



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

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

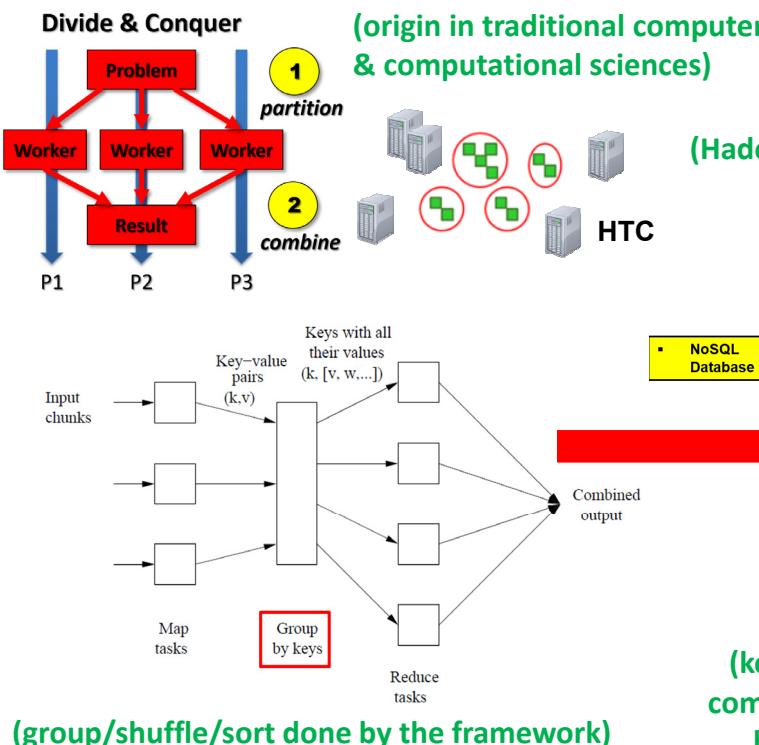


HELMHOLTZAI

ARTIFICIAL INTELLIGENCE
COOPERATION UNIT

Review of Lecture 5 – Map-Reduce Computing Paradigm

■ Map-Reduce Approach



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

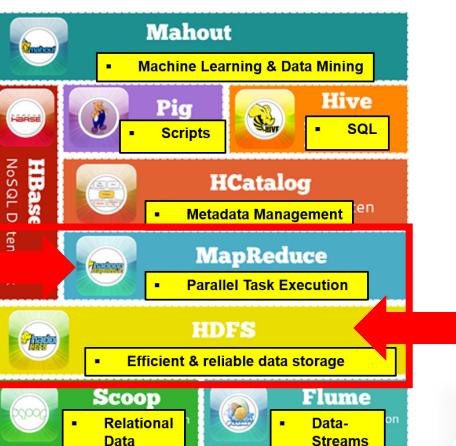
Lecture 5.1 – Understanding Map-Reduce in Cloud Applications

■ Hadoop Distributed File System (HDFS)



[2] Apache Hadoop

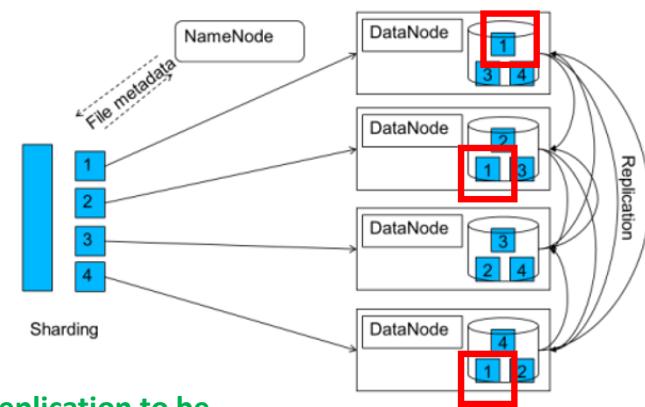
(Hadoop 'big data frameworks' ecosystem)



(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



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

Modified from [4] Stampede Virtual Workshop



Outline of the Course

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

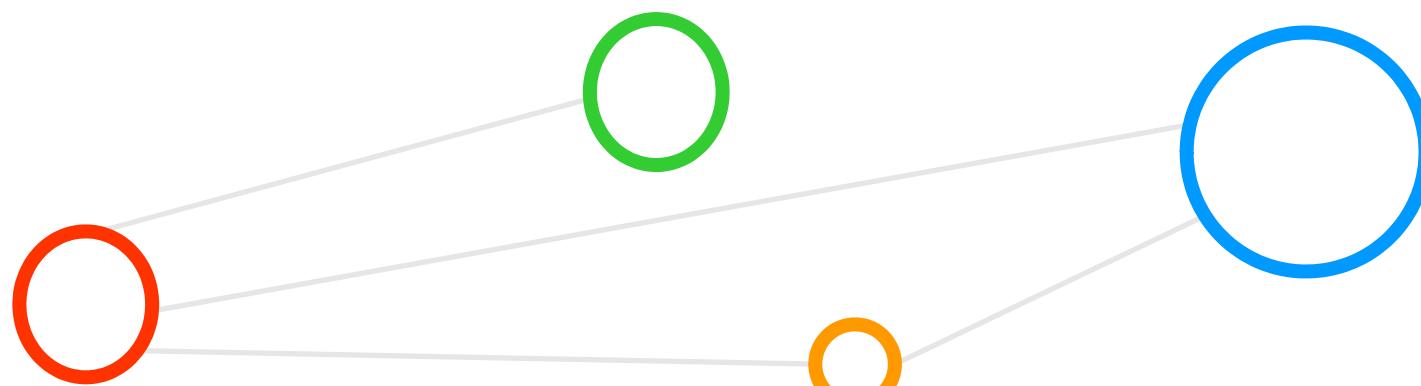
Outline

- 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*

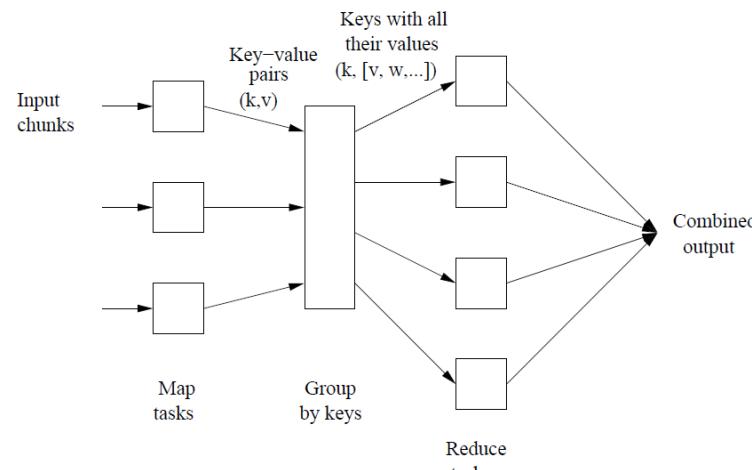


Amazon Web Services Cloud Environment



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



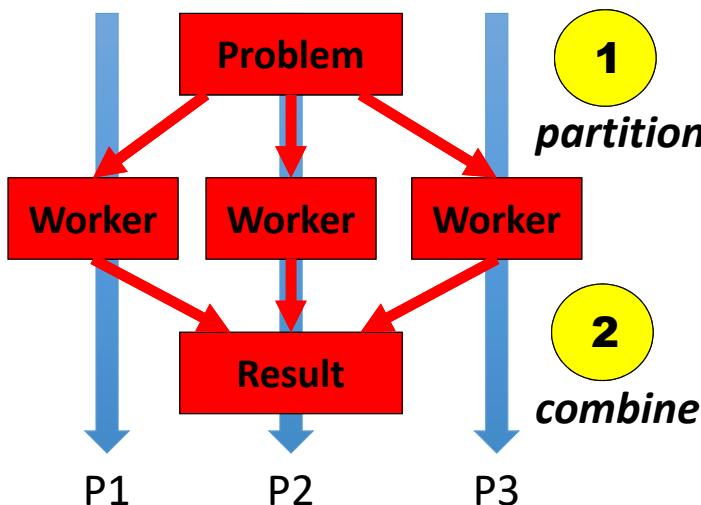
[2] Apache Hadoop

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

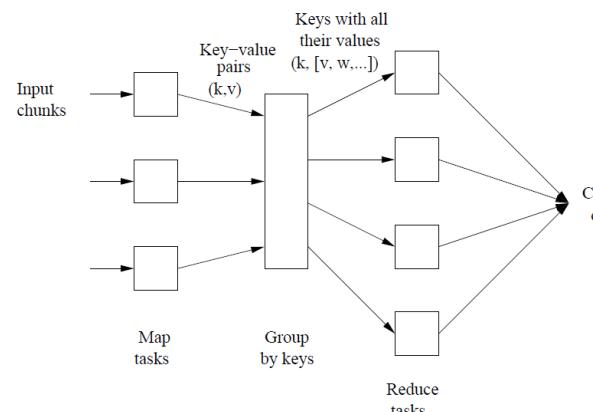
- Follows divide & conquer approach
- Injects a key element between this process: short/shuffle/group
- Open Source Implementation:
[Apache Hadoop](#)



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



[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



[9] AWS Web page

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

Amazon EMR

Easily run and scale Apache Spark, Hive, Presto, and other big data frameworks

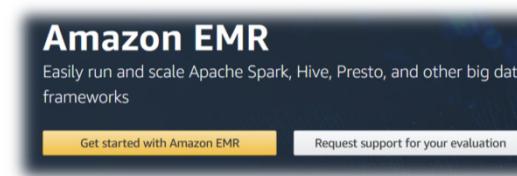
[Get started with Amazon EMR](#)

[Request support for your evaluation](#)

