Assignment 1 - Hadoop MapReduce using AWS

Name: Sourabh Gopal Parvatikar
UFID: 79325142

Github link: https://github.com/sourabhparvatikar/Hadoop-Projects-AWS

Task 1 - Single Word Count:

- Created a Maven project.
- Created WordCount.java in src/main/java/hadoop_wordcount.
- WordCount class contains Mapper and Reducer classes and a main program.
- Map class emits a <key, value> pair for each word in input files.
- Reduce class sums up the values, which are the occurrence counts for each key.
- Main program takes in command line arguments and creates jobs.
- Updated project dependencies in pom.xml
- Output files are written into S3 bucket by EC2 instances. Outputs are also uploaded to Github.
- Output format:
  Each line contains - <each_word_in_input> <number_of_occurrences>
- Note: More info about the logic and code details are mentioned as code comments.
- Output links:
  1. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output/part-r-00000
  2. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output/part-r-00001
  3. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output/part-r-00002
  4. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output/part-r-00003
  5. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output/part-r-00004

Task 2 - Double Word Count:

- Created a Maven project.
- Created WordCountDouble.java in src/main/java/hadoop_wordcount.
- WordCountDouble class contains Mapper and Reducer classes and a main program.
- DoubleWcMap class emits a <key, val> pair for every 2 word sequence in the input file.
- DoubleWcMap class sums up the values, which are the occurrence counts for each key.
- Main program takes in command line arguments and creates jobs.
- Updated project dependencies in pom.xml
- Output files are written into S3 bucket by EC2 instances. Outputs are also uploaded to Github.
- Output format:
  Each line contains - <word1>,<word2> <number_of_occurrences>
- Note: More info about the logic and code details are mentioned as code comments.
- Output links:

1. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_doublewordcounter/part-r-00000
2. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_doublewordcounter/part-r-00001
3. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_doublewordcounter/part-r-00002
4. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_doublewordcounter/part-r-00003
5. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_doublewordcounter/part-r-00004

Task 3 - Distributed Word Count:

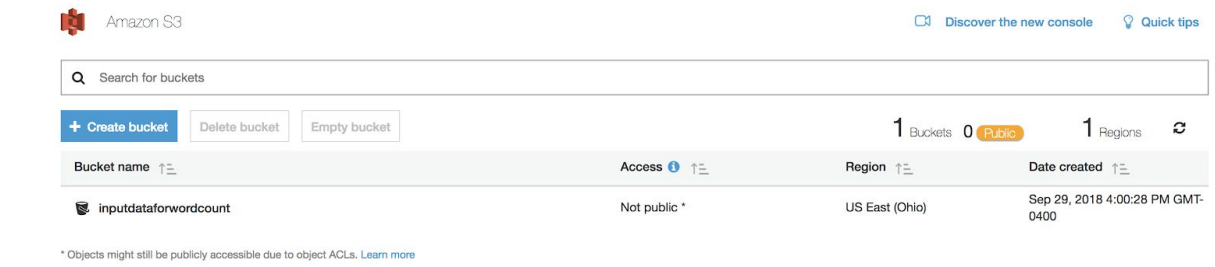
- Created a Maven project.
- Created DistributedWordCount.java in src/main/java/distributed_wordcount
- DistributedWordCount class contains Mapper and Reducer classes and a main program.
- Map class emits a <key, value> pair for each word in input files that is present in another file word-patterns.text.
- word-patterns.txt is accessible locally to each EC2 instance through distributed cache feature.
- Reduce class sums up the values, which are the occurrence counts for each key.
- Main program takes in command line arguments and creates jobs.
- Updated project dependencies in pom.xml
- Output files are written into S3 bucket by EC2 instances. Outputs are also uploaded to Github.
- Output format:
  Each line contains - <each_input_word_in_chache_file> <number_of_occurrences>
- Note: More info about the logic and code details are mentioned as code comments.
- Output links:
  1. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_distributed/part-r-00000
  2. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_distributed/part-r-00001
  3. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_distributed/part-r-00002
  4. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_distributed/part-r-00003
  5. https://s3.us-east-2.amazonaws.com/inputdataforwordcount/output_distributed/part-r-00004

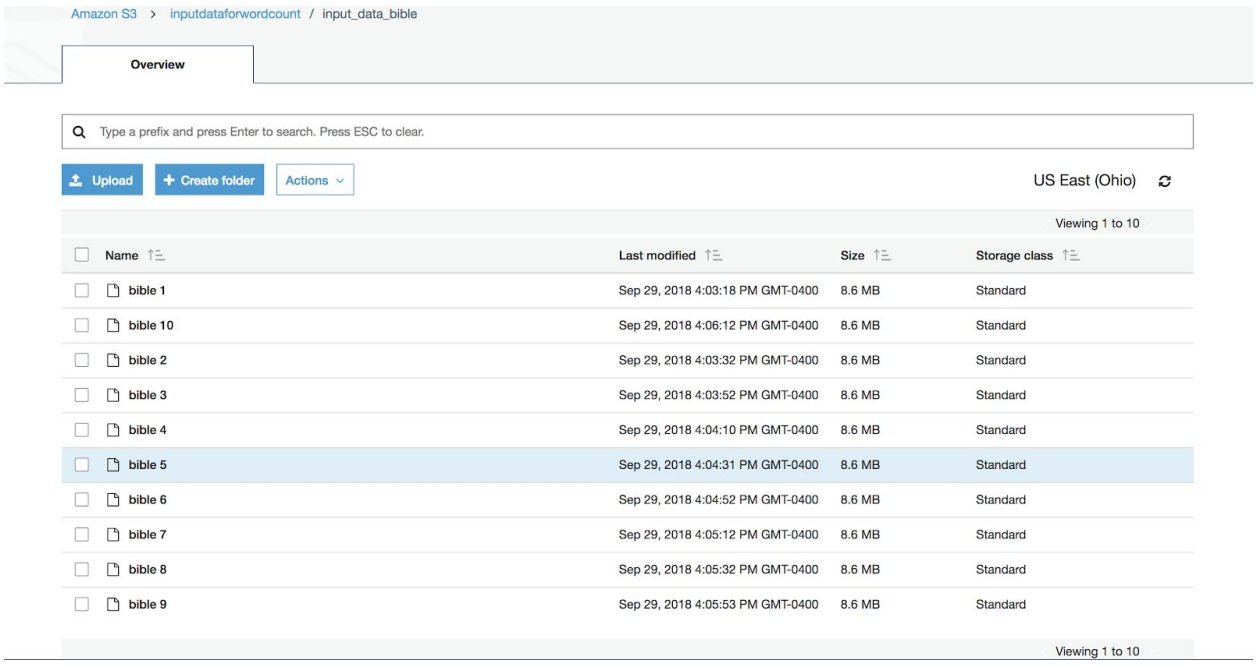
Steps to execute on AWS:

