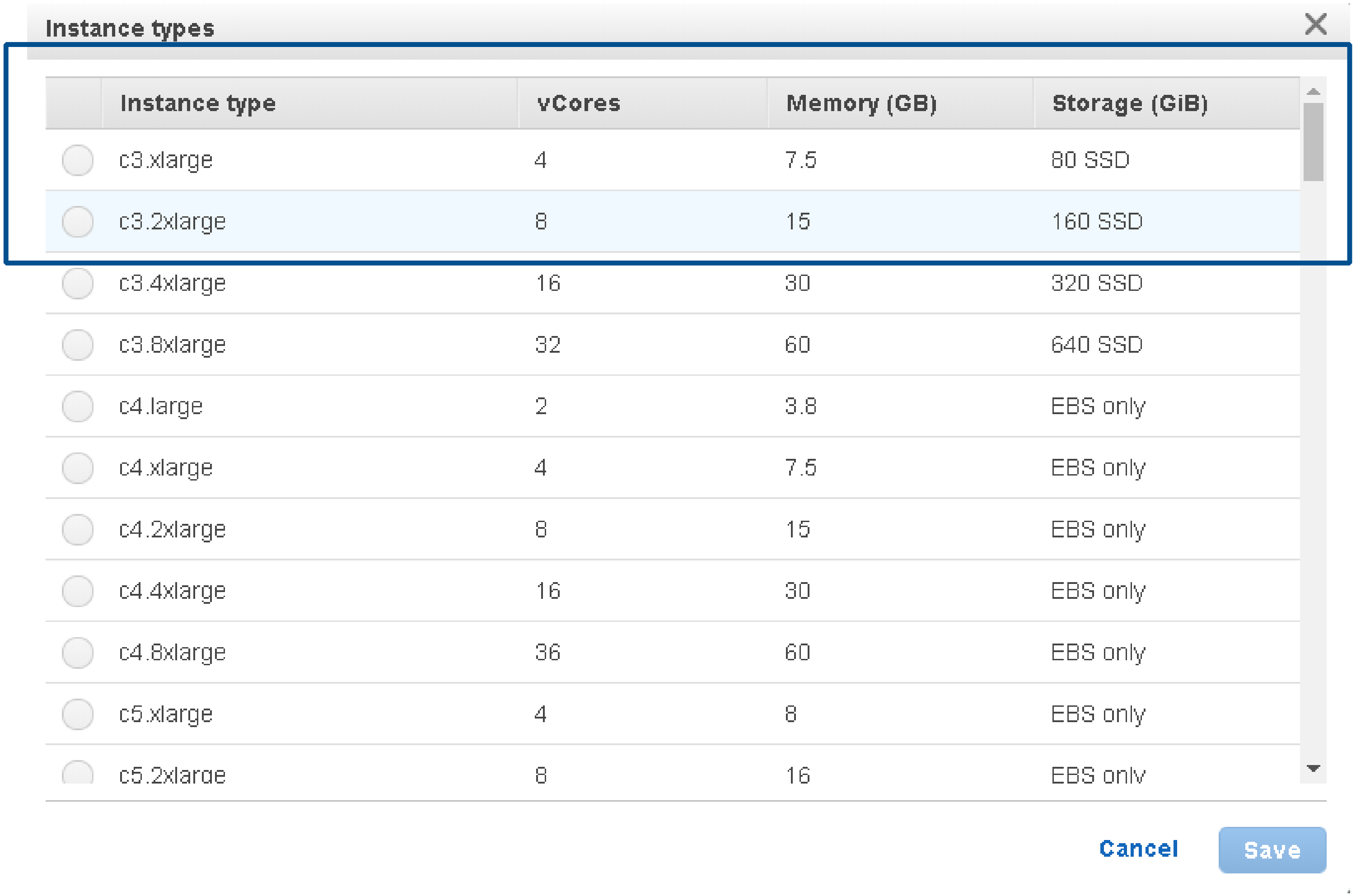
**EC2 Subnet** subnet-8ca58fc6 | Default in eu-central-1c

**Root device EBS volume size** 10 GiB

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. Learn more about instance purchasing options

| Node type | Instance type | Instance count | Purchasing option | Auto Scaling |
|---|---|---|---|---|
| **Master**<br>Master - 1 | **m4.large**<br>4 vCore, 8 GiB memory, EBS only storage<br>EBS Storage: 32 GiB<br>Add configuration settings | 1 Instances | ● **On-demand**<br>○ **Spot**<br>Use on-demand as max price | Not available for Master |
| **Core**<br>Core - 2 | **m4.large**<br>4 vCore, 8 GiB memory, EBS only storage<br>EBS Storage: 32 GiB<br>Add configuration settings | 2 Instances | ● **On-demand**<br>○ **Spot**<br>Use on-demand as max price | Not enabled |
| **Task** ✖<br>Task - 3 | **m4.large**<br>4 vCore, 8 GiB memory, EBS only storage<br>EBS Storage: 32 GiB<br>Add configuration settings | 0 Instances | ● **On-demand**<br>○ **Spot**<br>Use on-demand as max price | Not enabled |

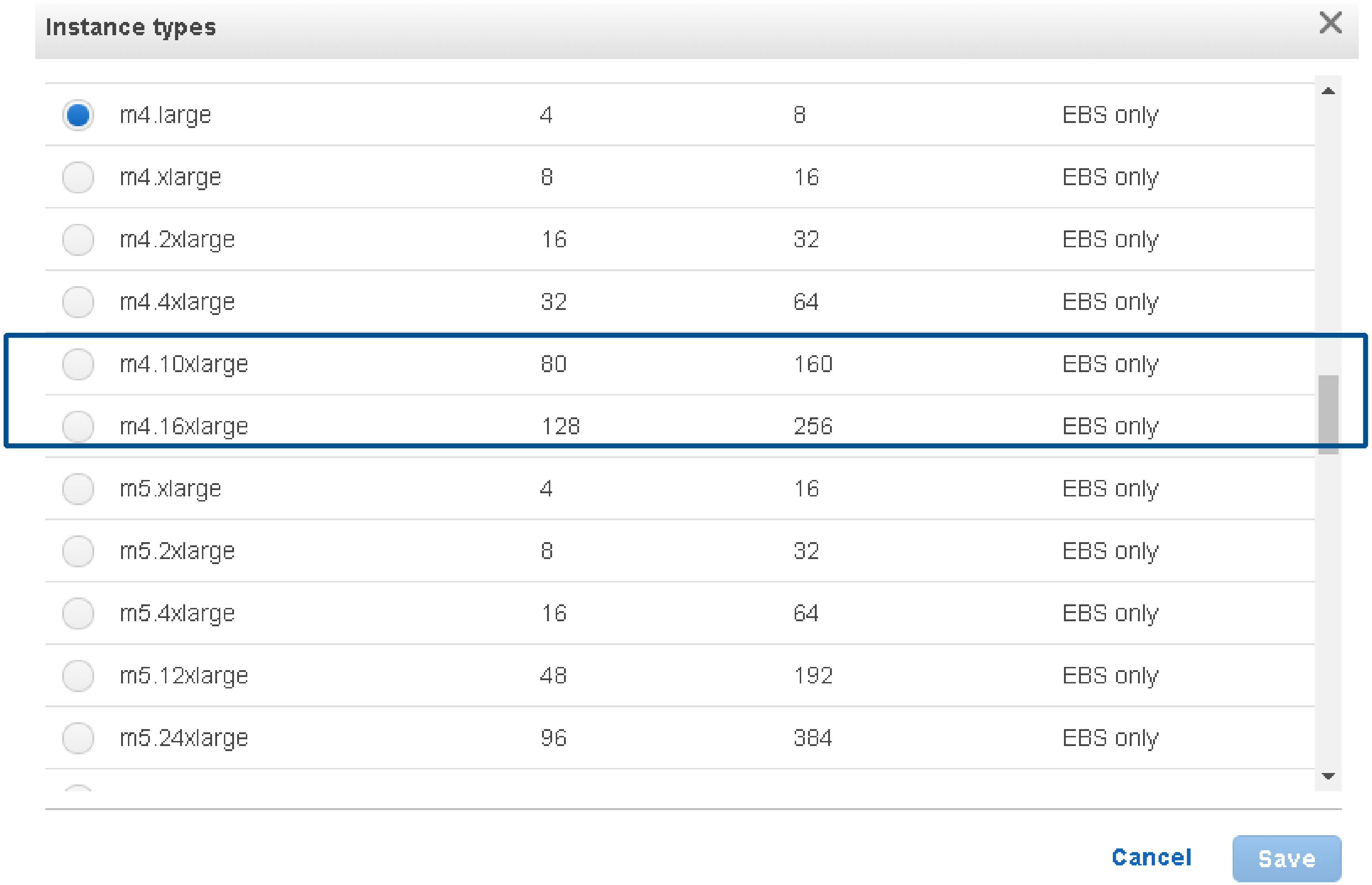
# Step 2: Hardware



Instance types

| | Instance type | vCores | Memory (GB) | Storage (GiB) |
|---|---|---|---|---|
| ○ | c3.xlarge | 4 | 7.5 | 80 SSD |
| ○ | c3.2xlarge | 8 | 15 | 160 SSD |
| ○ | c3.4xlarge | 16 | 30 | 320 SSD |
| ○ | c3.8xlarge | 32 | 60 | 640 SSD |
| ○ | c4.large | 2 | 3.8 | EBS only |
| ○ | c4.xlarge | 4 | 7.5 | EBS only |
| ○ | c4.2xlarge | 8 | 15 | EBS only |
| ○ | c4.4xlarge | 16 | 30 | EBS only |
| ○ | c4.8xlarge | 36 | 60 | EBS only |
| ○ | c5.xlarge | 4 | 8 | EBS only |
| ○ | c5.2xlarge | 8 | 16 | EBS only |