[8] AWS EMR Web page

Cloud Computing Approach with AWS EMR using AWS Educate

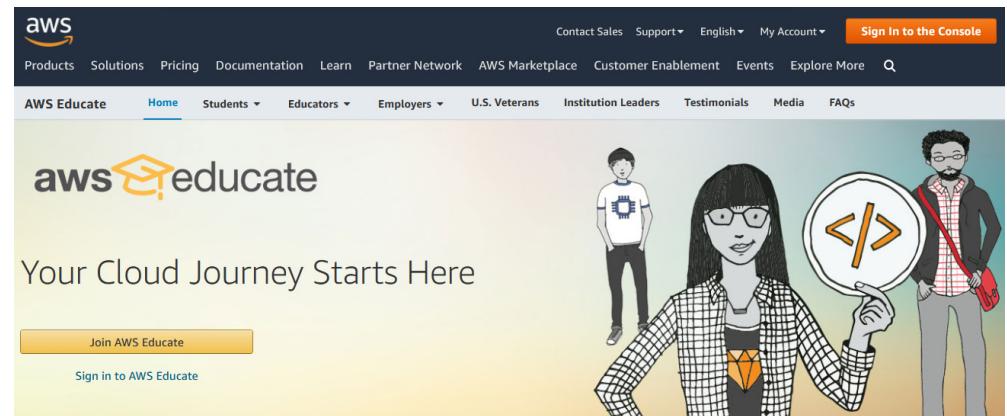
- Amazon Web Services (AWS) Cloud
 - Wide variety of different cloud-based services & resources for many application areas
 - Managed via AWS Management Console
 - Needs an AWS account
- Elastic Map Reduce (EMR) ‘Service’
 - Full managed services to deploy known open source ‘big data analytics’ frameworks
 - 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 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 verification 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	<p>Starter Accounts support only following instance types: t2.small, t2.micro, t2.nano, 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.</p> <p>The following are NOT supported in EC2/VPC:</p> <ul style="list-style-type: none">• 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 <p>*Additional restrictions may apply</p>
ECS	
ECR	
EKS	
ElastiCache	No reserved instance purchase supported
ElasticFilesystem	
ElasticLoadBalancing	
ElasticInference	
ElasticMapReduce	
events	
Execute-api	
ElasticBeanstalk	
Firehose	
Forecast	
Gamelift	
Ground Truth Labeling	
Glue	
Guard Duty	
GreenGrass	
Health	
IAM	<p>You can create users, but cannot associate login profile or access keys for them.</p> <p>* Additional restrictions may apply</p>

AWS Educate Starter Account – Educator Profile Example

■ Educator Profile

■ AWS Educate Classrooms

Classrooms, Promotional Credits & Tools

Get your class up and running quickly on AWS with tools to help you set up cloud assignments and class cloud infrastructure. Check out the AWS Educate LMS integration option, class promotional credit request, cloud tools for the classroom, and tutorials to get your students started on AWS in 10 minutes.

Request or go to an AWS Educate Classroom

Request or access your AWS Educate Classrooms. Follow a simple 3-step process to get started and provide your students with free and easy access to the AWS Console for homework, labs, and projects.

Request Credits for Your Class

AWS Educate provides an option for educators at member institutions to request a centralized AWS Promotional Credit code granting free usage to support setting up shared resources for homework, labs, and projects.

AWS Educate and Your LMS

Streamline the process students use to access AWS resources for your course through a simple integration with your LMS.

Request Registration Link For Your Class

If you are a high school / secondary school teacher, you can streamline the registration process by requesting a custom signup link for your students.



Content Classrooms & Credits Professional Resources AWS Account Profile



AWS Educate Starter Account

Your cloud journey has only just begun. Use your AWS Educate Starter Account to access the AWS Console and resources, and start building in the cloud!

[AWS Educate Starter Account](#)

Your account has an estimated **150** credits remaining and access will end on **Oct 12, 2021**.

Note: Clicking this button will take you to a third party site managed by Vocareum, Inc. ("Third Party Servicer"). In addition to the AWS Educate terms of service, your use of the AWS Educate Starter Account is governed by the Third Party Servicer's terms, including its Privacy Policy. AWS assumes no responsibility or liability and makes no representations or warranties regarding services provided by a Third Party Servicer.

You are eligible to renew your account and receive more credit!

[Start Renewal](#)

[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

The screenshot shows the AWS Educate student dashboard. At the top, there is a navigation bar with links for Portfolio, Career Pathways, Badges, Jobs, AWS Account, and Logout. The 'My Classrooms' link is highlighted with a red box. Below the navigation bar, there is a header with the user's name 'Rocco Sedona', consecutive days (1), pathways completed (0), badges earned (0), and a language preference dropdown set to English. The main content area is titled 'My Classrooms' and contains a sub-instruction: 'View your list of Classroom invitations and accept or decline the invitation. Access a Classroom by clicking Go to my classroom.' Below this, there is a table listing a single classroom invitation:

Course Name	Description	Educator	Course End Date	Credit Allocated Per Student	Status
Cloud Computing & Big Data - Parallel & Scalable Machine Learning & Deep Learning	Overview of high performance computing (HPC) and "Big Data", HPC environments with computing, network and storage resources, overview of parallel programming. Storage infrastructures and services for Big Data, Big Data analytics, the map-reduce paradigm, structured and unstructured data. Practical exercises: (A) Students will use the Amazon Web Services (AWS) cloud or equivalent to set up a multi-computer web service and an associated multi-computer testing application. (B) Students will get hands on experience of processing large data sets using map-reduce techniques with AWS. More information and initial lectures at instructors personal Web page: http://www.morrisriedel.de/cloud-computing-and-big-data-course-fall-2020	Morris Riedel	12/31/2020	\$50	Accept Invitation Decline

The screenshot shows the 'Welcome to the AWS Educate Community' page. It features a 'Set Your Password' form with fields for 'Your Login Credential' (set to 'ros21@hi.is'), 'New Password', 'Verify New Password', and a 'Set Password' button. Below the form, there is a note about password security and a list of password requirements:

- i. Password must be at least 8 characters long
- ii. Password must contain at least one letter
- iii. Password must contain at least one number
- iv. Password cannot equal or contain your user name
- v. Password must contain at least one of the following characters ! # % _ - + < >

At the bottom right of the page, there is a blue button labeled 'Go to classroom' with a small checkmark icon.

AWS Educate Starter Account – Account Status in Classrooms

■ Workbench & Example Classroom

▪ Cloud Computing & Big Data – Parallel and Scalable Machine Learning and Deep Learning

The screenshot shows the AWS Educate Starter Account interface. At the top, there's a navigation bar with icons for Home, My Classes, Manage, Help, and an email icon, along with the user's email address: morris@hi.is.

Welcome to your AWS Educate Account

AWS Educate provides you with access to a wide variety of AWS Services for you to get your hands on and build on AWS! To get started, click on the AWS Console button to log in to your AWS console.

Please read the FAQ below to help you get started on your Starter Account.

- What are the list of services supported?
- What regions are supported with Starter Accounts or Classroom Accounts?
- I can't start any resources. What happened?
- Can I create users within my Starter or Classroom Account for others to access?

Your AWS Account Status

	Active full access (morris@hi.is)
	\$50 remaining credits (estimated)
	2:60 session time

[Account Details](#) [AWS Console](#)

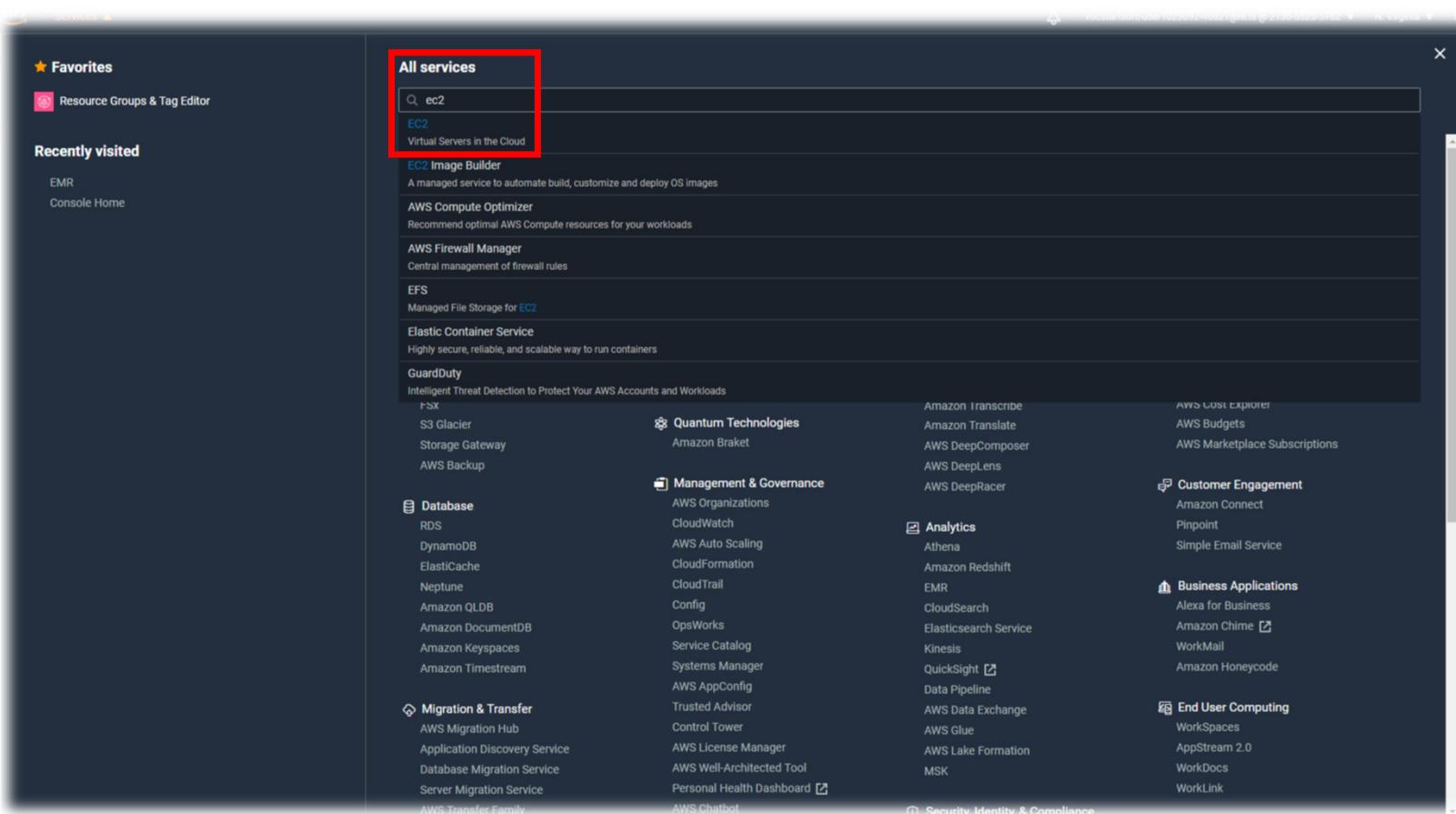
Please use AWS Educate Account responsibly. Remember to shut down your instances when not in use to make the best use of your credits. And, don't forget to logout once you are done with your work!