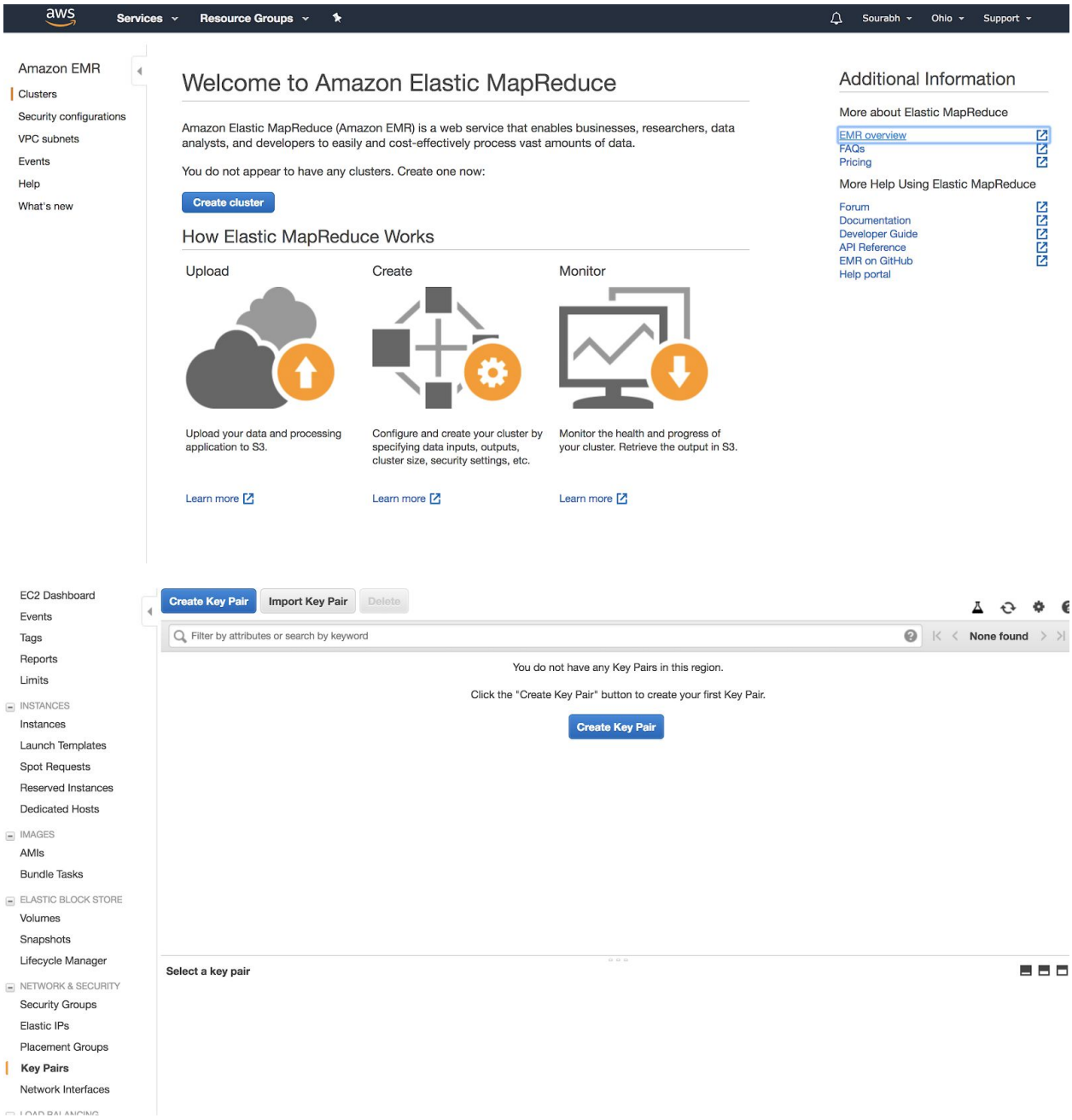
- Run the each maven project as Java application.
- Export the project as a Runnable JAR file.
- Create a Amazon S3 bucket and upload all input files and JAR's to it.

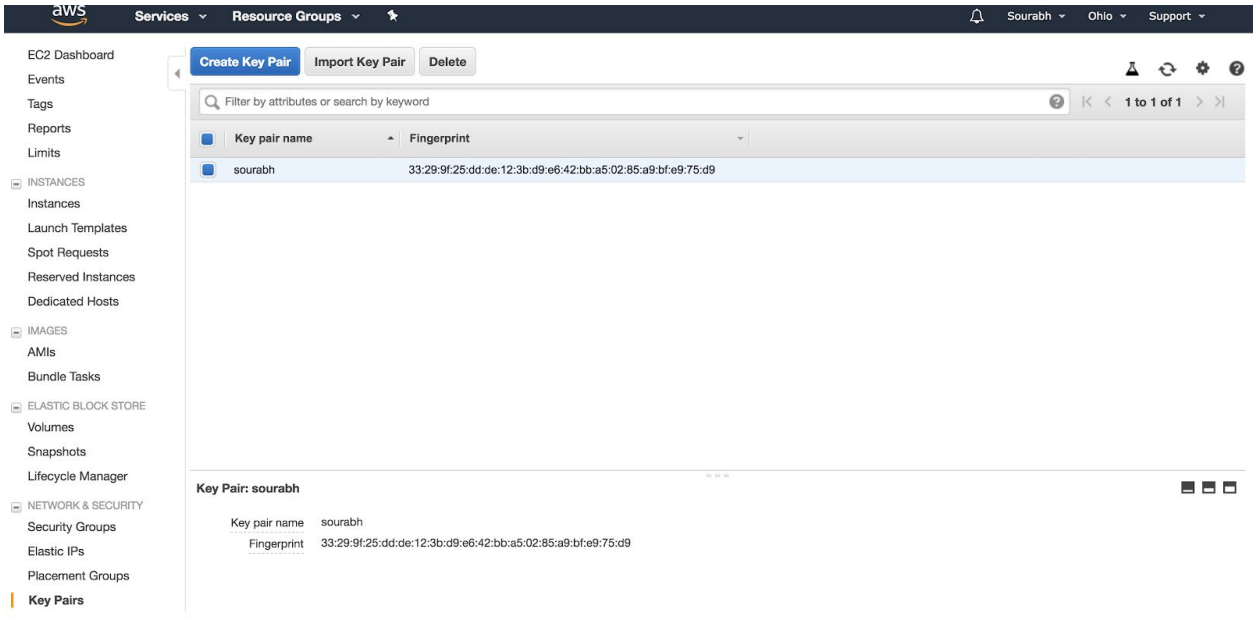
- Create a EC2 Key Pair.
- Create a cluster in Amazon EMR by selecting required EC2 instance type and number of instances. 1 will be a master instance and the others will be slave instances.
- After instances are running, create a task by selecting the appropriate JAR, input files and output folder that were uploaded previously in S3 bucket.
- After the task is completed, output will be written to the mentioned output folder in S3 bucket.
- Download the output files if required.
- Screenshots of above mentioned steps are attached below.

Screenshots:

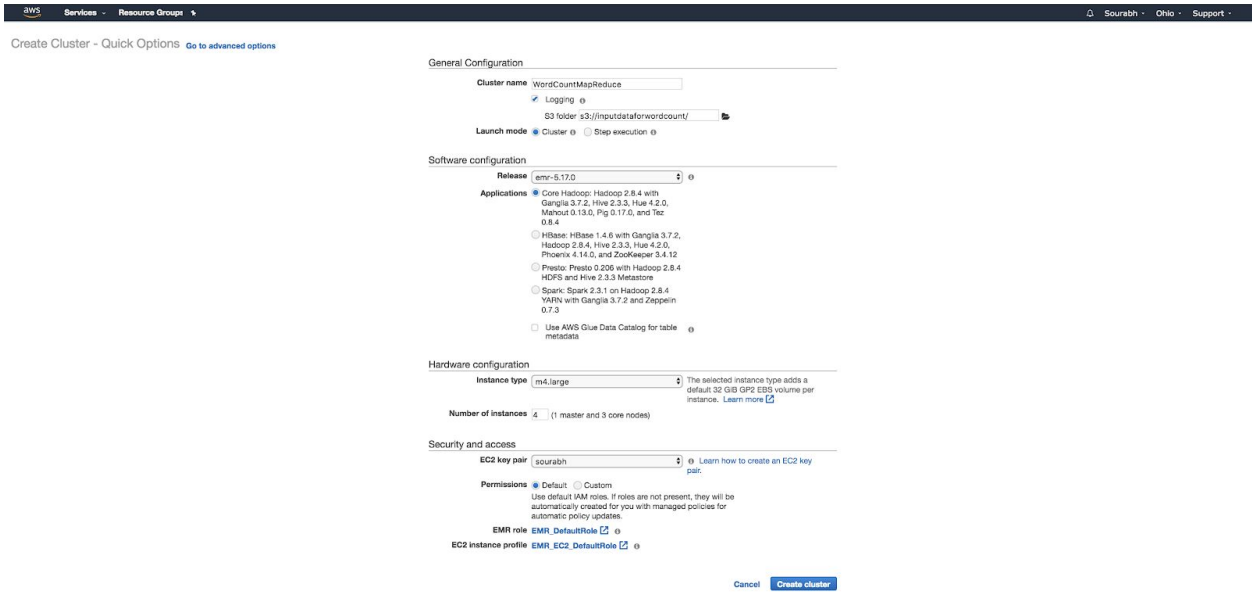


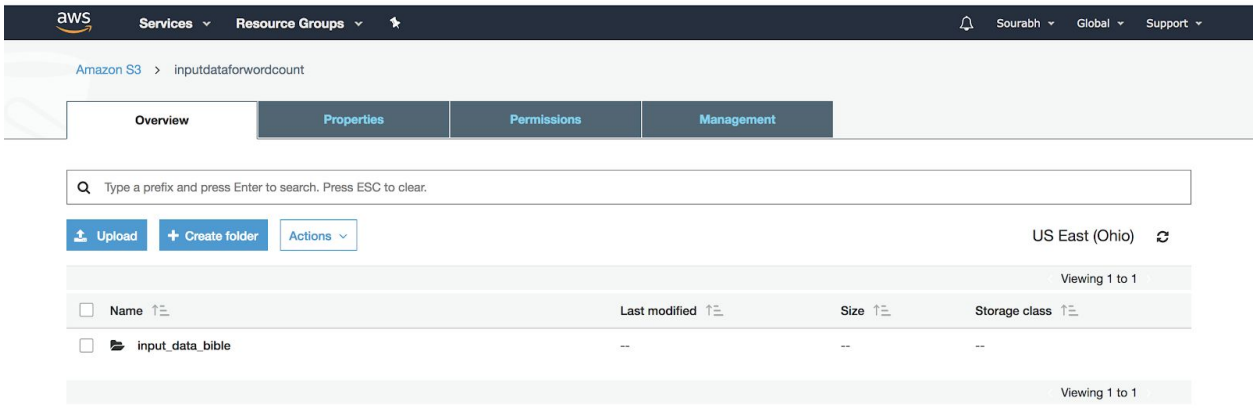



aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Welcome to Amazon Elastic MapReduce
Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and 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
Upload your data and processing application to S3.
Learn more
Create
Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.
Learn more
Monitor
Monitor the health and progress of your cluster. Retrieve the output in S3.
Learn more
Additional Information
More about Elastic MapReduce
EMR overview
FAQs
Pricing
More Help Using Elastic MapReduce
Forum
Documentation
Developer Guide
API Reference
EMR on GitHub
Help portal
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces
Create Key Pair
Import Key Pair
Delete
Filter by attributes or search by keyword
None found
You do not have any Key Pairs in this region.
Click the "Create Key Pair" button to create your first Key Pair.
Create Key Pair
Select a key pair

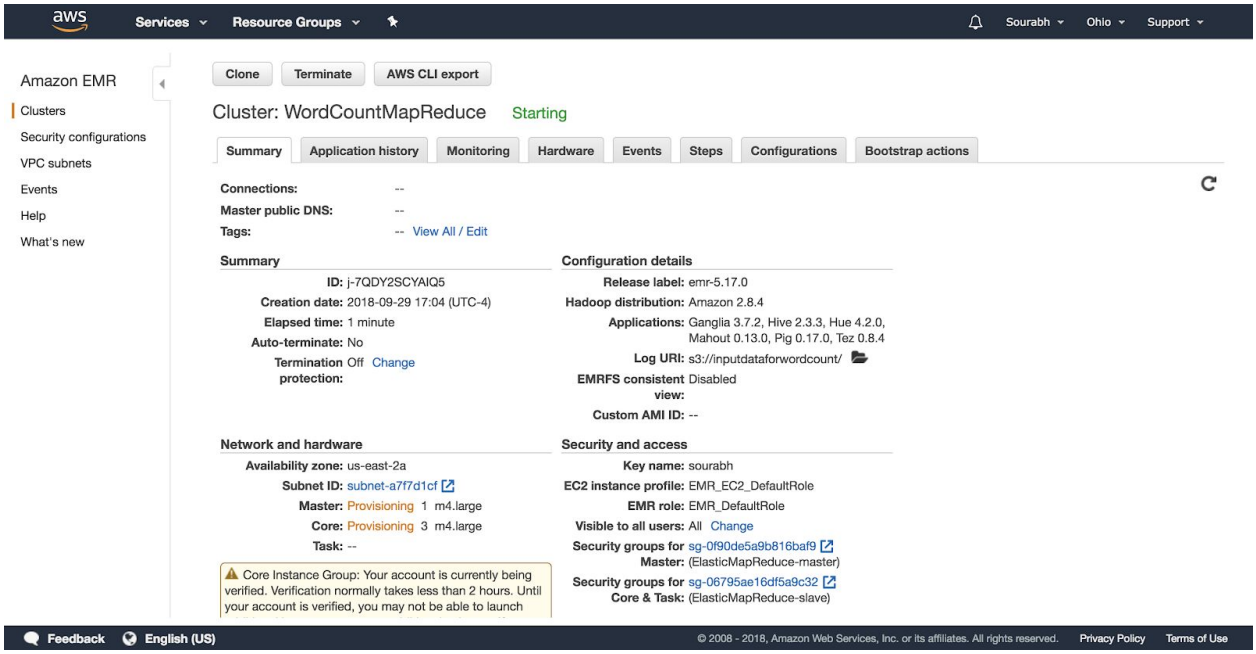

Services
Resource Groups
Sourabh
Ohio
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Create Key Pair
Import Key Pair
Delete
Filter by attributes or search by keyword
1 to 1 of 1
Key pair name
Fingerprint
sourabh
33:29:9f:25:dd:de:12:3b:d9:e6:42:bb:a5:02:85:a9:bf:e9:75:d9
Key Pair: sourabh
Key pair name sourabh
Fingerprint 33:29:9f:25:dd:de:12:3b:d9:e6:42:bb:a5:02:85:a9:bf:e9:75:d9
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

