Cloud Computing & Big Data

PARALLEL & SCALABLE MACHINE LEARNING & DEEP LEARNING

Prof. Dr. – Ing. Morris Riedel
Associated Professor
School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
Research Group Leader, Juelich Supercomputing Centre, Forschungszentrum Juelich, Germany

@Morris Riedel @MorrisRiedel @MorrisRiedel

PRACTICAL LECTURE 5.1

Understanding Map-Reduce in Cloud Applications

October 15, 2020
Online Lecture

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

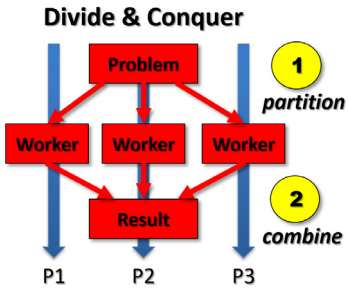
JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

DEEP Projects

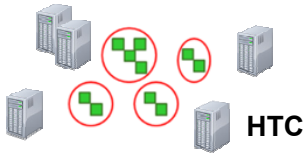
HELMHOLTZ AI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

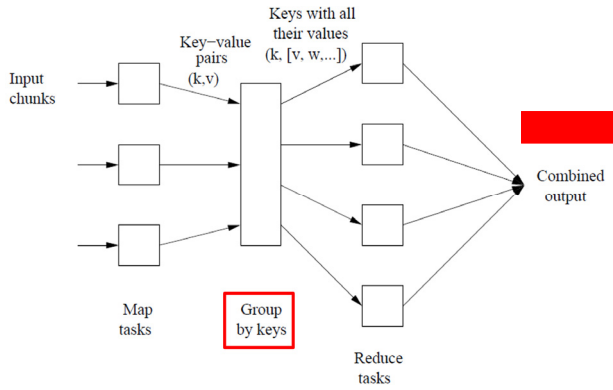
Review of Lecture 5 – Map-Reduce Computing Paradigm

- Map-Reduce Approach



(origin in traditional computer & computational sciences)





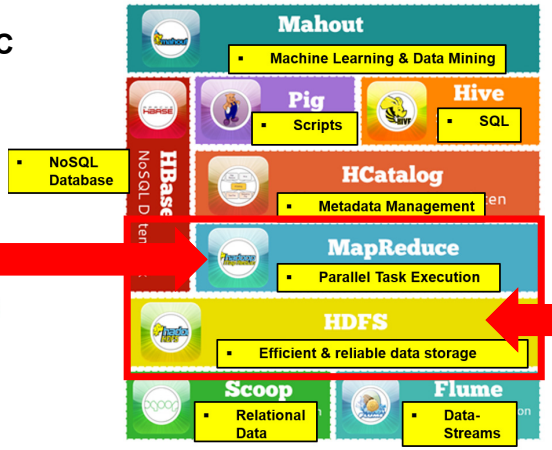
(group/shuffle/sort done by the framework)

Google

[3] MapReduce: Simplified Data Processing on Large Clusters, 2004

[2] Apache Hadoop

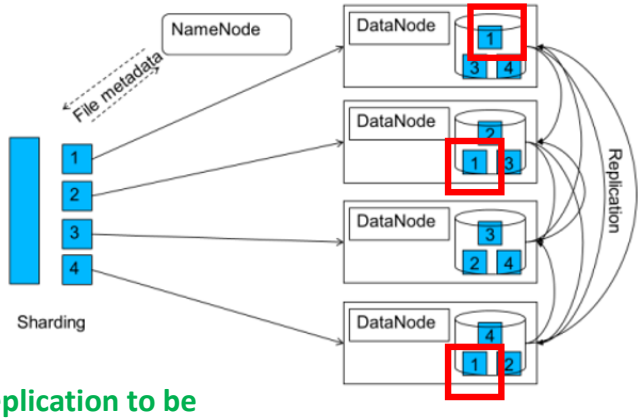
(Hadoop 'big data frameworks' ecosystem)



(key motivations: hide underlying computing & storage complexity and being fault tolerant by design)

Modified from [1] Mining of Massive Datasets
Modified from [6] Map-Reduce

- Hadoop Distributed File System (HDFS)



(replication to be more fault tolerant & data locality technique)

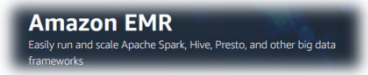
Modified form [4] Stampede Virtual Workshop

Google Cloud

[5] Google Dataproc service

Microsoft Azure

[7] Microsoft Azure HDInsight Service



[8] AWS EMR Web page

Outline of the Course

1. Cloud Computing & Big Data Introduction
2. Machine Learning Models in Clouds
3. Apache Spark for Cloud Applications
4. Virtualization & Data Center Design
5. Map-Reduce Computing Paradigm
6. Deep Learning driven by Big Data
7. Deep Learning Applications in Clouds
8. Infrastructure-As-A-Service (IAAS)
9. Platform-As-A-Service (PAAS)
10. Software-As-A-Service (SAAS)
11. Big Data Analytics & Cloud Data Mining
12. Docker & Container Management
13. OpenStack Cloud Operating System
14. Online Social Networking & Graph Databases
15. Big Data Streaming Tools & Applications
16. Epilogue

+ additional practical lectures & Webinars for our hands-on assignments in context

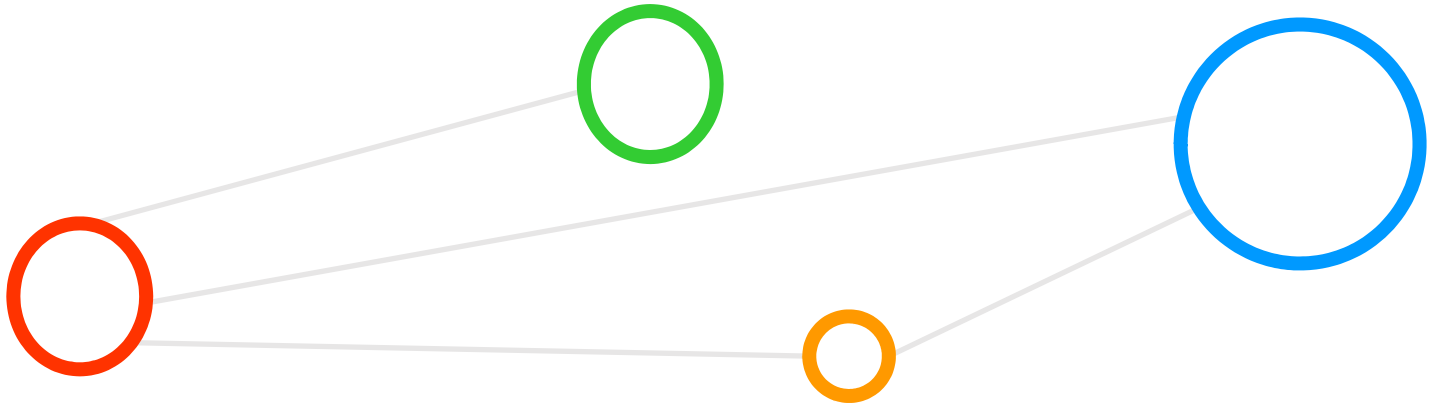
- Practical Topics
- Theoretical / Conceptual Topics

Outline

- Amazon Web Services (AWS) Cloud Computing Example
  - Understanding AWS Broad Service Portfolio & Pay-Per-Use Approach
  - AWS Elastic Map-Reduce (EMR) Service & AWS Educate Program
  - AWS Elastic Compute Cloud (EC2) Service & SSH Key-Pair Generation
  - AWS Marketplace & Airbnb and Spotify Application Examples
  - AWS Machine Learning Examples & Application Impacts
- Understanding Map-Reduce in Cloud Applications
  - Deploying AWS EMR Services in AWS Cloud Environments
  - Dependency of AWS EMR Services to EC2 and Key Pairs
  - Using AWS EMR & S3 for Simple Cloud Applications with WordCount
  - Understanding Streaming Map-Reduce Functionalities & Built-in Aggregations
  - Larger Map-Reduce Ecosystem of Services for Applications

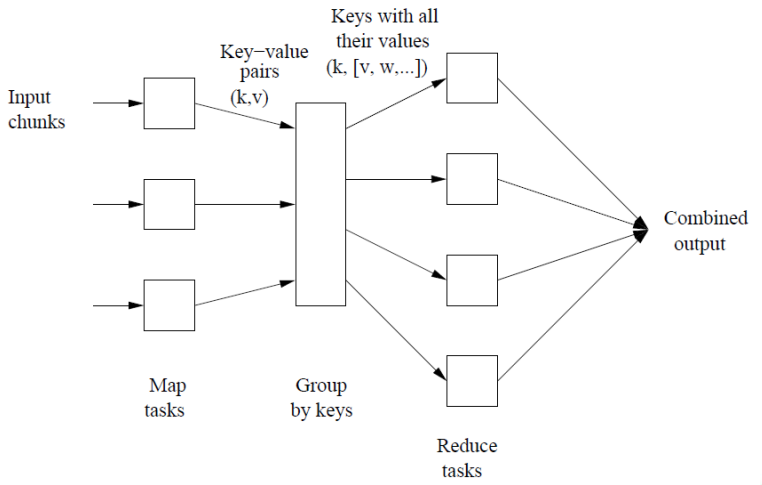
- **Promises from previous lecture(s):**
- ***Lecture 5:* Practical Lecture 5.1 will offer some concrete examples of map-reduce applications in modern Cloud computing environments today**
- ***Lecture 5:* Practical Lecture 5.1 will offer some concrete examples of map-reduce applications that leverage the group/sort/shuffle techniques**
- ***Lecture 5:* Practical Lecture 5.1 will explore a concrete Cloud application example of using map-reduce via Apache Hadoop & its ecosystem tools**

Amazon Web Services Cloud Environment



Origins of the Map-Reduce Programming Model – Revisited

- Origin: Invented via the proprietary Google technology by Google technologists
  - Drivers: Applications and 'data mining approaches' around the Web
  - Foundations go back to functional programming (e.g. LISP)
- Large 'open source community'
  - Apache Hadoop, versions 1/2/3
  - Open Source Implementation of the 'map-reduce' programming model
  - Based on Java programming language, e.g. map & reduce tasks objects
  - Offers a scheduler for distributed computing
  - Broadly used – also by commercial vendors within added-value software
  - Foundation for many higher-level algorithms, frameworks, and approaches



Google

[3] MapReduce: Simplified Dataset on Large Clusters, 2004

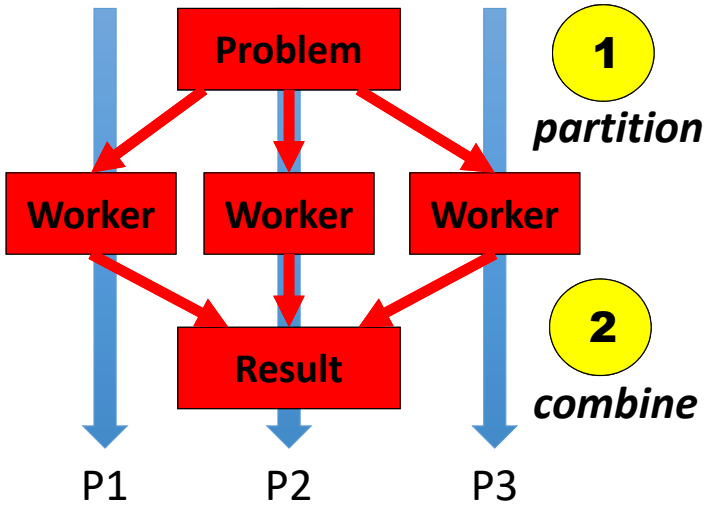
hadoop

[2] Apache Hadoop

Map-Reduce Computing Paradigm & Open Source Implementation – Revisited

- Idea not completely new
  - Derived divide & conquer strategy
  - Needs some smart addition (e.g., grouping intermediate results)

Divide & Conquer



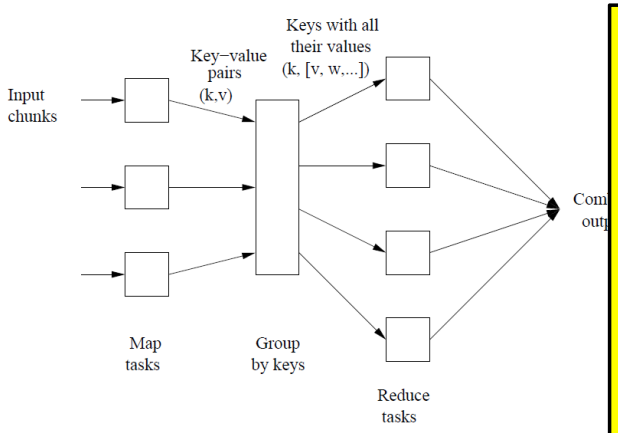
- Map-Reduce Paradigm
  - Follows divide & conquer approach
  - Injects a key element between this process: short/shuffle/group
  - Open Source Implementation: Apache Hadoop

Google

[3] MapReduce: Simplified Dataset on Large Clusters, 2004

hadoop

[2] Apache Hadoop



- **Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models**
- **Apache Hadoop is an open source implementation of the map-reduce computing paradigm**
- **Apache Hadoop is designed to scale up from single servers to thousands of machines, each offering local computation and storage**

Understand AWS Cloud Service Portfolio – Analytics Services – Revisited

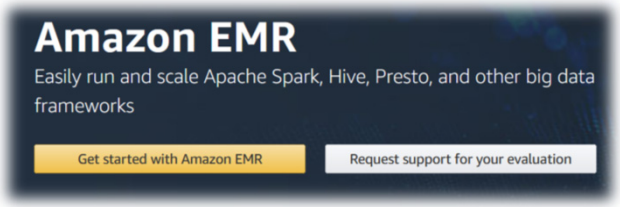
- Multiple analytics products
  - Extracting insights and actionable information from data requires technologies like analytics & machine learning
- Analytics Services
  - Amazon Athena: Serverless Query Service
  - Amazon ElasticMapReduce (EMR): Hadoop ecosystem
  - Amazon ElasticSearch Service: Elasticsearch on AWS
  - Amazon Kinesis: Streaming Data
  - Amazon QuickSight: Business Analytics
  - Amazon Redshift: Data Warehouse

- **AWS offers a wide variety of Big Data Analytics Cloud services such as Amazon Athena (interactive analytics), Amazon EMR (big data processing with Apache Hadoop & Spark), Amazon Redshift (data warehousing), Amazon Kinesis (real-time analytics), Amazon Elasticsearch Service (operational analytics), and Amazon Quicksight (dashboards and visualizations)**

[9] AWS Web page

AWS Analytics services

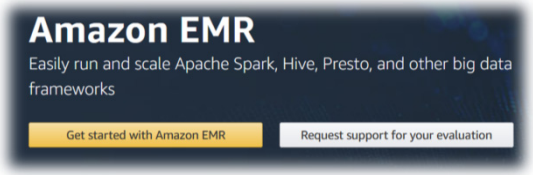
Category	Use cases	AWS service
Analytics	Interactive analytics	Amazon Athena
	Big data processing	Amazon EMR
	Data warehousing	Amazon Redshift
	Real-time analytics	Amazon Kinesis
	Operational analytics	Amazon Elasticsearch Service
	Dashboards and visualizations	Amazon Quicksight



[8] AWS EMR Web page

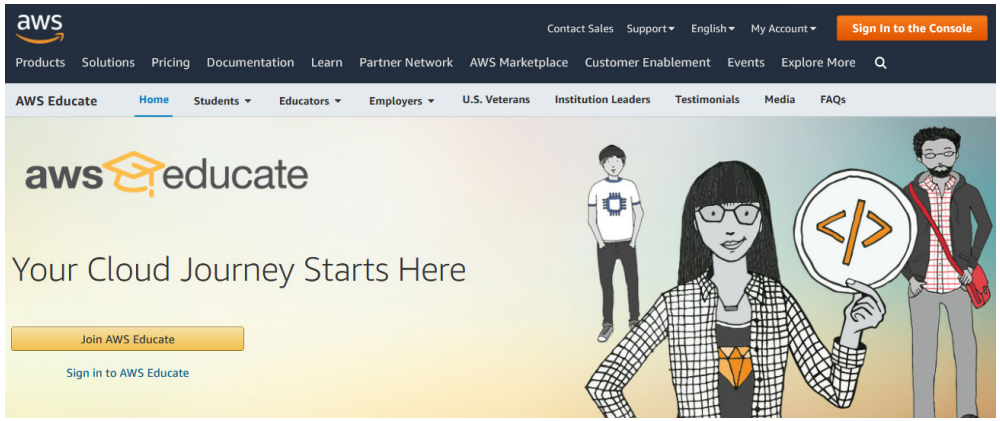
Cloud Computing Approach with AWS EMR using AWS Educate

- Amazon Web Services (AWS) Cloud
  - Wide variety of different cloud-based services & resources for many application areas
  - Managed via AWS Management Console
  - Needs an AWS account
- Elastic Map Reduce (EMR) 'Service'
  - Full managed services to deploy known open source 'big data analytics' frameworks
  - Easy deployment & configuration
  - Pay-per-use model
  - Part of AWS Educate package for universities
  - Unfortunately not in 12 month free tier
- AWS Educate
  - University of Iceland is officially registered



[8] AWS EMR Web page

[9] AWS Web page



[10] AWS Educate Web page

AWS Services Supported with AWS Educate Starter Account

- AWS Educate Starter Account
  - Only a limited set of services available
  - AWS Elastic Map-Reduce (EMR) available
  - Organization in Classrooms or student alone

aws educate

AWS Services Supported with AWS Educate Starter Account

Updated August 2020

Below is a list of all the services that are supported as part of AWS Educate Starter Account. AWS services that are not listed in the table below are not supported as part of Starter Accounts. Some services may have additional restrictions as described in the table below.