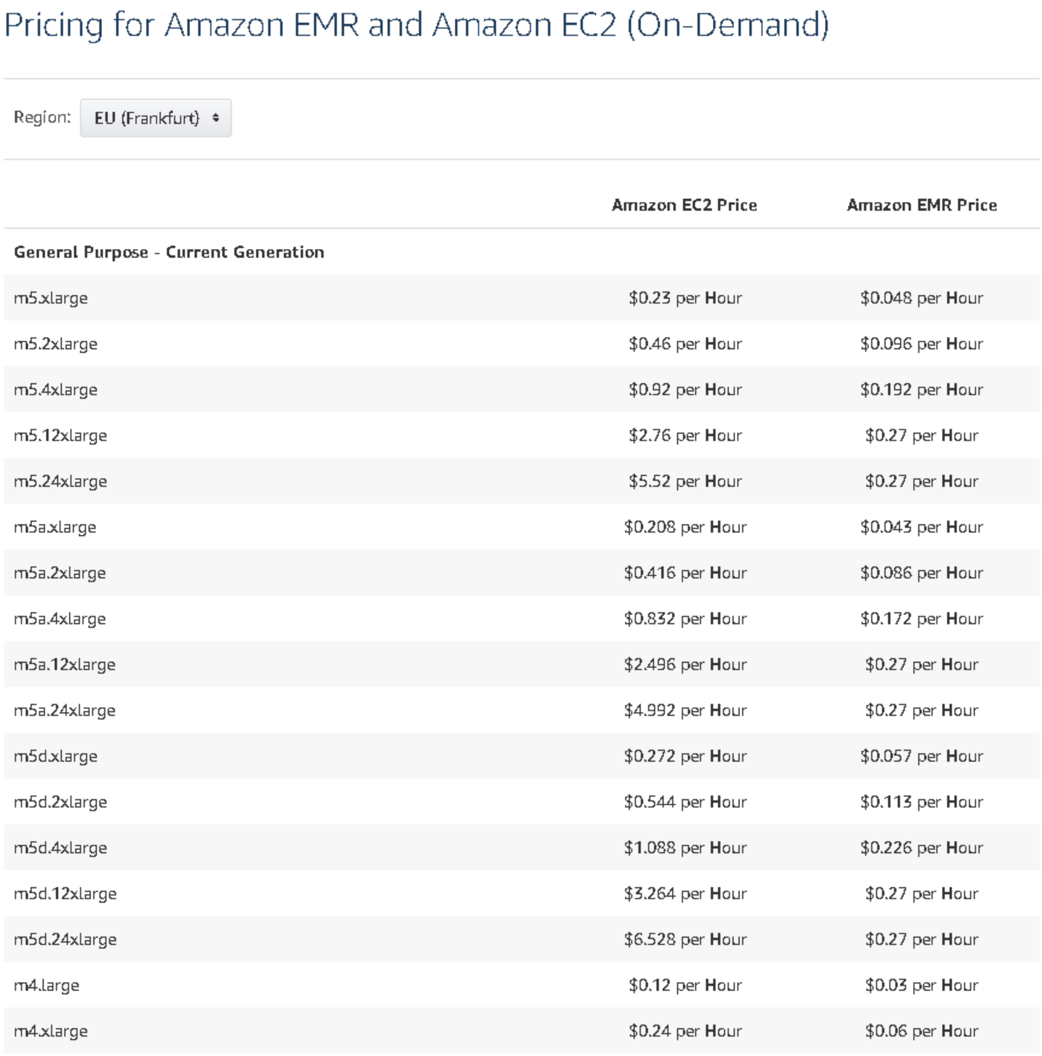
# Step 2: Hardware



Instance types

| | | | | |
|---|---|---|---|---|
| ● | m4.large | 4 | 8 | EBS only |
| ○ | m4.xlarge | 8 | 16 | EBS only |
| ○ | m4.2xlarge | 16 | 32 | EBS only |
| ○ | m4.4xlarge | 32 | 64 | EBS only |
| ○ | m4.10xlarge | 80 | 160 | EBS only |
| ○ | m4.16xlarge | 128 | 256 | EBS only |
| ○ | m5.xlarge | 4 | 16 | EBS only |
| ○ | m5.2xlarge | 8 | 32 | EBS only |
| ○ | m5.4xlarge | 16 | 64 | EBS only |
| ○ | m5.12xlarge | 48 | 192 | EBS only |
| ○ | m5.24xlarge | 96 | 384 | EBS only |

# Step 2: Hardware



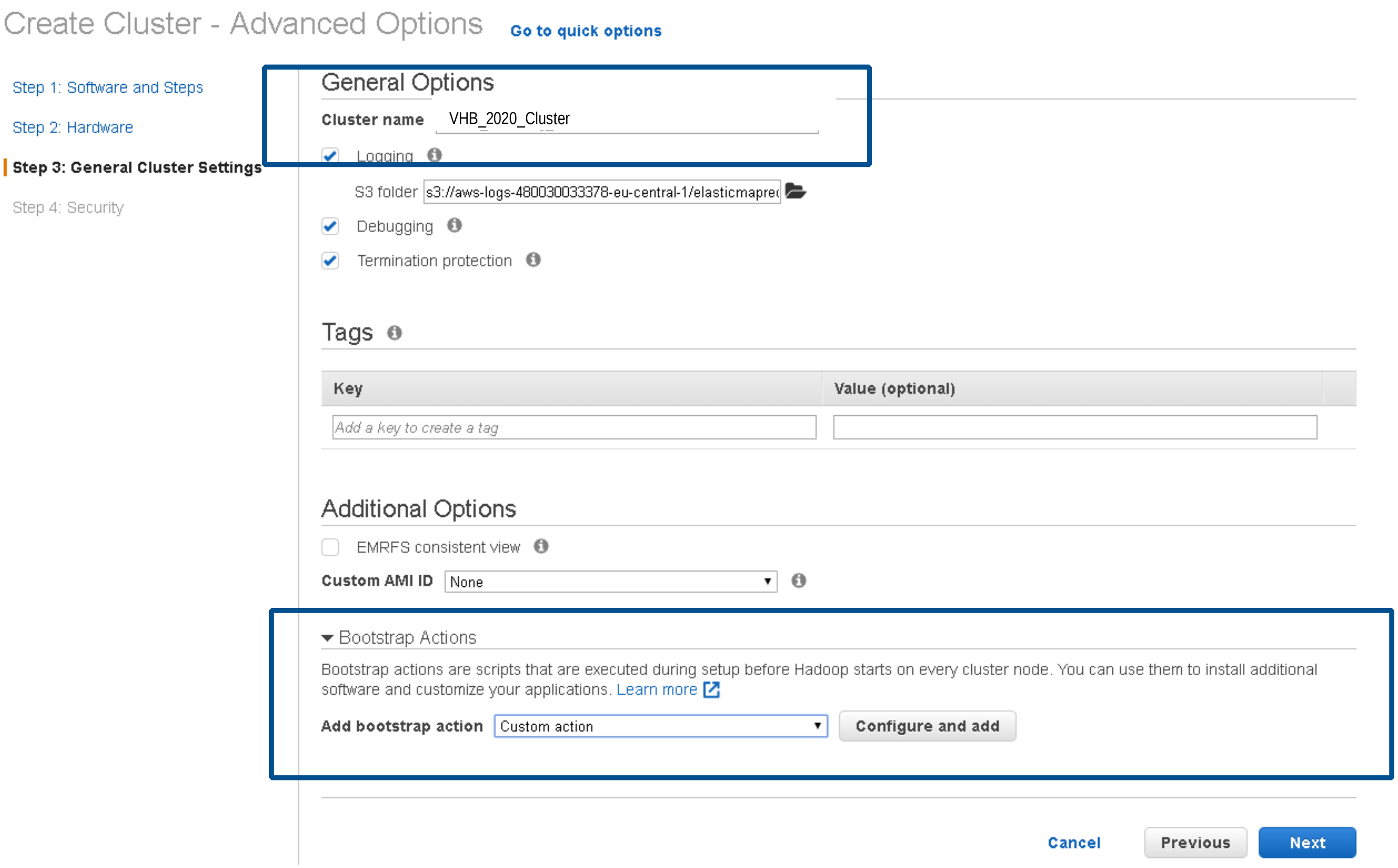
## Pricing for Amazon EMR and Amazon EC2 (On-Demand)