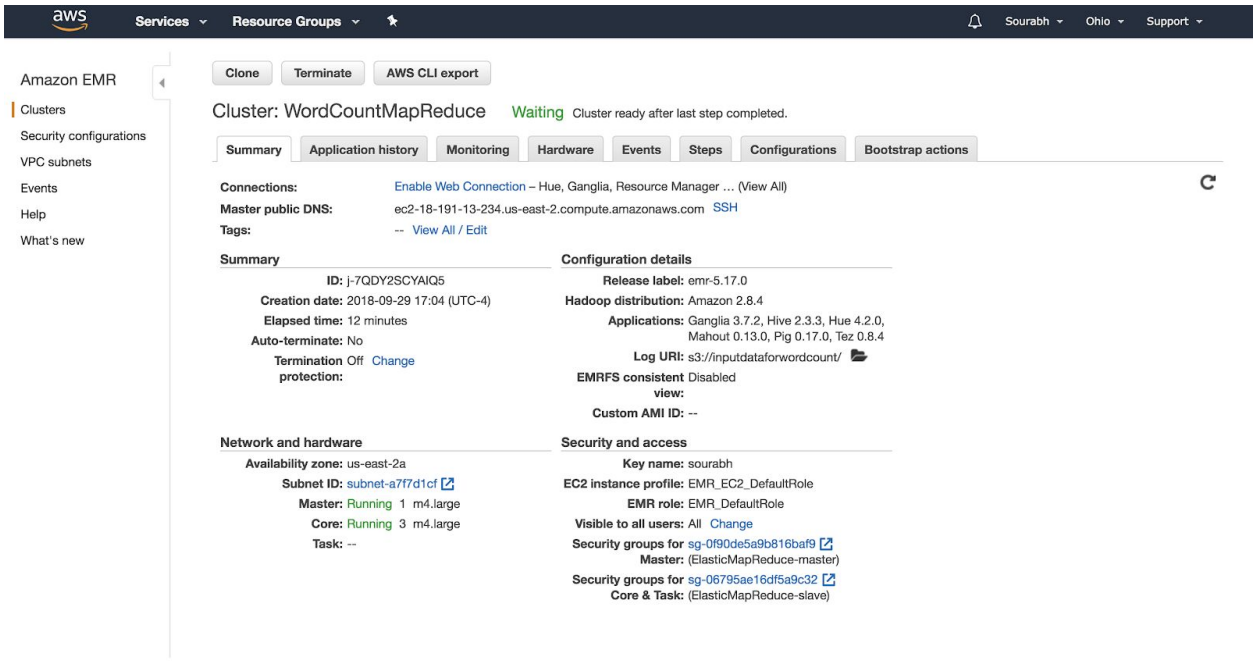

sourabh.pem
Show All

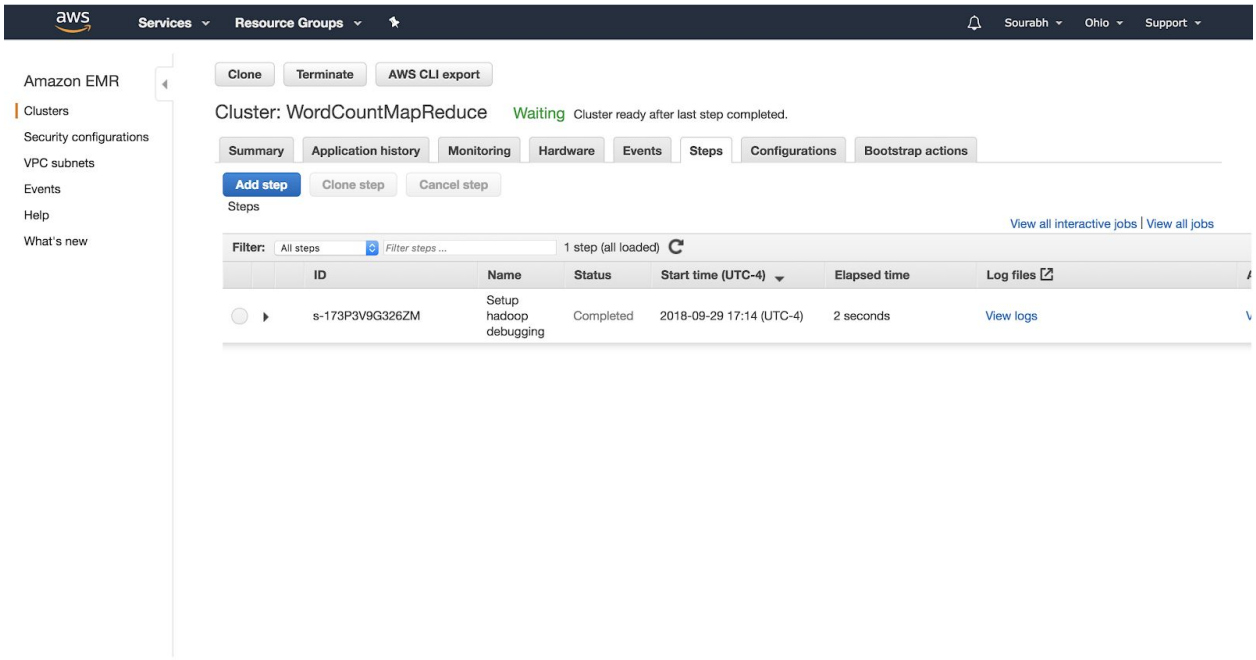

Services
Resource Groups
Sourabh
Ohio
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name WordCountMapReduce
Logging
S3 folder s3://inputdataforwordcount/
Launch mode Cluster Step execution
Software configuration
Release emr-5.17.0
Applications
Core Hadoop: Hadoop 2.8.4 with Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.8.4
HBase: HBase 1.4.6 with Ganglia 3.7.2, Hadoop 2.8.4, Hive 2.3.3, Hue 4.2.0, Phoenix 4.14.0, and ZooKeeper 3.4.12
Presto: Presto 0.206 with Hadoop 2.8.4 HDFS and Hive 2.3.3 Metastore
Spark: Spark 2.3.1 on Hadoop 2.8.4 YARN with Ganglia 3.7.2 and Zeppelin 0.7.3
Use AWS Glue Data Catalog for table metadata
Hardware configuration
Instance type m4.large
The selected instance type adds a default 32 GiB GP2 EBS volume per instance. Learn more
Number of instances 4 (1 master and 3 core nodes)
Security and access
EC2 key pair sourabh
Learn how to create an EC2 key pair.
Permissions Default Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role EMR_DefaultRole
EC2 instance profile EMR_EC2_DefaultRole
Cancel
Create cluster

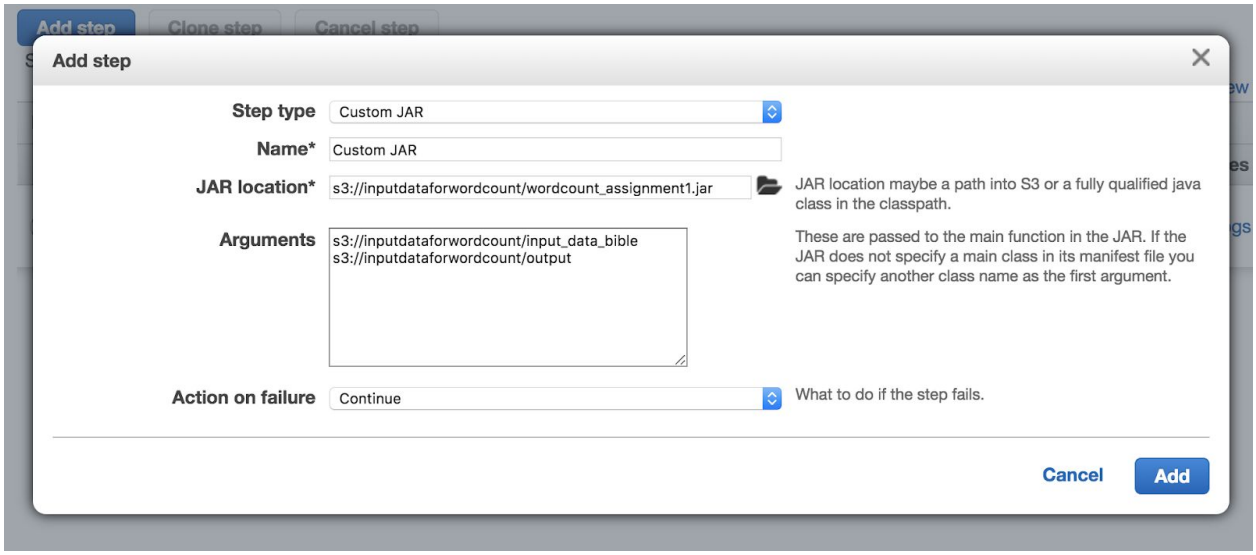

aws
Services
Resource Groups
Sourabh
Global
Support
Amazon S3 > inputdataforwordcount
Overview
Properties
Permissions
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
+ Create folder
Actions
US East (Ohio)
Viewing 1 to 1
Name
Last modified
Size
Storage class
input_data_bible
--
--
--
Viewing 1 to 1