Course Name	Request Date	Course Number	Start Date	Credit Allocated Per Student	# Invited Students	# Students Joined	Status
Cloud Computing & Big Data - Parallel & Scalable Machine Learning & Deep Learning	10/12/2020	REI504M	10/12/2020	\$50	54	0	Go to classroom

AWS Management Console – Central Access to all Services & Functionalities

The screenshot shows the AWS Management Console homepage. At the top, there's a navigation bar with a dropdown menu set to 'aws'. To the right of the menu are user details ('vocstarsoft/user178005:msmith@hi.la @ 6422-6180-5225'), a location indicator ('N. Virginia'), and a 'Support' link. Below the navigation bar, the title 'AWS Management Console' is displayed. On the left, there's a sidebar with sections for 'AWS services', 'Build a solution', 'Learn to build', and 'Machine Learning'. The main content area features several cards for quick access to services like EC2, S3, Lambda, and CloudWatch. On the right side, there are sections for 'Stay connected to your AWS resources on-the-go', 'Explore AWS', 'AWS Certification', 'Amazon EFS for AWS Lambda', 'Move to Managed File Storage', 'Amazon S3 Glacier', and 'Have feedback?'. At the bottom, there's a footer with links for 'AWS Support', 'AWS Marketplace', 'AWS Documentation', and 'AWS Community'.

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



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

Products / Compute / Amazon EC2 / Amazon EC2 Pricing / ...

Amazon EC2 On-Demand Pricing



PAGE CONTENT

On-Demand Pricing

Data Transfer

EBS-Optimized Instances

Elastic IP Addresses

Carrier IP Addresses

Elastic Load Balancing

On-Demand Capacity Reservations

T2/T3/T4g Unlimited Mode Pricing

On-Demand Pricing

On-Demand Instances let you pay for compute capacity by the hour or second (minimum of 60 seconds) with no long-term commitments. This frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs.

The pricing below includes the cost to run private and public AMIs on the specified operating system ("Windows Usage" prices apply to Windows Server 2003 R2, 2008, 2008 R2, 2012, 2012 R2, 2016, and 2019). Amazon also provides you with additional instances for Amazon EC2 running Microsoft Windows with SQL Server, Amazon EC2 running SUSE Linux Enterprise Server, Amazon EC2 running Red Hat Enterprise Linux and Amazon EC2 running IBM that are priced differently.

Linux	RHEL	SLES	Windows	Windows with SQL Standard	Windows with SQL Web
Windows with SQL Enterprise	Linux with SQL Standard	Linux with SQL Web	Linux with SQL Enterprise		

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

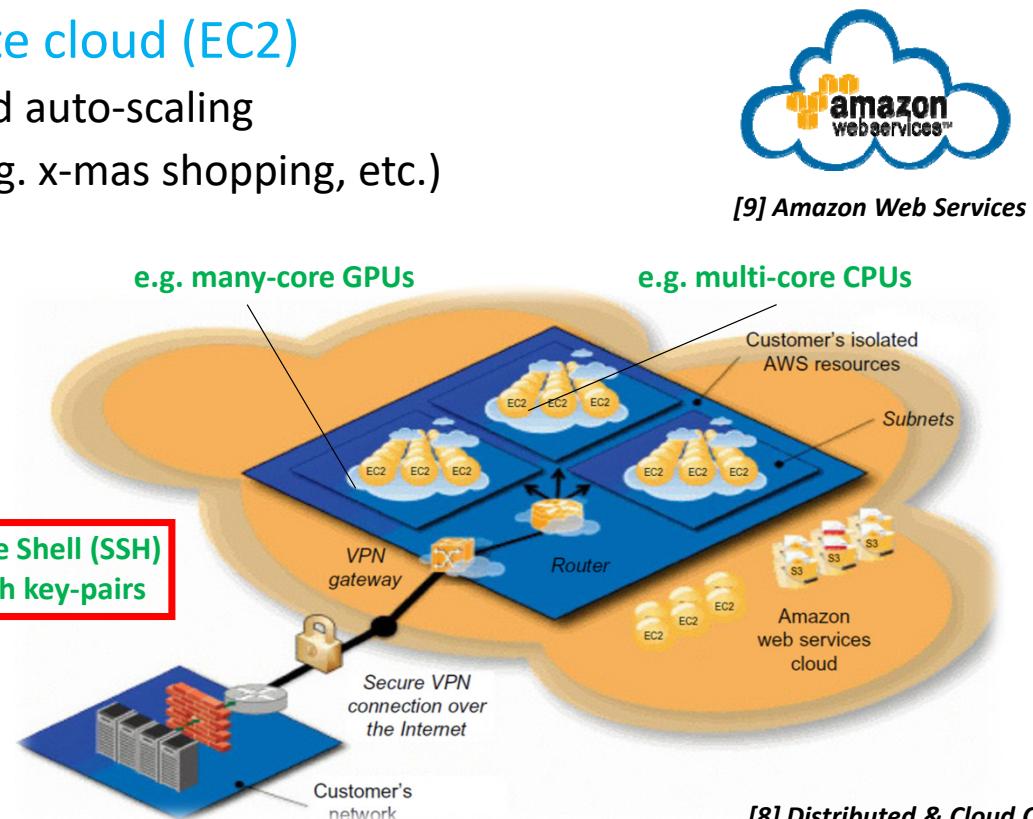
Amazon Web Services – Modern Infrastructure-as-a-Service (IAAS) Example

- 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

[8] Distributed & Cloud Computing Book

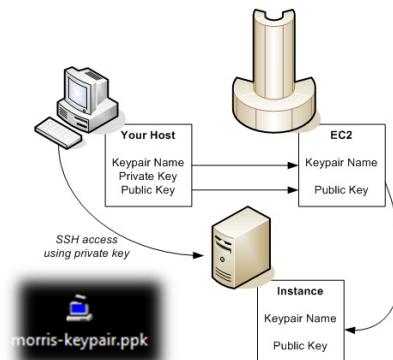
➤ Lecture 8 provides more details about Amazon Web Services and its Infrastructure-as-a-Service (IAAS) models & EC2 Service Elements

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

■ Secure Shell (SSH)

- Universal technique to securely access remote clusters & HPC machines

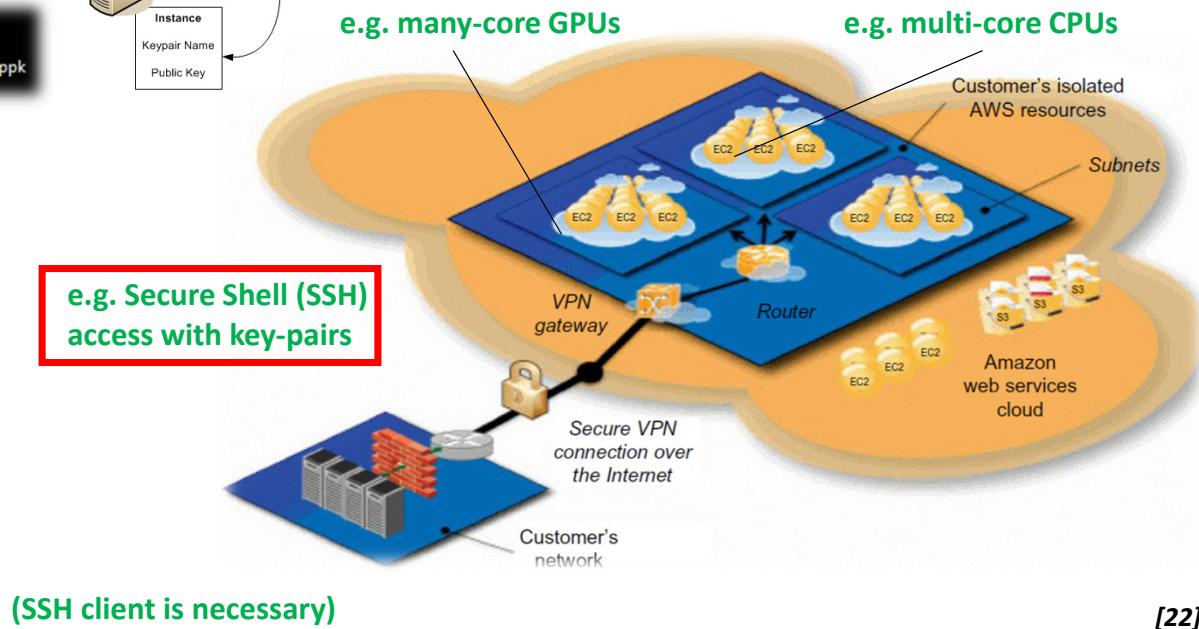
- The Secure Shell (SSH) is a technique to securely access remote AWS computing instances (e.g., AWS EC2) using a named key pair
- An SSH key pair consists of a public key that is known by the Amazon Cloud and a private key that remains only on the laptop of cloud users



- 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

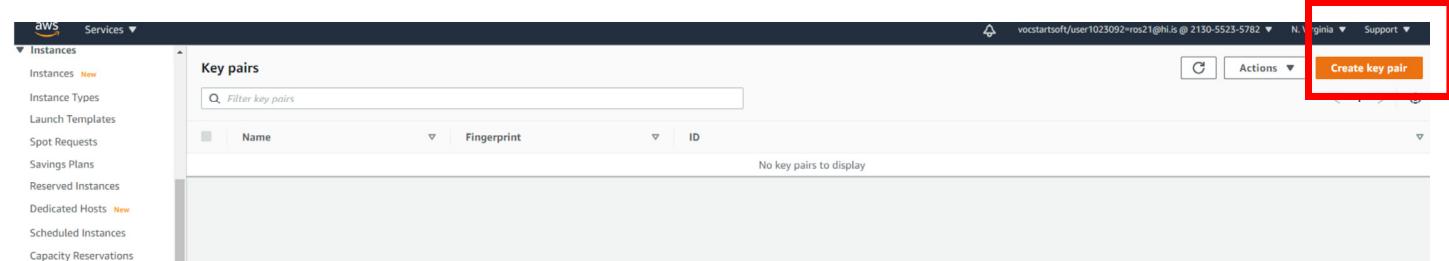
AWS Service Example – Elastic Compute Cloud (EC2) Virtual Servers & Key Pairs

The screenshot shows the AWS EC2 console interface. On the left, the navigation menu is open, with the 'Key Pairs' section highlighted by a red box. The main content area displays various resources and service status. The 'Resources' section shows counts for Running instances, Elastic IPs, Dedicated Hosts, Snapshots, Volumes, Load balancers, Key pairs (which is 1), Security groups, and Placement groups. Below this is a note about deploying Microsoft SQL Server Always On availability groups. The 'Service health' section indicates that the service is operating normally. The 'Zone status' section lists six zones (us-east-1a through us-east-1f) all operating normally. There are also sections for 'Launch instance', 'Scheduled events', 'Migrate a machine', and 'Explore AWS' with various promotional links.

