Big Data Analytics

with Amazon Web Services

Dr. Matt Wood

An Online Seminar for Partners. Wednesday 1st August.

Hello, and thank you.

amazon web services

Big Data Analytics

An introduction

Big Data Analytics

- An introduction
- The story of analytics on AWS

amazon web services

Big Data Analytics

- An introduction
- The story of analytics on AWS
- Integrating partners

amazon web services

Big Data Analytics

- An introduction
- The story of analytics on AWS
- Integrating partners
- Partner success stories

amazon web services

1

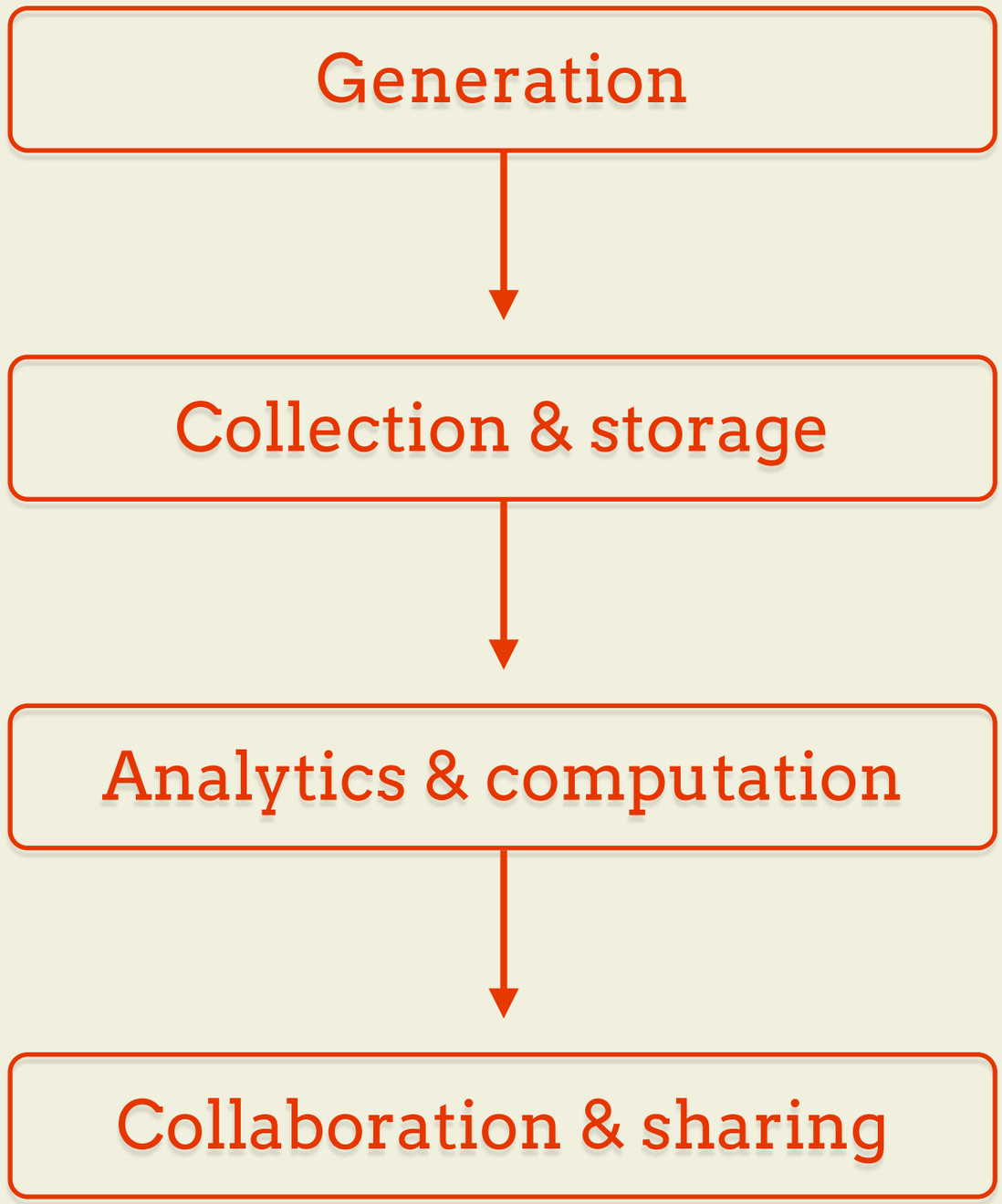
INTRODUCING BIG DATA

Data for competitive advantage.

amazon web services™

Using data

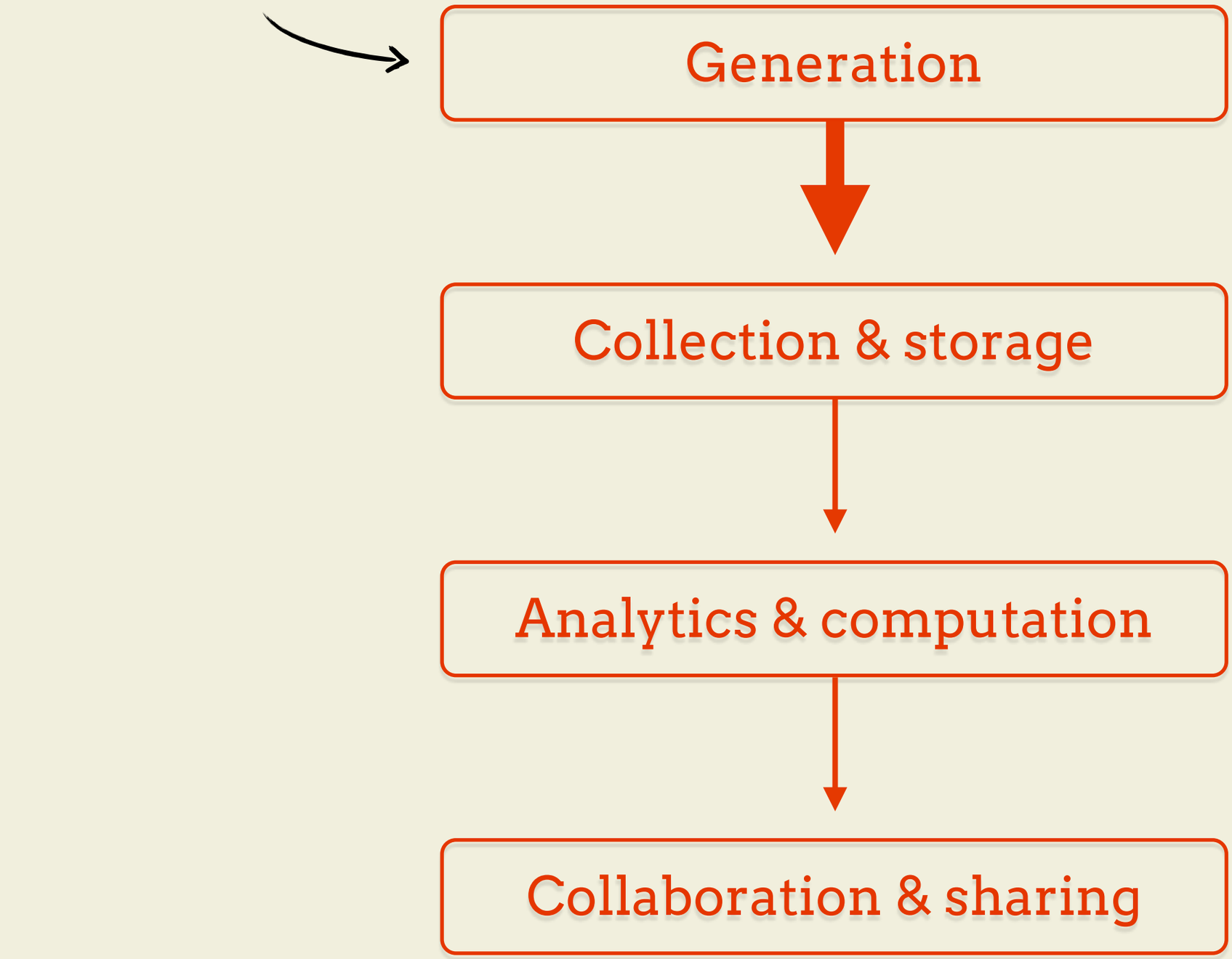
Customer segmentation,
financial modeling,
system analysis,
line-of-sight,
business intelligence.



amazon web services

Cost of data generation is falling.