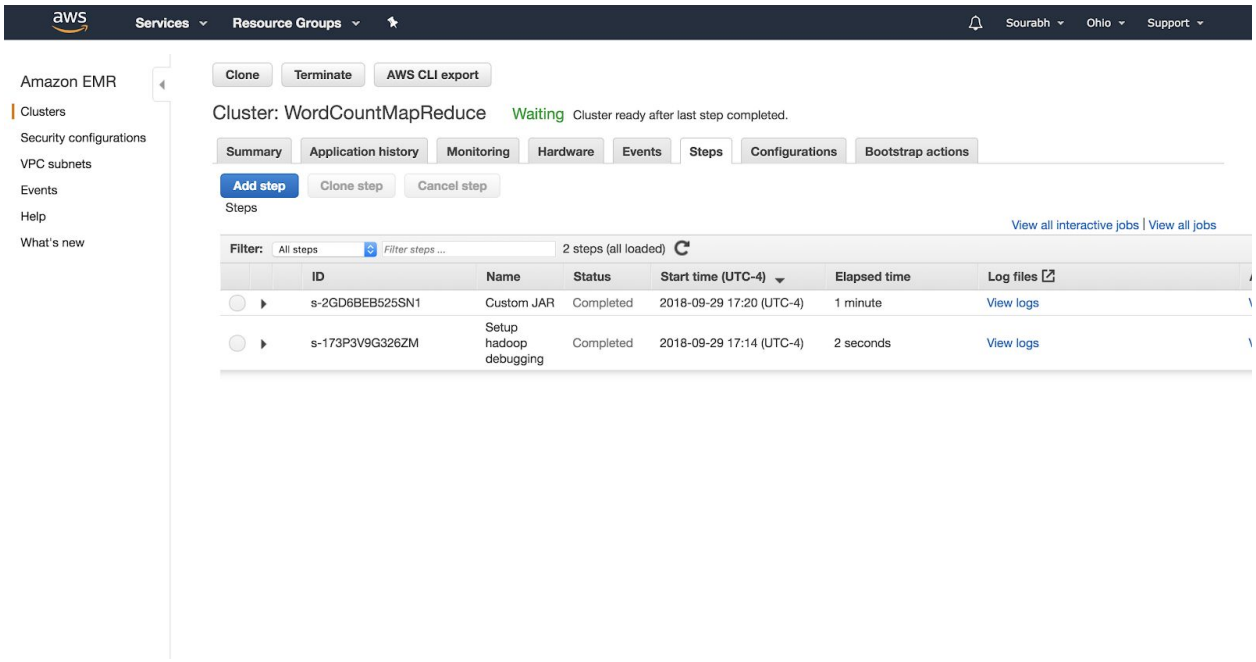

Upload (1.1 MB/s)
14.06% Successful
Uploading file: wordcount_assignment1.jar
Operations
1 In progress
3 Success
0 Error

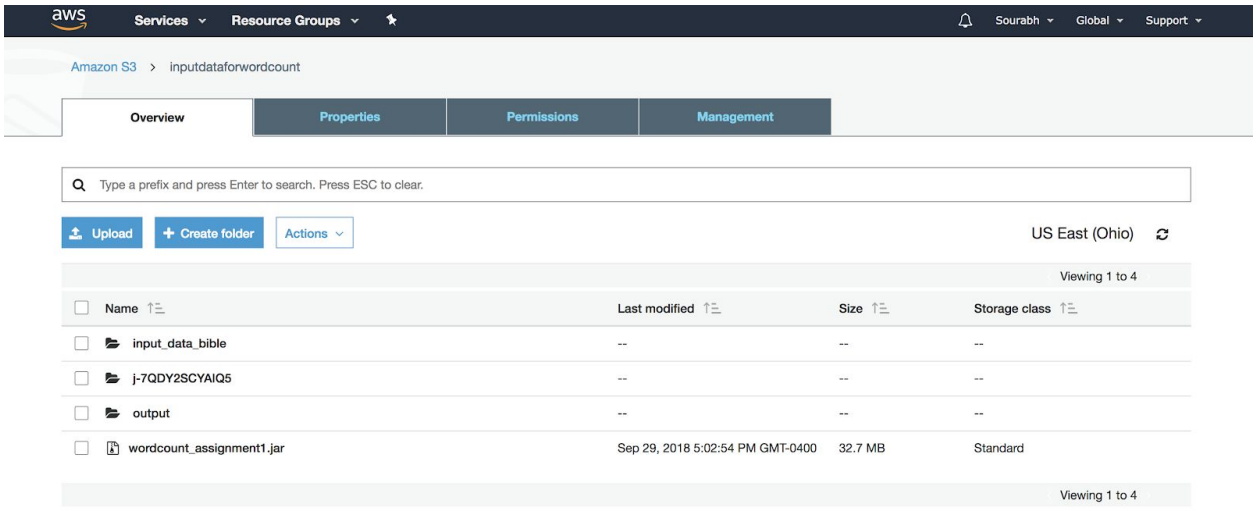

aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: WordCountMapReduce
Starting
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Connections: --
Master public DNS: --
Tags: -- View All / Edit
Summary
ID: j-7QDY2SCYAIQ5
Creation date: 2018-09-29 17:04 (UTC-4)
Elapsed time: 1 minute
Auto-terminate: No
Termination protection: Off Change
Configuration details
Release label: emr-5.17.0
Hadoop distribution: Amazon 2.8.4
Applications: Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://inputdataforwordcount/
EMRFS consistent view: Disabled
Custom AMI ID: --
Network and hardware
Availability zone: us-east-2a
Subnet ID: subnet-a7f7d1cf
Master: Provisioning 1 m4.large
Core: Provisioning 3 m4.large
Task: --
Core Instance Group: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch
Security and access
Key name: sourabh
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0f90de5a9b816baf9 (ElasticMapReduce-master)
Security groups for Core & Task: sg-06795ae16df5a9c32 (ElasticMapReduce-slave)
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

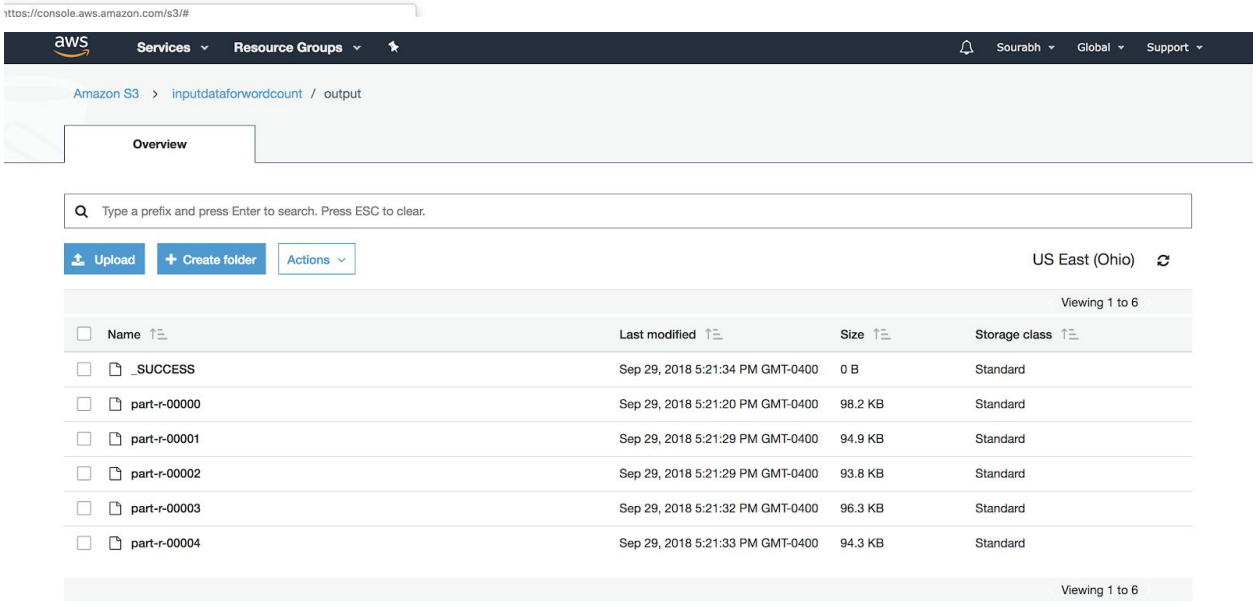

aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: WordCountMapReduce
Waiting
Cluster ready after last step completed.
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Connections: Enable Web Connection – Hue, Ganglia, Resource Manager … (View All)
Master public DNS: ec2-18-191-13-234.us-east-2.compute.amazonaws.com SSH
Tags: -- View All / Edit
Summary
ID: j-7QDY2SCYAIQ5
Creation date: 2018-09-29 17:04 (UTC-4)
Elapsed time: 12 minutes
Auto-terminate: No
Termination protection: Off Change
Configuration details
Release label: emr-5.17.0
Hadoop distribution: Amazon 2.8.4
Applications: Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://inputdataforwordcount/
EMRFS consistent view: Disabled
Custom AMI ID: --
Network and hardware
Availability zone: us-east-2a
Subnet ID: subnet-a7f7d1cf
Master: Running 1 m4.large
Core: Running 3 m4.large
Task: --
Security and access
Key name: sourabh
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0f90de5a9b816baf9 (ElasticMapReduce-master)
Security groups for Core & Task: sg-06795ae16df5a9c32 (ElasticMapReduce-slave)