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

EC2 > Key pairs > Create key pair

Create key pair

Key pair
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name A key pair name can contain up to 255 ASCII characters. It can't include leading or trailing spaces.

File format **ppk** For use with PuTTY

pem For use with OpenSSH

Tags (Optional)
No tags associated with the resource.

Add tag You can add 50 more tags

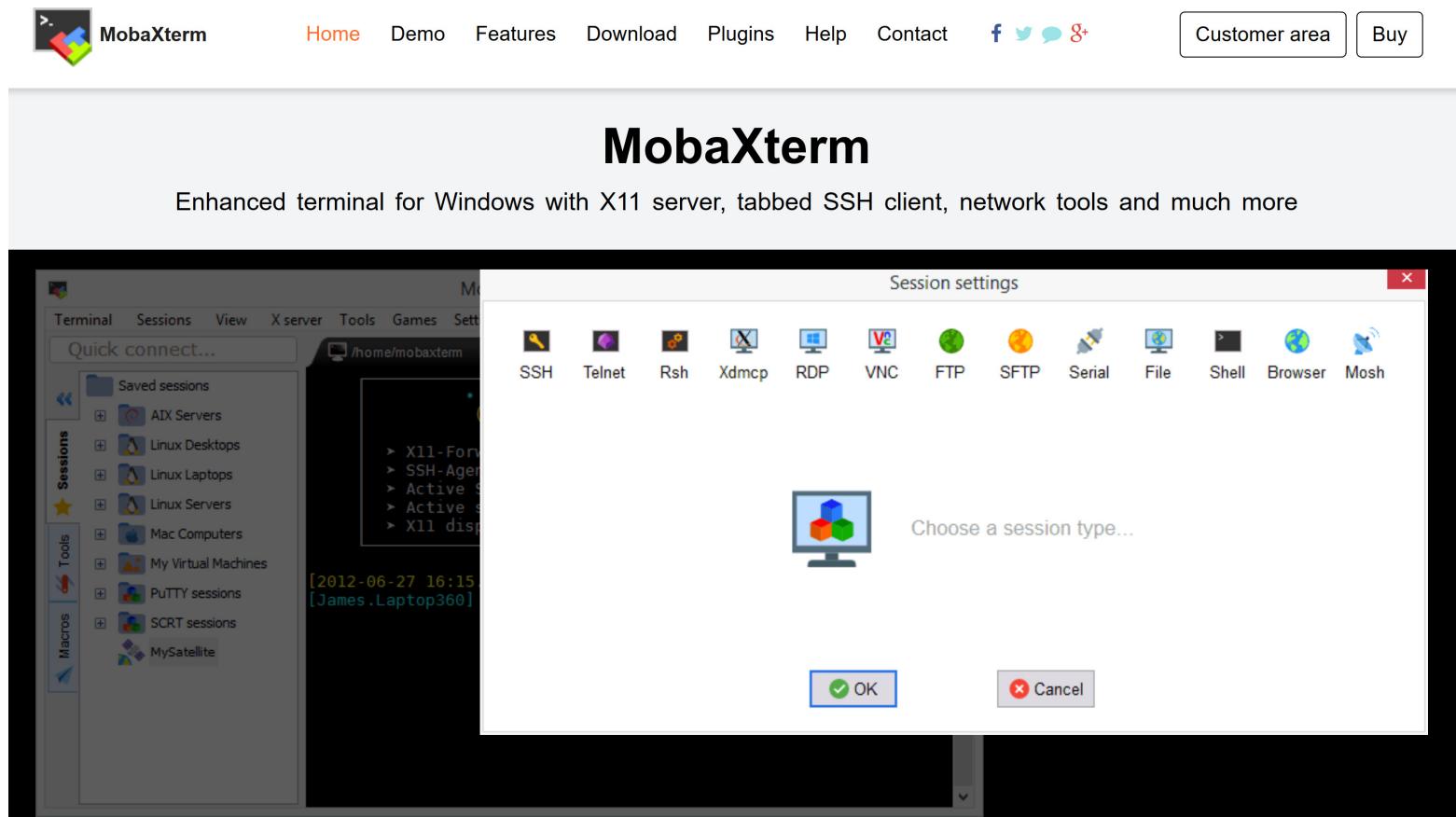
Cancel **Create key pair**

SuccessFully created key pair

Key pairs (1)

<input type="checkbox"/>	Name	Fingerprint	ID
<input type="checkbox"/>	rocco-keypair	e2:39:23:27:9f:e9:82:75:5e:34:6a:1d:2...	key-0364bf684f98e3972

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](#)



A screenshot of the AWS Marketplace website. At the top, there's a search bar with placeholder text "Find AWS Marketplace products that meet your needs." Below it are four dropdown menus: "Categories" (All categories), "Vendors" (All vendors), "Pricing Plans" (All pricing plans), and "Delivery Methods" (All delivery methods). A button "View all products" is located to the right of the search bar. Below these filters, a message says "Total results: 9425". To the right, there are sections for "Popular Categories" with icons for Operating Systems, Security, Networking, Storage, Data Analytics, Dev Ops, Machine Learning, and Data Products. A link "View all categories" is at the bottom of this section.

[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

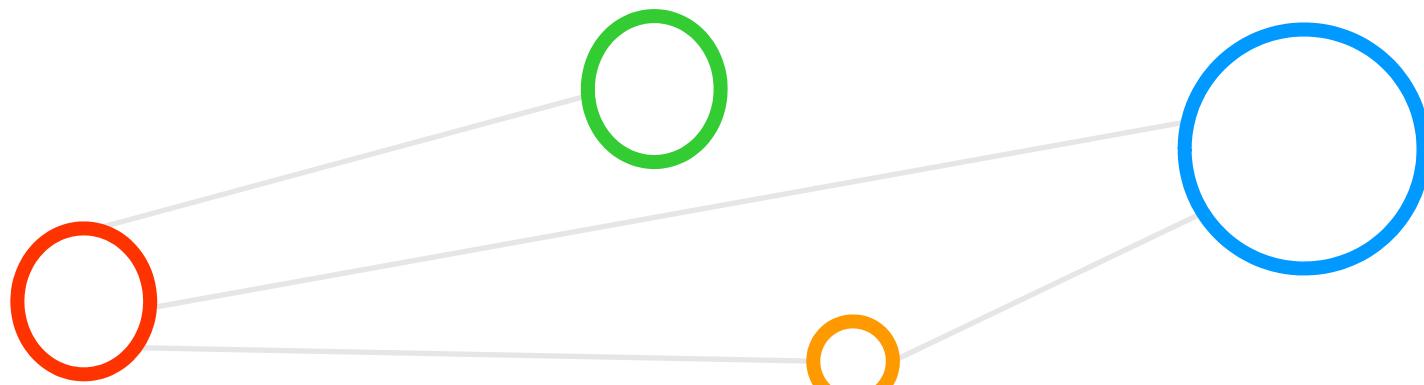
➤ Lecture 8 provides more details about Amazon Web Services with different applications using Infrastructure-As-A-Service (IAAS) models