LOWER COST,
INCREASED THROUGHPUT



Generation

Collection & storage

HIGHLY CONSTRAINED

Analytics & computation

Collaboration & sharing

amazon web services

Very high barrier to turning **data** into **information**.

amazon web services™

Move from a **data generation** challenge to **analytics** challenge.

Enter the **Cloud**.

amazon web services™

Remove the constraints.

Enable data-driven innovation.

Move to a **distributed** data approach.

Maturation of two things.

Software for distributed storage and analysis

Maturation of two things.

amazon web services

Software for distributed storage and analysis

Maturation of two things.

Infrastructure for distributed storage and analysis

amazon web services

Software

Frameworks for
data-intensive workloads.

Distributed by design.

Infrastructure

Platform for
data-intensive workloads.

Distributed by design.

amazon web services™

Support the data timeline.

Generation

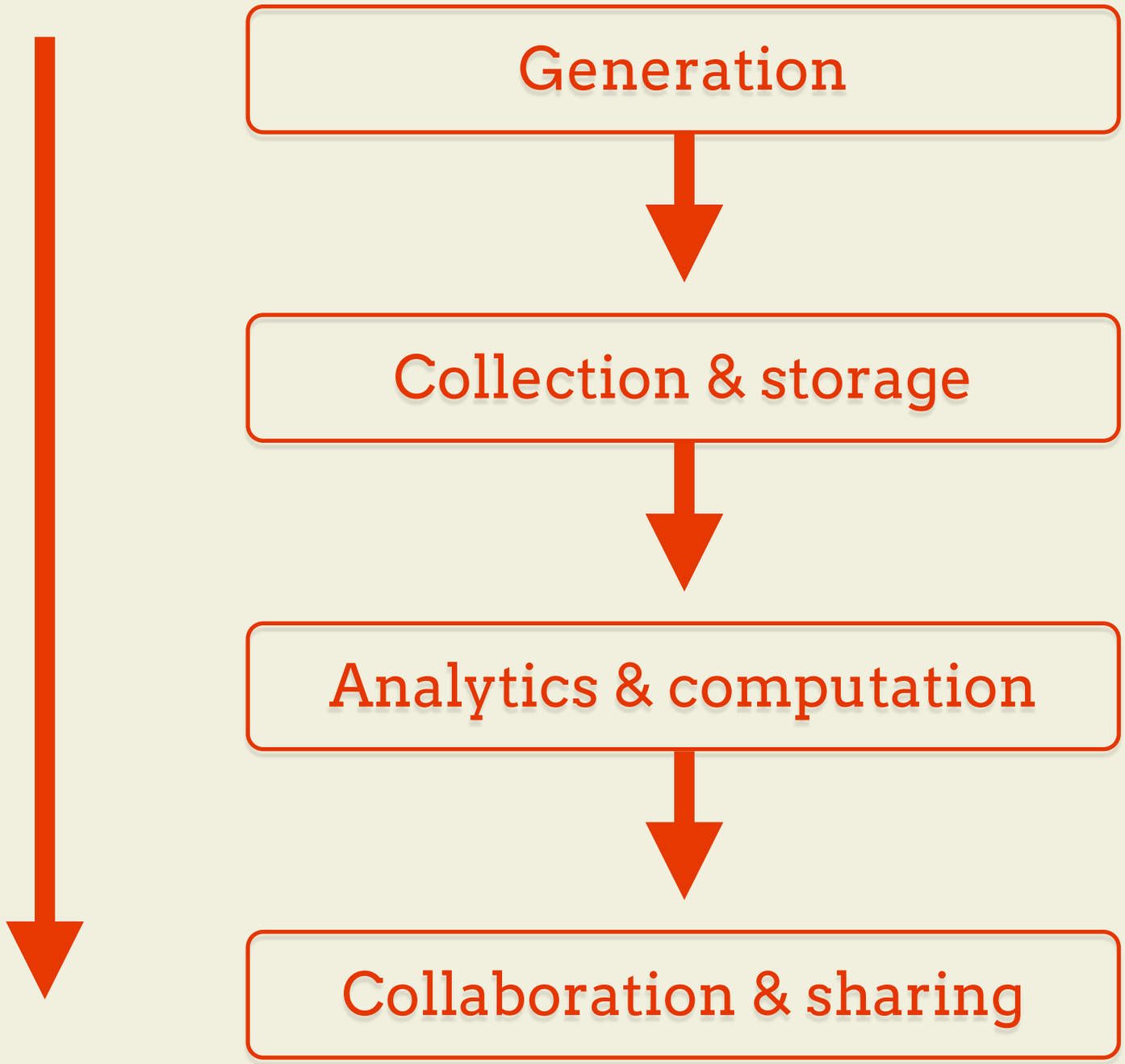
Collection & storage

HIGHLY CONSTRAINED

Analytics & computation

Collaboration & sharing

amazon web services



Lower the barrier to entry.

Accelerate time to market and increase agility.

amazon web services™

Enable new business opportunities.

amazon web services

Washington Post

Pinterest

NASA

"AWS enables Pfizer to explore difficult or deep scientific questions in a **timely**, **scalable** manner and helps us make **better** decisions more **quickly**"

Michael Miller, Pfizer

amazon web services

2

THE STORY OF ANALYTICS

EC2

Utility computing.
6 years young.

amazon web services™

Scale out systems

Embarrassingly parallel problems.

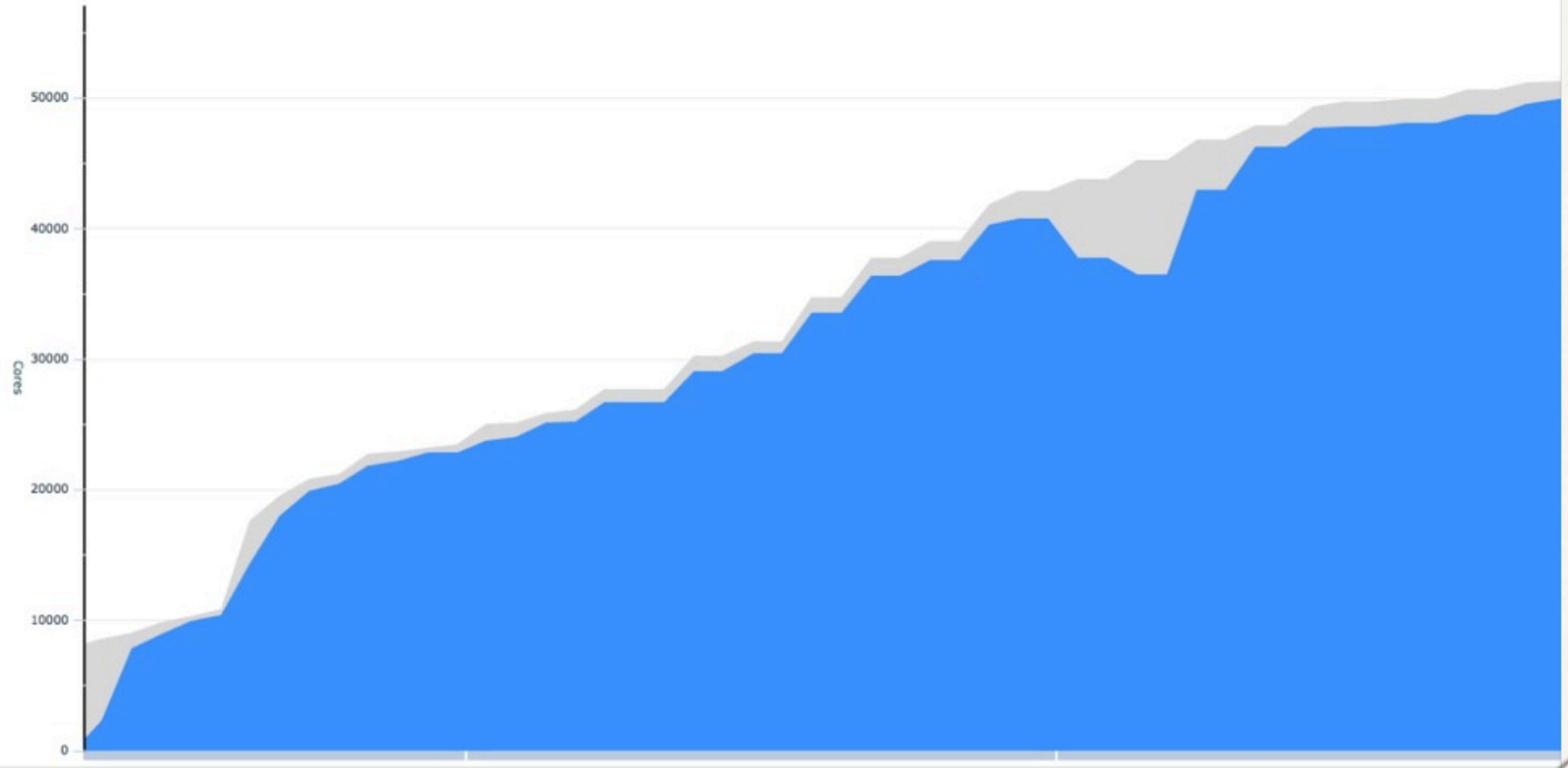
Queue based distribution.