Region: EU (Frankfurt)

| | Amazon EC2 Price | Amazon EMR Price |
|---|---|---|
| **General Purpose - Current Generation** | | |
| m5.xlarge | $0.23 per Hour | $0.048 per Hour |
| m5.2xlarge | $0.46 per Hour | $0.096 per Hour |
| m5.4xlarge | $0.92 per Hour | $0.192 per Hour |
| m5.12xlarge | $2.76 per Hour | $0.27 per Hour |
| m5.24xlarge | $5.52 per Hour | $0.27 per Hour |
| m5a.xlarge | $0.208 per Hour | $0.043 per Hour |
| m5a.2xlarge | $0.416 per Hour | $0.086 per Hour |
| m5a.4xlarge | $0.832 per Hour | $0.172 per Hour |
| m5a.12xlarge | $2.496 per Hour | $0.27 per Hour |
| m5a.24xlarge | $4.992 per Hour | $0.27 per Hour |
| m5d.xlarge | $0.272 per Hour | $0.057 per Hour |
| m5d.2xlarge | $0.544 per Hour | $0.113 per Hour |
| m5d.4xlarge | $1.088 per Hour | $0.226 per Hour |
| m5d.12xlarge | $3.264 per Hour | $0.27 per Hour |
| m5d.24xlarge | $6.528 per Hour | $0.27 per Hour |
| m4.large | $0.12 per Hour | $0.03 per Hour |
| m4.xlarge | $0.24 per Hour | $0.06 per Hour |

Prices range between $.03 - $.27 for EMR

# Step 3: Cluster Settings



Create Cluster - Advanced Options

General Options

Cluster name: VHB_2020_Cluster

☑ Logging

S3 folder: s3://aws-logs-480030033378-eu-central-1/elasticmapred

☑ Debugging

☑ Termination protection

Tags

| Key | Value (optional) |
|---|---|
| Add a key to create a tag | |

Additional Options

□ EMRFS consistent view

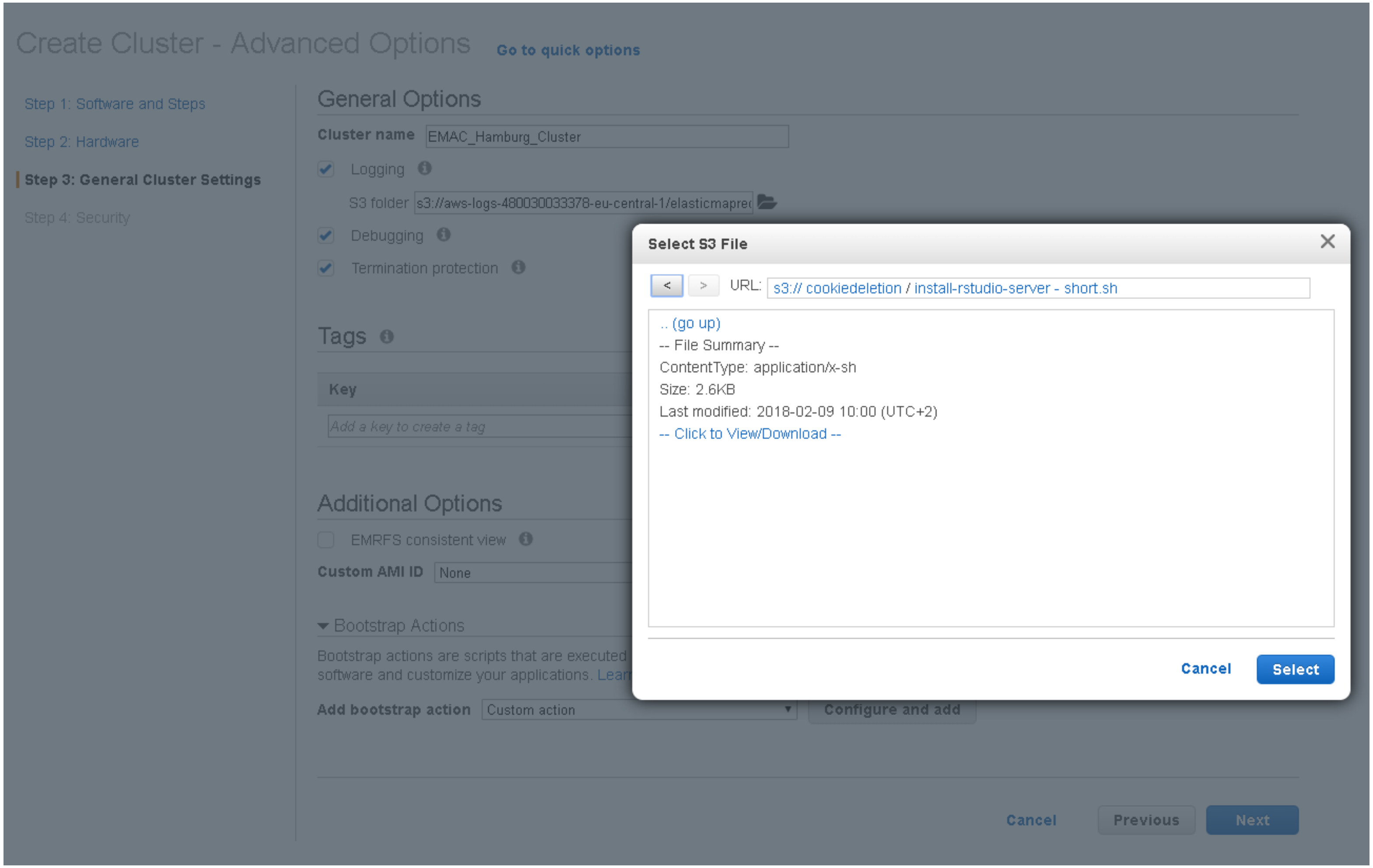
Custom AMI ID: None

Bootstrap Actions

Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. Learn more

Add bootstrap action: Custom action

# Step 3: Cluster Settings



Create Cluster - Advanced Options

General Options

Cluster name: EMAC_Hamburg_Cluster

☑ Logging

S3 folder: s3://aws-logs-480030033378-eu-central-1/elasticmapred

☑ Debugging

☑ Termination protection

Tags

Key

Add a key to create a tag

Additional Options

□ EMRFS consistent view

Custom AMI ID: None

Bootstrap Actions

Bootstrap actions are scripts that are executed
software and customize your applications. Lear

Add bootstrap action: Custom action

Select S3 File

URL: s3:// cookiedeletion / install-rstudio-server - short.sh

.. (go up)