[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));
```

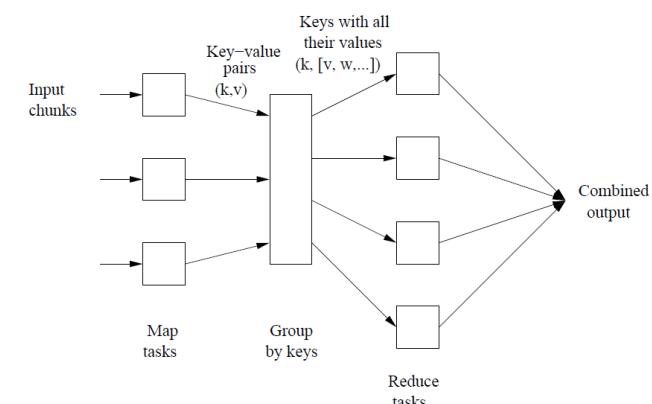
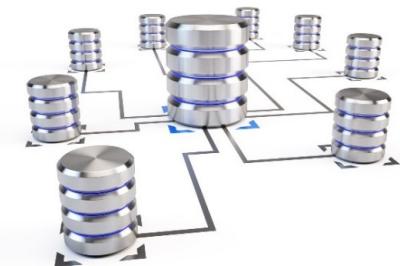
- Goal: Counting the number of each word appearing in a document (or text-stream more general)
- Wordcount in documents or Twitter textstreams are excellent examples for independent ‘nicely parallel’

- 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



[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

vocareum

Welcome to your AWS Educate Account

AWS Educate provides you with access to a wide variety of AWS Services for you to get your hands on and build on AWS! To get started, click on the AWS Console button to log in to your AWS console.

Please read the FAQ below to help you get started on your Starter Account.

- What are the list of services supported?
- What regions are supported with Starter Accounts or Classroom Accounts?
- I can't start any resources. What happened?
- Can I create users within my Starter or Classroom Account for others to access?
- Can I create my own IAM policy within Starter Account or Classroom?
- Can I use marketplace software with my Starter Account or Classrooms?
- Are there any restrictions on AWS services in my AWS Educate Account?
- Are FPGA Instances Supported?
- How do I share image with my students?
- Can I access the billing and cost console?

Your AWS Account Status

Active full access (ros21@hi.is)
\$50 remaining credits (estimated)
2:60 session time

Account Details **AWS Console**

Please use AWS Educate Account responsibly. Remember to shut down your instances when not in use to make the best use of your credits. And, don't forget to logout once you are done with your work!



▪ **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

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:

Create cluster

How Elastic MapReduce Works

- Upload
- Create
- Monitor

Upload your data and processing application to S3.

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

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

[Learn more](#)

[Learn more](#)

[Learn more](#)

Create Cluster - Quick Option

Go to advanced options

General Configuration

Cluster name: My cluster

Logging: S3 folder: `s3://aws-logs-213055235782-us-east-1/elasticmapr/`

Launch mode: Cluster Step execution

Software configuration

Release: emr-5.31.0

Applications:
 Core Hadoop: Hadoop 2.10.0, Hive 2.3.7, Hue 4.7.1, MapReduce 0.13.0, Pig 0.13.0, and Tez 0.9.0
 HBase: HBase 1.4.15, HDFS 2.10.5, Hive 2.3.7, Hue 4.7.1, Phoenix 4.14.3, and Zookeeper 3.4.14
 Presto: Presto 0.238.3 with Hadoop 2.10.0 HDFS and Hive 2.3.7 Metastore
 Spark: Spark 2.4.6 on Hadoop 2.10.0 YARN and Zeppelin 0.8.2

Use AWS Glue Data Catalog for table metadata

Hardware configuration

Instance type: m5.xlarge

Number of instances: 3 (1 master and 2 core nodes)

Cluster scaling: scale cluster nodes based on workload

Security and access

EC2 key pair: No key pairs found

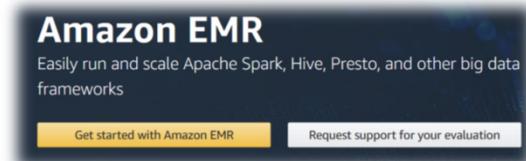
Permissions: Default Custom

EMR role: `EMR_DefaultRole`

EC2 instance profile: `EMR_EC2_DefaultRole`



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

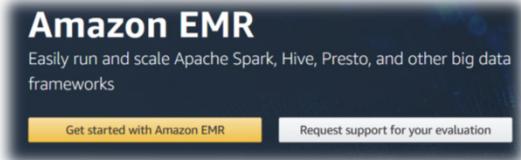


[8] AWS EMR Web page

This screenshot shows the 'Create Cluster - Advanced Options' page in the AWS EMR console. The page is divided into several sections:

- Step 1: Software and Steps**: Shows the selected release 'emr-5.31.0'. Under 'Software Configuration', several boxes are checked and highlighted with red boxes:
 - Hadoop 2.10.0
 - JupyterHub 1.1.0
 - Hive 2.3.7
 - Hue 4.7.1
 - Spark 2.4.6
- Multiple master nodes (optional)**: Contains a checkbox for using multiple master nodes to improve cluster availability.
- AWS Glue Data Catalog settings (optional)**: Contains a checked checkbox for 'Use for Hive table metadata'.
- Edit software settings**: Offers options to 'Enter configuration' (selected) or 'Load JSON from S3', with a text input field containing the configuration JSON.
- Concurrency**: A checkbox for running multiple steps at the same time.
- After last step completes**: Radio buttons for 'Clusters enters waiting state' (selected) and 'Cluster auto-terminates'.
- Steps (optional)**: A section for defining work units for the cluster.
- Step type**: A dropdown menu set to 'Select a step'.
- Buttons**: 'Cancel' and 'Next' buttons at the bottom right.

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 amount 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)

Create Cluster - Advanced Options

Hardware Configuration

Specify the networking and hardware configuration for your cluster. Request Spot instances (unused EC2 capacity) to save money.

Cluster Composition

Specify the configuration of the master, core and task nodes as an Instances group or instance fleet. This choice applies to all nodes for the lifetime of the cluster. Instance fleets and instance groups cannot coexist in a cluster. [See this topic](#)

Instance group configuration

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

Instance fleets
Specify the instance type and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#)

Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Launch the cluster into a VPC with a public, private, or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Network: vpc-19f33864 (172.31.0.0/16) (default) [Create a VPC](#)

EC2 Subnet: subnet-060fb28 (Default in us-east-1b)

Cluster Nodes and Instances

Choose the instance type, number of instances, and a purchasing option. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option
Master	m5.xlarge	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
Master - 1	4 vCore, 16 GB memory, EBS only storage EBS Storage: 32 GB	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
	Add configuration settings		
Core	m5.xlarge	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
Core - 2	4 vCore, 16 GB memory, EBS only storage EBS Storage: 32 GB	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
	Add configuration settings		
Task	m5.xlarge	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
Task - 3	4 vCore, 16 GB memory, EBS only storage EBS Storage: 32 GB	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price
	Add configuration settings		

+ Add task instance group

Total core and task units: 2 Total units

Cluster scaling

Adjust the number of Amazon EC2 instances available to an EMR cluster via EMR-managed scaling or a custom automatic scaling policy. [Learn more](#)

Cluster scaling Enable Cluster Scaling

Amazon EC2 On-Demand Pricing

On-Demand Pricing

On-Demand Instances let you pay for compute capacity by the hour or second (minimum of 60 seconds) with no long-term commitments. This frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs.

The pricing below includes the cost to run private and public AMIs on the specified operating system ("Windows Usage" prices apply to Windows Server 2003 R2, 2008, 2008 R2, 2012, 2012 R2, 2016, and 2019). Amazon also provides you with additional instances for Amazon EC2 running Microsoft Windows with SQL Server, Amazon EC2 running SUSE Linux Enterprise Server, Amazon EC2 running Red Hat Enterprise Linux and Amazon EC2 running RHEL that are priced differently.

Instance Type	Cores	Threads	Memory (GiB)	Storage (GiB)	Price
m5.large	2	10	8 GiB	EBS Only	\$0.096 per Hour
m5.xlarge	4	16	16 GiB	EBS Only	\$0.192 per Hour
m5.2xlarge	8	37	32 GiB	EBS Only	\$0.384 per Hour