Small, medium and high scale.

CycleServer Usage Jobs Grid Settings

Show: Historical grid usage in Naga-RC1 pool

Time Frame: 3 Hours | Day | Week | Month



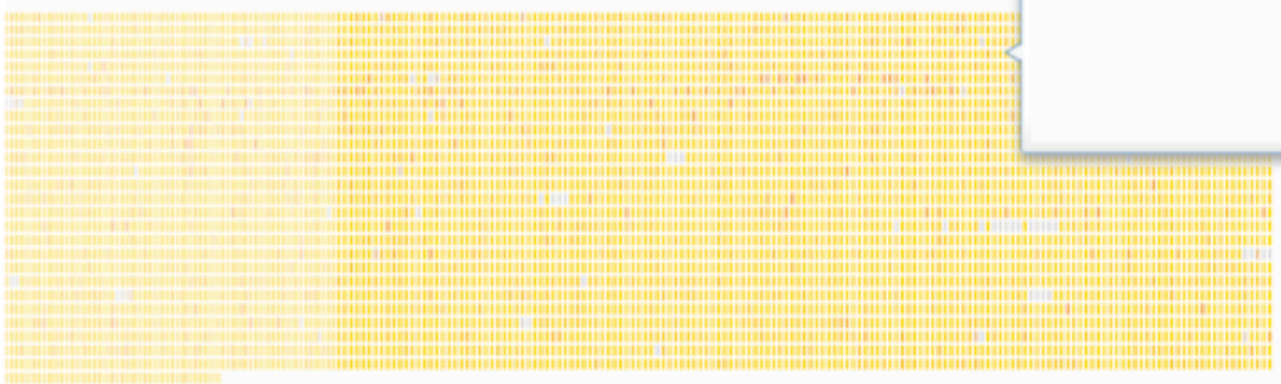
amazon web services

Logged in as admin | Admin ▾ | Log Out

CycleServer | Clusters | Jobs | Queues | Hosts | Reports | Chef

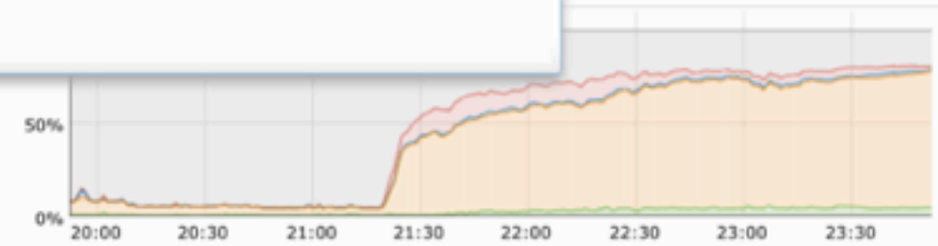
Hosts in all clusters

Show: CPU Usage by Host



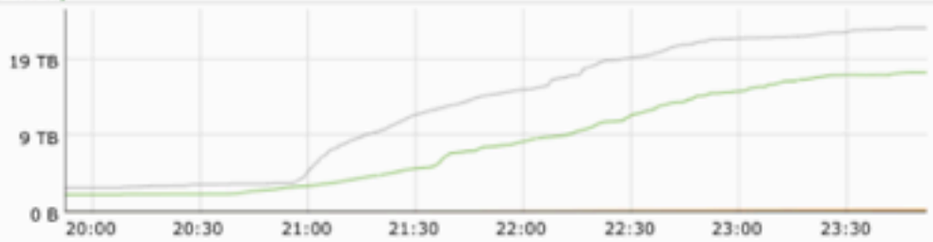
Host: ip-10-108-242-201.ec2.internal

OS:	Linux	Idle CPU:	1%
OS Ver.:	2.6.18-238.12.1.el5xen	User CPU:	4%
# CPUs:	8	System CPU:	2%
Clock Speed:	2 GHz	Nice CPU:	94%
		I/O Wait CPU:	0%



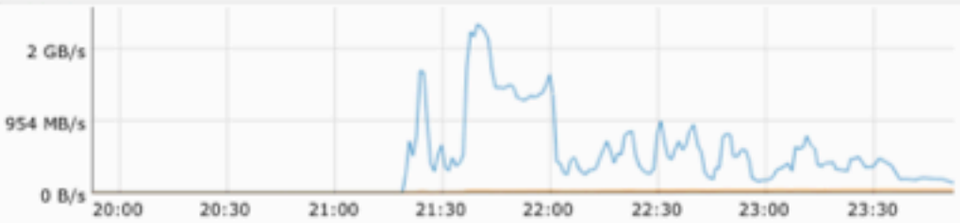
cpu_idle	22%
cpu_wio	4%
cpu_system	1%
cpu_nice	69%
cpu_user	5%

Memory



mem_total	21 TB
mem_swapped	2 GB
mem_buffers	209 GB
mem_cached	15 TB
mem_shared	0 B

Network



bytes_in	491 MB/s
bytes_out	36 MB/s

Host Listing

View: Performance ▾ Save Refresh | View Detail | Search:

Host	Cluster	Clock Speed	Memory	CPU Usage	Mem Usage	Net In	Net Out	Last Reported ▾
ip-10-252-3-6.sa-east-1.	SAEast1a	2 GHz	15 GB	7%	32%	10 kB/s	6 kB/s	11:56 PM
ip-10-252-1-135.sa-east	SAEast1a	2 GHz	7 GB	0%	54%	18 kB/s	2 kB/s	11:56 PM
ip-10-252-0-173.sa-east	SAEast1a	2 GHz	7 GB	100%	95%	146 B/s	4 kB/s	11:56 PM
ip-10-252-64-46.sa-east	SAEast1a	3 GHz	68 GB	0%	11%	227 B/s	3 kB/s	11:56 PM
ip-10-252-0-219.sa-east	SAEast1a	2 GHz	7 GB	100%	95%	528 B/s	8 kB/s	11:56 PM
ip-10-252-1-154.sa-east	SAEast1a	2 GHz	7 GB	100%	98%	290 B/s	2 kB/s	11:56 PM
ip-10-252-2-68.sa-east-	SAEast1a	2 GHz	7 GB	100%	98%	3 kB/s	11 kB/s	11:56 PM
ip-10-252-0-166.sa-east	SAEast1a	2 GHz	7 GB	100%	96%	247 B/s	2 kB/s	11:56 PM
ip-10-252-1-151.sa-east	SAEast1a	2 GHz	7 GB	100%	96%	138 B/s	2 kB/s	11:56 PM
ip-10-119-41-201.ec2.in	Naga-RC1	3 GHz	34 GB	27%	50%	8 MB/s	2 kB/s	11:56 PM
ip-10-102-23-171.ec2.in	Naga-RC1	3 GHz	34 GB	98%	48%	116 kB/s	10 kB/s	11:56 PM
ip-10-252-13-66.sa-east	SAEast1a	2 GHz	7 GB	100%	96%	198 B/s	4 kB/s	11:56 PM
ip-10-252-1-137.sa-east	SAEast1a	2 GHz	7 GB	100%	96%	2 kB/s	2 kB/s	11:56 PM
ip-10-252-3-233.sa-east	SAEast1a	2 GHz	15 GB	0%	35%	16 kB/s	548 B/s	11:56 PM
ip-10-252-2-102.sa-east	SAEast1a	2 GHz	7 GB	88%	95%	3 kB/s	10 kB/s	11:56 PM
ip-10-252-8-119.sa-east	SAEast1a	2 GHz	15 GB	0%	22%	8 kB/s	4 kB/s	11:56 PM
ip-10-252-2-46.sa-east-	SAEast1a	2 GHz	15 GB	0%	22%	152 B/s	159 B/s	11:56 PM
ip-10-248-69-41.us-west	USWest2a	2 GHz	7 GB	86%	94%	1,020 ...	8 kB/s	11:56 PM
ip-10-248-69-42.us-west	USWest2a	2 GHz	7 GB	100%	96%	3 kB/s	2 kB/s	11:56 PM
ip-10-252-81-4.us-west-	USWest2a	2 GHz	15 GB	0%	22%	198 B/s	216 B/s	11:56 PM
ip-10-252-76-184.us-we	USWest2a	2 GHz	15 GB	0%	34%	18 kB/s	2 kB/s	11:56 PM