All services are only supported in **US East (N. Virginia) [us-east-1] region**. IAM restrictions apply on all services. All services may have additional restrictions not listed below.

Note about Cloud9:
When you launch Cloud9 Desktop for the first time, you may see a warning message like below-
"The environment cannot be created AWS is currently verifying your account. Until this verificatior is complete, you cannot create environments. This verification could take up to 2 hours or longer. Try creating the environment again later. If you are still receiving this message more than 2 hours, email aws-verification@amazon.com"

If the error does not go away after 24 hours, please delete the environment you were trying to launch and try again. Note – This applies *only* to new environment you are trying to launch. If you already have an existing environment running, then you should not be seeing the error message above.

[11] AWS Educate Supported Services

Comprehend	
Comprehend Medical	
Config	
Data Pipeline	
Deep Lens	
Deep Racer	Coming Soon
Deep Comoser	
DynamoDB	
EC2	Starter Accounts support only following instance types: t2.small, t2.micro, t2.nanao, m4.large, c4.large, c5.large, m5.large, t2.medium, m4.xlarge, c4.xlarge, c5.xlarge, t2.2xlarge, m5.2xlarge, t2.large, t2.xlarge, m5.xlarge. The following are NOT supported in EC2/VPC: • Spot Instances • Reserved Instances purchases • VPN Gateways, VPN Connections or Customer Gateways • Marketplace EC2 purchases or software, including free Marketplace AMIs • Scheduled EC2 instances supported • Spot Fleet supported *Additional restrictions may apply
ECS	
ECR	
EKS	
Elasticache	No reserved instance purchase supported
Elasticfilesystem	
Elasticloadbalancing	
Elastic Inference	
ElasticMapReduce	
Events	
Execute-api	
Elasticbeanstalk	
Firehose	
Forecast	
GameLift	
Ground Truth Labeling	
Glue	
Guard Duty	
GreenGrass	
Health	
IAM	You can create users, but cannot associate login profile or access keys for them. * Additional restrictions may apply

AWS Educate Starter Account – Educator Profile Example

- Educator Profile
  - AWS Educate Classrooms

[11] AWS Educate Supported Services

AWS Educate Starter Account – Student Invitation Email for Workbench Access

Mi 14.10.2020 16:14

Rocco Sedona

Fw: You have been invited to join an AWS Educate Classroom

An Morris Riedel

From: AWS Educate Support <support@awseducate.com>
Sent: Wednesday, 14 October 2020 4:30 PM
To: Rocco Sedona <ros21@hi.is>
Subject: You have been invited to join an AWS Educate Classroom

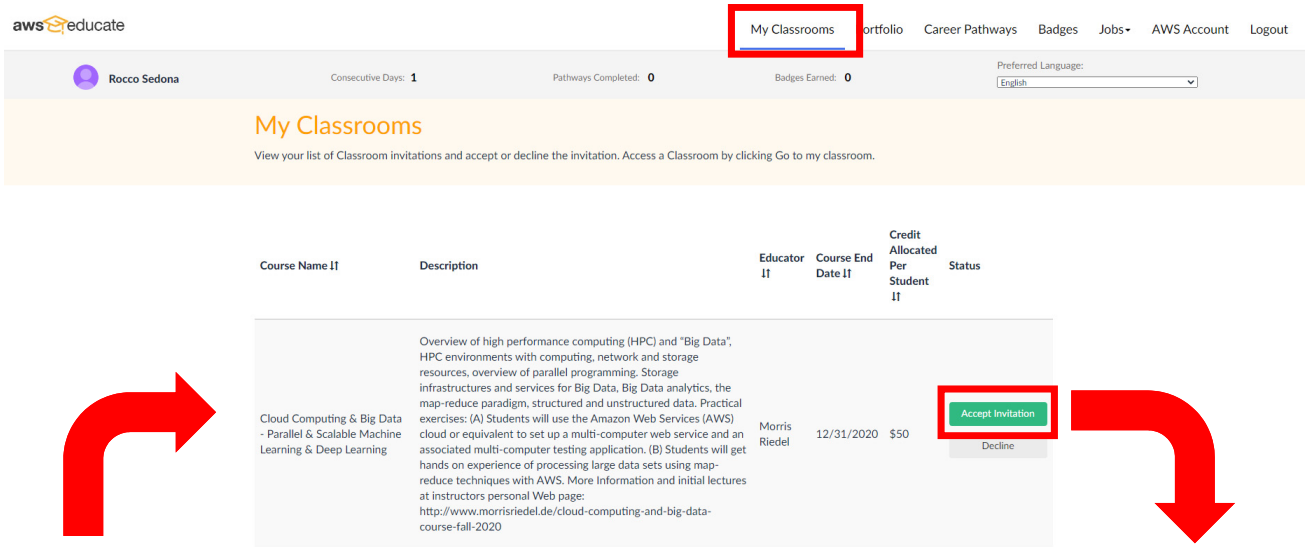
Hi,

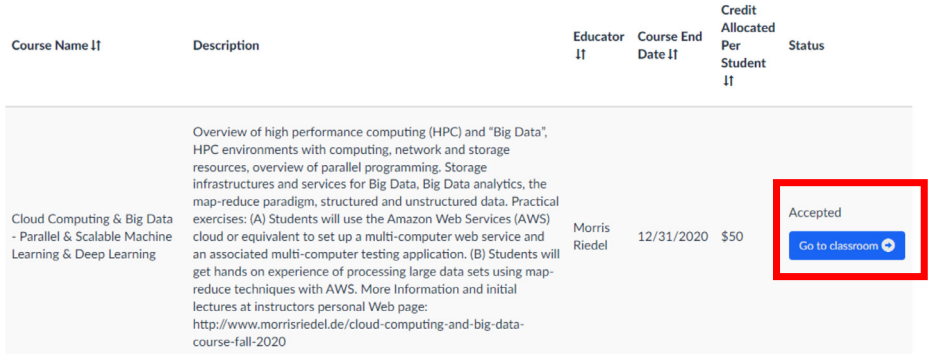
Your educator has invited you to join AWS Educate and access a "Classroom" for Cloud Computing & Big Data - Parallel & Scalable Machine Learning & Deep Learning . A "Classroom" is a hands-on learning environment for you to access AWS services and practice AWS. There are no costs or fees to access a Classroom.

Classrooms are managed by a third-party content and service provider, Vocareum ("Third-Party Content Provider"), and use of the Classroom feature is governed by the Third-Party Content Provider's terms and conditions (including its Privacy Policy) in addition to the AWS Educate Terms & Conditions.

If you accept the Classroom invitation, the Third-Party Content Provider may allow your educator to view your Classroom account and activity, including the AWS console in your Classroom account, the number of EC2 instances running and any Content running in the services, and your access activity. Click to sign in to AWS Educate to Accept or Decline the invitation under the "My Classrooms" menu option.

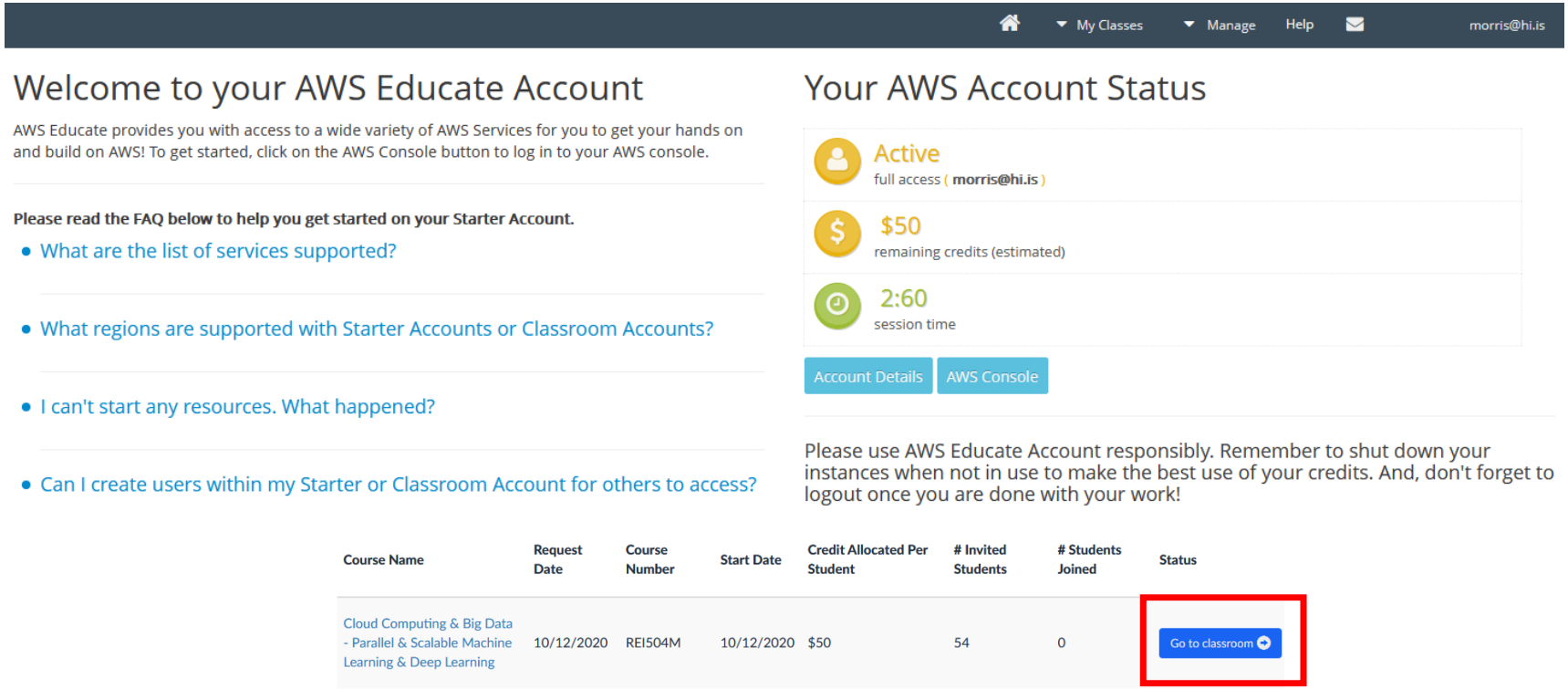
AWS Educate



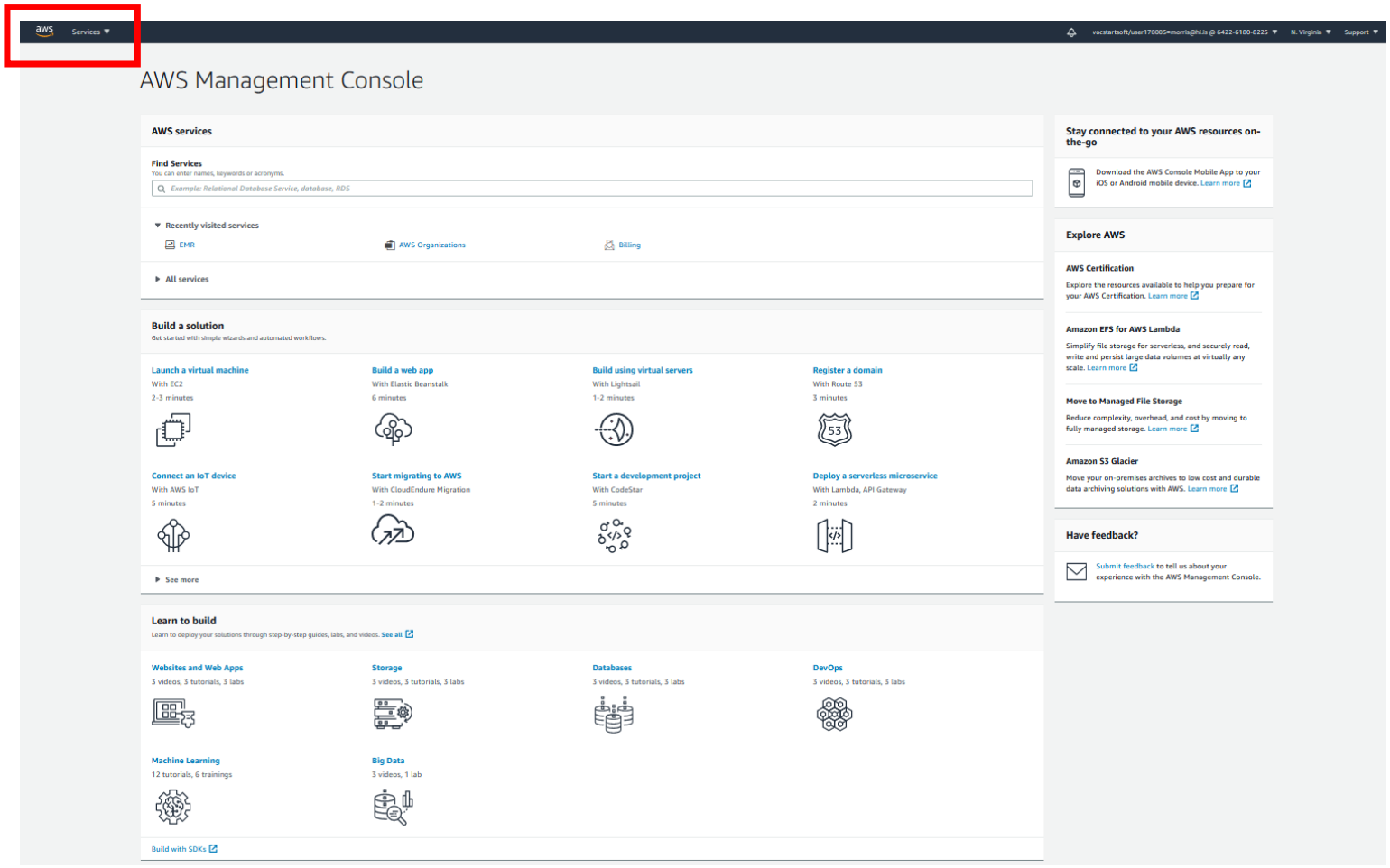


AWS Educate Starter Account – Account Status in Classrooms

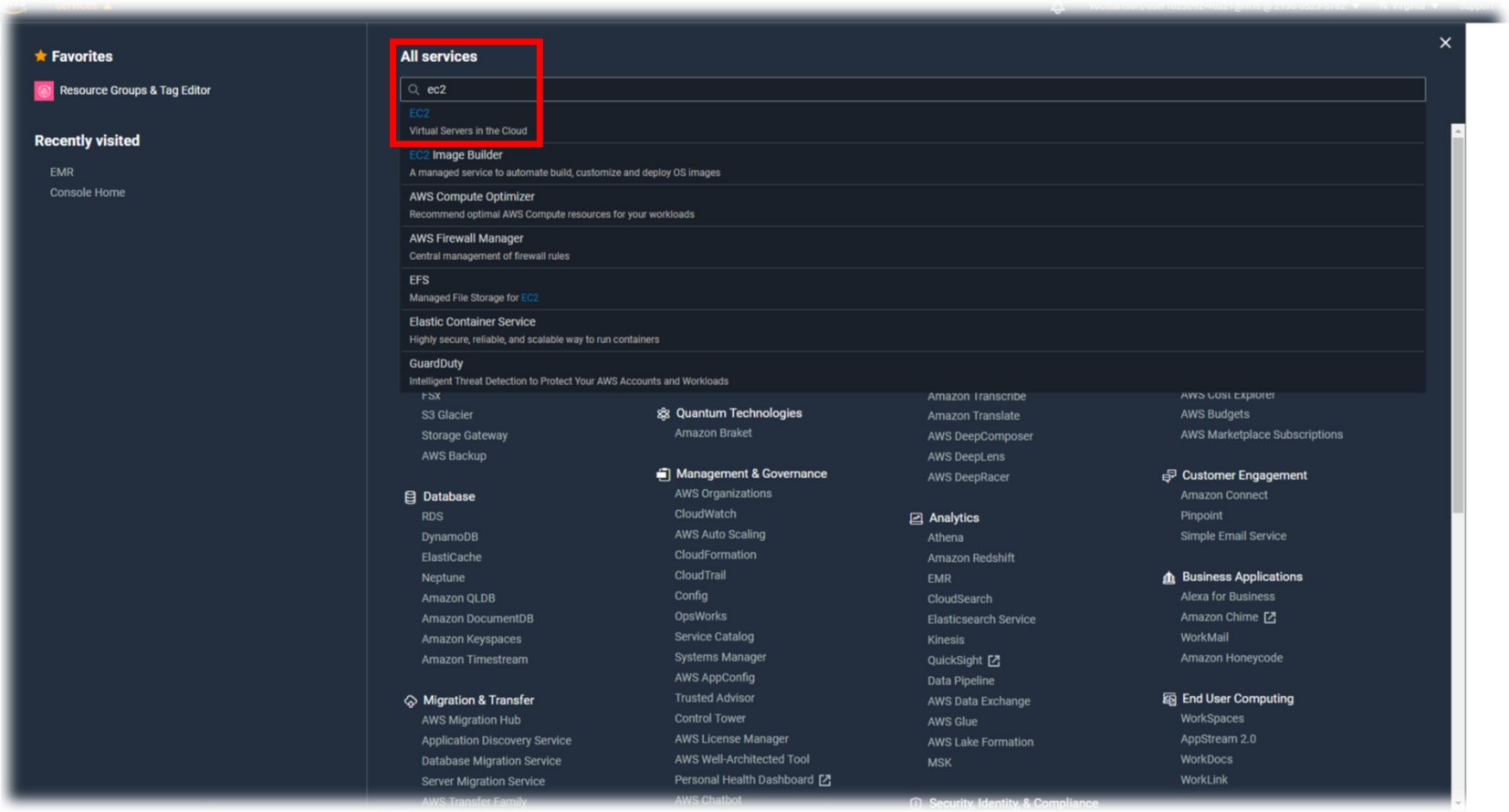
- Workbench & Example Classroom
  - Cloud Computing & Big Data – Parallel and Scalable Machine Learning and Deep Learning