-- File Summary --

ContentType: application/x-sh

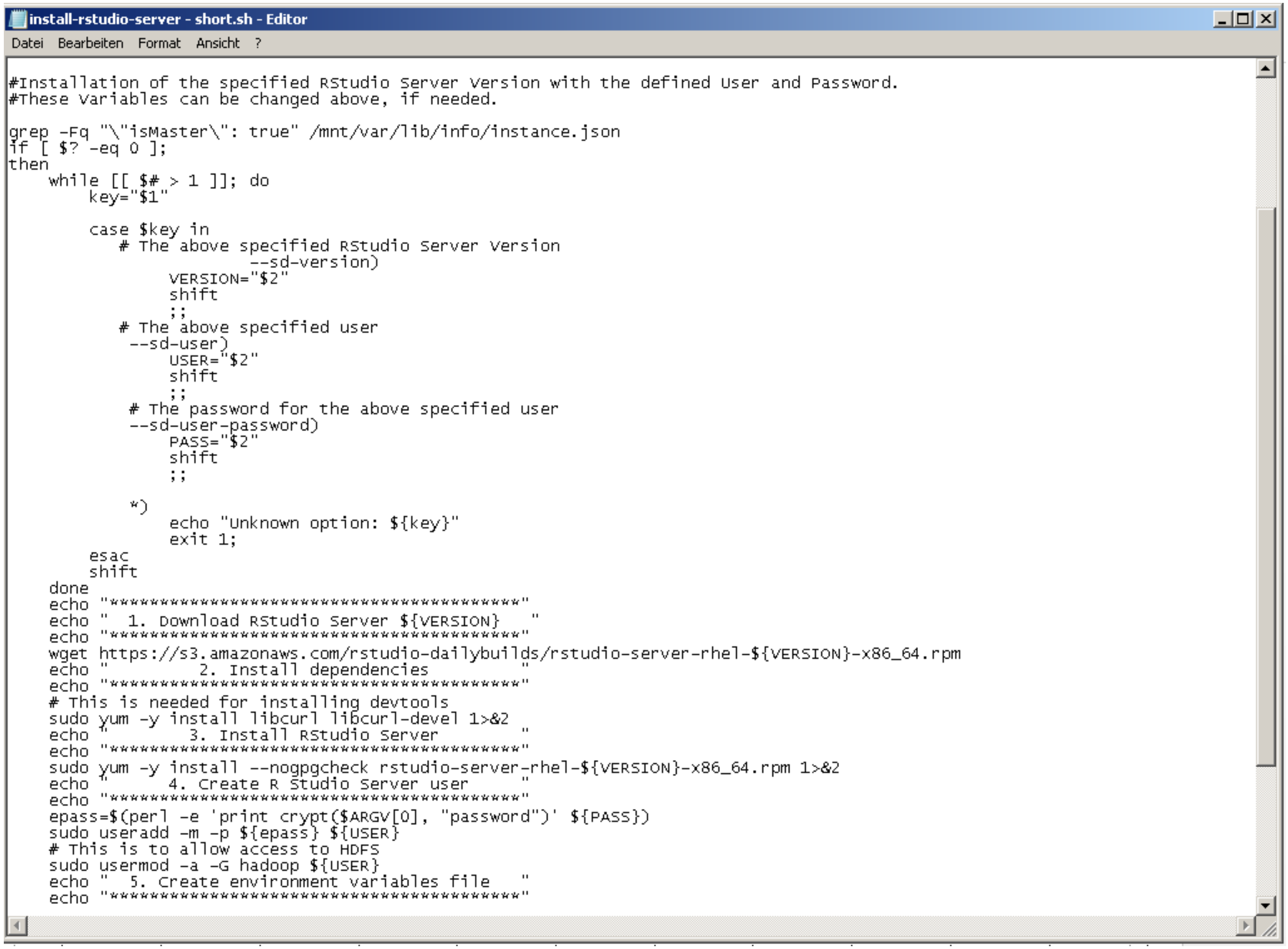
Size: 2.6KB

Last modified: 2018-02-09 10:00 (UTC+2)

-- Click to View/Download --

Install RStudioServer on Cluster

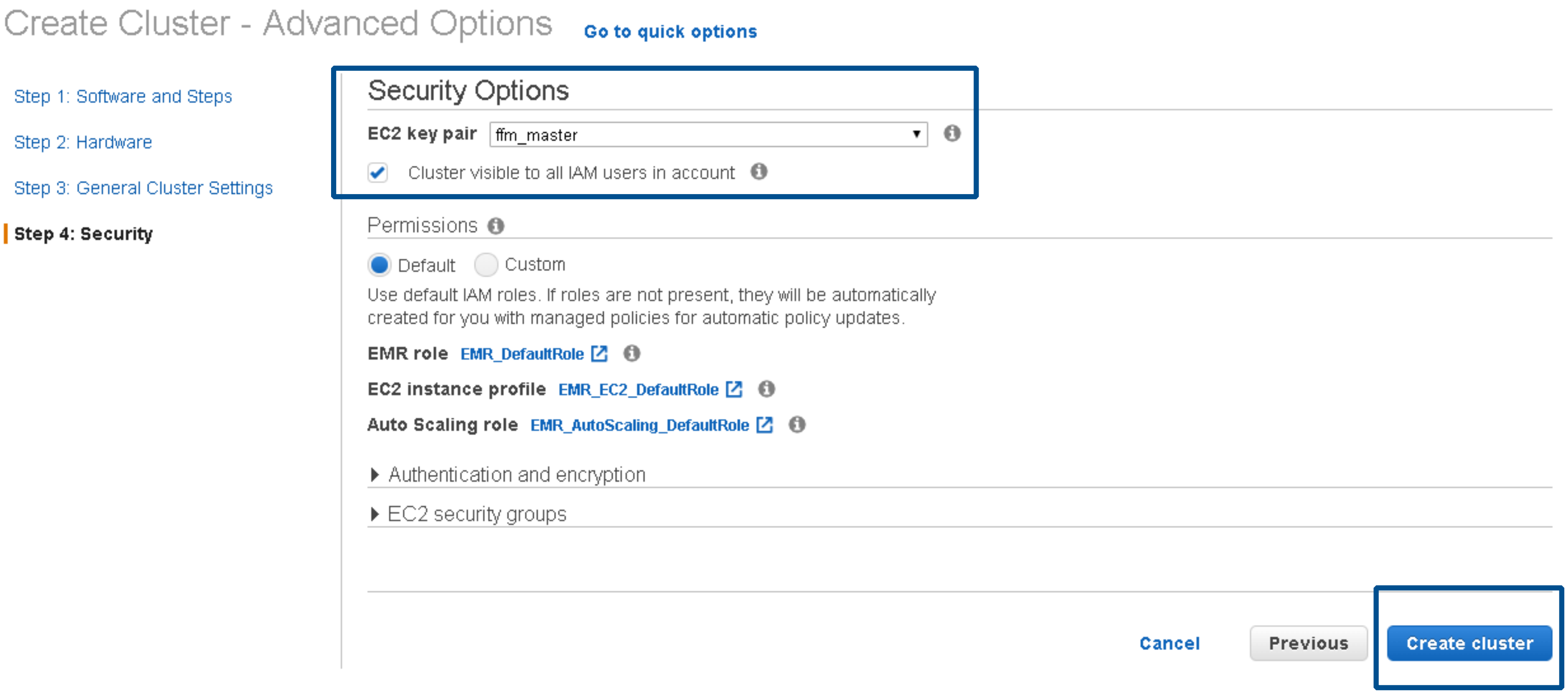
# Step 3: Cluster Settings



```
#Installation of the specified RStudio Server Version with the defined User and Password.
#These Variables can be changed above, if needed.

grep -Fq "\"isMaster\": true" /mnt/var/lib/info/instance.json
if [ $? -eq 0 ];
then
    while [[ $# > 1 ]]; do
        key="$1"

        case $key in
           # The above specified RStudio Server Version
                        --sd-version)
                VERSION="$2"
                shift
                ;;
           # The above specified user
            --sd-user)
                USER="$2"
                shift
                ;;
            # The password for the above specified user
            --sd-user-password)
                PASS="$2"
                shift
                ;;

            *)
                echo "Unknown option: ${key}"
                exit 1;
        esac
        shift
    done
    echo "*****************************************"
    echo "  1. Download RStudio Server ${VERSION}   "
    echo "*****************************************"
    wget https://s3.amazonaws.com/rstudio-dailybuilds/rstudio-server-rhel-${VERSION}-x86_64.rpm
    echo "         2. Install dependencies         "
    echo "*****************************************"
    # This is needed for installing devtools
    sudo yum -y install libcurl libcurl-devel 1>&2
    echo "        3. Install RStudio Server        "
    echo "*****************************************"
    sudo yum -y install --nogpgcheck rstudio-server-rhel-${VERSION}-x86_64.rpm 1>&2
    echo "      4. Create R Studio Server user     "
    echo "*****************************************"
    epass=$(perl -e 'print crypt($ARGV[0], "password")' ${PASS})
    sudo useradd -m -p ${epass} ${USER}
    # This is to allow access to HDFS
    sudo usermod -a -G hadoop ${USER}
    echo "  5. Create environment variables file   "
    echo "*****************************************"
```

Bootstrap Script to Install RStudioServer

# Step 4: Security



## Create Cluster - Advanced Options

Go to quick options

- Step 1: Software and Steps
- Step 2: Hardware
- Step 3: General Cluster Settings
- **Step 4: Security**

### Security Options

**EC2 key pair** ffm_master

☑ Cluster visible to all IAM users in account

Permissions

● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

**EMR role** EMR_DefaultRole

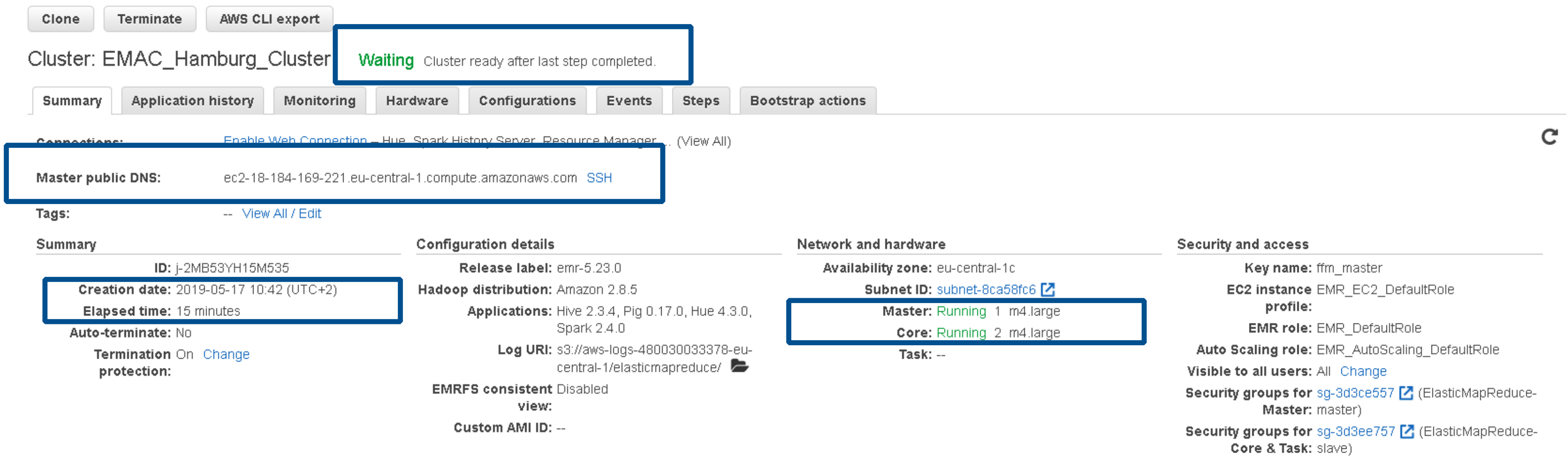
**EC2 instance profile** EMR_EC2_DefaultRole

**Auto Scaling role** EMR_AutoScaling_DefaultRole

▸ Authentication and encryption

▸ EC2 security groups

# Cluster Ready to Use



Cluster: EMAC_Hamburg_Cluster Waiting Cluster ready after last step completed.