My Requests ▲

amazon web services™

CycleServer | Chef Hosts | CycleServer | Hosts in all clu

184.73.45.242:8080/chef/hosts

Google | Fun | News | Chef | Wikipedia | Note in Reader | Amazon Cloud Drive | TuneIn | Other Bookmarks

Logged in as admin | Admin | Log Out

CycleServer — Clusters | Jobs | Queues | Hosts | Reports | Chef

Show: Hosts that have converged over the last 12 hours

Save | Show Legend | Refresh | Mark Persistent | Mark Transient | Search:

Host Name	Persistent	Status	Total Converges	Last Completed Converge	Longest Cor
ec2-50-17-96-156.compute-1.am	no		8	1:51 AM	3:39.044
ec2-50-16-133-111.compute-1.a	no		13	1:51 AM	56:02.671
ec2-50-16-9-198.compute-1.ama	no		11	1:51 AM	4:58.855
ec2-184-73-98-167.compute-1.a	no		11	1:51 AM	3:44.581
ec2-50-16-109-110.compute-1.a	no		9	1:51 AM	4:41.413
ec2-107-20-8-36.compute-1.ama	no		11	1:50 AM	3:36.304
ec2-23-20-52-196.compute-1.am	no		19	1:50 AM	3:23.085
ec2-184-73-111-171.compute-1.	no		9	1:50 AM	3:27.248
ec2-50-16-66-42.compute-1.ama	no		10	1:50 AM	4:13.445
ec2-50-16-80-67.compute-1.ama	no		11	1:50 AM	3:24.926
ec2-50-19-168-196.compute-1.a	no		8	1:50 AM	4:29.131
ec2-184-73-79-153.compute-1.a	no		8	1:50 AM	3:30.288
ec2-50-17-5-157.compute-1.ama	no		11	1:50 AM	3:34.119
ec2-75-101-184-84.compute-1.a	no		22	1:50 AM	35:12.254
ec2-50-17-173-2.compute-1.ama	no		10	1:50 AM	3:24.671
ec2-107-22-159-135.compute-1.	no		11	1:50 AM	3:28.050
ec2-107-20-103-77.compute-1.a	no		14	1:50 AM	32:01.404
ec2-174-129-114-157.compute-1	no		14	1:50 AM	21:29.432
ec2-184-72-211-137.compute-1.	no		11	1:50 AM	3:20.657
ec2-50-19-57-29.compute-1.ama	no		11	1:50 AM	3:27.477
ec2-107-20-44-231.compute-1.a	no		8	1:49 AM	4:11.138
ec2-107-20-76-143.compute-1.a	no		8	1:49 AM	3:13.357
ec2-184-73-44-173.compute-1.a	no		14	1:49 AM	14:04.485
ec2-204-236-246-202.compute-1	no		13	1:49 AM	1:00:11.055
ec2-107-21-81-168.compute-1.a	no		8	1:49 AM	4:17.398
ec2-107-22-83-239.compute-1.a	no		11	1:49 AM	3:24.760
ec2-23-20-99-118.compute-1.am	no		11	1:49 AM	3:28.829
ec2-50-19-188-13.compute-1.am	no		14	1:49 AM	45:24.865
ec2-174-129-51-28.compute-1.a	no		11	1:49 AM	3:28.779
ec2-23-20-69-53.compute-1.ama	no		18	1:49 AM	3:35.584
ec2-184-72-148-166.compute-1.	no		19	1:49 AM	3:28.499
ec2-50-17-52-37.compute-1.ama	no		8	1:49 AM	4:25.534
ec2-107-22-133-51.compute-1.a	no		10	1:49 AM	3:20.627
ec2-107-22-152-186.compute-1.	no		8	1:49 AM	3:30.764
ec2-23-20-74-53.compute-1.ama	no		8	1:49 AM	3:31.580
ec2-174-129-119-140.compute-1	no		13	1:49 AM	38:42.571
ec2-184-73-26-196.compute-1.a	no		10	1:49 AM	3:15.726
ec2-184-73-77-52.compute-1.am	no		12	1:48 AM	3:27.873
ec2-50-17-165-34.compute-1.am	no		8	1:48 AM	3:56.134
ec2-23-20-72-99.compute-1.ama	no		14	1:48 AM	3:37.595
ec2-107-22-73-144.compute-1.a	no		11	1:48 AM	3:28.744
ec2-50-19-8-13.compute-1.amaz	no		7	1:48 AM	4:14.421
ec2-23-20-216-197.compute-1.a	no		9	1:48 AM	3:28.298
ec2-107-21-150-245.compute-1.	no		14	1:48 AM	43:15.090
ec2-107-20-59-96.compute-1.am	no		10	1:48 AM	3:22.870
ec2-50-16-160-163.compute-1.a	no		11	1:48 AM	3:24.694

Viewing top 2000 of 7394 sets of hosts. View All

Completed Converges every 10 minutes over the last 12 hours

My Requests ▲

amazon web services

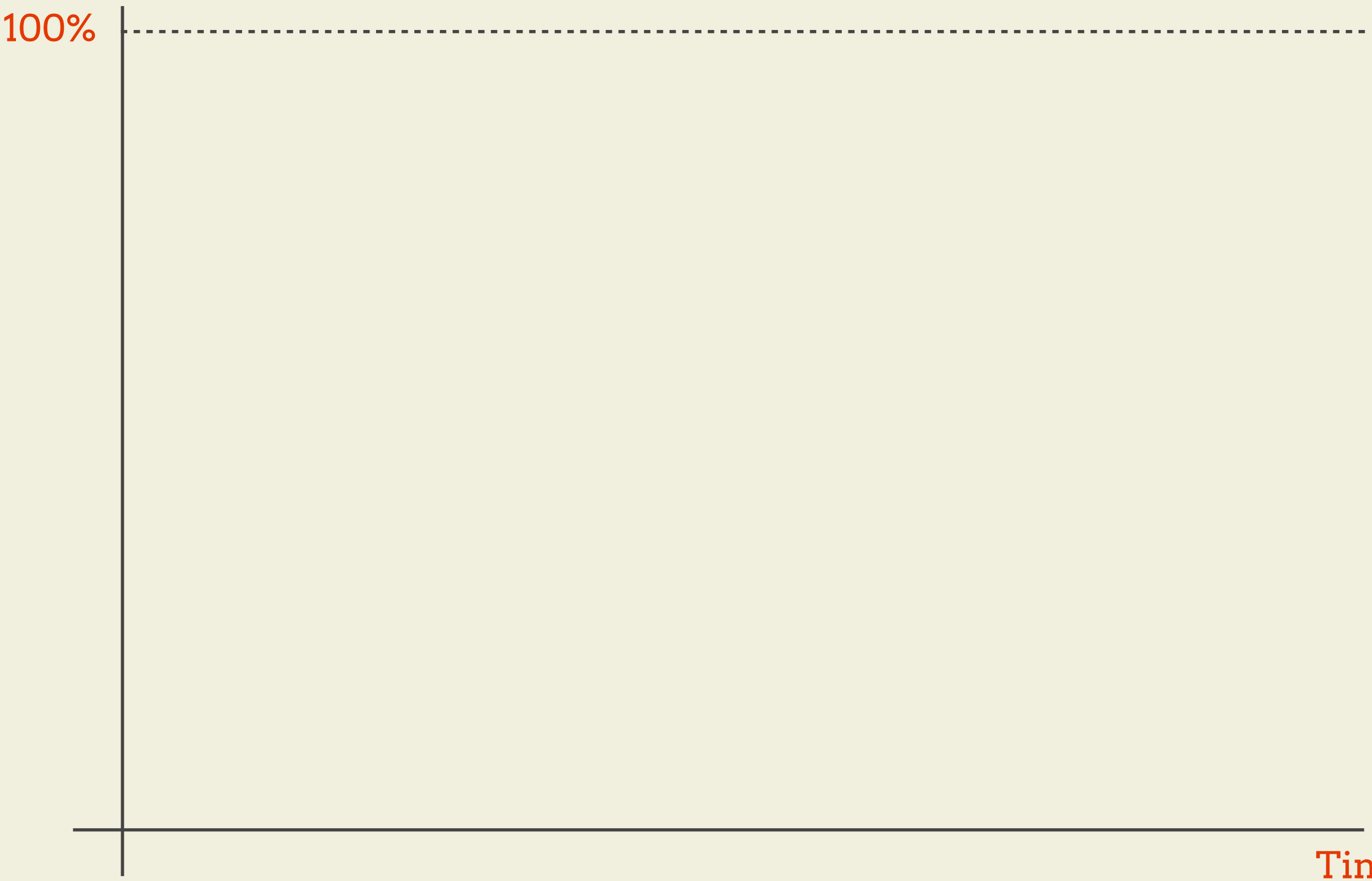
Cost optimization.