AWS Management Console – Central Access to all Services & Functionalities



AWS Service Example – Elastic Compute Cloud (EC2) Virtual Servers



Pay-Per-Use Approach – AWS EC2 Examples of Underlying Costs

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
m6gd.medium	1	N/A	4 GiB	1 x 59 NVMe SSD	$0.0452 per Hour
m6gd.large	2	N/A	8 GiB	1 x 118 NVMe SSD	$0.0904 per Hour
m6gd.xlarge	4	N/A	16 GiB	1 x 237 NVMe SSD	$0.1808 per Hour
m6gd.2xlarge	8	N/A	32 GiB	1 x 475 NVMe SSD	$0.3616 per Hour
m6gd.4xlarge	16	N/A	64 GiB	1 x 950 NVMe SSD	$0.7232 per Hour
m6gd.8xlarge	32	N/A	128 GiB	1 x 1900 NVMe SSD	$1.4464 per Hour
m6gd.12xlarge	48	N/A	192 GiB	2 x 1425 NVMe SSD	$2.1696 per Hour
m6gd.16xlarge	64	N/A	256 GiB	2 x 1900 NVMe SSD	$2.8928 per Hour

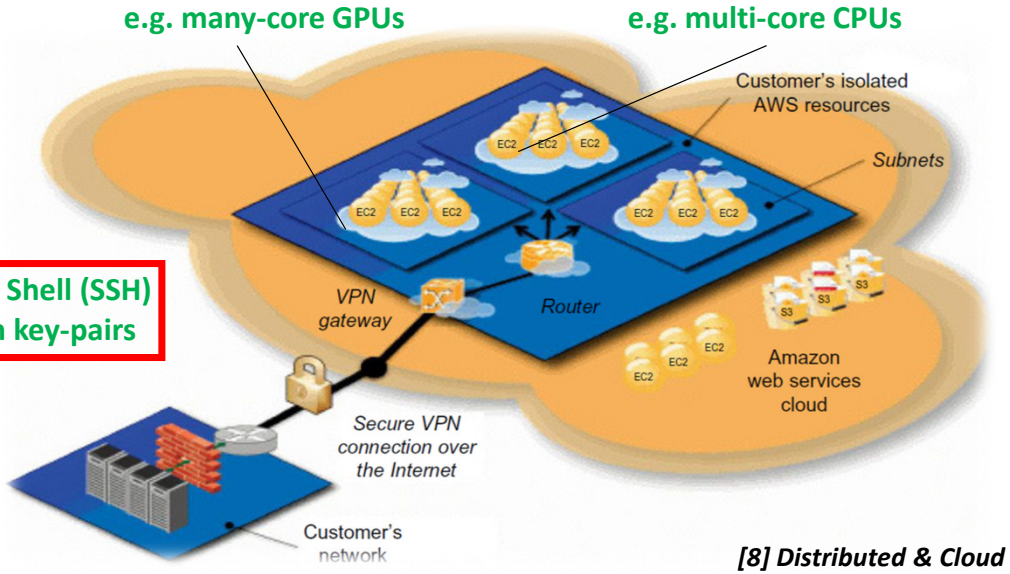
Memory Optimized - Current Generation					
x1.16xlarge	64	174.5	976 GiB	1 x 1920 SSD	$6.669 per Hour
x1.32xlarge	128	349	1,952 GiB	2 x 1920 SSD	$13.338 per Hour
x1e.xlarge	4	12	122 GiB	1 x 120 SSD	$0.834 per Hour
x1e.2xlarge	8	23	244 GiB	1 x 240 SSD	$1.668 per Hour
x1e.4xlarge	16	47	488 GiB	1 x 480 SSD	$3.336 per Hour
x1e.8xlarge	32	91	976 GiB	1 x 960 SSD	$6.672 per Hour
x1e.16xlarge	64	179	1,952 GiB	1 x 1920 SSD	$13.344 per Hour
x1e.32xlarge	128	340	3,904 GiB	2 x 1920 SSD	$26.688 per Hour
General Purpose - Current Generation					
a1.medium	1	N/A	2 GiB	EBS Only	$0.0255 per Hour
a1.large	2	N/A	4 GiB	EBS Only	$0.051 per Hour
a1.xlarge	4	N/A	8 GiB	EBS Only	$0.102 per Hour
a1.2xlarge	8	N/A	16 GiB	EBS Only	$0.204 per Hour
a1.4xlarge	16	N/A	32 GiB	EBS Only	$0.408 per Hour
a1.metal	16	N/A	32 GiB	EBS Only	$0.408 per Hour

[13] AWS EC2 Pricing

Amazon Web Services – Modern Infrastructure-as-a-Service (IAAS) Example

- Amazon EC2 provides an elastic compute cloud (EC2)
  - Elastic load balancing services and so-called auto-scaling
  - E.g. great during peak times in business (e.g. x-mas shopping, etc.)
  - Ensures that a sufficient number of EC2 instances are provisioned to meet expected performance
  - E.g. New York Times use it to quickly retrieve pictorial information from millions of articles
- Amazon Web Services (AWS)
  - Offers infrastructure used for Amazon shopping also for computing customers
  - Ideal situation for Amazon
  - Offers high number of resources & services

[9] Amazon Web Services



e.g. Secure Shell (SSH) access with key-pairs

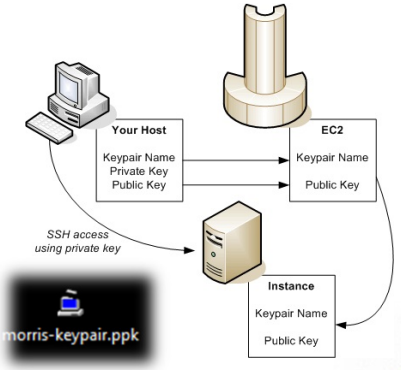
[8] Distributed & Cloud Computing Book

➢ **Lecture 8 provides more details about Amazon Web Services and its Infrastructure-as-a-Service (IAAS) models & EC2 Service Elements**

AWS Elastic Compute Cloud (EC2) Virtual Servers & Using SSH with Key Pairs

- Secure Shell (SSH)
  - Universal technique to securely access remote clusters & HPC machines

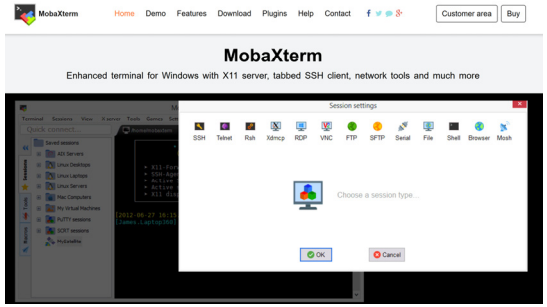
- **The Secure Shell (SSH) is a technique to securely access remote AWS computing instances (e.g., AWS EC2) using a named key pair**
- **An SSH key pair consists of a public key that is known by the Amazon Cloud and a private key that remains only on the laptop of cloud users**



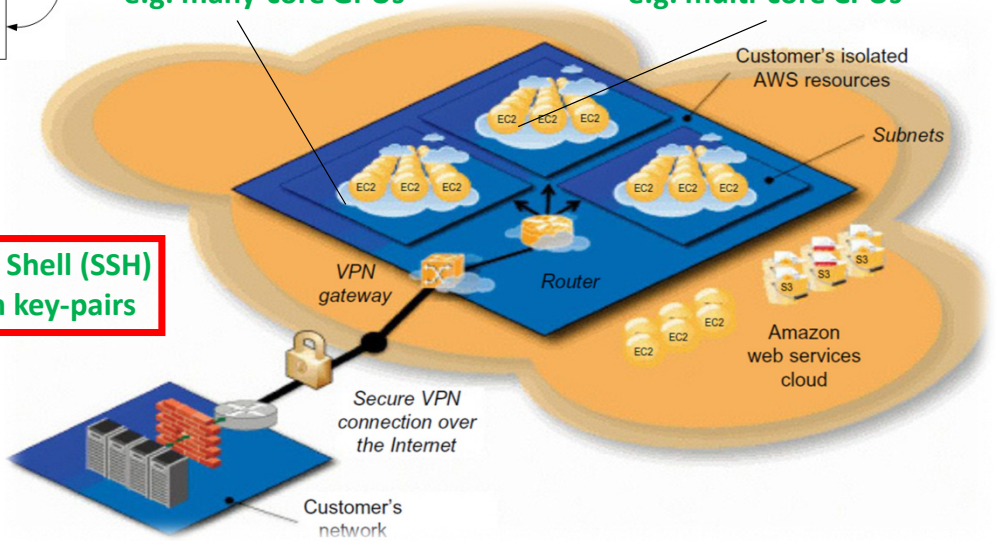
- **Generated AWS key pairs are created per region (e.g., Virginia) in the AWS Cloud**
- **Switching regions means new and/or other SSH keys needs to be used as before**

[9] Amazon Web Services

e.g. Secure Shell (SSH) access with key-pairs



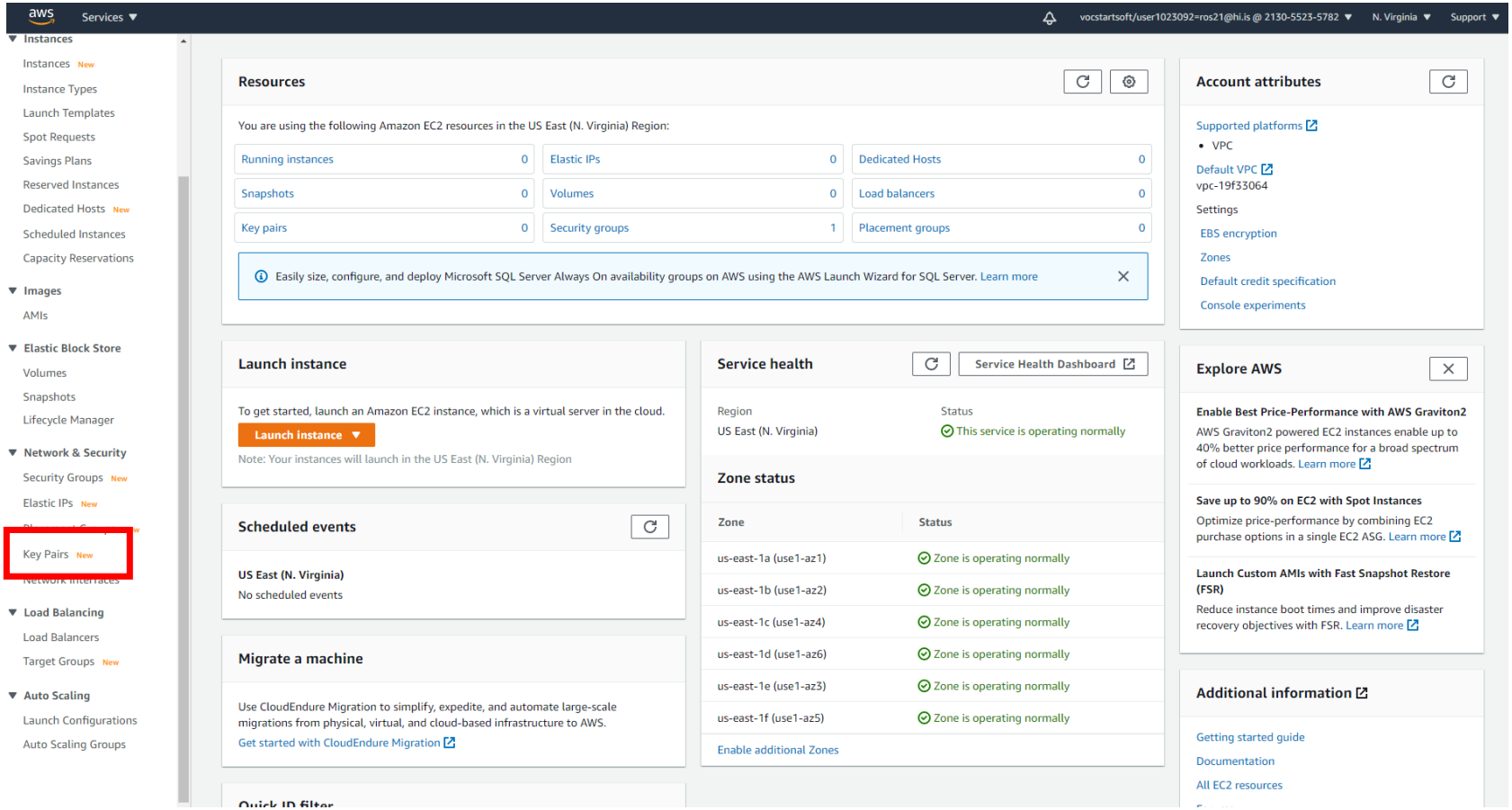
(SSH client is necessary)



[14] MobaXterm Web page

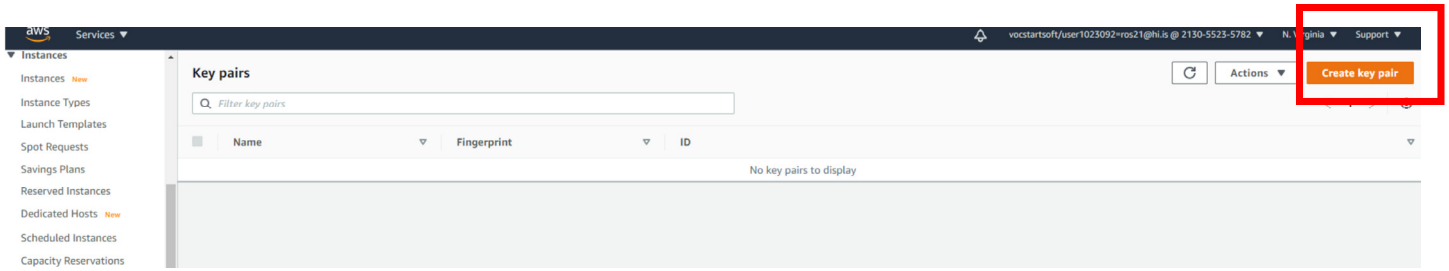
[22] Key Concepts from the AWS Cloud

AWS Service Example – Elastic Compute Cloud (EC2) Virtual Servers & Key Pairs

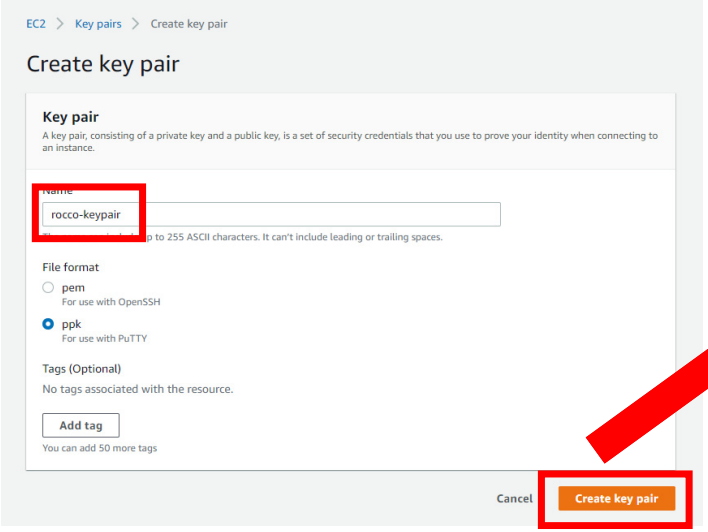


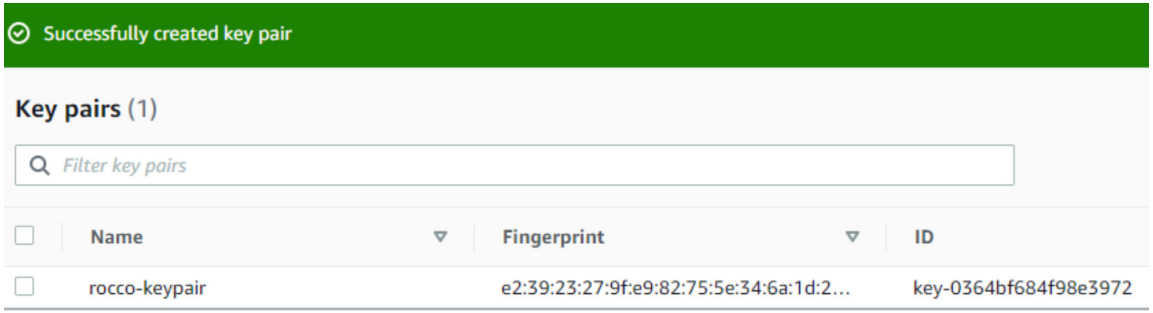
AWS Key Pair – Key Pair Generation

- Usage
  - Public Key remains in Cloud
  - Private Key on Laptop
  - Use SSH Client tool with private key to access remote cloud with matching public key