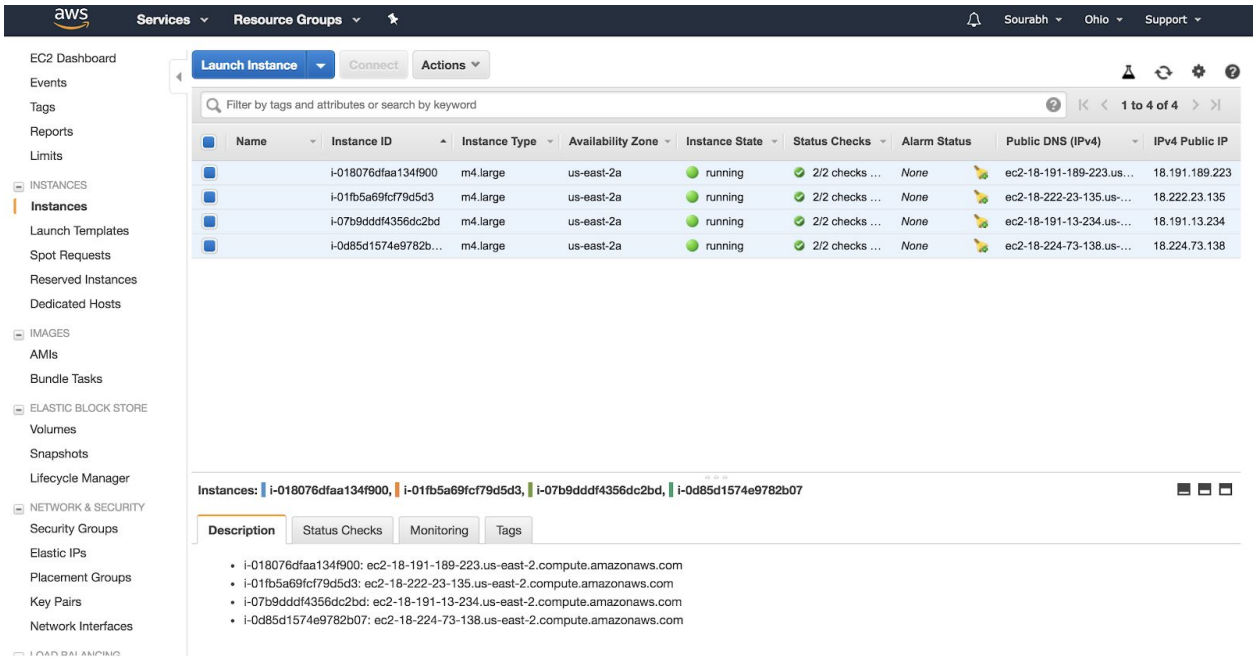

aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: WordCountMapReduce
Waiting
Cluster ready after last step completed.
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add step
Clone step
Cancel step
Steps
View all interactive jobs | View all jobs
Filter: All steps Filter steps ... 1 step (all loaded)
ID
Name
Status
Start time (UTC-4)
Elapsed time
Log files
s-173P3V9G326ZM
Setup hadoop debugging
Completed
2018-09-29 17:14 (UTC-4)
2 seconds
View logs

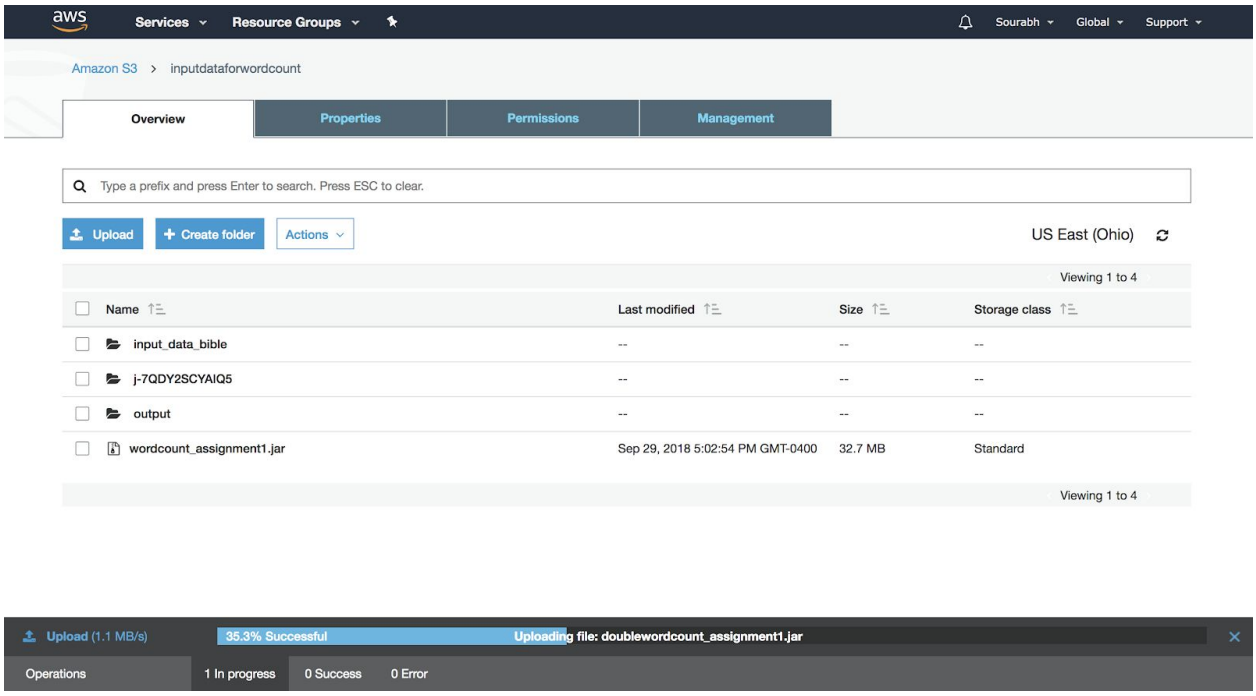

Add step
Clone step
Cancel step
Add step
Step type
Custom JAR
Name*
Custom JAR
JAR location*
s3://inputdataforwordcount/wordcount_assignment1.jar
JAR location maybe a path into S3 or a fully qualified java class in the classpath.
Arguments
s3://inputdataforwordcount/input_data_bible
s3://inputdataforwordcount/output
These are passed to the main function in the JAR. If the JAR does not specify a main class in its manifest file you can specify another class name as the first argument.
Action on failure
Continue
What to do if the step fails.
Cancel
Add


aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: WordCountMapReduce
Waiting
Cluster ready after last step completed.
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add step
Clone step
Cancel step
Steps
View all interactive jobs | View all jobs
Filter:
All steps
Filter steps ...
2 steps (all loaded)
ID
Name
Status
Start time (UTC-4)
Elapsed time
Log files
s-2GD6BEB525SN1
Custom JAR
Completed
2018-09-29 17:20 (UTC-4)
1 minute
View logs
s-173P3V9G326ZM
Setup hadoop debugging
Completed
2018-09-29 17:14 (UTC-4)
2 seconds
View logs