EC2

Utility computing.
6 years young.

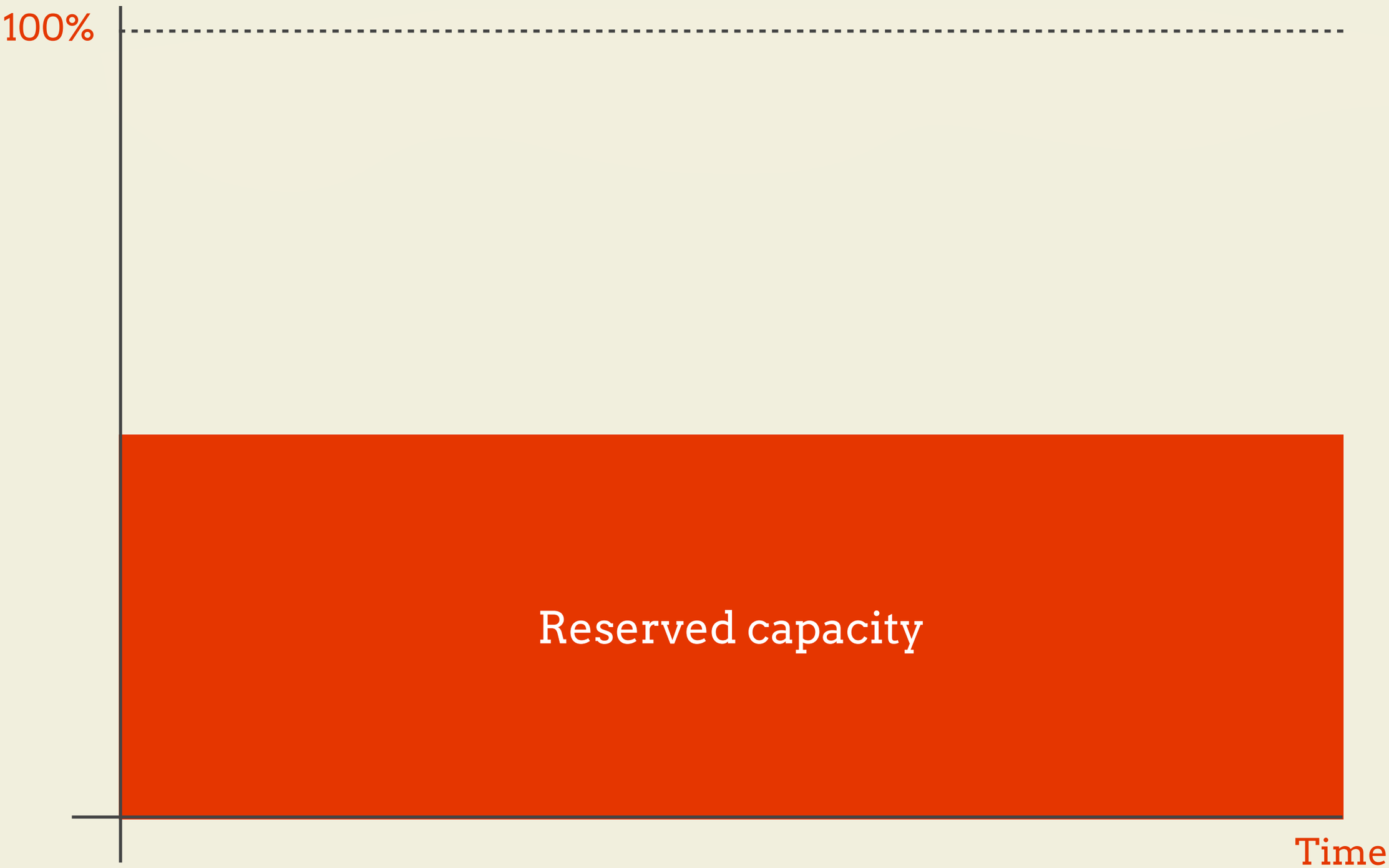
amazon web services™

Achieving economies of scale



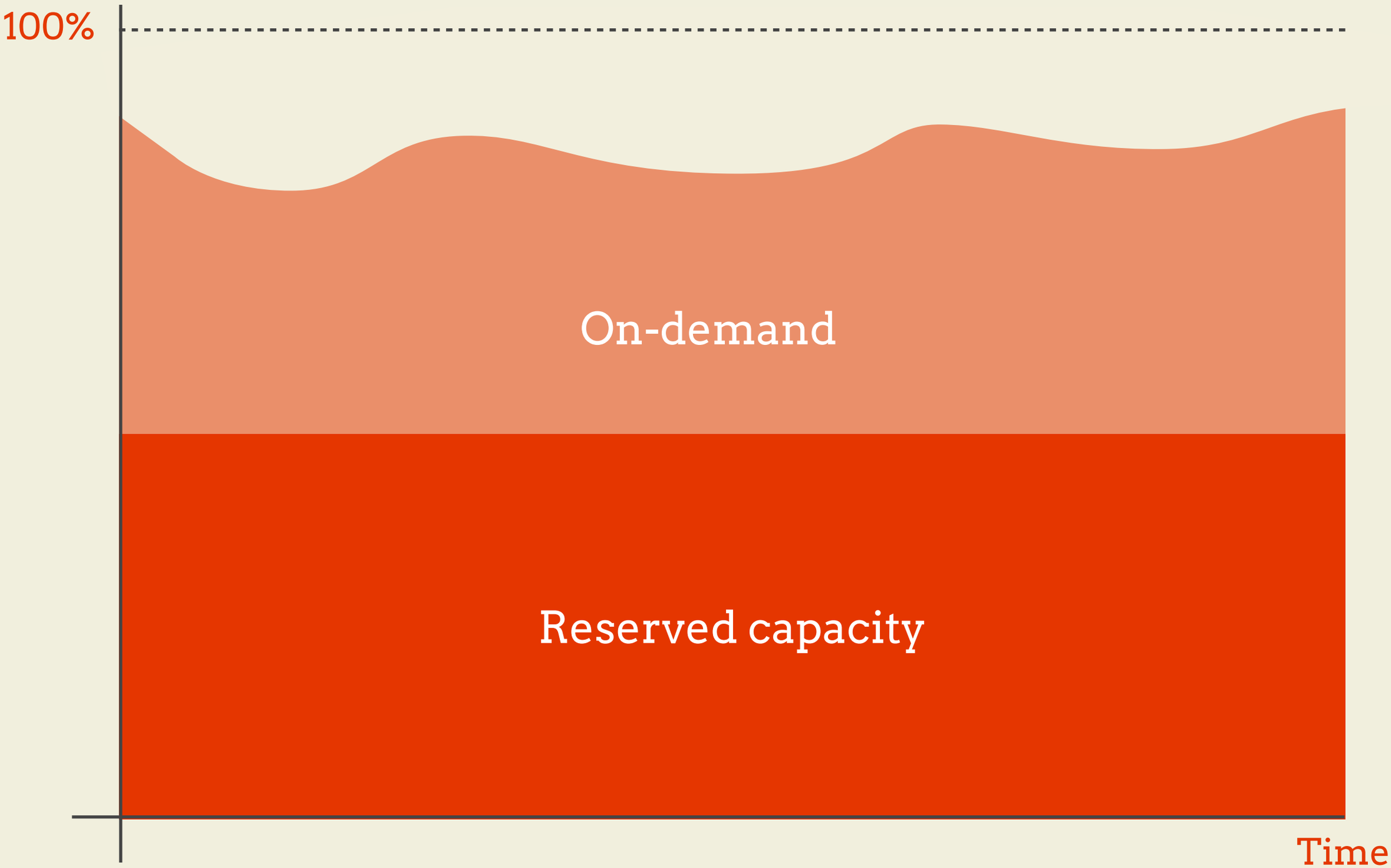
amazon web services™

Achieving economies of scale



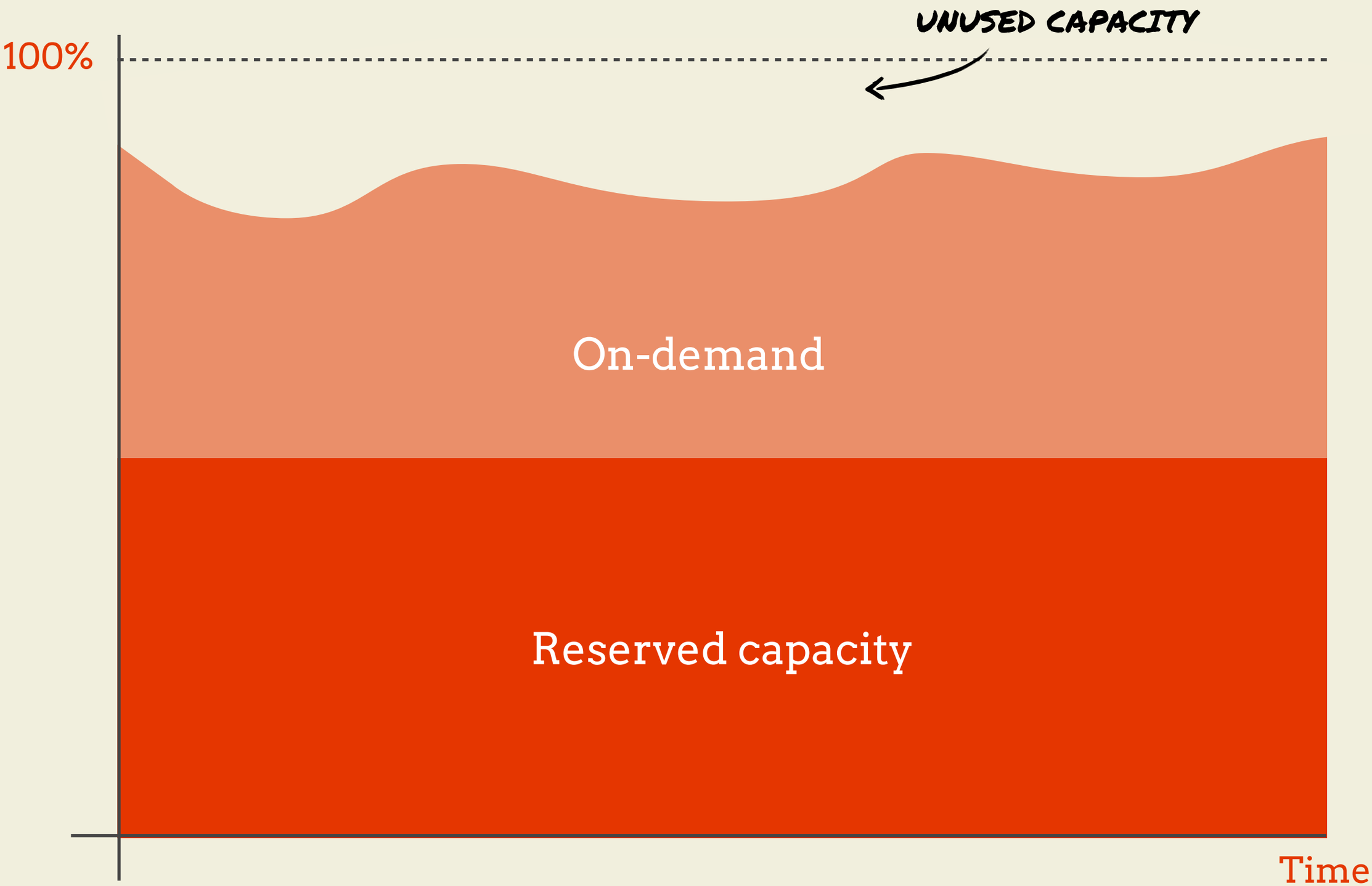
amazon web services

Achieving economies of scale



amazon web services™

Achieving economies of scale



amazon web services™

Spot Instances

Bid on unused EC2 capacity.

Very large discount.

Perfect for batch runs.

Balance cost and scale.

$650 per hour

Map/reduce

Pattern for distributed computing.

Software frameworks such as Hadoop.

Write two functions. Scale up.

Map/reduce

Pattern for distributed computing.

Software frameworks such as Hadoop.

Write two functions. Scale up.

Complex cluster configuration and management.

amazon web services

Amazon Elastic MapReduce

Managed Hadoop clusters.

Easy to provision and monitor.

Write two functions. Scale up.

Optimized for S3 access.