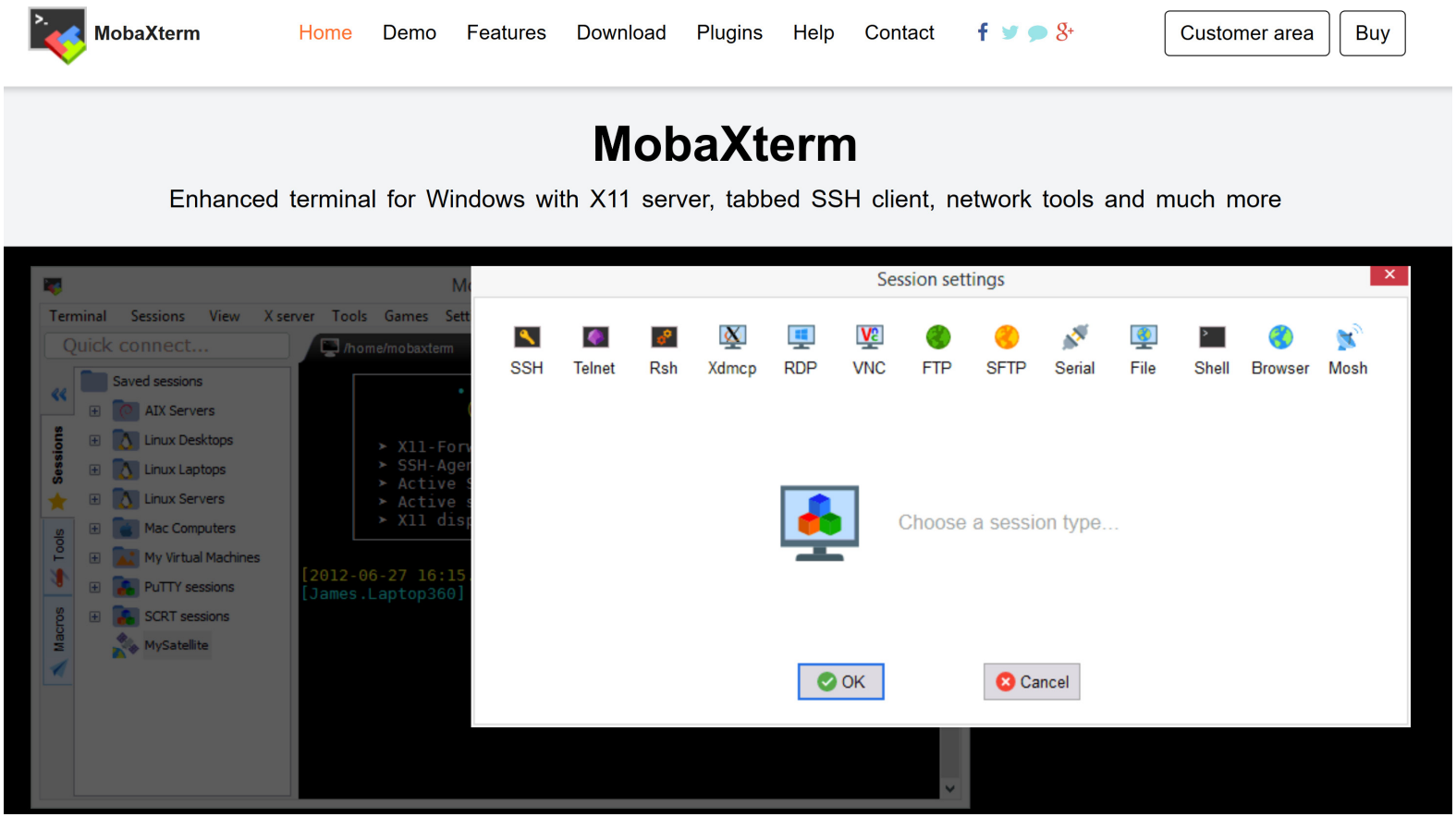


- **After the AWS Key Pair generation, the name of the key is known in many AWS service configuration deployment options such as within the Elastic Compute Cloud (EC2) service or Elastic Map-Reduce (EMR) service**





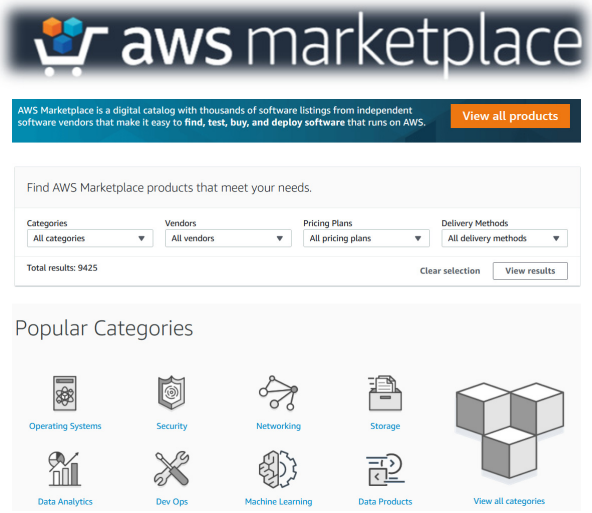
SSH Client for Windows – MobaXTerm



[14] MobaXterm Web page

AWS Marketplace & Selected Amazon Elastic Map Reduce Applications

- AWS Marketplace
  - E.g. collection of community and Amazon created pre-installed images
  - Software infrastructure, developer tools, business & desktop software
  - User success stories and details of how AWS was adopted in solutions
- 'Startup company' Airbnb (travel)
  - Scales infrastructure automatically using AWS
  - Uses 200 Amazon EC2 instances for its application
  - Uses elastic load balancing with Amazon EC2 instances
  - Analyzes 50 GB of data daily via Amazon Elastic MapReduce (Amazon EMR)
- 'Startup company' Spotify (music)
  - Instant access to over 16 million licensed songs
  - Stores its huge volume of content in Amazon S3



[21] AWS Marketplace

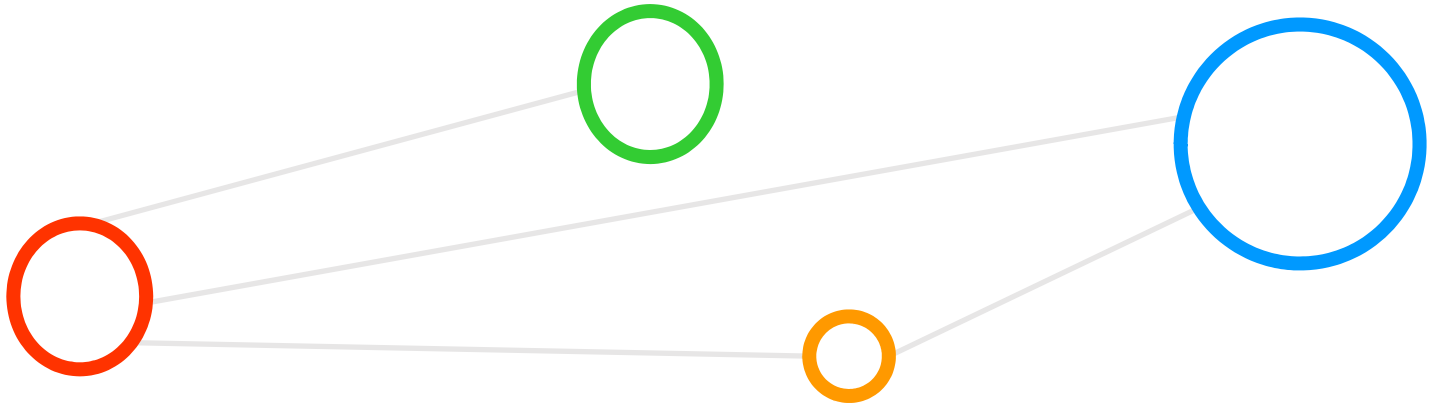
- **AWS Elastic Map Reduce (EMR) enables Hadoop & Spark on AWS cloud & is used in Airbnb applications (50 GB/day) & Spotify applications to access 16 million songs**

➢ **Lecture 8 provides more details about Amazon Web Services with different applications using Infrastructure-As-A-Service (IAAS) models**

[Video] AWS Machine Learning Examples & Application Impact

[20] YouTube video, AWS Machine Learning

Understanding Map-Reduce in Cloud Applications



Key-Value Data Structure – Simple 'Wordcount' Application Example – Example

```
// counting words example

map(String key, String value):
  // key: document name
  // value: document contents
for each word w in value:
EmitIntermediate(w, "1");

// the framework performs sort/shuffle
// with the specified keys

reduce(String key, Iterator values):
// key: a word
// values: a list of counts
int result = 0;
for each v in values:
result += ParseInt(v);
Emit(AsString(result));
```

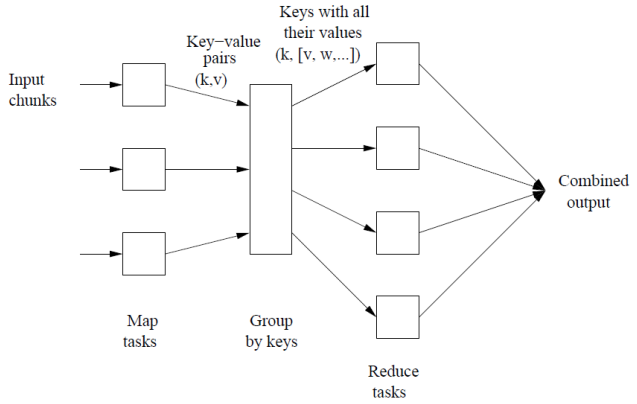
- Goal: Counting the number of each word appearing in a document (or text-stream more general)
- Wordcount in documents or Twitter textstreams are excellent examples for independent 'nicely parallel'

- Key-Value pairs are implemented as Strings in this text-processing example for each function and as 'Iterator' over a list

- Map (docname, doctext) →list (wordkey, 1), …

- Reduce (wordkey, list (1, …)) →list (numbercounted)

- map (k1,v1) → list(k2,v2)
- reduce (k2,list(v2)) → list(v2)



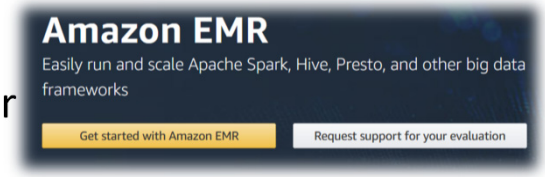
[3] MapReduce: Simplified Dataset on Large Clusters, 2004

Cloud Computing using Step-wise Approach via Amazon Web Services

- Big Data Challenges (cf. Lecture 2)
  - Jupyter, Anaconda and scikit-learn are great – but only for small data sets
  - Issues: Laptop too slow runtime (less CPUs/GPUs) and/or errors due to memory problems/limits

- Cloud Computing Step-wise Approach
  1. Check and/or create subscription (setup pay-per-use of resources, free in course)
  2. Deploy a Spark Cluster in AWS Elastic Map-Reduce (EMR) & prepare computing infrastructure
  3. Optional Startup Jupyter notebook to use Apache Spark and Apache Hadoop map reduce
  4. Create AWS S3 service 'buckets' & check application dataset locations (inputs/outputs)
  5. Deploy a Spark Cluster in AWS EMR with application WordCount using Map Reduce steps
  6. Perform automated WordCount on AWS EMR cluster
  7. WordCount result data analysis & Terminate cluster

[12] Jupyter

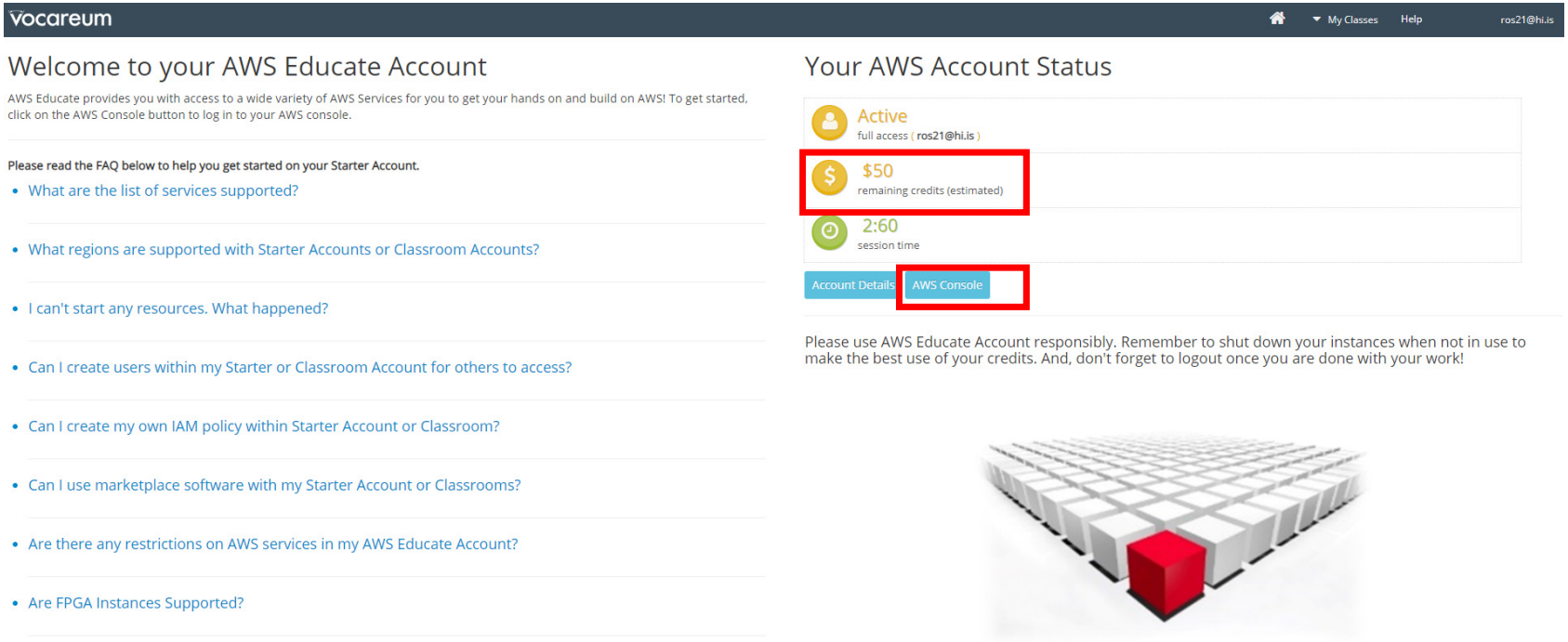


[8] AWS EMR Web page

[9] AWS Web page

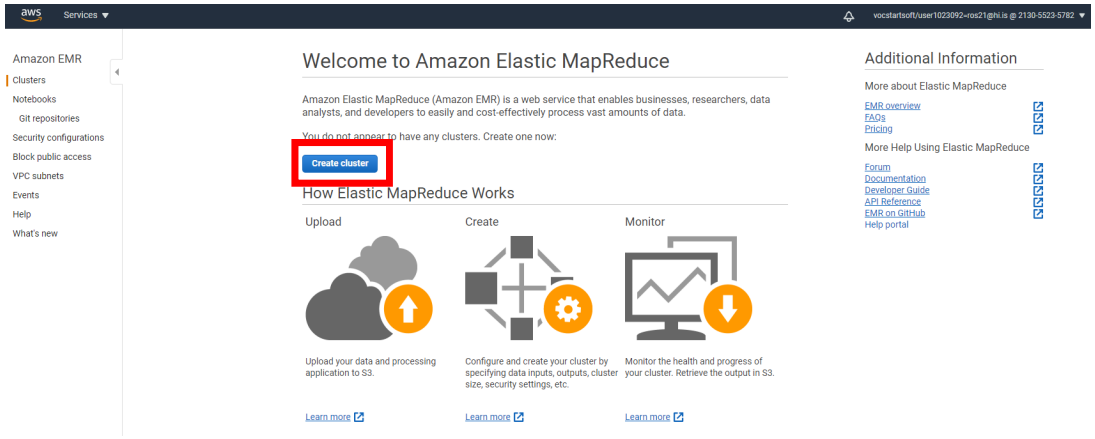
Step 1: Check and/or Create Subscription & Move to AWS Management Console