**Connections:** Enable Web Connection – Hue, Spark History Server, Resource Manager ... (View All)

**Master public DNS:** ec2-18-184-169-221.eu-central-1.compute.amazonaws.com SSH

**Tags:** -- View All / Edit

**Summary**

**ID:** j-2MB53YH15M535
**Creation date:** 2019-05-17 10:42 (UTC+2)
**Elapsed time:** 15 minutes
**Auto-terminate:** No
**Termination protection:** On Change

**Configuration details**

**Release label:** emr-5.23.0
**Hadoop distribution:** Amazon 2.8.5
**Applications:** Hive 2.3.4, Pig 0.17.0, Hue 4.3.0, Spark 2.4.0
**Log URI:** s3://aws-logs-480030033378-eu-central-1/elasticmapreduce/
**EMRFS consistent view:** Disabled
**Custom AMI ID:** --

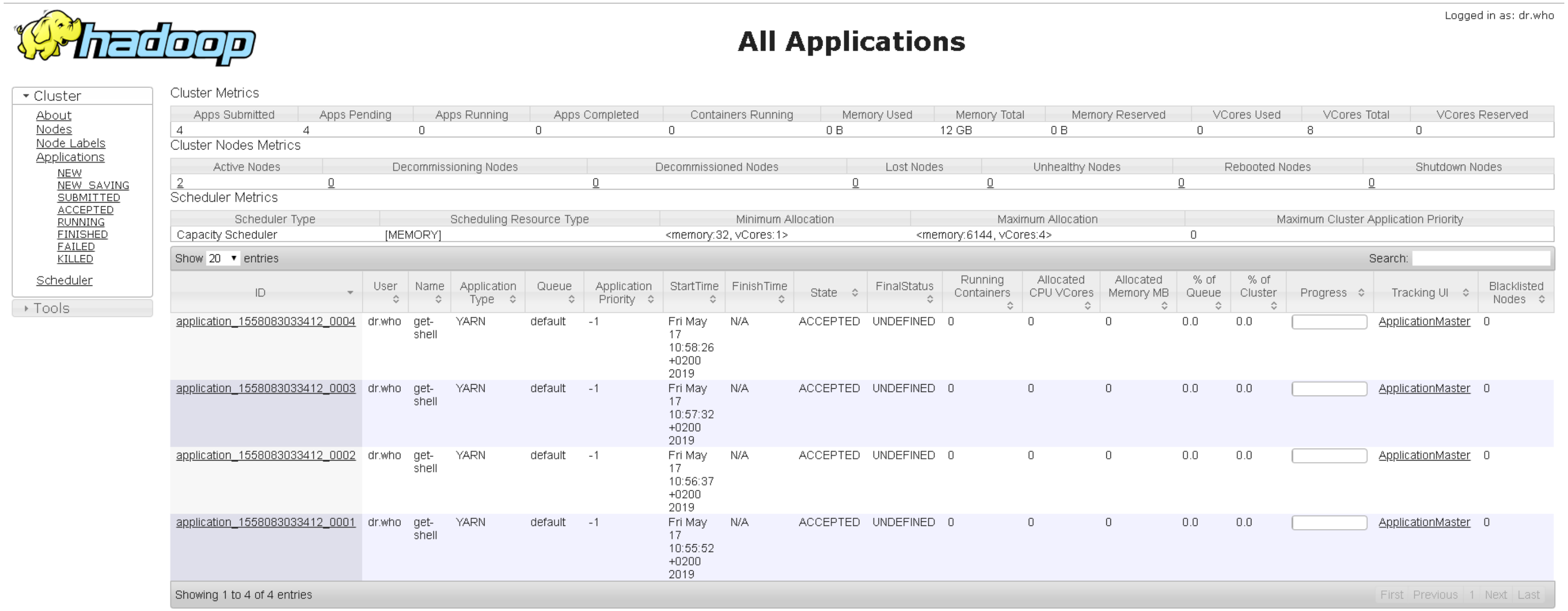
**Network and hardware**

**Availability zone:** eu-central-1c
**Subnet ID:** subnet-8ca58fc6
**Master:** Running 1 m4.large
**Core:** Running 2 m4.large
**Task:** --

**Security and access**

**Key name:** ffm_master
**EC2 instance profile:** EMR_EC2_DefaultRole
**EMR role:** EMR_DefaultRole
**Auto Scaling role:** EMR_AutoScaling_DefaultRole
**Visible to all users:** All Change
**Security groups for Master:** sg-3d3ce557 (ElasticMapReduce-master)
**Security groups for Core & Task:** sg-3d3ee757 (ElasticMapReduce-slave)

# Hadoop Cluster Manager Overview



Logged in as: dr.who

## All Applications

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

| Apps Submitted | Apps Pending | Apps Running | Apps Completed | Containers Running | Memory Used | Memory Total | Memory Reserved | VCores Used | VCores Total | VCores Reserved |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 4 | 0 | 0 | 0 | 0 B | 12 GB | 0 B | 0 | 8 | 0 |

Cluster Nodes Metrics

| Active Nodes | Decommissioning Nodes | Decommissioned Nodes | Lost Nodes | Unhealthy Nodes | Rebooted Nodes | Shutdown Nodes |
|---|---|---|---|---|---|---|
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |

Scheduler Metrics

| Scheduler Type | Scheduling Resource Type | Minimum Allocation | Maximum Allocation | Maximum Cluster Application Priority |
|---|---|---|---|---|
| Capacity Scheduler | [MEMORY] | <memory:32, vCores:1> | <memory:6144, vCores:4> | 0 |

Show 20 ▾ entries

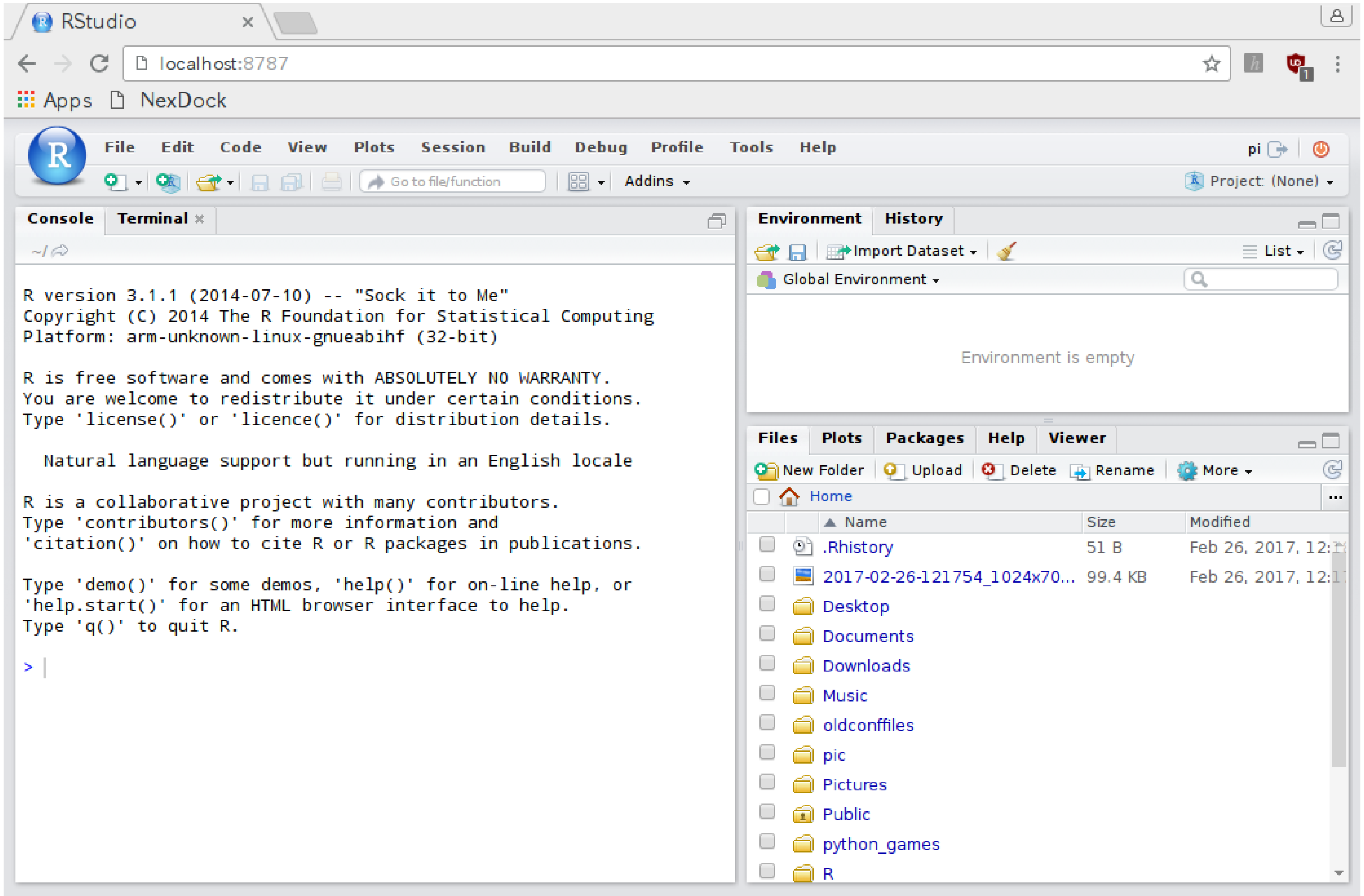
Search:

| ID | User | Name | Application Type | Queue | Application Priority | StartTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | % of Queue | % of Cluster | Progress | Tracking UI | Blacklisted Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| application_1558083033412_0004 | dr.who | get-shell | YARN | default | -1 | Fri May 17 10:58:26 +0200 2019 | N/A | ACCEPTED | UNDEFINED | 0 | 0 | 0 | 0.0 | 0.0 | | ApplicationMaster | 0 |
| application_1558083033412_0003 | dr.who | get-shell | YARN | default | -1 | Fri May 17 10:57:32 +0200 2019 | N/A | ACCEPTED | UNDEFINED | 0 | 0 | 0 | 0.0 | 0.0 | | ApplicationMaster | 0 |
| application_1558083033412_0002 | dr.who | get-shell | YARN | default | -1 | Fri May 17 10:56:37 +0200 2019 | N/A | ACCEPTED | UNDEFINED | 0 | 0 | 0 | 0.0 | 0.0 | | ApplicationMaster | 0 |
| application_1558083033412_0001 | dr.who | get-shell | YARN | default | -1 | Fri May 17 10:55:52 +0200 2019 | N/A | ACCEPTED | UNDEFINED | 0 | 0 | 0 | 0.0 | 0.0 | | ApplicationMaster | 0 |

Showing 1 to 4 of 4 entries

http://ec2-18-184-169-221.eu-central-1.compute.amazonaws.com:8088

# Accessing RStudioServer



```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: arm-unknown-linux-gnueabihf (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

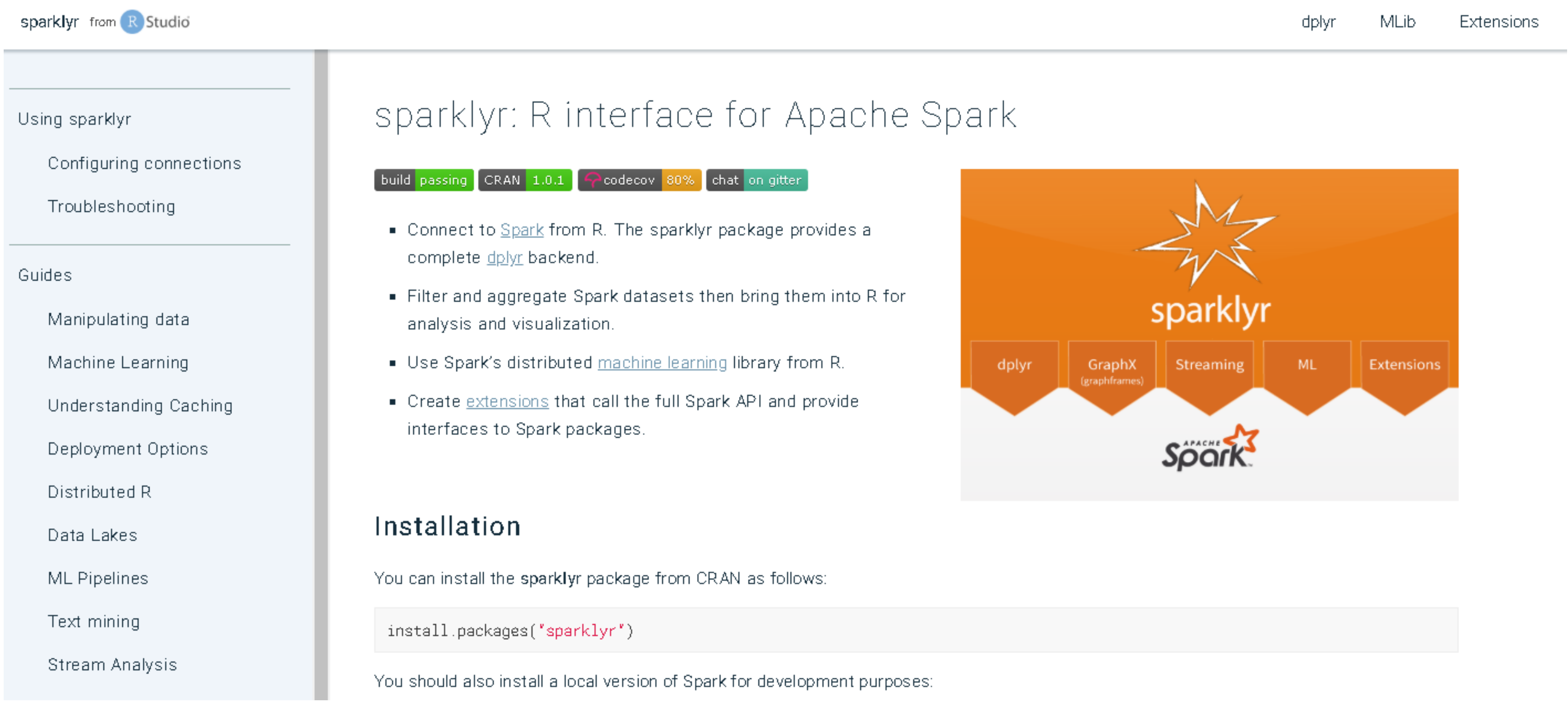
| Name | Size | Modified |
|---|---|---|
| .Rhistory | 51 B | Feb 26, 2017, 12:1 |
| 2017-02-26-121754_1024x70... | 99.4 KB | Feb 26, 2017, 12:1 |
| Desktop | | |
| Documents | | |
| Downloads | | |
| Music | | |
| oldconffiles | | |
| pic | | |
| Pictures | | |
| Public | | |
| python_games | | |
| R | | |

http://ec2-18-184-169-221.eu-central-1.compute.amazonaws.com:8787

# Run Analysis

sparklyr
www.rstudio.com

# Sparklyr: R Interface for Apache Spark



sparklyr from RStudio

dplyr MLib Extensions

Using sparklyr

- Configuring connections
- Troubleshooting

Guides

- Manipulating data
- Machine Learning
- Understanding Caching
- Deployment Options
- Distributed R
- Data Lakes
- ML Pipelines
- Text mining
- Stream Analysis

## sparklyr: R interface for Apache Spark

build passing CRAN 1.0.1 codecov 80% chat on gitter

- Connect to Spark from R. The sparklyr package provides a complete dplyr backend.
- Filter and aggregate Spark datasets then bring them into R for analysis and visualization.
- Use Spark's distributed machine learning library from R.
- Create extensions that call the full Spark API and provide interfaces to Spark packages.

### Installation

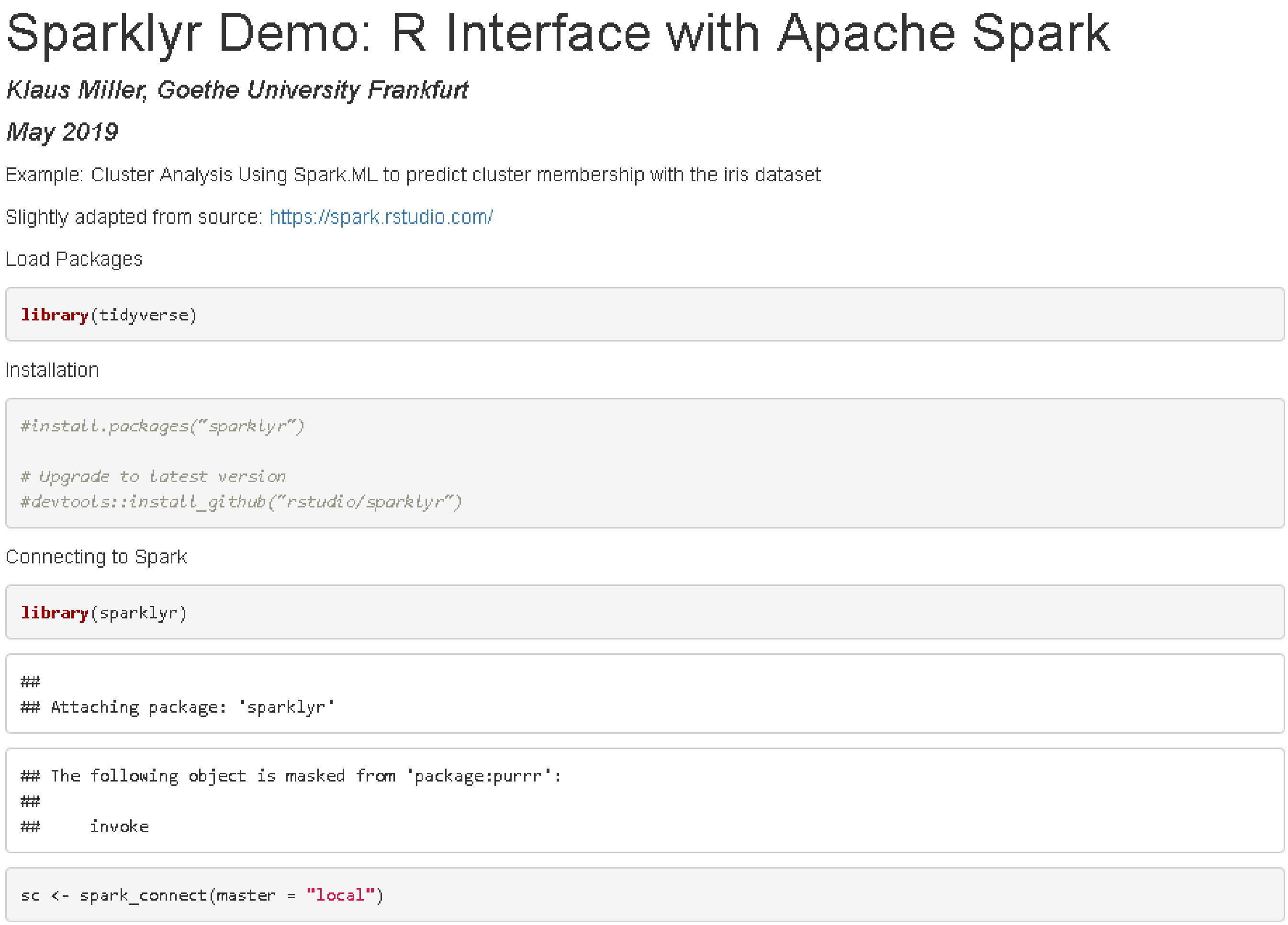
You can install the **sparklyr** package from CRAN as follows:

```
install.packages("sparklyr")
```

You should also install a local version of Spark for development purposes:

Source: https://spark.rstudio.com/

# Sparklyr Demo



## Sparklyr Demo: R Interface with Apache Spark

*Klaus Miller, Goethe University Frankfurt*

*May 2019*

Example: Cluster Analysis Using Spark.ML to predict cluster membership with the iris dataset

Slightly adapted from source: https://spark.rstudio.com/

Load Packages

```
library(tidyverse)
```

Installation

```
#install.packages("sparklyr")

# Upgrade to latest version
#devtools::install_github("rstudio/sparklyr")
```

Connecting to Spark

```
library(sparklyr)
```

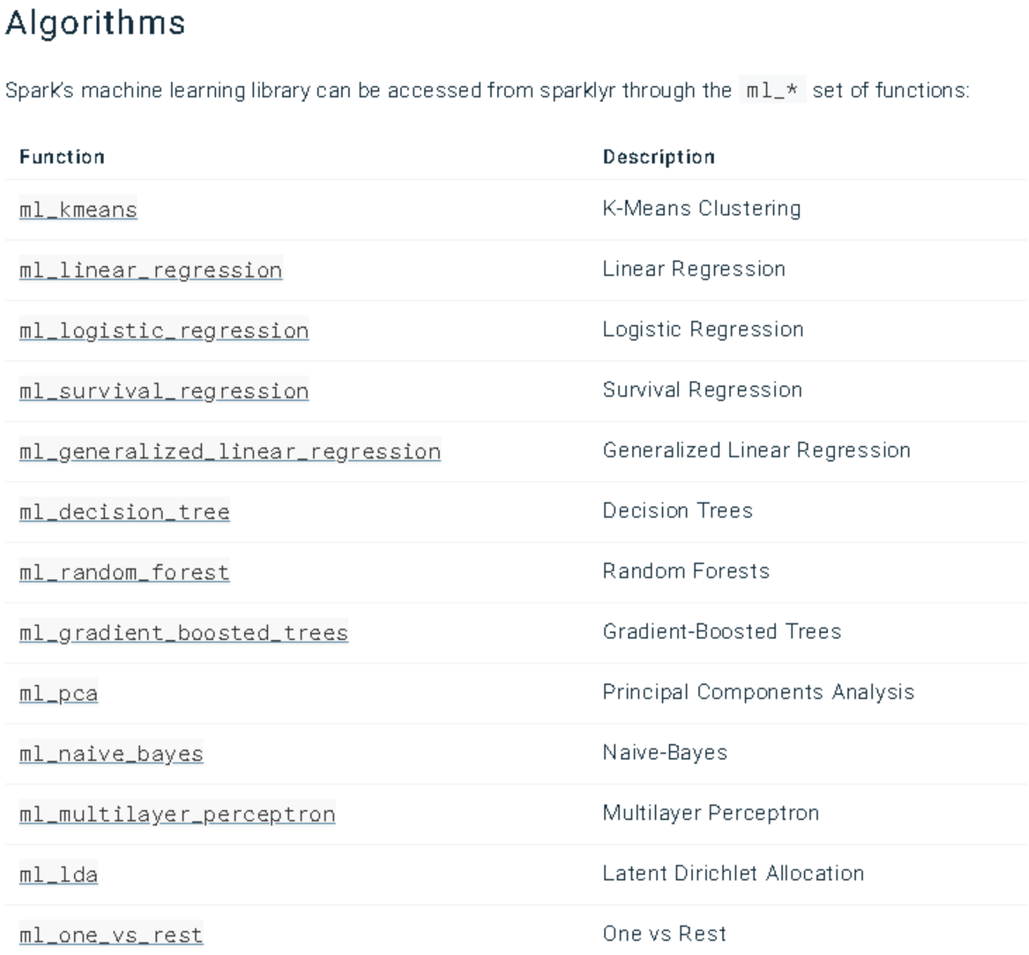
```
##
## Attaching package: 'sparklyr'
```