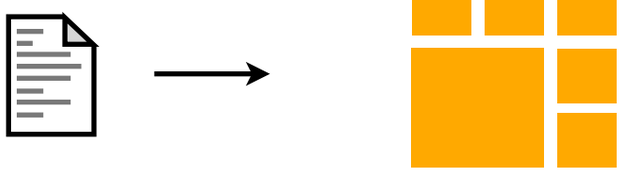
amazon web services


S3
Input data


UNDER
THE
HOOD

S3

Input data



Code

Elastic
MapReduce

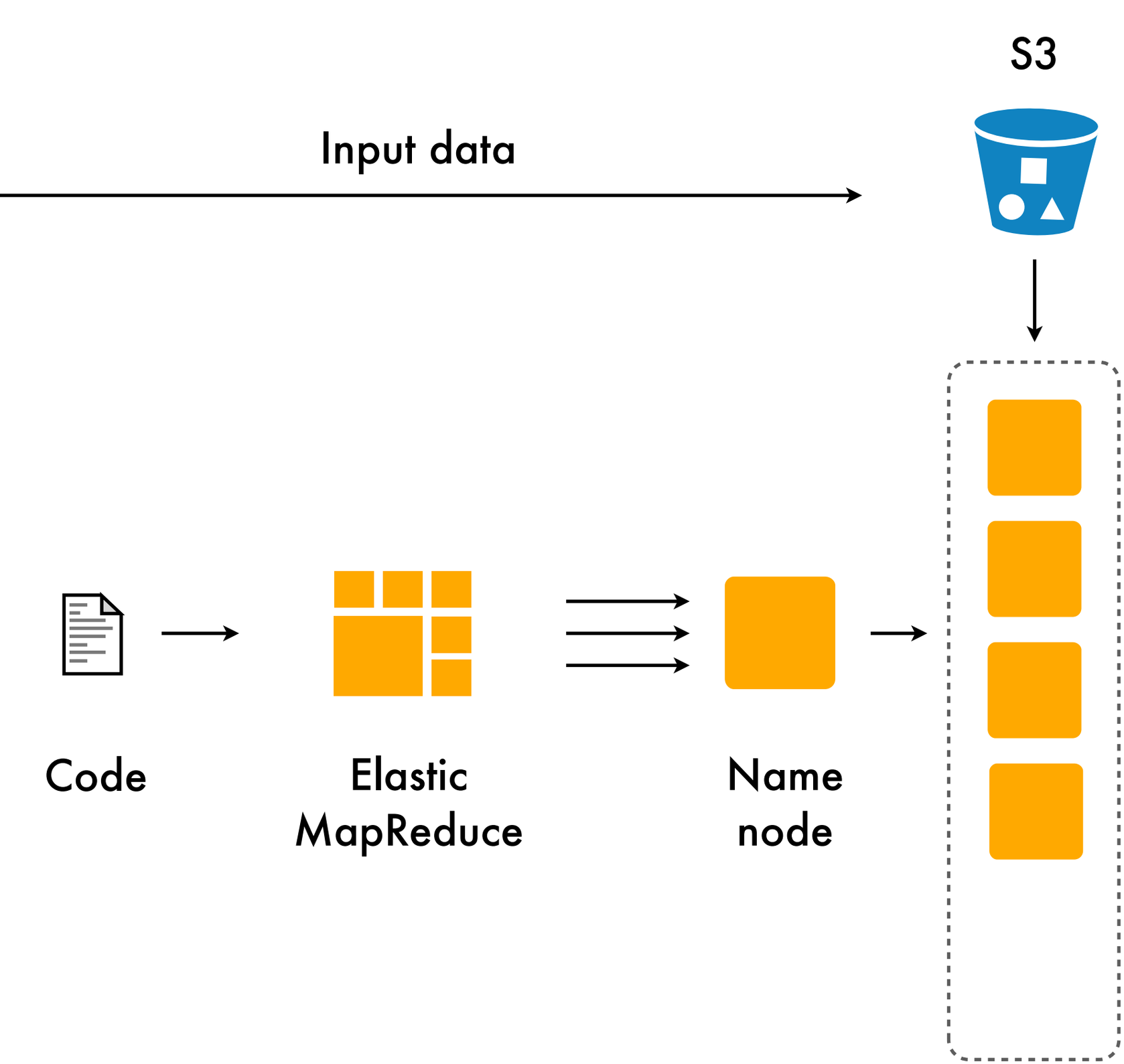
S3

Input data

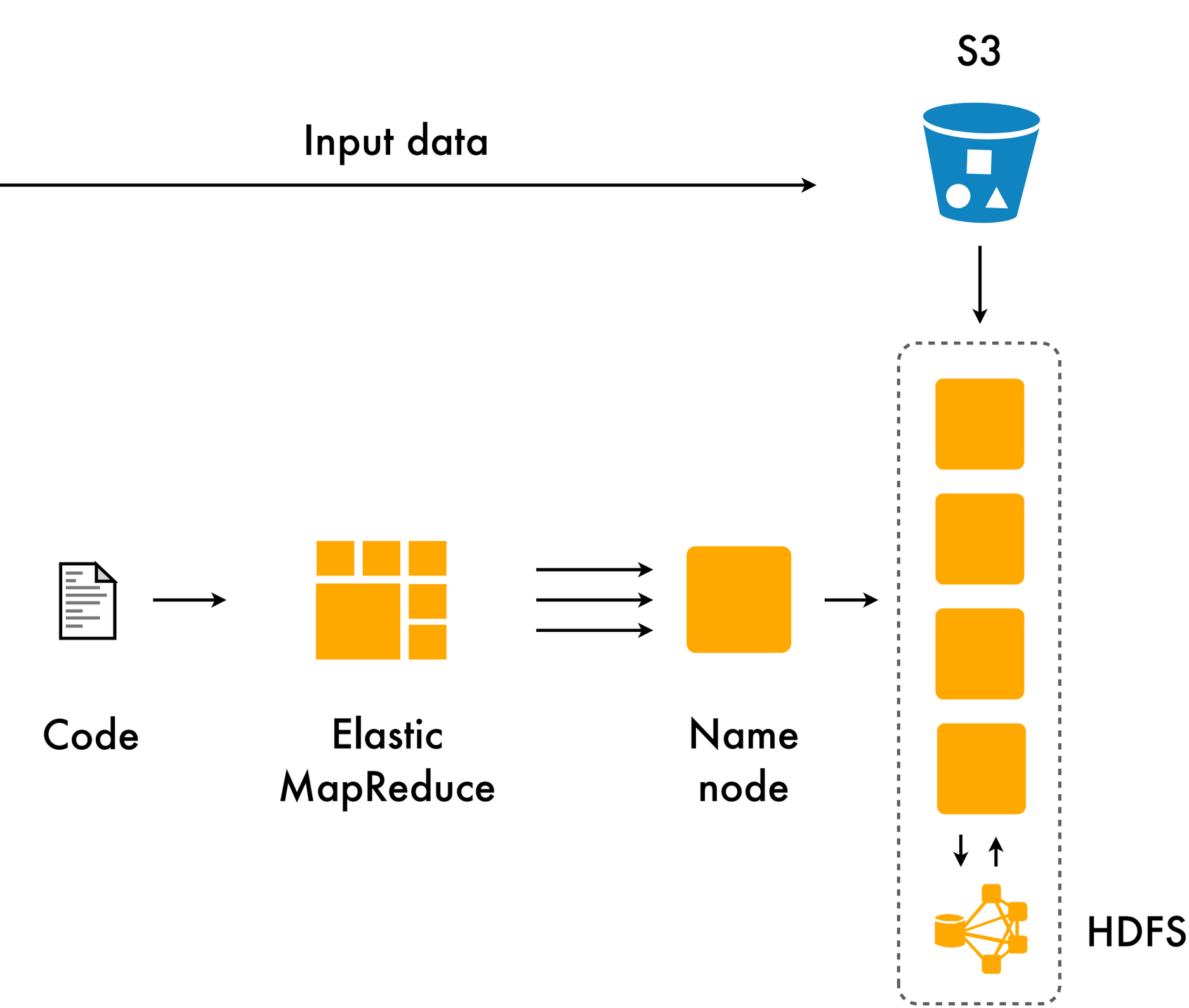
Code

Elastic MapReduce

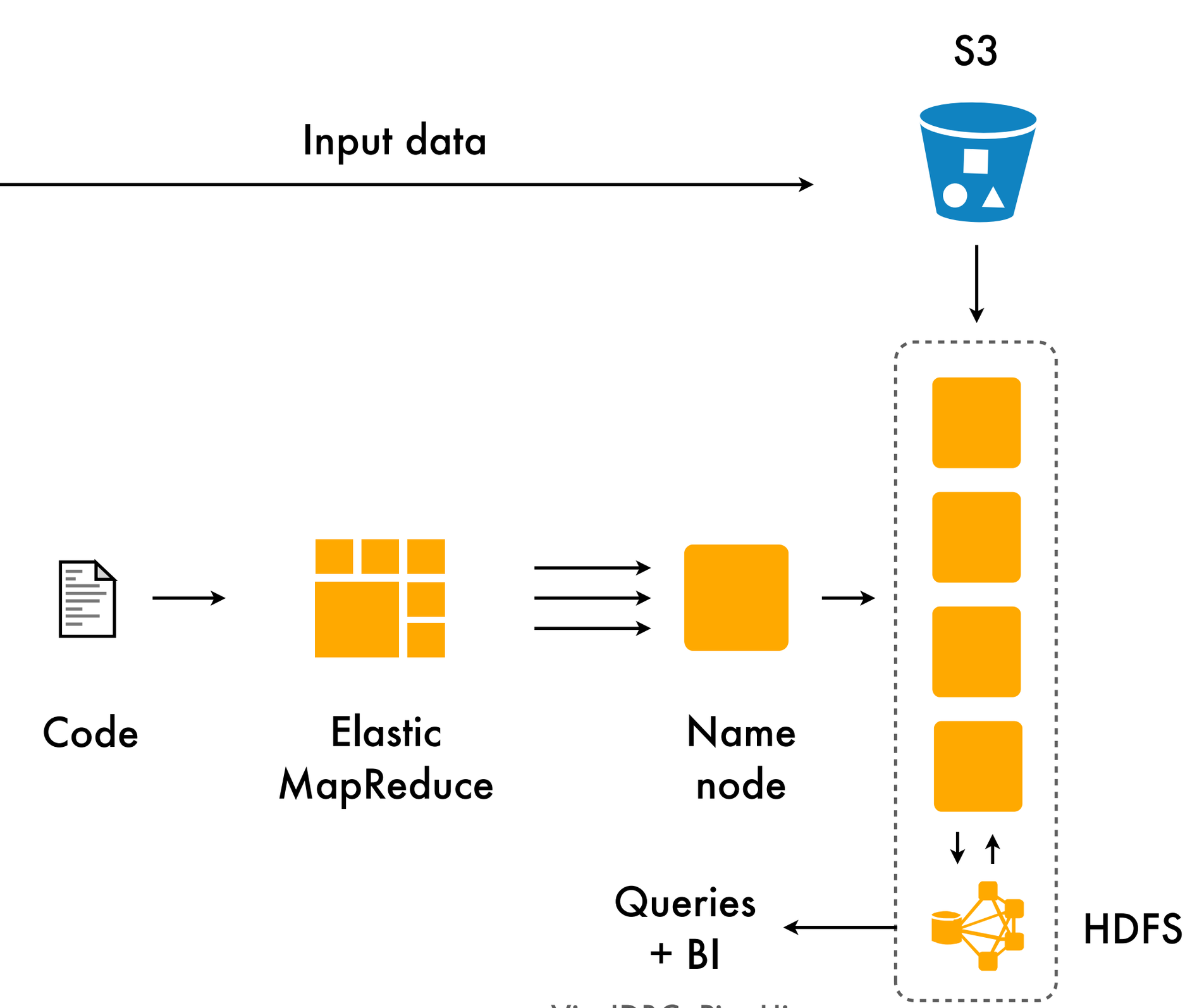
Name node


S3
Input data
Code
Elastic
MapReduce
Name
node
Elastic
cluster


UNDER
THE
HOOD


S3
Input data
Code
Elastic
MapReduce
Name
node
HDFS
Elastic
cluster


UNDER
THE
HOOD


S3
Input data
Code
Elastic
MapReduce
Name
node
Queries
+ BI
Via JDBC, Pig, Hive
HDFS
Elastic
cluster