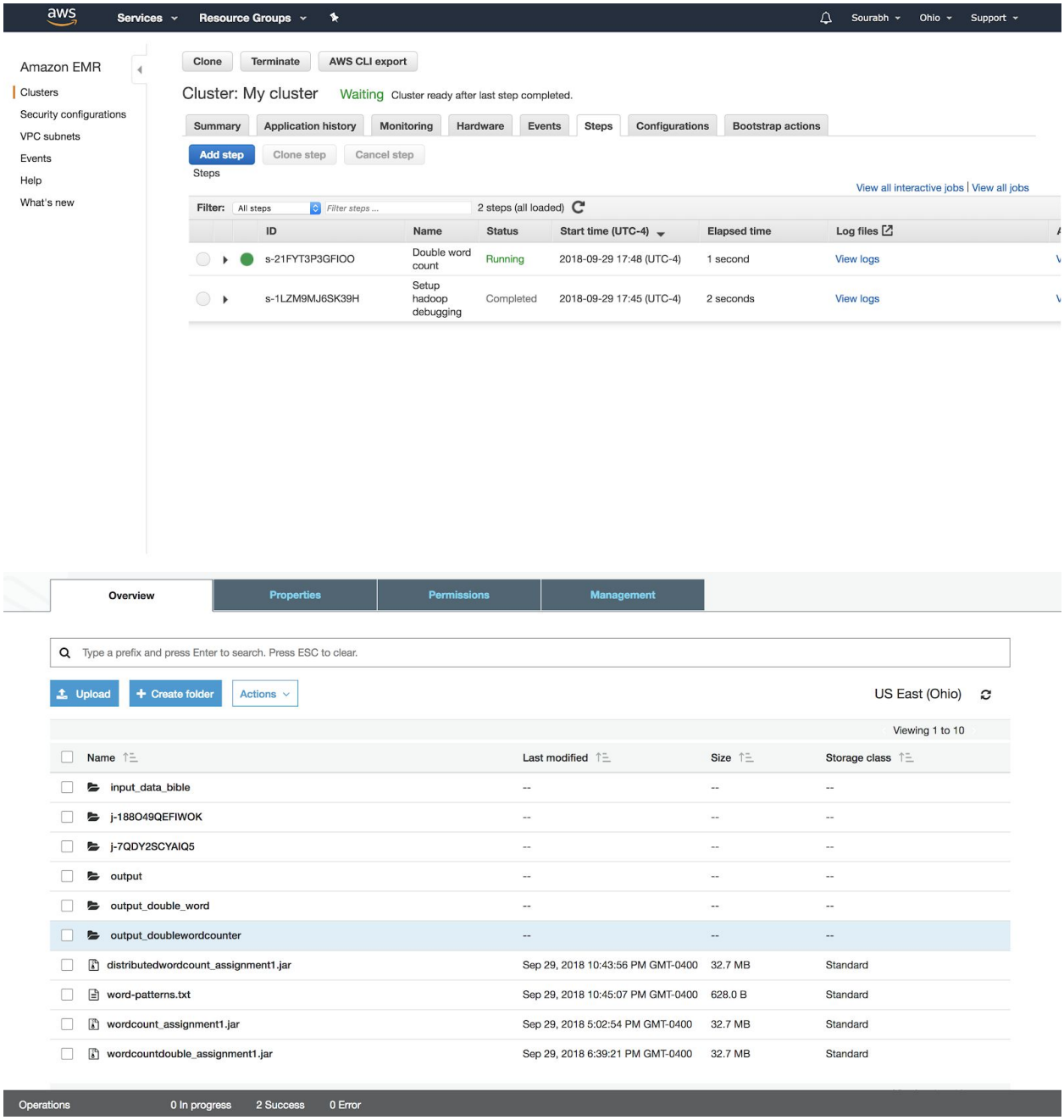

aws
Services
Resource Groups
Sourabh
Global
Support
Amazon S3 > inputdataforwordcount
Overview
Properties
Permissions
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
+ Create folder
Actions
US East (Ohio)
Viewing 1 to 4
Name
Last modified
Size
Storage class
input_data_bible -- -- --
j-7QDY2SCYAIQ5 -- -- --
output -- -- --
wordcount_assignment1.jar Sep 29, 2018 5:02:54 PM GMT-0400 32.7 MB Standard
Viewing 1 to 4

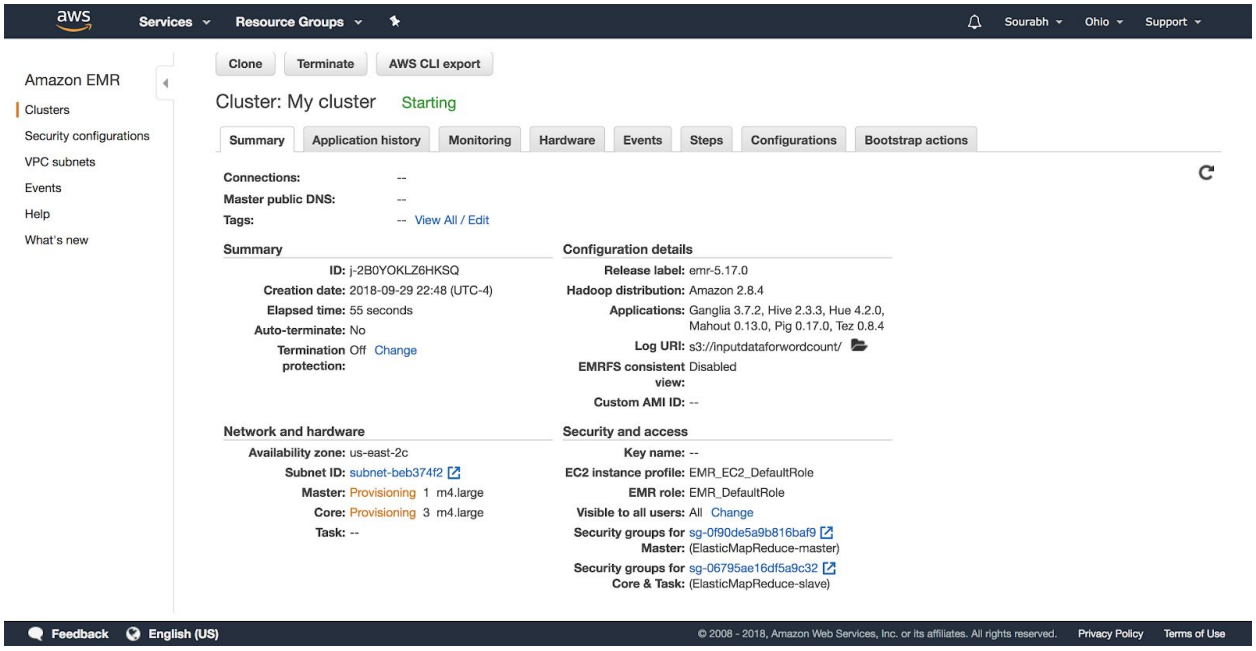

https://console.aws.amazon.com/s3/#
aws
Services
Resource Groups
Sourabh
Global
Support
Amazon S3 > inputdataforwordcount / output
Overview
Type a prefix and press Enter to search. Press ESC to clear.
Upload
+ Create folder
Actions
US East (Ohio)
Viewing 1 to 6
Name
Last modified
Size
Storage class
_SUCCESS Sep 29, 2018 5:21:34 PM GMT-0400 0 B Standard
part-r-00000 Sep 29, 2018 5:21:20 PM GMT-0400 98.2 KB Standard
part-r-00001 Sep 29, 2018 5:21:29 PM GMT-0400 94.9 KB Standard
part-r-00002 Sep 29, 2018 5:21:29 PM GMT-0400 93.8 KB Standard
part-r-00003 Sep 29, 2018 5:21:32 PM GMT-0400 96.3 KB Standard
part-r-00004 Sep 29, 2018 5:21:33 PM GMT-0400 94.3 KB Standard
Viewing 1 to 6