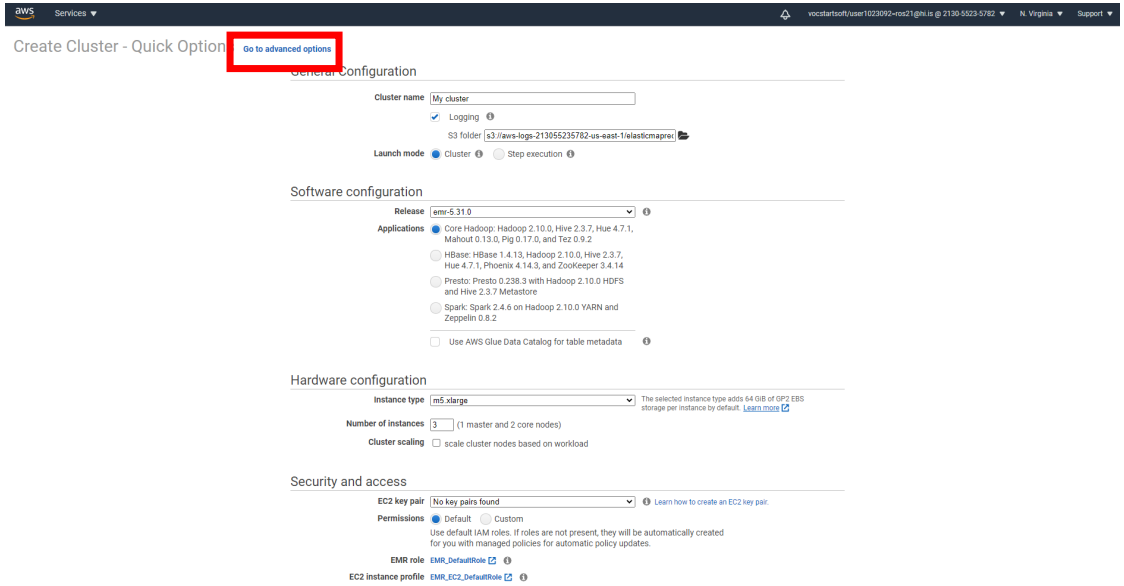
- Three approaches
  - AWS Educate
  - AWS Free Tier
  - AWS Pay-Per-Use



- **Running services in the cloud despite being not really used actively still can generate massive costs – be careful what you spend – track your spending – terminate clusters & services!**

Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR – Advanced Options



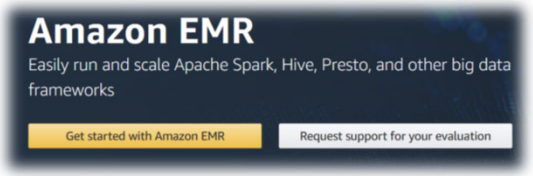


AWS Analytics services

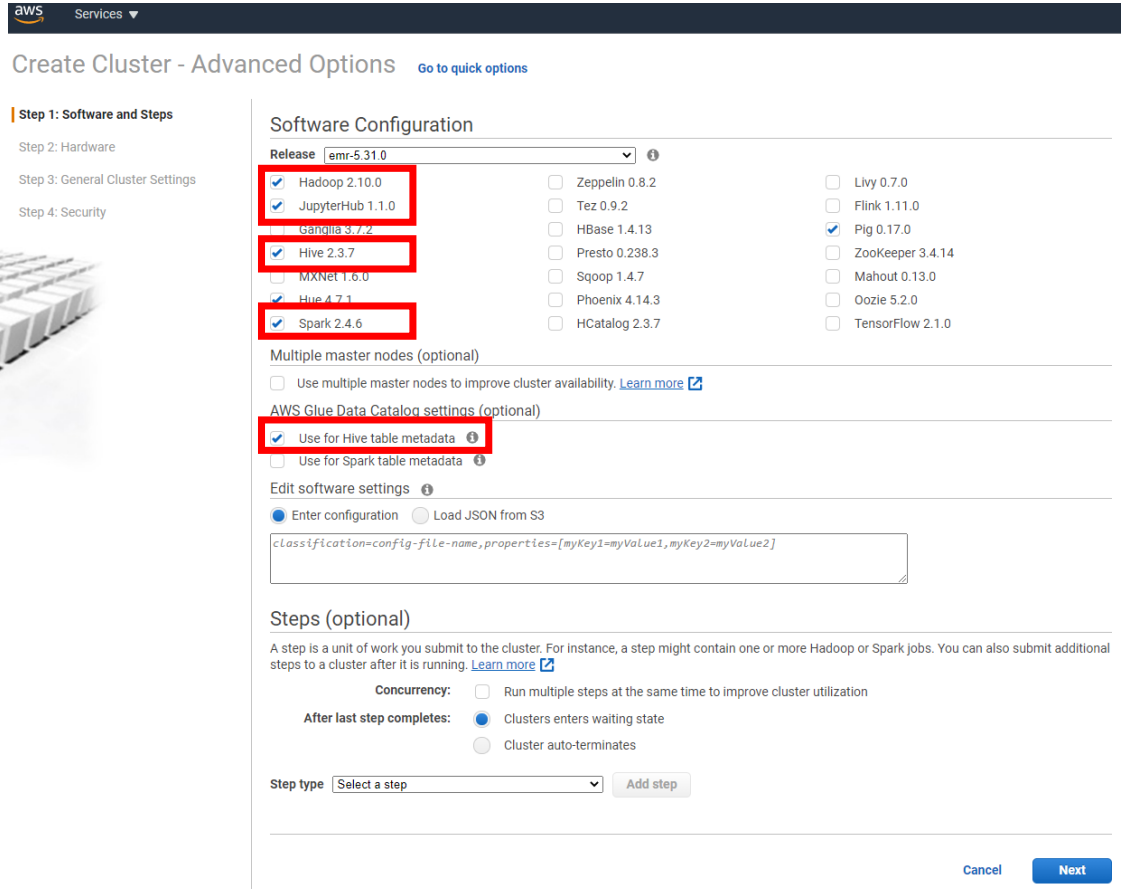
Category	Use cases	AWS service
Analytics	Interactive analytics	Amazon Athena
	Big data processing	Amazon EMR
	Data warehousing	Amazon Redshift
	Real-time analytics	Amazon Kinesis
	Operational analytics	Amazon Elasticsearch Service
	Dashboards and visualizations	Amazon Quicksight

[9] AWS Web page

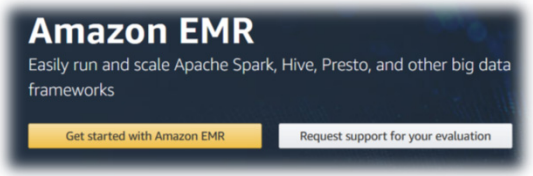
Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR (1)



[8] AWS EMR Web page

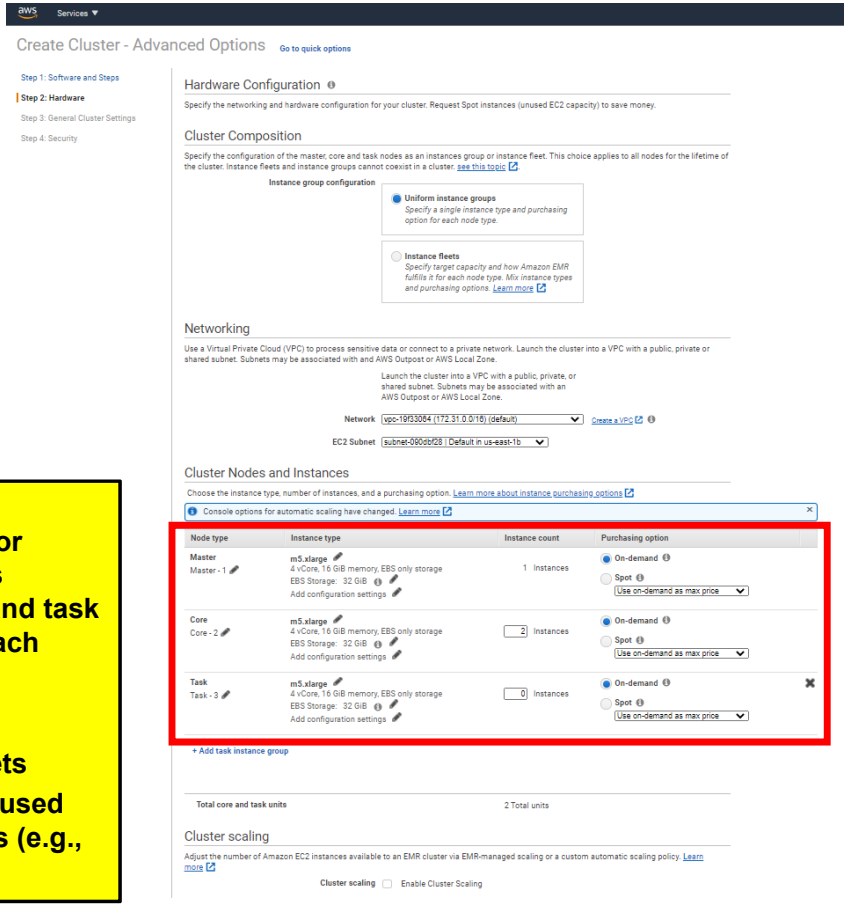


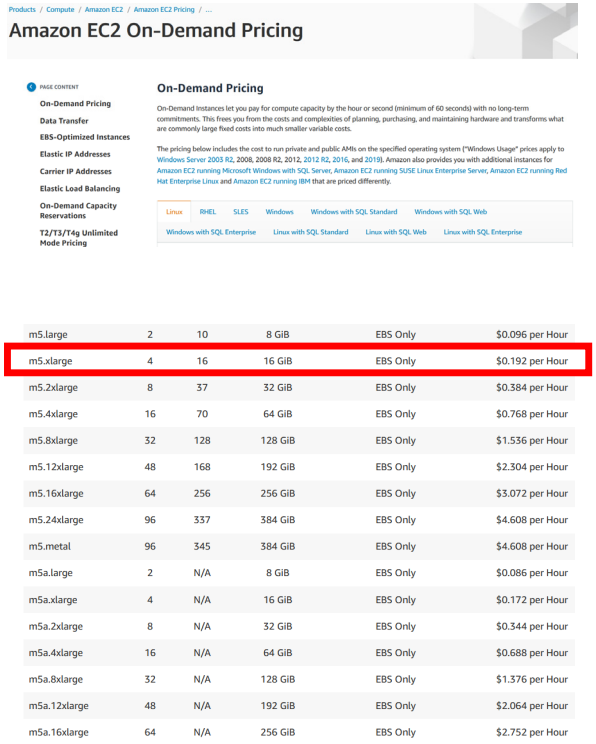
Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR (2)



[8] AWS EMR Web page

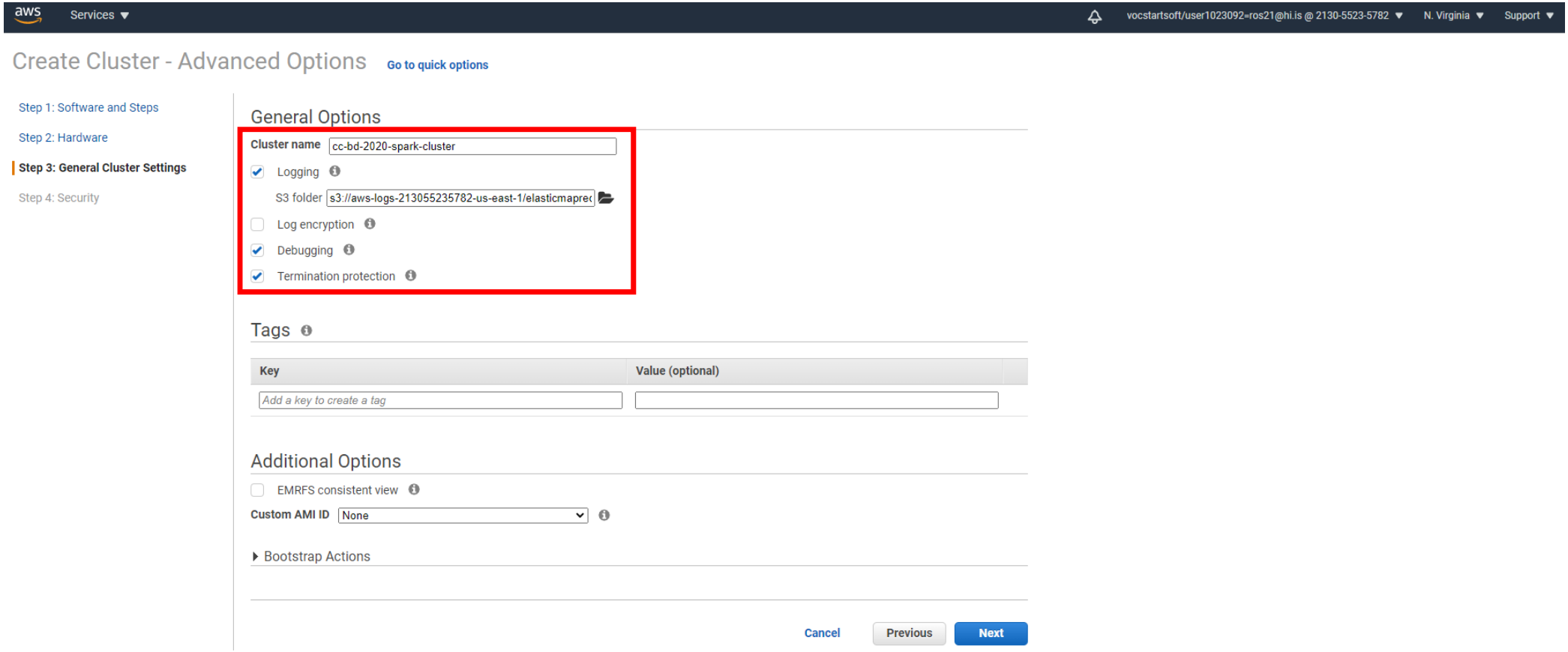
- **The Cloud deployment of AWS Elastic Map Reduce (EMR) offers various possibilities for optimizations in the configurations such as configuring the ammount of master, core, and task nodes that follow the master-worker approach**
- **Master node: scheduler and name node**
- **Core node: computing and storage data**
- **Task node: computing w/o requiring datasets**
- **Each of the nodes can be configured to be used with different types of computing resources (e.g., m5.xlarge**



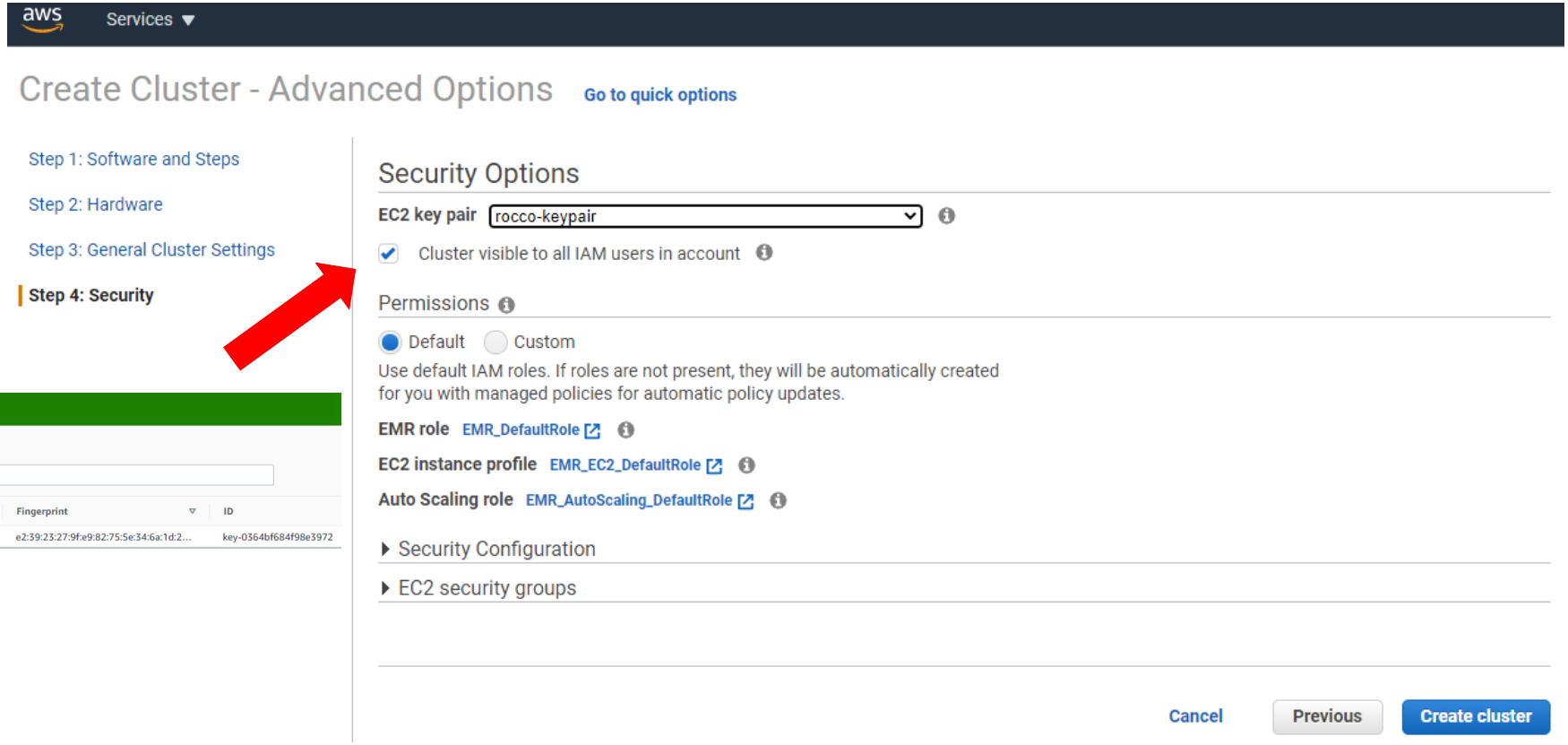


[13] AWS EC2 Pricing

Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR (3)



Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR (4)



Step 2: Deploy Apache Hadoop & Spark Cluster via AWS EMR – Starting

aws Services

Amazon EMR

- Clusters
- Notebooks
- Git repositories
- Security configurations
- Block public access
- VPC subnets
- Events
- Help
- What's new

Clone | Terminate | AWS CLI export

Cluster: cc-bd-2020-spark-cluster Starting

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID:	j-3FQEIU7LI2FTM
Creation date:	2020-10-14 18:41 (UTC+2)
Elapsed time:	0 seconds
After last step completes:	Cluster waits
Termination protection:	On Change
Tags:	– View All / Edit
Master public DNS:	–

Configuration details

Release label:	emr-5.31.0
Hadoop distribution:	Amazon 2.10.0
Applications:	Hive 2.3.7, Pig 0.17.0, Hue 4.7.1, Spark 2.4.6, JupyterHub 1.1.0
Log URI:	s3://aws-logs-213055235782-us-east-1/elasticmapreduce/
EMRFS consistent view:	Disabled
Custom AMI ID:	–

Application user interfaces

Persistent user interfaces:	–
On-cluster user interfaces:	–

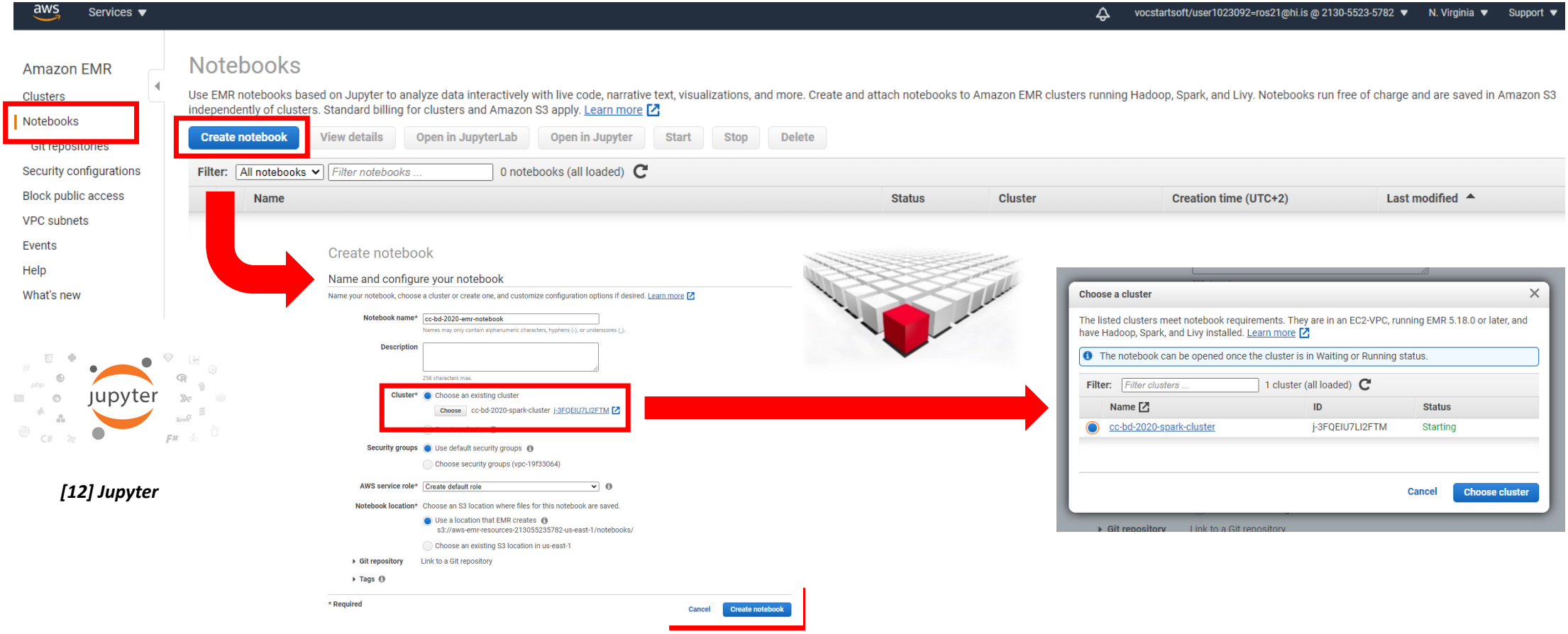
Network and hardware

Availability zone:	–
Subnet ID:	subnet-090dbf28
Master:	Provisioning 1 m5.xlarge
Core:	Provisioning 2 m5.xlarge
Task:	–
Cluster scaling:	Not enabled

Security and access

Key name:	rocco-keypair
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:	
Security groups for Core & Task:	

Optional Step 3: Startup Jupyter Notebook (1)



[12] Jupyter

Optional Step 3: Startup Jupyter Notebook (2)

[12] Jupyter

(all Spark application examples from the MS Azure Cloud from our previous lectures can also be executed in the AWS Cloud)

Notebook: cc-bd-2020-emr-notebook Starting Starting workspace(notebook). Cluster j-3FQEIU7LI2FTM.

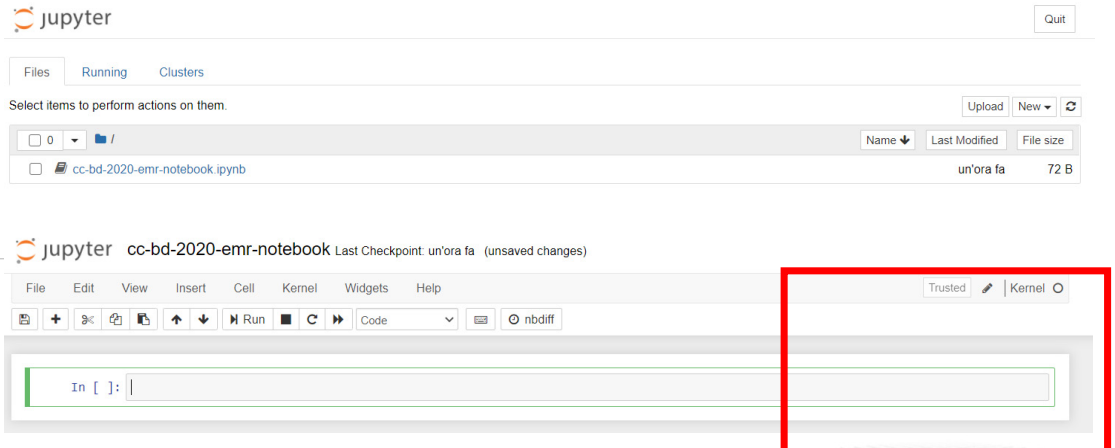
Open in JupyterLab | Open in Jupyter | Stop | Delete

Notebook

Notebook ID: e-5X8GLS6OCGHM3IYTGIRWN6WKV
Description: –
Last modified: 0 seconds ago
Last modified by: ...assumed-role/vocstartsoft/user1023092=ros21@hi.is
Created on: 2020-10-14 18:46 (UTC+2)
Created by: ...assumed-role/vocstartsoft/user1023092=ros21@hi.is
Service IAM role: EMR_Notebooks_DefaultRole
Notebook tags: creatorUserId = AROATDGYU43DJ6XHBANK2:user1023092=ros21@hi.is View All / Edit
Notebook location: s3://aws-emr-resources-213055235782-us-east-1/notebooks/

Cluster

Cluster: cc-bd-2020-spark-cluster
Cluster Id: j-3FQEIU7LI2FTM
Cluster status: Starting Configuring cluster software
Cluster tags: –
Step logs: s3://aws-logs-213055235782-us-east-1/elasticmapreduce/



Git repositories

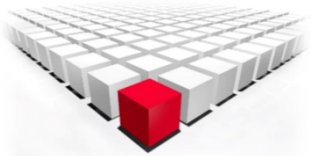
The repository can be linked to a notebook once the notebook is ready. Make sure your cluster, service role and security groups have the required settings. Learn more

Link new repository | Unlink repository

Repository name	URL	Branch	Link status	Failure reason

Step 4: Amazon Web Services – Broadly used Storage-as-a-Service (S3) Example

- S3 is ‘storage as a service’ with a Web messaging interface
  - Using API with Representational State Transfer (REST)
  - Using API with Simple Object Access Protocol (SOAP)
- Remote object storage
  - Data considered objects to be named by end users
  - Objects alongside metadata are stored in bucket containers
  - Buckets enable the organization with namespace for user identification & accounting
  - (Automatically) scalable



Expedia airbnb NETFLIX Allergan NASA JPL CATHAY PACIFIC

SCALABLE

Amazon S3 automatically scales to high request rates. It is simple to increase your read or write performance exponentially.

DURABLE

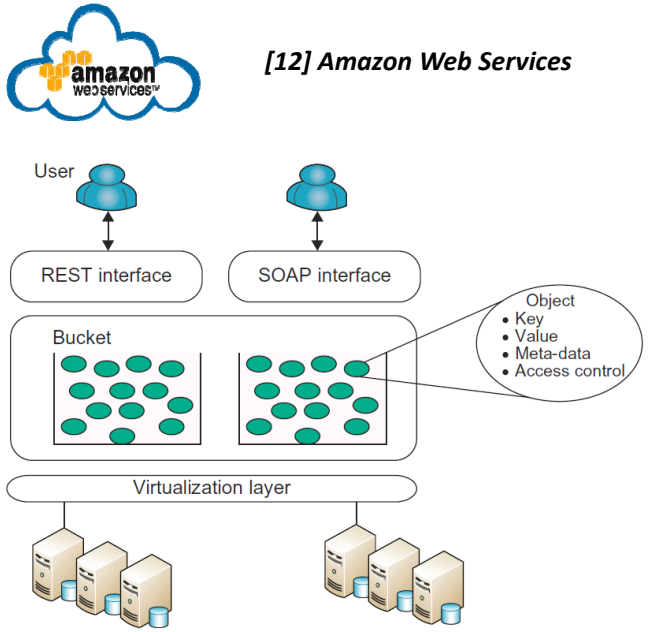
Amazon S3 runs on the world's largest global cloud infrastructure, and is designed to deliver 99.999999999% of durability.

SECURE

Amazon S3's highly durable, secure, global infrastructure offers a robust disaster recovery solution designed to provide superior data protection.

EASY MIGRATION

You can choose from the widest range of options to transfer your data into (or out of) Amazon S3. Our simple APIs make it easy to transfer data over the Internet.

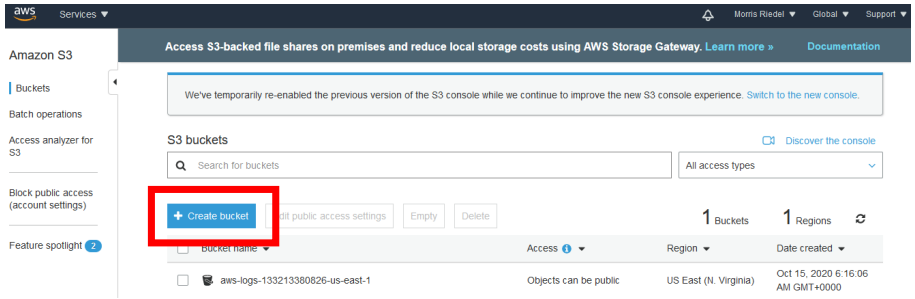


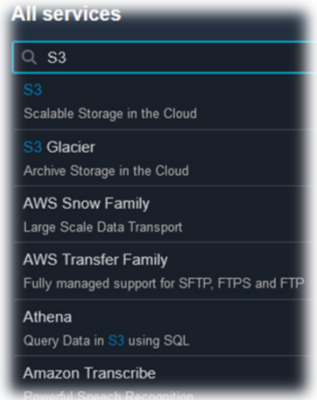
[12] Amazon Web Services

[8] Distributed & Cloud Computing Book

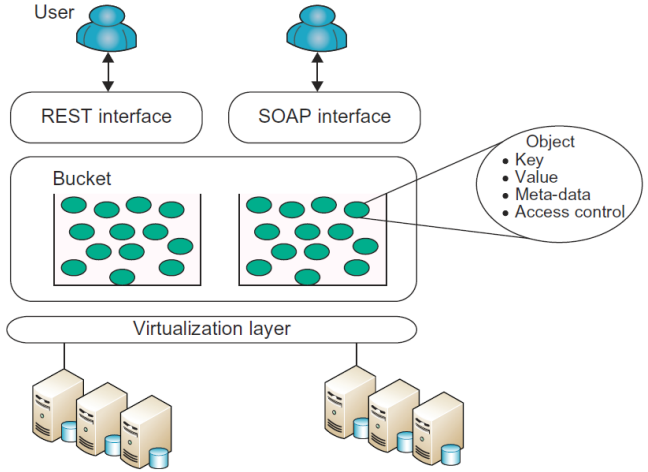
Step 4: Create AWS S3 service ‘bucket‘ & Check Application Dataset

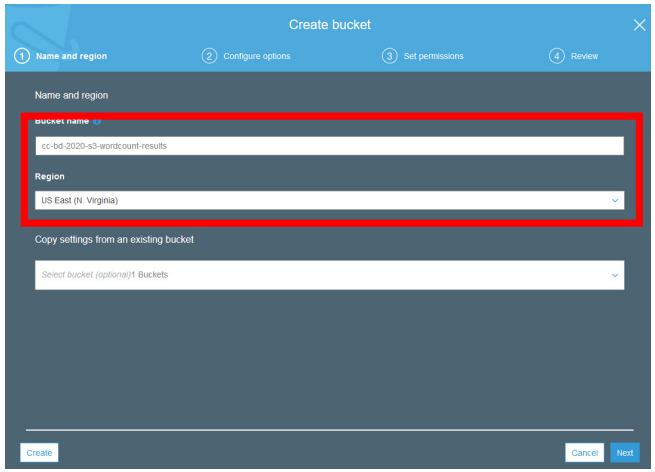
- User Bucket
  - Create own S3 bucket for WordCount results

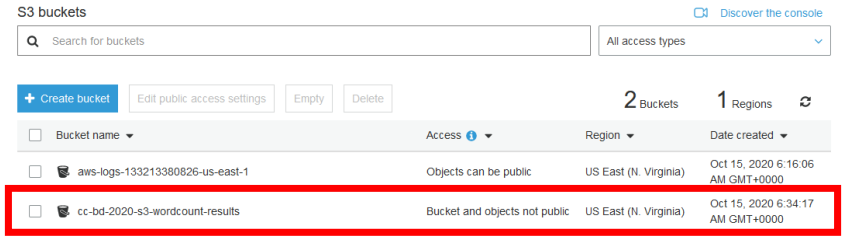




[12] Amazon Web Services







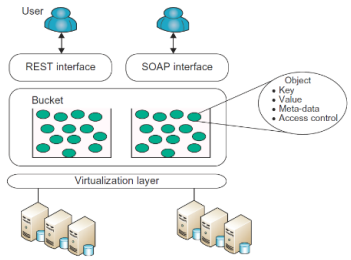
[8] Distributed & Cloud Computing Book

Step 4: Create AWS S3 service ‘bucket‘ & Check Application Dataset

- AWS Sample Bucket
  - Existing Datasets & Applications
- Input dataset
  - s3://elasticmapreduce/samples/wordcount/input
- Browser view to understand ‘flat objects‘ in buckets
  - http://s3.amazonaws.com/elasticmapreduce?prefix=samples/wordcount/input/
  - One document: http://s3.amazonaws.com/elasticmapreduce/samples/wordcount/input/0001

[12] Amazon Web Services

[8] Distributed & Cloud Computing Book



Nicht sicher | s3.amazonaws.com/elasticmapreduce?prefix=samples/wordcount/input/

This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
<ListBucketResult xmlns="http://s3.amazonaws.com/doc/2006-03-01/">
  <Name>elasticmapreduce</Name>
  <Prefix>samples/wordcount/input/</Prefix>
  <Marker/>
  <MaxKeys>1000</MaxKeys>
  <IsTruncated>false</IsTruncated>
  <Contents>
    <Key>samples/wordcount/input/0001</Key>
    <LastModified>2009-04-02T02:55:30.000Z</LastModified>
    <ETag>"e233f19d98472f37076d009d9658997a"</ETag>
    <Size>2392524</Size>
    <StorageClass>STANDARD</StorageClass>
  </Contents>
  <Contents>
    <Key>samples/wordcount/input/0002</Key>
    <LastModified>2009-04-02T02:55:32.000Z</LastModified>
    <ETag>"7ecfdd010570a0501c2a68e70cd891d4"</ETag>
    <Size>2396618</Size>
    <StorageClass>STANDARD</StorageClass>
  </Contents>
```


(0001 document)


```
CIA -- The World Factbook -- Country Listing
      World Factbook Home
    The World Factbook
       