m5.4xlarge	16	70	64 GiB	EBS Only	\$0.768 per Hour
m5.8xlarge	32	128	128 GiB	EBS Only	\$1.536 per Hour
m5.12xlarge	48	168	192 GiB	EBS Only	\$2.304 per Hour
m5.16xlarge	64	256	256 GiB	EBS Only	\$3.072 per Hour
m5.24xlarge	96	337	384 GiB	EBS Only	\$4.608 per Hour
m5.metal	96	345	384 GiB	EBS Only	\$4.608 per Hour
m5a.large	2	N/A	8 GiB	EBS Only	\$0.086 per Hour
m5a.xlarge	4	N/A	16 GiB	EBS Only	\$0.172 per Hour
m5a.2xlarge	8	N/A	32 GiB	EBS Only	\$0.344 per Hour
m5a.4xlarge	16	N/A	64 GiB	EBS Only	\$0.688 per Hour
m5a.8xlarge	32	N/A	128 GiB	EBS Only	\$1.376 per Hour
m5a.12xlarge	48	N/A	192 GiB	EBS Only	\$2.064 per Hour
m5a.16xlarge	64	N/A	256 GiB	EBS Only	\$2.752 per Hour

[13] AWS EC2 Pricing

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

Screenshot of the AWS EMR "Create Cluster - Advanced Options" step. The "General Options" section is highlighted with a red box.

General Options

- Cluster name: cc-bd-2020-spark-cluster
- Logging: S3 folder `s3://aws-logs-213055235782-us-east-1/elasticmapre` (with a file icon)
- Log encryption
- 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

Cancel Previous Next

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

aws Services ▾

Create Cluster - Advanced Options [Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Successfully created key pair

Key pairs (1)

Name	Fingerprint	ID
rocco-keypair	e2:39:23:27:9fe9:82:75:5e:34:6a:1d:2...	key-0364bf684f98e3972

Filter key pairs

Security Options

EC2 key pair **rocco-keypair**

Cluster visible to all IAM users in account [i](#)

Permissions [i](#)

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](#) [i](#)

EC2 instance profile [EMR_EC2_DefaultRole](#) [i](#)

Auto Scaling role [EMR_AutoScaling_DefaultRole](#) [i](#)

▶ Security Configuration

▶ EC2 security groups

[Cancel](#) [Previous](#) **Create cluster**

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

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

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)

The screenshot shows the AWS Amazon EMR Notebooks interface. On the left sidebar, the 'Notebooks' option is selected and highlighted with a red box. A large red arrow points from this sidebar to the 'Create notebook' button at the top of the main content area. The main content area displays a 'Create notebook' form. The 'Cluster*' field is also highlighted with a red box and has another red arrow pointing to it from the right side of the screen, which is pointing to a separate 'Choose a cluster' modal window. The modal window lists a single cluster: 'cc-bd-2020-spark-cluster j-3FQEIU7L12FTM'. The status of this cluster is 'Starting'. At the bottom of the main form, there is a red box around the 'Create notebook' button.

Amazon EMR

Clusters

Notebooks **Selected**

Git repositories

Security configurations

Block public access

VPC subnets

Events

Help

What's new

jupyter

[12] Jupyter

Notebooks

Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create and attach notebooks to Amazon EMR clusters running Hadoop, Spark, and Livy. Notebooks run free of charge and are saved in Amazon S3 independently of clusters. Standard billing for clusters and Amazon S3 apply. [Learn more](#)

Create notebook View details Open in JupyterLab Open in Jupyter Start Stop Delete

Filter: All notebooks Filter notebooks ... 0 notebooks (all loaded) C

Name	Status	Cluster	Creation time (UTC+2)	Last modified
------	--------	---------	-----------------------	---------------

Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. [Learn more](#)

Notebook name* cc-bd-2020-emr-notebook
Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description

256 characters max.

Cluster* Choose an existing cluster [Choose](#) cc-bd-2020-spark-cluster j-3FQEIU7L12FTM C

Security groups Use default security groups [Edit](#)
 Choose security groups (vpc-19f33064)

AWS service role* [Create default role](#)

Notebook location* Choose an S3 location where files for this notebook are saved.
 Use a location that EMR creates [Edit](#) s3://aws-emr-resources-213055235782-us-east-1/notebooks/
 Choose an existing S3 location in us-east-1

Git repository [Link to a Git repository](#)

Tags [Edit](#)

* Required

Cancel **Create notebook**

Choose a cluster

The listed clusters meet notebook requirements. They are in an EC2-VPC, running EMR 5.18.0 or later, and have Hadoop, Spark, and Livy installed. [Learn more](#)

The notebook can be opened once the cluster is in Waiting or Running status.

Filter: Filter clusters ... 1 cluster (all loaded) C

Name	ID	Status
cc-bd-2020-spark-cluster	j-3FQEIU7L12FTM	Starting

Cancel **Choose cluster**

Optional Step 3: Startup Jupyter Notebook (2)

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-5X8GLS60CGHM3IYTGIRWN6WKV

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 = AROATDGYU43DJ6XBANK2: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/

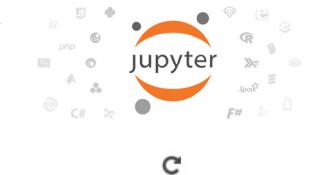
Git repositories

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
-----------------	-----	--------	-------------	----------------

[12] **Jupyter**



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

jupyter

Files Running Clusters

Select items to perform actions on them.

0 /

cc-bd-2020-emr-notebook.ipynb

Name Last Modified File size

un'ora fa 72 B

Quit

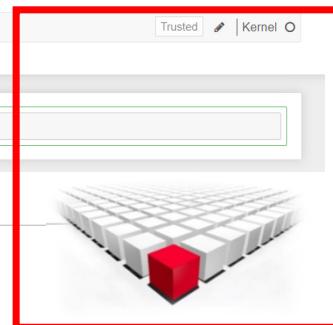
jupyter cc-bd-2020-emr-notebook Last Checkpoint: un'ora fa (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

nbdiff

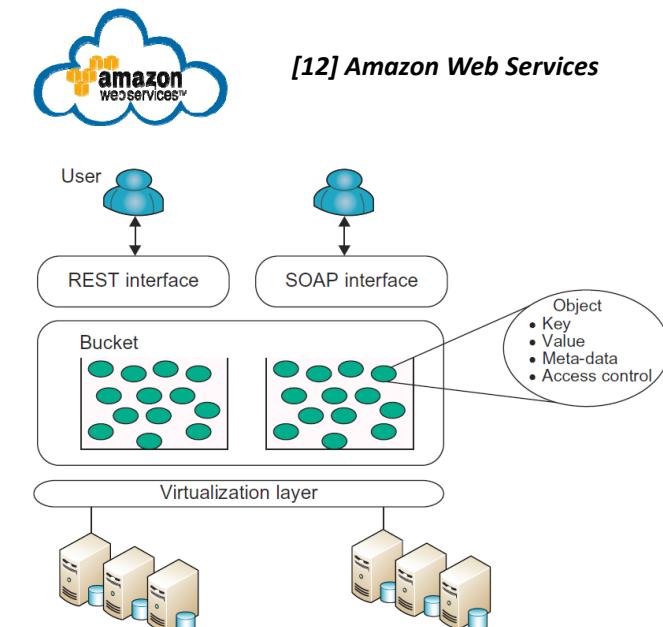
In []:

Trusted Kernel



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



[8] Distributed & Cloud Computing Book

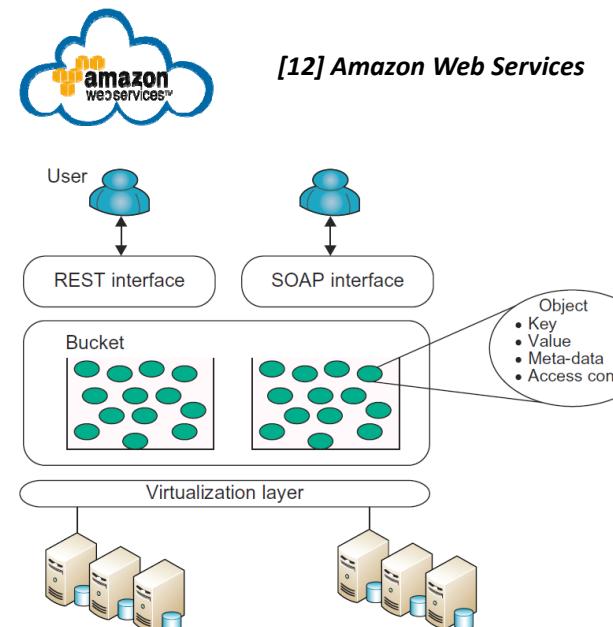
Step 4: Create AWS S3 service ‘bucket’ & Check Application Dataset

User Bucket

- Create own S3 bucket for WordCount results

The screenshot shows the AWS S3 console interface. At the top, there's a message about the temporary re-enablement of the previous version of the S3 console. Below it, the 'S3 buckets' section shows one existing bucket ('aws-logs-133213380826-us-east-1') and a 'Create bucket' button highlighted with a red box. The 'Create bucket' dialog box is open, showing the 'Name and region' step. It has a 'Bucket name' field containing 'cc-bd-2020-s3-wordcount-results' and a 'Region' dropdown set to 'US East (N. Virginia)', both also highlighted with a red box. The 'Configure options' and 'Set permissions' steps are visible at the top of the dialog.

This screenshot shows the 'All services' page in the AWS Management Console. The 'S3' service is selected, which is described as 'Scalable Storage in the Cloud'. Other services listed include S3 Glacier, AWS Snow Family, AWS Transfer Family, Athena, and Amazon Transcribe. The 'Discover the console' link is visible at the top right of the S3 section.



[8] Distributed & Cloud Computing Book

This screenshot shows the 'S3 buckets' list after the new bucket was created. It now displays two buckets: 'aws-logs-133213380826-us-east-1' and 'cc-bd-2020-s3-wordcount-results'. The new bucket is highlighted with a red box. The 'Discover the console' link is visible at the top right of the list.

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>

← → ⌂ ⌂ 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.

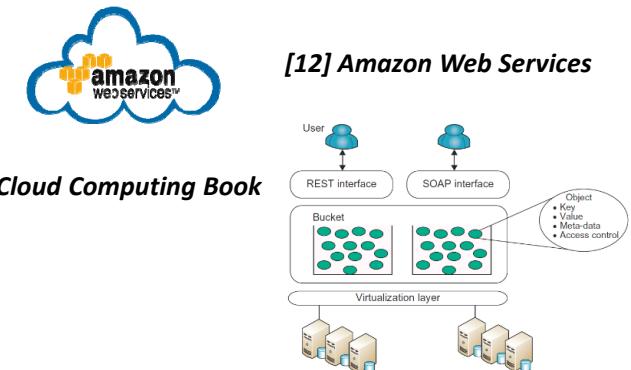
```
<ListBucketResult xmlns="http://s3.amazonaws.com/doc/2006-03-01/">
  <Name>elasticmapreduce</Name>
  <Prefix>samples/wordcount/input/<Prefix>
  <Marker/>