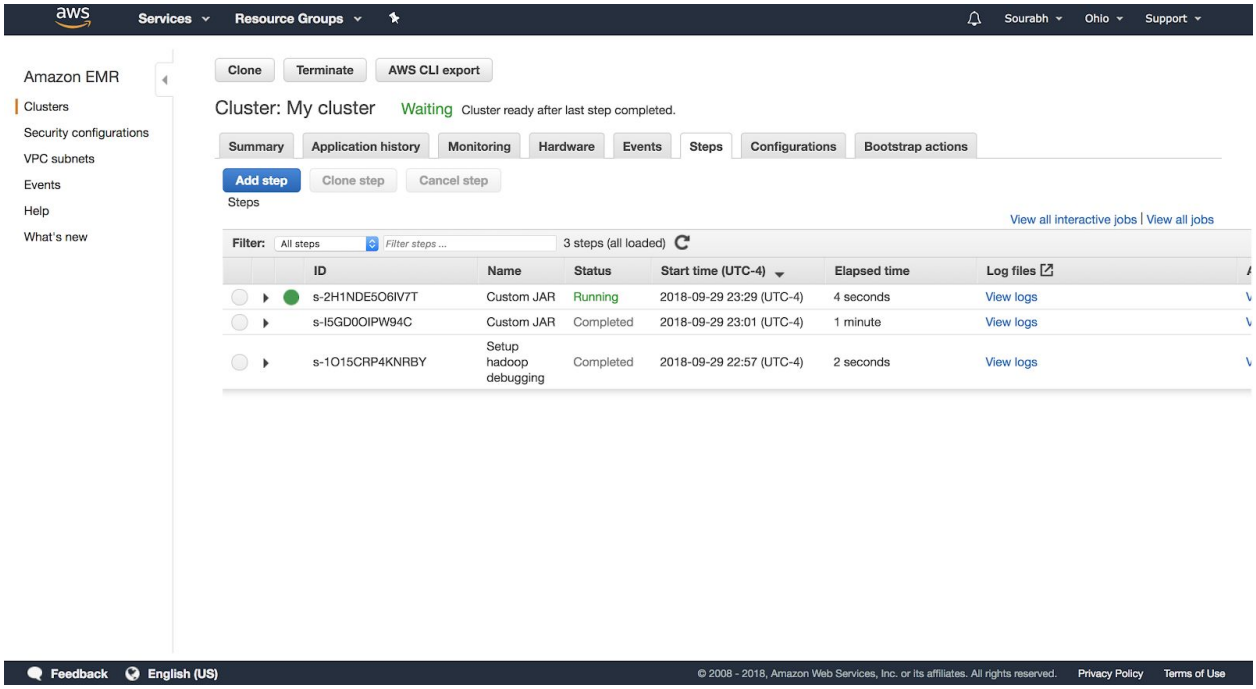

aws
Services
Resource Groups
Sourabh
Ohio
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 4 of 4
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
Alarm Status
Public DNS (IPv4)
IPv4 Public IP
i-018076dfaa134f900 m4.large us-east-2a running 2/2 checks ... None ec2-18-191-189-223.us... 18.191.189.223
i-01fb5a69fcf79d5d3 m4.large us-east-2a running 2/2 checks ... None ec2-18-222-23-135.us-... 18.222.23.135
i-07b9dddf4356dc2bd m4.large us-east-2a running 2/2 checks ... None ec2-18-191-13-234.us-... 18.191.13.234
i-0d85d1574e9782b... m4.large us-east-2a running 2/2 checks ... None ec2-18-224-73-138.us-... 18.224.73.138
Instances: i-018076dfaa134f900, i-01fb5a69fcf79d5d3, i-07b9dddf4356dc2bd, i-0d85d1574e9782b07
Description
Status Checks
Monitoring
Tags
i-018076dfaa134f900: ec2-18-191-189-223.us-east-2.compute.amazonaws.com
i-01fb5a69fcf79d5d3: ec2-18-222-23-135.us-east-2.compute.amazonaws.com
i-07b9dddf4356dc2bd: ec2-18-191-13-234.us-east-2.compute.amazonaws.com
i-0d85d1574e9782b07: ec2-18-224-73-138.us-east-2.compute.amazonaws.com


aws
Services
Resource Groups
Sourabh
Global
Support
Amazon S3 > inputdataforwordcount
Overview
Properties
Permissions
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
Actions
US East (Ohio)
Viewing 1 to 4
Name
Last modified
Size
Storage class
input_data_bible -- -- --
j-7QDY2SCYAIQ5 -- -- --
output -- -- --
wordcount_assignment1.jar Sep 29, 2018 5:02:54 PM GMT-0400 32.7 MB Standard
Viewing 1 to 4

Upload (1.1 MB/s)
35.3% Successful
Uploading file: doublewordcount_assignment1.jar
Operations
1 In progress
0 Success
0 Error


aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: My cluster
Waiting
Cluster ready after last step completed.
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add step
Clone step
Cancel step
Steps
View all interactive jobs | View all jobs
Filter:
All steps
Filter steps ...
2 steps (all loaded)
ID
Name
Status
Start time (UTC-4)
Elapsed time
Log files
s-21FYT3P3GFIOO
Double word count
Running
2018-09-29 17:48 (UTC-4)
1 second
View logs
s-1LZM9MJ6SK39H
Setup hadoop debugging
Completed
2018-09-29 17:45 (UTC-4)
2 seconds
View logs
Overview
Properties
Permissions
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
Actions
US East (Ohio)
Viewing 1 to 10
Name
Last modified
Size
Storage class
input_data_bible
j-188O49QEFIWOK
j-7QDY2SCYAIQ5
output
output_double_word
output_doublewordcounter
distributedwordcount_assignment1.jar
Sep 29, 2018 10:43:56 PM GMT-0400
32.7 MB
Standard
word-patterns.txt
Sep 29, 2018 10:45:07 PM GMT-0400
628.0 B
Standard
wordcount_assignment1.jar
Sep 29, 2018 5:02:54 PM GMT-0400
32.7 MB
Standard
wordcountdouble_assignment1.jar
Sep 29, 2018 6:39:21 PM GMT-0400
32.7 MB
Standard
Operations
0 In progress
2 Success
0 Error


aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: My cluster
Starting
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Connections: --
Master public DNS: --
Tags: -- View All / Edit
Summary
ID: j-2B0YOKLZ6HKSQ
Creation date: 2018-09-29 22:48 (UTC-4)
Elapsed time: 55 seconds
Auto-terminate: No
Termination protection: Off Change
Configuration details
Release label: emr-5.17.0
Hadoop distribution: Amazon 2.8.4
Applications: Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://inputdataforwordcount/
EMRFS consistent view: Disabled
Custom AMI ID: --
Network and hardware
Availability zone: us-east-2c
Subnet ID: subnet-beb374f2
Master: Provisioning 1 m4.large
Core: Provisioning 3 m4.large
Task: --
Security and access
Key name: --
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0f90de5a9b816baf9 (ElasticMapReduce-master)
Security groups for Core & Task: sg-06795ae16df5a9c32 (ElasticMapReduce-slave)
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
Sourabh
Ohio
Support
Amazon EMR
Clusters
Security configurations
VPC subnets
Events
Help
What's new
Clone
Terminate
AWS CLI export
Cluster: My cluster
Waiting
Cluster ready after last step completed.
Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions
Add step
Clone step
Cancel step
Steps
View all interactive jobs | View all jobs
Filter: All steps
Filter steps ...
3 steps (all loaded)
ID
Name
Status
Start time (UTC-4)
Elapsed time
Log files
s-2H1NDE5O6IV7T
Custom JAR
Running
2018-09-29 23:29 (UTC-4)
4 seconds
View logs
s-I5GD0OIPW94C
Custom JAR
Completed
2018-09-29 23:01 (UTC-4)
1 minute
View logs
s-1O15CRP4KNRBY
Setup hadoop debugging
Completed
2018-09-29 22:57 (UTC-4)
2 seconds
View logs
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use