       Country Listing
       
A
  B C D E
  F G H I
  J K L M
  N O P Q
  R S T U
  V W X Y Z
     
  World
    A
  Afghanistan
  Akrotiri
  Albania
  Algeria
  American Samoa
  Andorra
  Angola
  Anguilla
  Antarctica
  Antigua and Barbuda
  Arctic Ocean
  Argentina
  Armenia
  Aruba
  Ashmore and Cartier Islands
  Atlantic Ocean
  Australia
  Austria
  Azerbaijan
    B
  Bahamas, The
  Bahrain
  Baker Island
description under United States Pacific Island Wildlife Refuges
```


(0002 document)

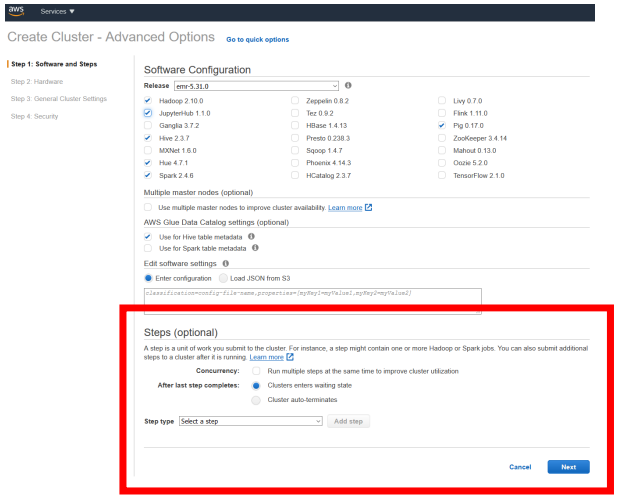

```
            Transnational Issues
            Nepal
                        Disputes - international:
                        joint border commission continues to work on contested sections of boundary with
activities; approximately 106,000 Bhutanese Lhotshampas (Hindus) have been confined in refugee camps in s
                                  Refugees and internally displaced persons:
                                        refugees (country of origin): 107,803 (Bhutan); 20,153 (Tibet/Chi
                                  IDPs: 50,000-70,000 (remaining from ten-year Maoist insurgency that offic
                                  Illicit drugs:
                                  illicit producer of cannabis and hashish for the domestic and internatior
This page was last updated on 1 January 2003
                                  This page was last updated on 19 March, 2009
 
CIA - The World Factbook -- Fiji
a {  font-family: Verdana, Arial, Helvetica, sans-serif; font-size: 12px; color: #000000; text-decoratior
 
      Country
         List | World Factbook Home
    The World Factbook
       
       Fiji
            Introduction
            Fiji
                        Background:
                        Fiji became independent in 1970, after nearly a century as a British colony. Demc
British in the 19th century). The coups and a 1990 constitution that cemented native Melanesian control c
peaceful elections in 1999 resulted in a government led by an Indo-Fijian, but a civilian-led coup in May
May 2006, QARASE was ousted in a December 2006 military coup led by Commodore Voreqe BAINIMARAMA, who ini
to hold elections.
```

Step 5: Deploy a Spark Cluster in AWS EMR with Sample Application WordCount

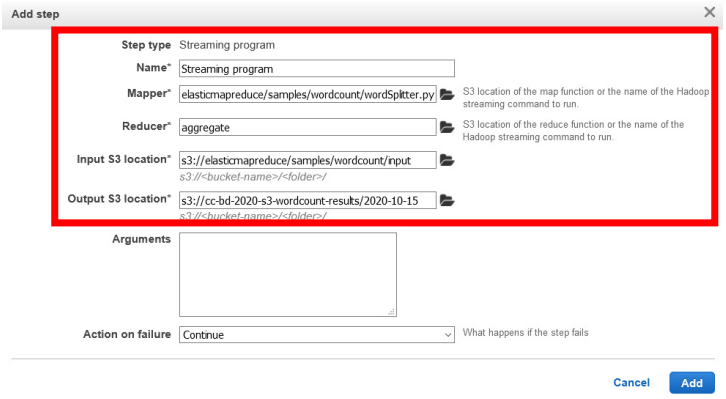
- AWS Example Map Application (+ Built-In Reduce)
  - s3://elasticmapreduce/samples/wordcount/wordSplitter.py
- Input dataset
  - s3://elasticmapreduce/samples/wordcount/input

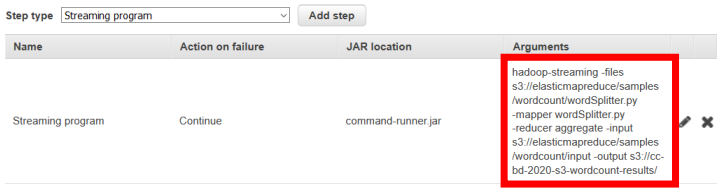
```
./elastic-mapreduce -j
JobFlowID
--stream \
--mapper
s3://elasticmapreduce/samples/wordcount/wordSplitter.py
\
--input
s3://elasticmapreduce/samples/wordcount/input
\
--output
s3n://myawsbucket/output
\
--reducer
aggregate
```

(alternative approach is possible via the command-line interface and log into the cluster with SSH keys)



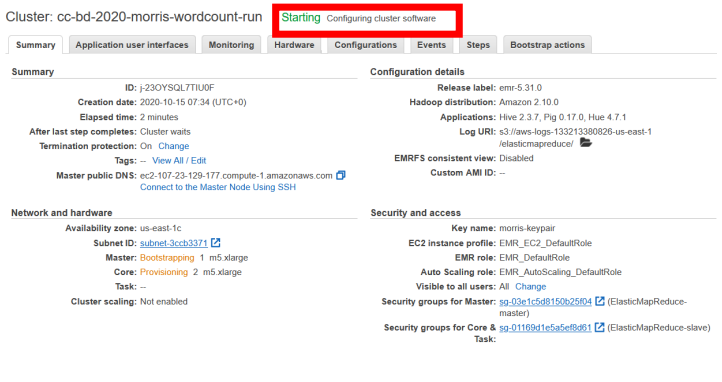
- **Hadoop streaming is a feature of Apache Hadoop that enables to create & run job flows using any executable program/script as Apache Hadoop mappers (i.e., map phase) & reducers (i.e., reduce phase)**
- **Example uses the built-in reducer called 'aggregate' that adds up counts of words from the map step of the wordsplitter-py mapper function**
- **Add to our 'bucket' a new(!) folder specific for the results that will be created by the framework itself**

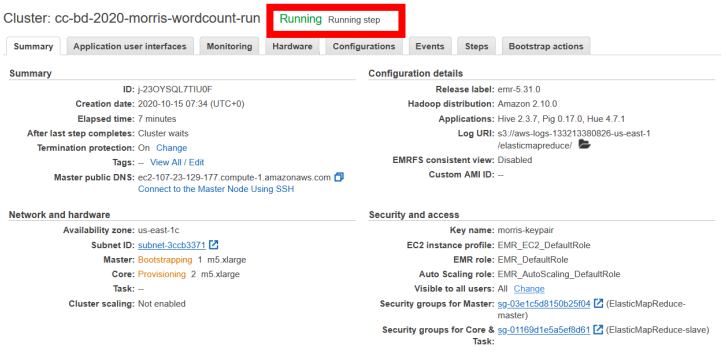




Step 6: Perform Automated WordCount on AWS EMR cluster & Debugging

- Application is executed as ‘streaming‘ step
  - Using the EMR cluster & map-reduce phases



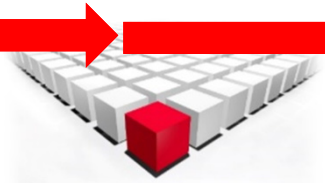



```
#!/usr/bin/python
import sys
import re

def main(argv):
    pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
    for line in sys.stdin:
        for word in pattern.findall(line):
            print "LongValueSum:" + word.lower() + "\t" + "1"


if __name__ == "__main__":
    main(sys.argv)
```

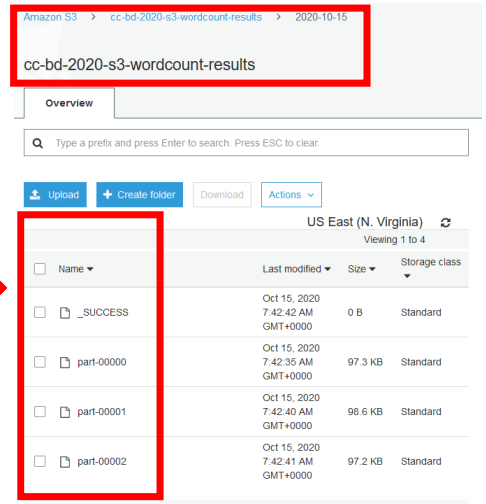
(‘mapper function‘ in wordsplitter.py)

(map phase)