  <MaxKeys>1000</MaxKeys>
  <IsTruncated>false</IsTruncated>
  <Contents>
    <Key>samples/wordcount/input/0001</Key>
    <LastModified>2009-04-02T02:55:30.000Z</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>
</ListBucketResult>
```

(0001 document)

CIA -- The World Factbook -- Country Listing
World Factbook Home
The World Factbook
About
Country Listing
nbsp;
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
nbsp;
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
Bahamas, The
Bahrain
Baker Island
description under United States Pacific Island Wildlife Refuge

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 internati
This page was last updated on 1 January 2003
This page was last updated on 19 March, 2009
nbsp;
CIA - The World Factbook -- Fiji
a (font-family: Verdana, Arial, Helvetica, sans-serif; font-size: 12px; color: #000000; text-decoration
nbsp;
Country
List | World Factbook Home
The World Factbook
nbsp;
Fiji
Introduction
Fiji
Background:
Fiji became independent in 1970, after nearly a century as a British colony. Democ
British in the 19th century). The coup 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.

(0002 document)



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)

The screenshot shows the 'Create Cluster - Advanced Options' page. Under 'Software Configuration', the 'Release' dropdown is set to 'emr-5.3.0'. The 'Hadoop' checkbox is checked. Under 'Steps (optional)', there is a note about submitting multiple steps and a 'Next' button at the bottom.

- 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

The 'Add step' dialog is shown for a 'Streaming program'. It has fields for 'Name' (set to 'Streaming program'), 'Mapper' (set to 'elasticmapreduce/samples/wordcount/wordSplitter.py'), 'Reducer' (set to 'aggregate'), 'Input S3 location' (set to 's3://elasticmapreduce/samples/wordcount/input s3://<bucket-name>/<folder>/'), and 'Output S3 location' (set to 's3://cc-bd-2020-s3-wordcount-results/2020-10-15 s3://<bucket-name>/<folder>/'). The 'Arguments' field is empty. The 'Action on failure' dropdown is set to 'Continue'.

Name	Action on failure	JAR location	Arguments
Streaming program	Continue	command-runner jar	hadoop-streaming -files s3://elasticmapreduce/samples/wordcount/wordSplitter.py -mapper wordSplitter.py -reducer aggregate -input s3://elasticmapreduce/samples/wordcount/input -output s3://cc-bd-2020-s3-wordcount-results/

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

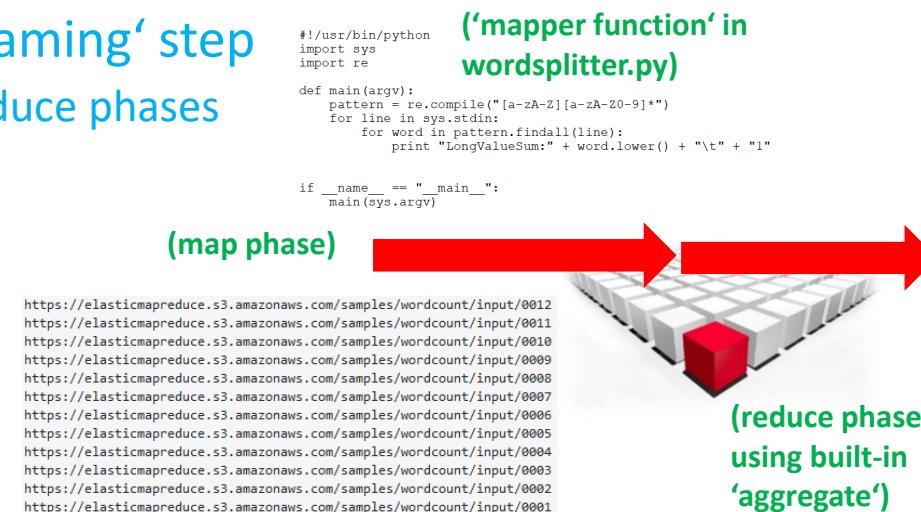
- Application is executed as ‘streaming’ step
 - Using the EMR cluster & map-reduce phases

Cluster: cc-bd-2020-morris-wordcount-run Starting Configuring cluster software

Summary	Application user interfaces	Monitoring	Hardware	Configurations	Events	Steps	Bootstrap actions
Summary ID: j230YSQL77IU0F Creation date: 2020-10-15 07:34 (UTC+0) Elapsed time: 2 minutes After last step completes: Cluster waits Termination protection: On Change Tags: -- View All / Edit Master public DNS: ec2-107-23-129-177.compute-1.amazonaws.com Connect to the Master Node Using SSH							
Configuration details Release label: emr-3.1.0 Hadoop distribution: Amazon 2.10.0 Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.7.1 Log URI: s3://aws-logs-133213380826-us-east-1/elasticmapreduce/ EMRFS consistent view: Enabled Custom AMI ID: --							
Network and hardware Availability zone: us-east-1c Subnet ID: subnet-3ccb3371 View Master: Bootstrapping 1 m5.xlarge Core: Provisioning 2 m5.xlarge Task: --							
Security and access Key name: morris-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: sg-03etc5db150b2504 (ElasticMapReduce-master) Security groups for Core & Task: sg-01160d1e5a5ef8d61 (ElasticMapReduce-slave)							

Cluster: cc-bd-2020-morris-wordcount-run Running Running step

Summary	Application user interfaces	Monitoring	Hardware	Configurations	Events	Steps	Bootstrap actions
Summary ID: j230YSQL77IU0F Creation date: 2020-10-15 07:34 (UTC+0) Elapsed time: 7 minutes After last step completes: Cluster waits Termination protection: On Change Tags: -- View All / Edit Master public DNS: ec2-107-23-129-177.compute-1.amazonaws.com Connect to the Master Node Using SSH							
Configuration details Release label: emr-3.1.0 Hadoop distribution: Amazon 2.10.0 Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.7.1 Log URI: s3://aws-logs-133213380826-us-east-1/elasticmapreduce/ EMRFS consistent view: Disabled Custom AMI ID: --							
Network and hardware Availability zone: us-east-1c Subnet ID: subnet-3ccb3371 View Master: Bootstrapping 1 m5.xlarge Core: Provisioning 2 m5.xlarge Task: --							
Security and access Key name: morris-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: sg-03etc5db150b2504 (ElasticMapReduce-master) Security groups for Core & Task: sg-01160d1e5a5ef8d61 (ElasticMapReduce-slave)							



Amazon S3 > cc-bd-2020-s3-wordcount-results > 2020-10-15

cc-bd-2020-s3-wordcount-results

Overview

Type a prefix and press Enter to search. Press ESC to clear.

Upload Create folder Download Actions

US East (N. Virginia) Viewing 1 to 4

Name	Last modified	Size	Storage class
_SUCCESS	Oct 15, 2020 7:42:40 AM GMT+0000	0 B	Standard
part-00000	Oct 15, 2020 7:42:35 AM GMT+0000	97.3 KB	Standard
part-00001	Oct 15, 2020 7:42:40 AM GMT+0000	98.6 KB	Standard
part-00002	Oct 15, 2020 7:42:41 AM GMT+0000	97.2 KB	Standard

Viewing 1 to 4

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

Filter: All steps Filter steps ... 2 steps (all loaded) C

ID	Name	Status	Start time (UTC+0)	Elapsed time	Log files
s-1VMQJ5AxE/BFO	Streaming program	Failed	2020-10-15 07:16 (UTC+0)	6 seconds	controller syslog stderr stdout
Status : FAILED Reason : Output directory already exists. Log File : s3://aws-logs-133213380826-us-east-1/elasticmapreduce/j-2DJZBFHP2BN1Z/steps/s-1VMQJ5AxE/BFO/syslog.gz View Details : 2020-10-15 07:16:12,901 ERROR org.apache.hadoop.streaming.StreamJob (main): Error Launching job : Output directory s3://cc-bd-2020-s3-wordcount-results/ already exists JAR location : command-runner.jar Main class : None Arguments : hadoop-streaming -files s3://elasticmapreduce/samples/wordcount/wordSplitter.py -mapper wordSplitter.py -reducer aggregate -input s3://elasticmapreduce/samples/wordcount/input -output s3://cc-bd-2020-s3-wordcount-results/ Action on failure: Continue					
s-291H5LMH5KR5K	Setup hadoop debugging	Completed	2020-10-15 07:15 (UTC+0)	4 seconds	View logs

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

Cluster: cc-bd-2020-morris-wordcount-run Waiting Cluster ready after last step completed.

[Summary](#) [Application user interfaces](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

Summary

ID: j-23OYSQL7TIU0F
Creation date: 2020-10-15 07:34 (UTC+0)
Elapsed time: 9 minutes
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: ec2-107-23-129-177.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

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
Log URI: s3://aws-logs-133213380826-us-east-1/elastictmapreduce/

EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-east-1c
Subnet ID: [subnet-3ccb3371](#)
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --
Cluster scaling: Not enabled

Security and access

Key name: morris-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: [sg-03e1c5d8150b25f04](#) (ElasticMapReduce-master)
Security groups for Core & [sg-01169d1e5a5ef8df61](#) (ElasticMapReduce-slave)
Task:

- 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