```
## The following object is masked from 'package:purrr':
##
##     invoke
```

```
sc <- spark_connect(master = "local")
```

Source: https://github.com/stm/vhb_2020

# Sparklyr Machine Learning Library

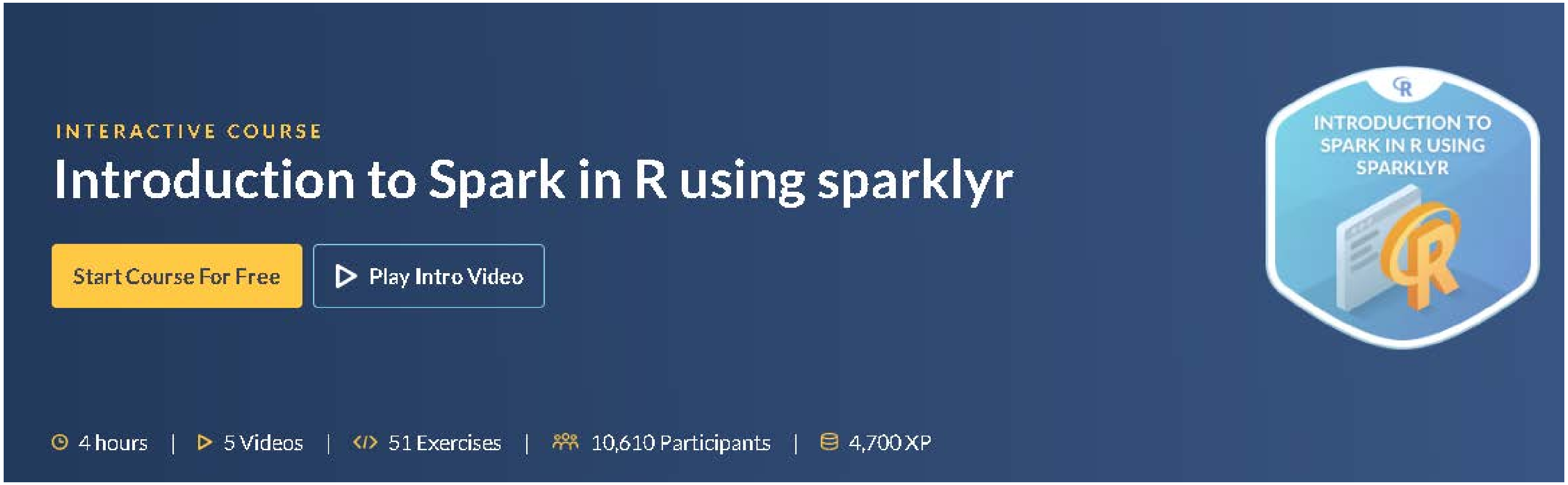


## Algorithms

Spark's machine learning library can be accessed from sparklyr through the `ml_*` set of functions:

| Function | Description |
|---|---|
| `ml_kmeans` | K-Means Clustering |
| `ml_linear_regression` | Linear Regression |
| `ml_logistic_regression` | Logistic Regression |
| `ml_survival_regression` | Survival Regression |
| `ml_generalized_linear_regression` | Generalized Linear Regression |
| `ml_decision_tree` | Decision Trees |
| `ml_random_forest` | Random Forests |
| `ml_gradient_boosted_trees` | Gradient-Boosted Trees |
| `ml_pca` | Principal Components Analysis |
| `ml_naive_bayes` | Naive-Bayes |
| `ml_multilayer_perceptron` | Multilayer Perceptron |
| `ml_lda` | Latent Dirichlet Allocation |
| `ml_one_vs_rest` | One vs Rest |

# Introduction to Spark in R

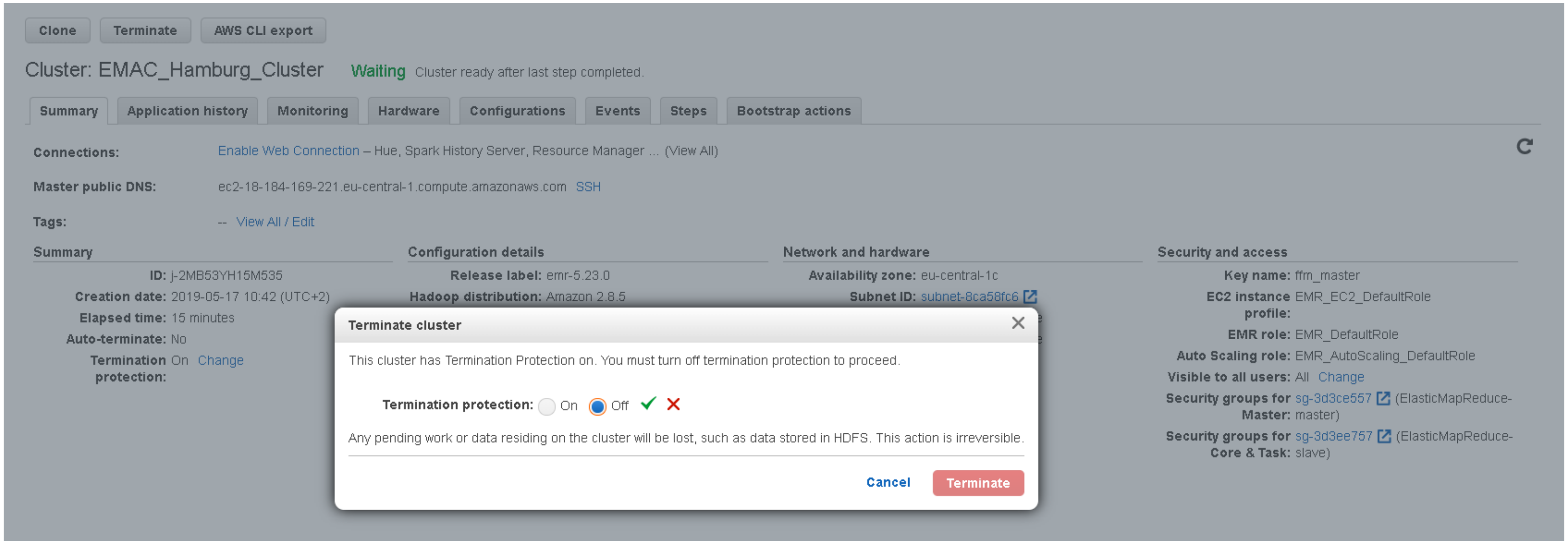


INTERACTIVE COURSE

Introduction to Spark in R using sparklyr

4 hours | 5 Videos | 51 Exercises | 10,610 Participants | 4,700 XP

# Terminating Cluster



Cluster: EMAC_Hamburg_Cluster Waiting Cluster ready after last step completed.

**Connections:** Enable Web Connection – Hue, Spark History Server, Resource Manager ... (View All)

**Master public DNS:** ec2-18-184-169-221.eu-central-1.compute.amazonaws.com SSH

**Tags:** -- View All / Edit

**Summary**

**ID:** j-2MB53YH15M535
**Creation date:** 2019-05-17 10:42 (UTC+2)
**Elapsed time:** 15 minutes
**Auto-terminate:** No
**Termination protection:** On Change

**Configuration details**

**Release label:** emr-5.23.0
**Hadoop distribution:** Amazon 2.8.5

**Network and hardware**

**Availability zone:** eu-central-1c
**Subnet ID:** subnet-8ca58fc6

**Security and access**

**Key name:** ffm_master
**EC2 instance profile:** EMR_EC2_DefaultRole
**EMR role:** EMR_DefaultRole
**Auto Scaling role:** EMR_AutoScaling_DefaultRole
**Visible to all users:** All Change
**Security groups for Master:** sg-3d3ce557 (ElasticMapReduce-master)
**Security groups for Core & Task:** sg-3d3ee757 (ElasticMapReduce-slave)

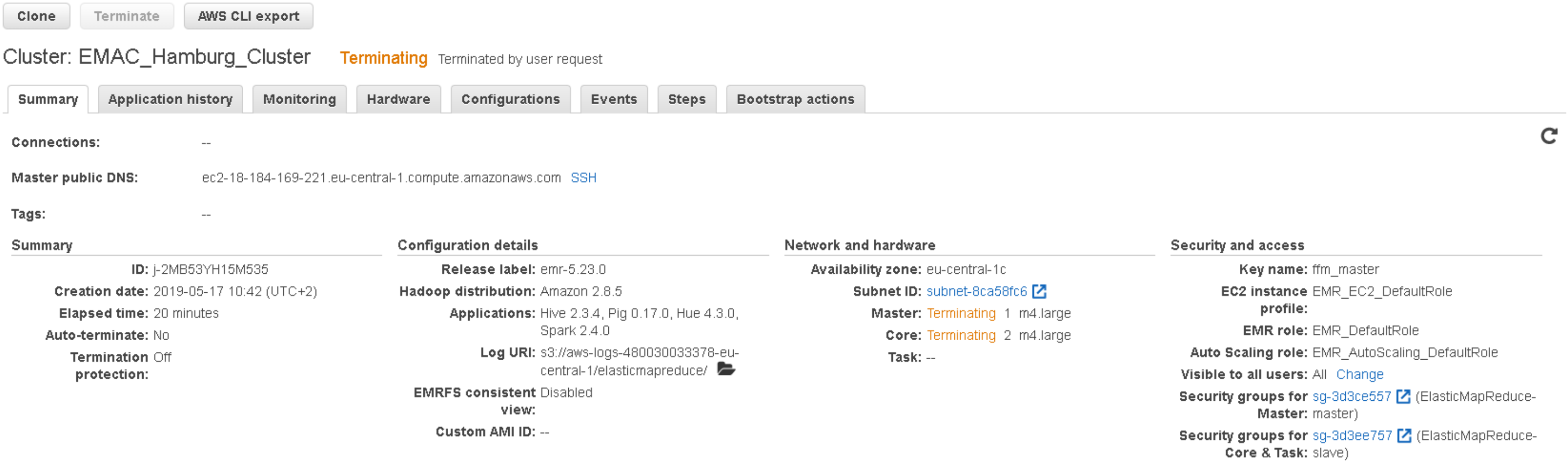
**Terminate cluster**

This cluster has Termination Protection on. You must turn off termination protection to proceed.

**Termination protection:** ○ On ● Off ✓ ✗

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.

# Terminating Cluster



Cluster: EMAC_Hamburg_Cluster Terminating Terminated by user request

**Connections:** --

**Master public DNS:** ec2-18-184-169-221.eu-central-1.compute.amazonaws.com SSH

**Tags:** --

**Summary**

**ID:** j-2MB53YH15M535
**Creation date:** 2019-05-17 10:42 (UTC+2)
**Elapsed time:** 20 minutes
**Auto-terminate:** No
**Termination protection:** Off

**Configuration details**

**Release label:** emr-5.23.0
**Hadoop distribution:** Amazon 2.8.5
**Applications:** Hive 2.3.4, Pig 0.17.0, Hue 4.3.0, Spark 2.4.0
**Log URI:** s3://aws-logs-480030033378-eu-central-1/elasticmapreduce/
**EMRFS consistent view:** Disabled
**Custom AMI ID:** --

**Network and hardware**

**Availability zone:** eu-central-1c
**Subnet ID:** subnet-8ca58fc6
**Master:** Terminating 1 m4.large
**Core:** Terminating 2 m4.large
**Task:** --

**Security and access**

**Key name:** ffm_master
**EC2 instance profile:** EMR_EC2_DefaultRole
**EMR role:** EMR_DefaultRole
**Auto Scaling role:** EMR_AutoScaling_DefaultRole
**Visible to all users:** All Change
**Security groups for Master:** sg-3d3ce557 (ElasticMapReduce-master)
**Security groups for Core & Task:** sg-3d3ee757 (ElasticMapReduce-slave)

# Big Data Not So „Big“ After All

# Thank You for Your Attention!

Klaus Miller
Theodor-W.-Adorno-Platz 4
60323 Frankfurt am Main

Tel: +49-(0)69-798-33-865
Email: klaus.miller@wiwi.uni-frankfurt.de
Web: fromdatatodecisions.com
🐦 @klausmiller