```
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0012
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0011
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0010
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0009
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0008
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0007
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0006
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0005
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0004
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0003
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0002
https://elasticmapreduce.s3.amazonaws.com/samples/wordcount/input/0001
```



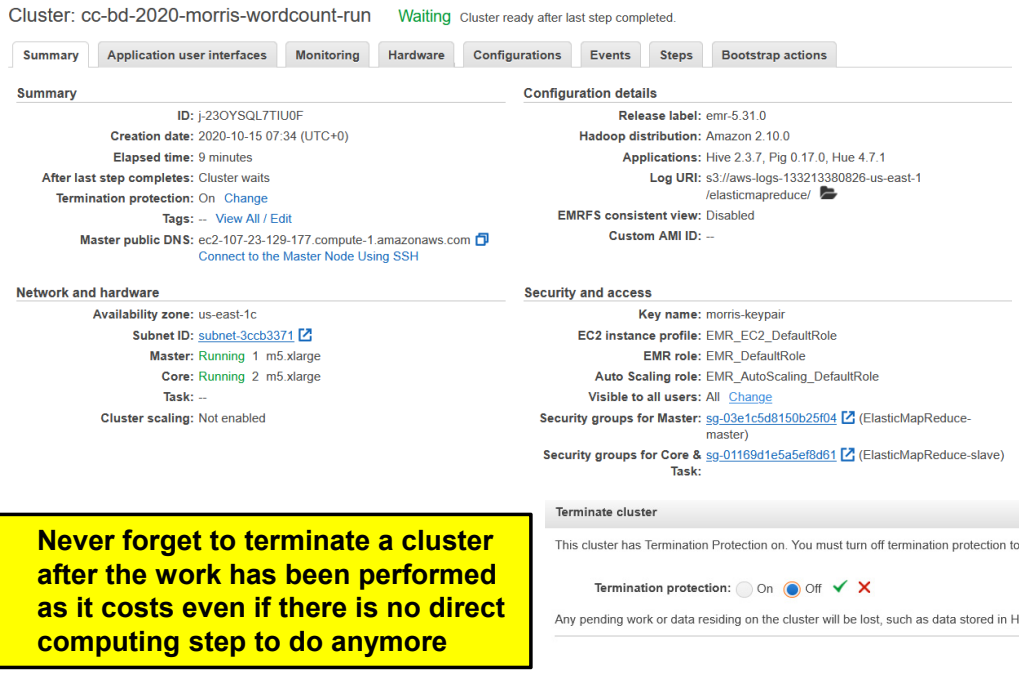
(reduce phase using built-in ‘aggregate‘)

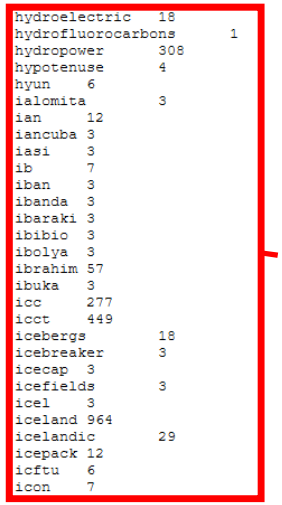


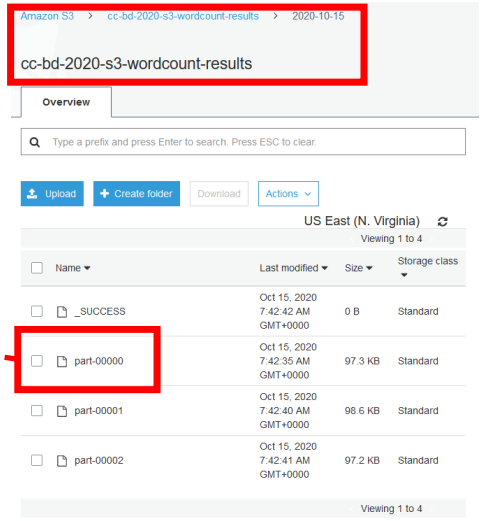
(very good debugging view example of a failed map reduce job)

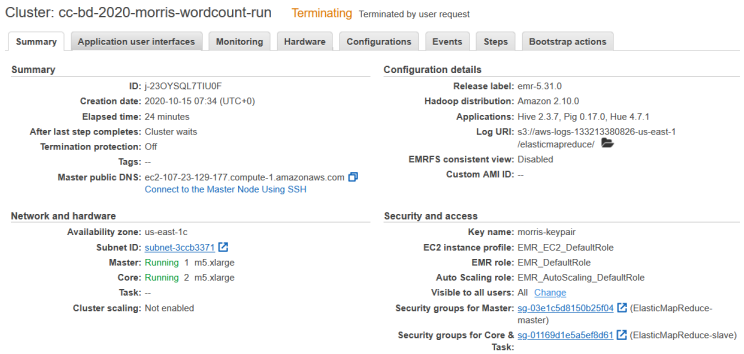
Step 7: WordCount Result Data Analysis & Terminate AWS EMR Cluster

- Application is completed after ‘last step‘
  - In this example was only one step
  - The cluster remains ‘waiting‘ as we configured





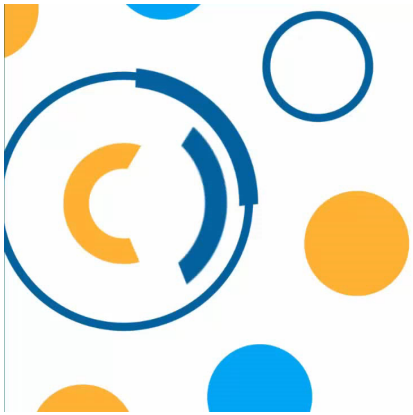




- **Never forget to terminate a cluster after the work has been performed as it costs even if there is no direct computing step to do anymore**

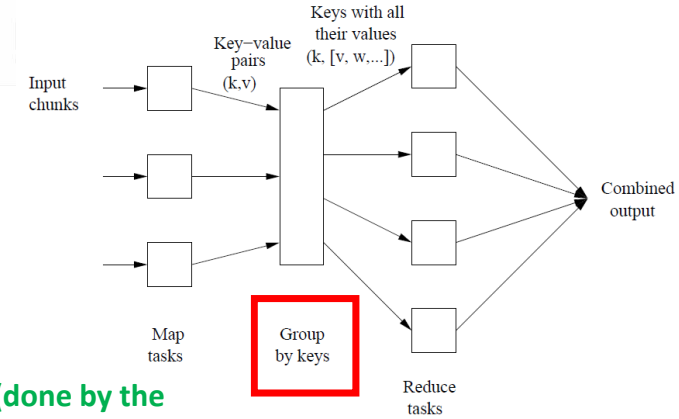
Understanding Map-[Sort/Shuffle/Group]-Reduce & Key-Value Data Structures

- Programming Model
  - Known as 'two phase approach', but actually 'three phases'
  - Two key functions to program by user: map and reduce
  - Third phase 'sort/shuffle' works with keys and sorts/groups them
  - Input keys and values **(k1,v1)** are drawn from a different domain than the output keys and values **(k2,v2)**
  - Intermediate keys and values **(k2,v2)** are from the same domain as the output keys and values
  - Definition of 'generic' keys is dependent on application problem



- **map (k1,v1) → list(k2,v2)**
- **reduce (k2,list(v2)) → list(v2)**

[18] YouTube video, Map-Reduce explained

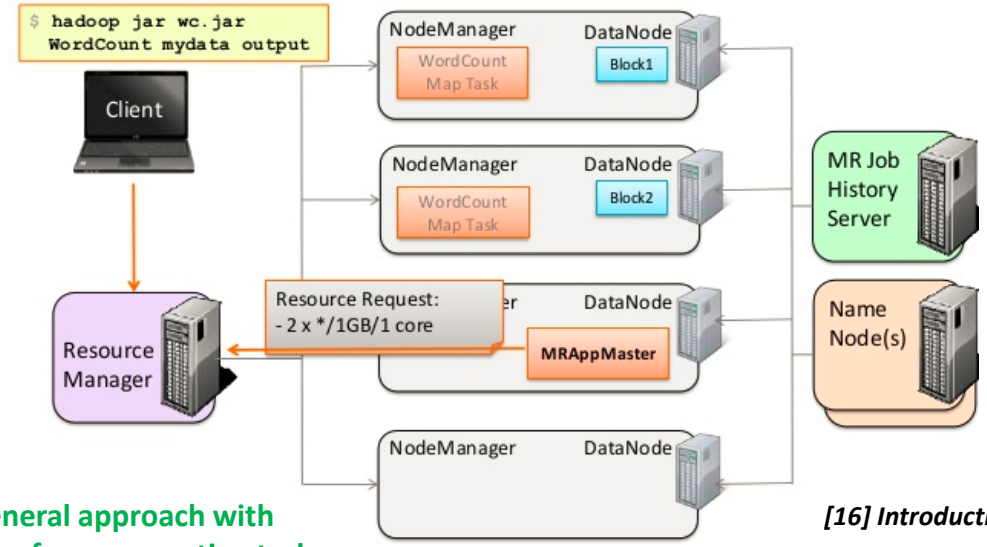


(done by the framework)

Modified from [1] Mining of Massive Datasets

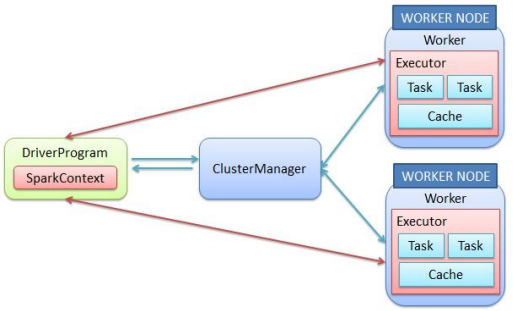
Hadoop Resource Manager YARN – Wordcount Example as Application Blueprint

- Apache Hadoop offers the Yet Another Resource Negotiator (YARN) scheduler for map-reduce jobs
- Idea of Yarn is to split up the functionalities of resource management & job scheduling/monitoring
- The ResourceManager is a scheduler that controls resources among all applications in the system
- The NodeManager is the per-machine system that is reponsible for containers, monitoring of their resource uage (cpu, memory, disk, network) and reporting to the ResourceManager



[16] Introduction to Yarn

[2] Apache Hadoop Web page

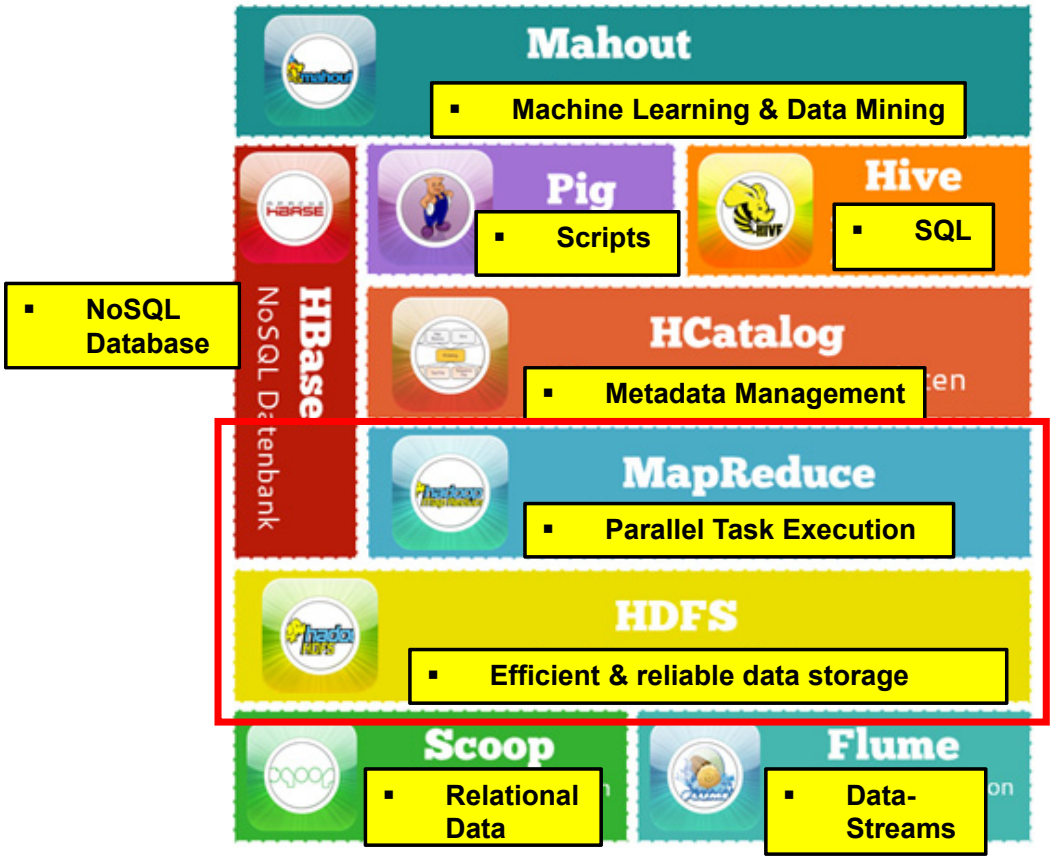


[17] Understanding Parallelization of Machine Learning Algorithms in Apache Spark

(the framework general approach with map & reduce to perform computing tasks on data can be used by many applications)

(the map & reduce functions from wordcount represent the general framework blueprint of how to define tasks)

Larger Hadoop Ecosystem with Big Data Analytics 'Tools' – Revisited

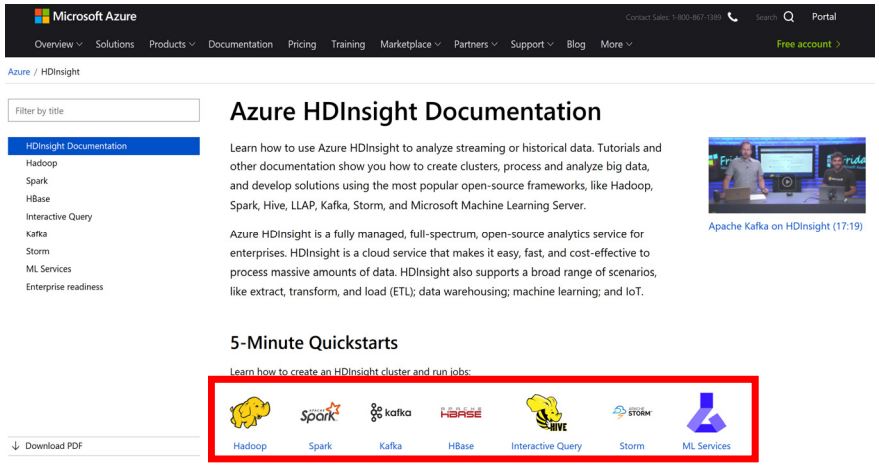


[5] Google Dataproc service

[7] Microsoft Azure HDInsight Service

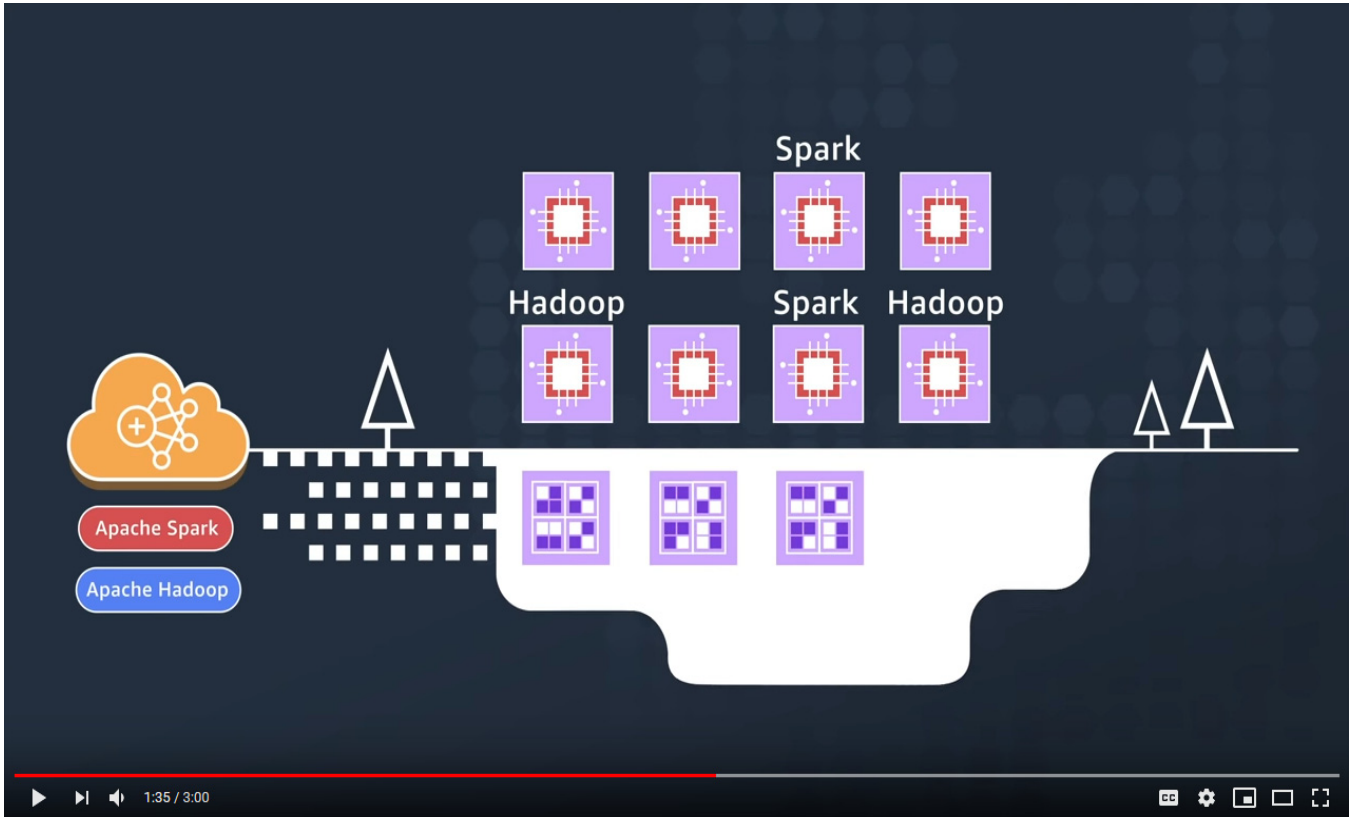
Modified from [6] Map-Reduce

HDInsight



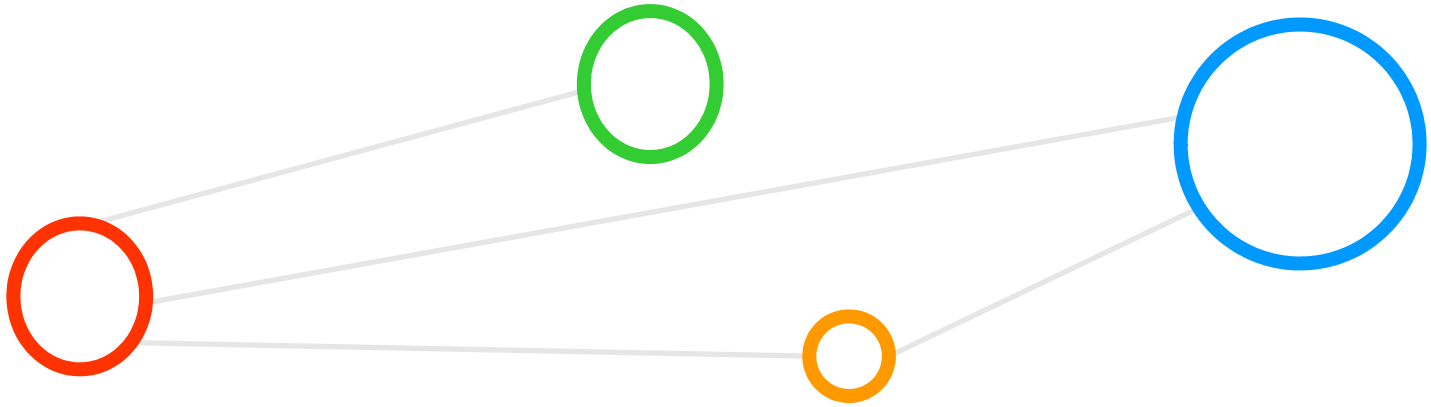
(known open source analytics frameworks)

[Video] Map-Reduce Summary with Examples



[19] YouTube video, An Introduction to Amazon EMR

Lecture Bibliography



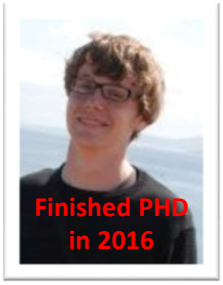
Lecture Bibliography (1)

- [1] Mining of Massive Datasets, Online: http://infolab.stanford.edu/~ullman/mmds/book.pdf
- [2] Apache Hadoop Web page, Online: http://hadoop.apache.org/
- [3] J. Dean, S. Ghemawat, 'MapReduce: Simplified Data Processing on Large Clusters', OSDI'04: Sixth Symposium on Operating System Design and Implementation, December, 2004
- [4] Stampede Virtual Workshop, Online: http://www.cac.cornell.edu/Ranger/MapReduce/dfs.aspx
- [5] Google DataProc Service, Online: https://cloud.google.com/dataproc/
- [6] Einführung in Hadoop (German language), Online: http://blog.codecentric.de/2013/08/einfuhrung-in-hadoop-die-wichtigsten-komponenten-von-hadoop-teil-3-von-5/
- [7] Microsoft Azure HDInsight Service, Online: https://azure.microsoft.com/en-us/services/hdinsight/
- [8] Amazon Web Services (AWS) Elastic Map-Reduce (EMR), Online: https://aws.amazon.com/emr
- [9] Amazon Web Services Web Page, Online: https://aws.amazon.com
- [10] Amazon Web Services Educate Web Page, Online: https://aws.amazon.com/education/awseducate/
- [11] AWS Services Supported with AWS Educate Starter Account, Online: https://awseducate-starter-account-services.s3.amazonaws.com/AWS_Educate_Starter_Account_Services_Supported.pdf

Lecture Bibliography (2)

- [12] K. Hwang, G. C. Fox, J. J. Dongarra, 'Distributed and Cloud Computing', Book, Online: http://store.elsevier.com/product.jsp?locale=en_EU&isbn=9780128002049
- [13] Amazon Web Services EC2 On-Demand Pricing models, Online: https://aws.amazon.com/ec2/pricing/on-demand/
- [14] SSH Client MobaXterm**,** Online: https://mobaxterm.mobatek.net/
- [15] Jupyter Web page, Online: http://jupyter.org/
- [16] SlideShare, ,'Introduction to Yarn and MapReduce', Online: https://www.slideshare.net/cloudera/introduction-to-yarn-and-mapreduce-2
- [17] Understanding Parallelization of Machine Learning Algorithms in Apache Spark, Online: https://www.slideshare.net/databricks/understanding-parallelization-of-machine-learning-algorithms-in-apache-spark/11
- [18] YouTube, Tutorial: MapReduce explained, Online: https://www.youtube.com/watch?v=lgWy7BwIKKQ
- [19] YouTube, An introduction to Amazon EMR - Amazon Web Services, Online: https://www.youtube.com/watch?v=QuwaBOESGiU
- [20] YouTube, AWS Machine Learning Solves Unique Problems, Online: https://www.youtube.com/watch?v=N0aFuCHUXpo
- [21] AWS Marketplace, Online: https://aws.amazon.com/marketplace/
- [22] Key Concepts from the AWS Cloud, Online: https://blogs.sap.com/2010/01/28/key-concepts-from-the-aws-cloud/

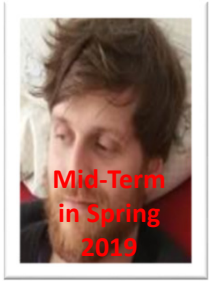
Acknowledgements – High Productivity Data Processing Research Group



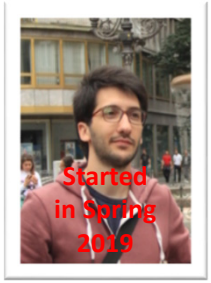
PD Dr.
G. Cavallaro

Senior PhD
Student A.S. Memon

Senior PhD
Student M.S. Memon



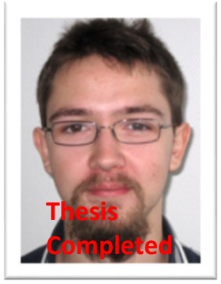
PhD Student
E. Erlingsson



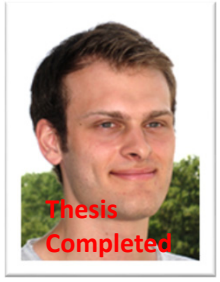
PhD Student
S. Bakarat

PhD Student
R. Sedona

Dr. M. Goetz
(now KIT)



MSc M.
Richerzhagen
(now other division)



MSc
P. Glock
(now INM-1)

MSc
C. Bodenstein
(now
Soccerwatch.tv)

MSc Student
G.S. Guðmundsson
(Landsverkjun)

This research group has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 763558 (DEEP-EST EU Project)


services
measurement
within
fusion
new
funding
Policy-based
concepts
device
Analysis
forms
cross-disciplinary
resources
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
different
Cross-Disciplinary
methods
High
big
nodes
PB
community
Environment
Modelling
EU
Science
centers
Key
scientific
important
files
etc
DKRZ
Use
storage
basis
Infrastructure
model
international
SMAQ
engineering
References
formats
simulations
ENES
technologies
brain
Big
Project
Data
increasing
via
EUDAT
physical
processing
IPCC
Common
TB
MapReduce
create
Methods
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
Large
Summary
data-intensive
Service
Scientific
levels
often
stored
Resources
analyze
DLCL
Level
Research
external
use
Matters
Concrete
performance
JSC
computational
research
e.g
science
HPC
Juelich
system
datasets
acquisition
compute
Landscape
management
System
Climate
modelling
AR5
Hardware
distributed
large-scale
cores
access
hundreds
Nutshell
Services
Technologies
Earth
Structure
current
network
DLCLs
Understanding
structures
directory
project
General
Germany
Simulation
Provide
NASA
Energy
systems
day
Processing
Health