```

hydroelectric 18
hydrofluorocarbons 1
hydropower 308
hypotenuse 4
hyun 6
ialomita 3
ian 12
iancuba 3
iasi 3
ib 7
iban 3
ibanda 3
ibaraki 3
ibibio 3
ibolya 3
ibrahim 57
ibuka 3
icc 277
icct 449
icebergs 18
icebreaker 3
icecap 3
icefields 3
icel 3
iceland 964
icelandic 29
icepack 12
icftu 6
icon 7

```

Amazon S3 > cc-bd-2020-s3-wordcount-results > 2020-10-15

cc-bd-2020-s3-wordcount-results

[Overview](#) [Actions](#)

Type a prefix and press Enter to search. Press ESC to clear.

[Upload](#) [Create folder](#) [Download](#) [Actions](#)

US East (N. Virginia) Viewing 1 to 4

Name	Last modified	Size	Storage class
_SUCCESS	Oct 15, 2020 7:42:40 AM GMT+0000	0 B	Standard
part-00000	Oct 15, 2020 7:42:35 AM GMT+0000	97.3 KB	Standard
part-00001	Oct 15, 2020 7:42:40 AM GMT+0000	98.6 KB	Standard
part-00002	Oct 15, 2020 7:42:41 AM GMT+0000	97.2 KB	Standard

Viewing 1 to 4

Cluster: cc-bd-2020-morris-wordcount-run Terminating Terminated by user request

[Summary](#) [Application user interfaces](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

Summary

ID: j-23OYSQL7TIU0F
Creation date: 2020-10-15 07:34 (UTC+0)
Elapsed time: 24 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: --
Master public DNS: ec2-107-23-129-177.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

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
Log URI: s3://aws-logs-133213380826-us-east-1/elastictmapreduce/

EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware

Availability zone: us-east-1c
Subnet ID: [subnet-3ccb3371](#)
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --
Cluster scaling: Not enabled

Security and access

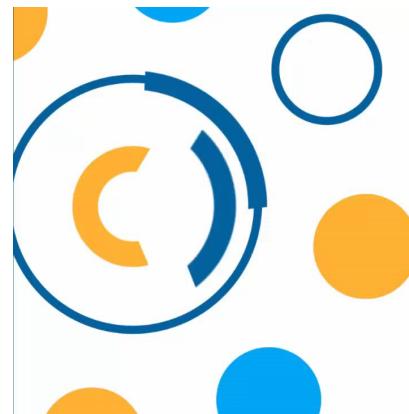
Key name: morris-keypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All [Change](#)

Security groups for Master: [sg-03e1c5d8150b25f04](#) (ElasticMapReduce-master)
Security groups for Core & [sg-01169d1e5a5ef8df61](#) (ElasticMapReduce-slave)
Task:

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

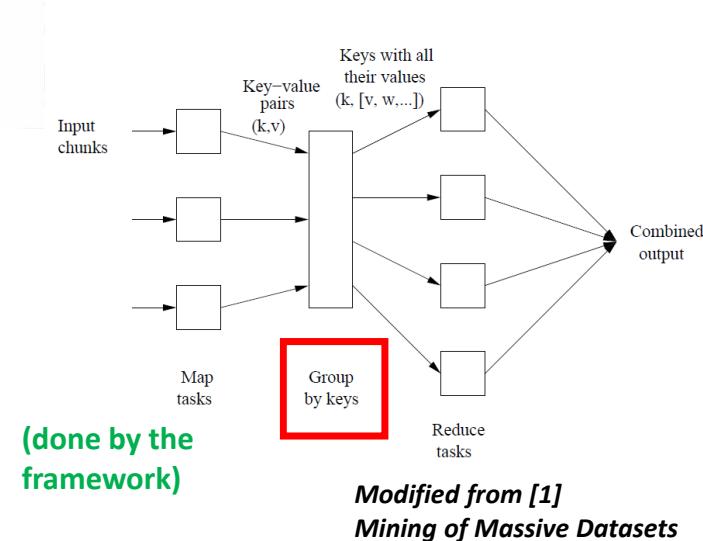
■ Programming Model

- 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 (k_1, v_1) are drawn from a different domain than the output keys and values (k_2, v_2)
- [Intermediate keys](#) and values (k_2, v_2) are from the same domain as the [output keys](#) and values
- Definition of 'generic' keys is dependent on application problem



- $\text{map } (k_1, v_1) \rightarrow \text{list}(k_2, v_2)$
- $\text{reduce } (k_2, \text{list}(v_2)) \rightarrow \text{list}(v_2)$

[18] YouTube video, [Map-Reduce explained](#)

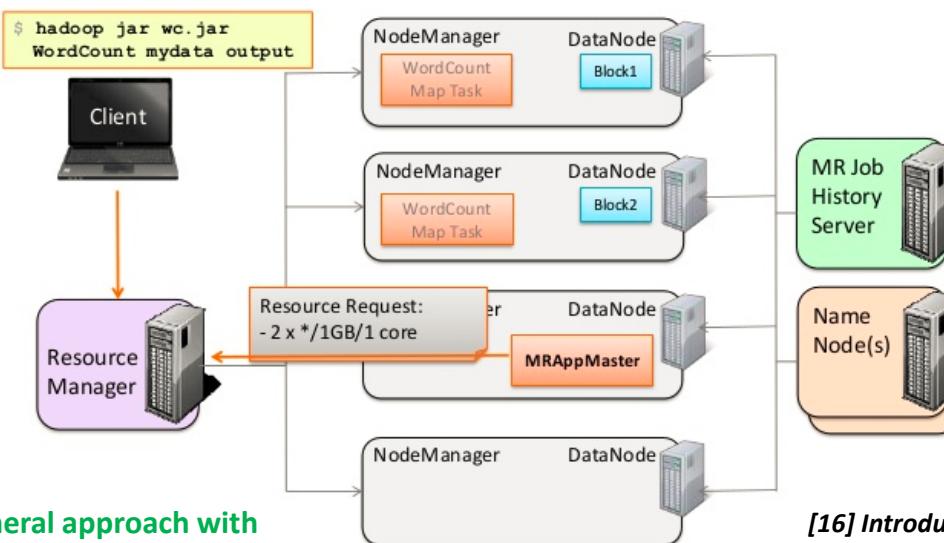


(done by the framework)

Modified from [1]
Mining of Massive Datasets

Hadoop Resource Manager YARN – Wordcount Example as Application Blueprint

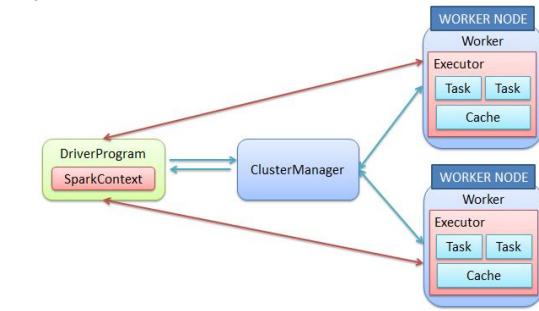
- Apache Hadoop offers the Yet Another Resource Negotiator (YARN) scheduler for map-reduce jobs
- Idea of Yarn is to split up the functionalities of resource management & job scheduling/monitoring
- The ResourceManager is a scheduler that controls resources among all applications in the system
- The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



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

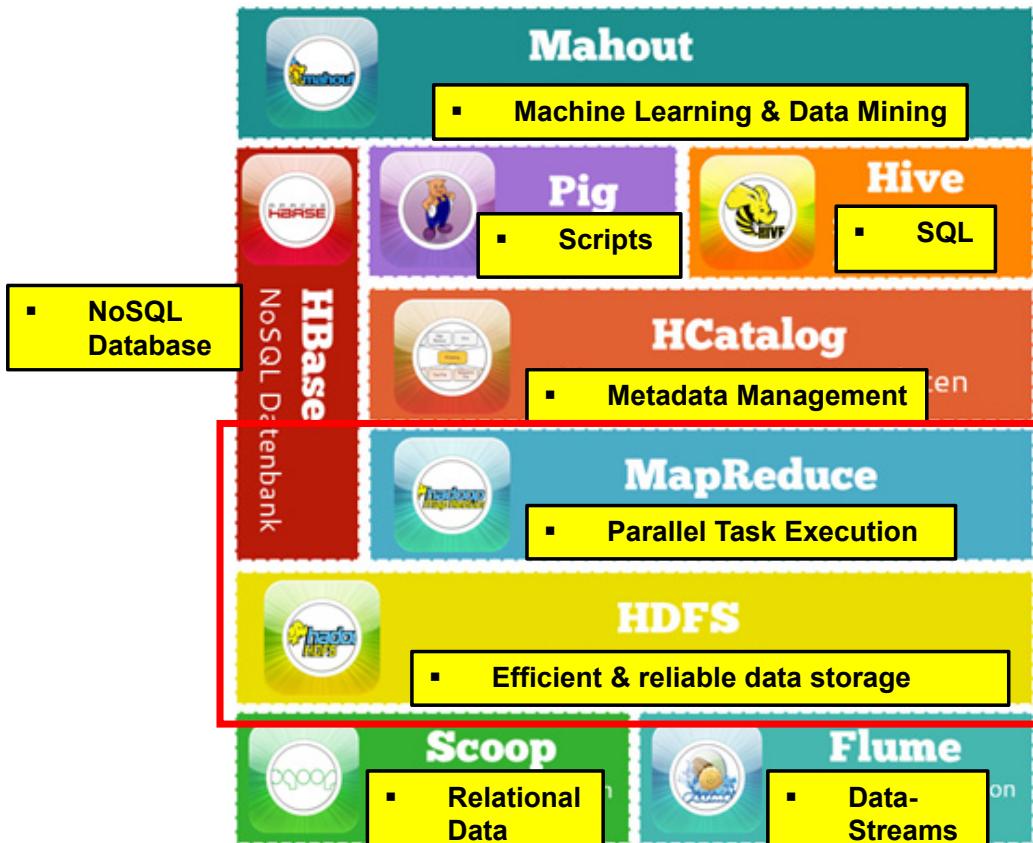
[16] Introduction to Yarn

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



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

Larger Hadoop Ecosystem with Big Data Analytics ‘Tools’ – Revisited



[5] Google Dataproc service



Microsoft Azure



[7] Microsoft Azure HDInsight Service

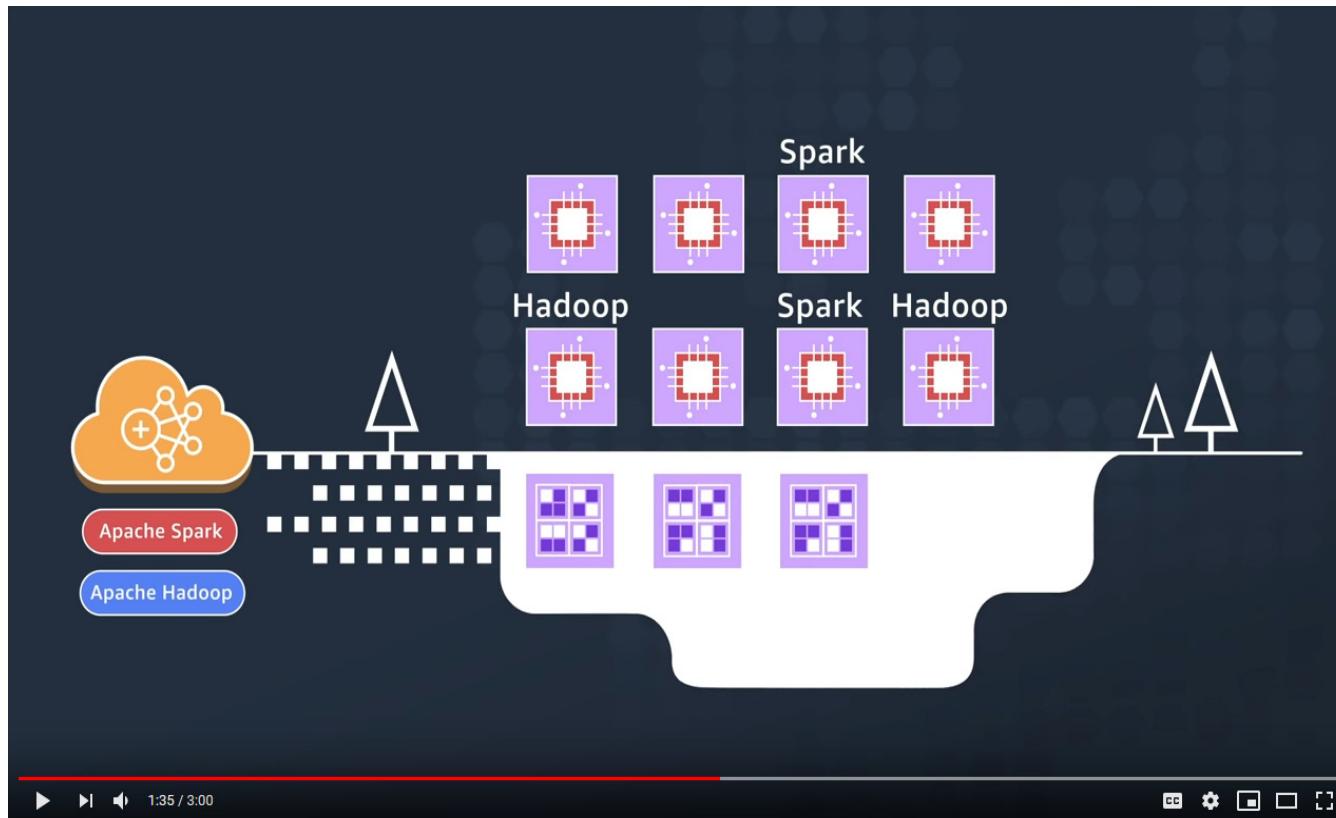
Modified from [6] Map-Reduce

Azure HDInsight Documentation page. The page includes a navigation bar, a sidebar with links to Hadoop, Spark, HBase, Interactive Query, Kafka, Storm, ML Services, and Enterprise readiness, and a main content area with sections for Azure HDInsight Documentation, 5-Minute Quickstarts, and a list of known open source analytics frameworks.

Known Open Source Analytics Frameworks

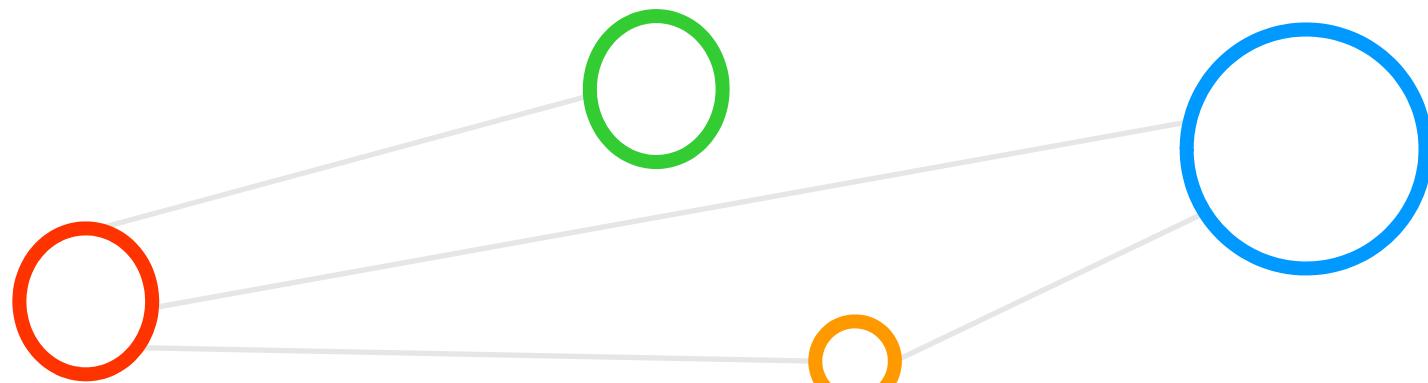
- Hadoop
- Spark
- Kafka
- HBase
- Interactive Query
- Storm
- ML Services

[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/einführung-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/>

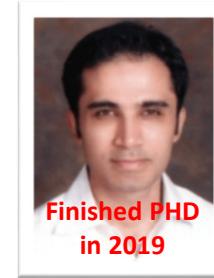
Acknowledgements – High Productivity Data Processing Research Group



Finished PhD
in 2016



Finishing
in Winter
2019



Finished PhD
in 2019



Mid-Term
in Spring
2019



Started
in Spring
2019

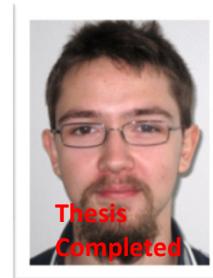


Started
in Spring
2019

Morris Riedel @MorrisRiedel · Feb 10
Enjoying our yearly research group dinner 'Iceland Section' to celebrate our productive collaboration of @uni_iceland @uisens @Haskell_Islands & @fz_jsc @fz_juelich & E.Erlingsson @emrie passed mid-term in modular supercomputing driven by @DEEPprojects - morrisriedel.de/research

A photograph showing a group of approximately ten people seated around tables in a restaurant. They are dressed in casual to semi-formal attire. The restaurant has a warm, ambient lighting with red and white decorations on the walls.

Finished PhD
in 2018



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)