UNDER
THE
HOOD

S3

Input data

Code

Elastic MapReduce

Name node

Output

S3 + SimpleDB

Queries + BI

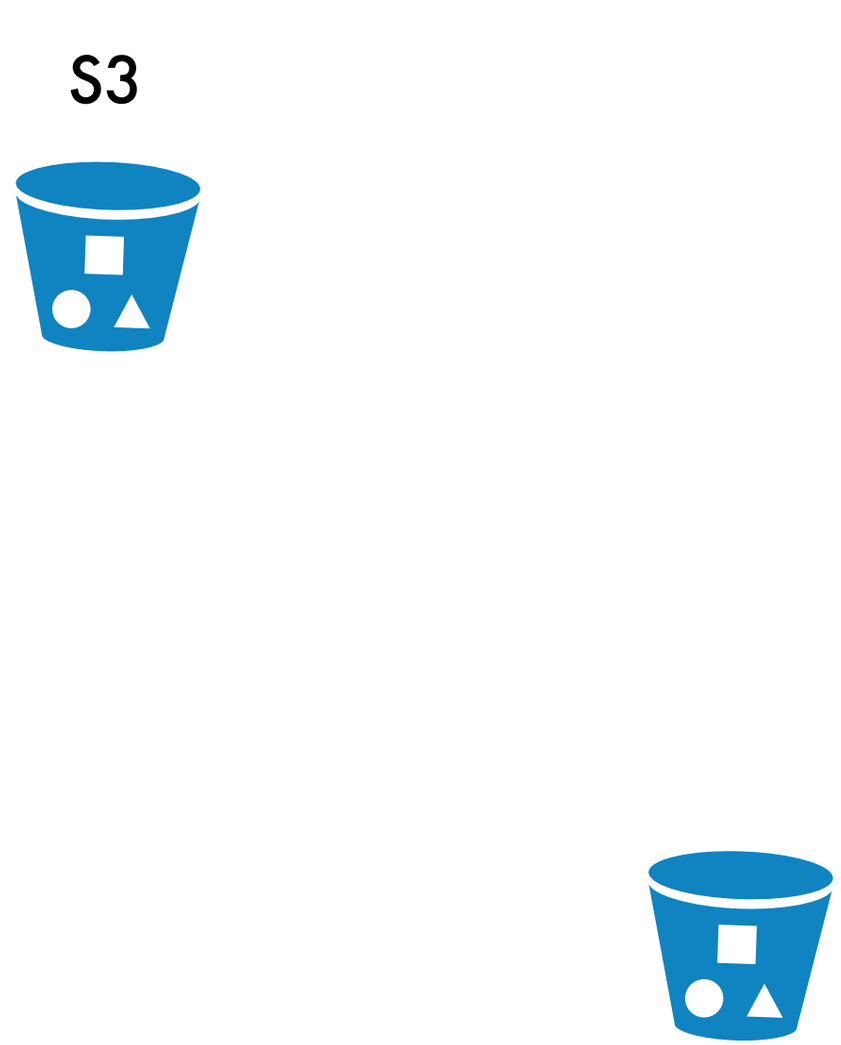
Via JDBC, Pig, Hive

HDFS

Elastic cluster

S3

Input data



Output

S3 + SimpleDB

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

No Job Flows

Create a Job Flow

You have not yet created any job flows. Click the button below to create your first job flow.
Sample processing applications are available to help get you started.

Create New Job Flow

How do I create one?

1. **Upload data to Amazon S3 Bucket**

Upload your application and any data you wish to process to an Amazon S3 bucket, using client tools available here.

2. **Create a job flow on Amazon Elastic MapReduce**

To create your job flow, simply specify data inputs, outputs, the processing application, and number of EC2 instances.

3. **Get results from Amazon S3 Bucket**

Monitor the status of your job flow and, when complete, pick up your results from the Amazon S3 bucket.

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=NewJobFlowWizard

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

Create a New Job Flow

Cancel

DEFINE JOB FLOW — SPECIFY PARAMETERS — CONFIGURE EC2 INSTANCES — BOOTSTRAP ACTIONS — REVIEW

Creating a job flow to process your data using Amazon Elastic MapReduce is simple and quick. Let's begin by giving your job flow a name and selecting its type. If you don't already have an application you'd like to run on Amazon Elastic MapReduce, samples are available to help you get started.

Job Flow Name*: My Job Flow

Job Flow Name doesn't need to be unique. We suggest you give it a descriptive name.

Create a Job Flow*:

- ○ Run your own application
- ◉ Run a sample application

Word Count (Streaming)

Word count is a Python application that counts occurrences of each word in provided documents.
Learn More

Continue

* Required field

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=NewJobFlowWizard

azon.com | AWS | Products | Developers | Community | Support | Account | **Welcome, Matt Wood** | Setting

Create a New Job Flow

Cancel

DEFINE JOB FLOW — SPECIFY PARAMETERS — CONFIGURE EC2 INSTANCES — BOOTSTRAP ACTIONS — REVIEW

Specify Mapper and Reducer functions to run within the Job Flow. The mapper and reducers may be either (i) class names referring to a mapper or reducer class in Hadoop or (ii) locations in Amazon S3. (Click Here for a list of available tools to help you upload and download files from Amazon S3.) The format for specifying a location in Amazon S3 is bucket_name/path_name. The location should point to an executable program, for example a python program. Extra arguments are passed to the Hadoop streaming program and can specify things such as additional files to be loaded into the distributed cache.

Input Location*: eu-west-1.elasticmapreduce/samples/wordcount/input
The URL of the Amazon S3 Bucket that contains the input files.

Output Location*: mza/wordcount/output/2011-05-12
The URL of the Amazon S3 Bucket to store output files. Should be unique.

Mapper*: eu-west-1.elasticmapreduce/samples/wordcount/wordSpli
The mapper Amazon S3 location or streaming command to execute.

Reducer*: aggregate
The reducer Amazon S3 location or streaming command to execute.

Extra Args:

< Back

Continue

* Required field

Create a New Job Flow

Cancel

DEFINE JOB FLOW ✓ — SPECIFY PARAMETERS ✓ — CONFIGURE EC2 INSTANCES — BOOTSTRAP ACTIONS — REVIEW

Enter the number and type of EC2 instances you'd like to run your job flow on.

Number of Instances*: 10

If you wish to run more than 20 instances, please complete the limit request form.

Type of Instance*: Small (m1.small) Learn more about instance types.

Amazon EC2 Key Pair: aws

Use an existing Key Pair to SSH into the master node of the Amazon EC2 cluster as the user "hadoop".

Configure your debugging options. Learn more.

Enable Debugging: ○ Yes ◉ No

Amazon S3 Log Path:

An Amazon S3 Log Path is required if you are enabling debugging.

Enable Hadoop Debugging: ○ Yes ◉ No

To enable Hadoop Debugging you will need to sign up for Amazon SimpleDB.

‹ Back

Continue

* Required field

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=NewJobFlowWizard

Create a New Job Flow

Cancel

DEFINE JOB FLOW ✓ — SPECIFY PARAMETERS ✓ — CONFIGURE EC2 INSTANCES ✓ — BOOTSTRAP ACTIONS — REVIEW

Proceed with no Bootstrap Actions

Configure your Bootstrap Actions

Use the table below to define the name, location and optional arguments for any Bootstrap Actions you want associated with this Job Flow.

Bootstrap Action

Action Type

Configure Hadoop Learn More

Name

Configure Hadoop

Amazon S3 Location

s3n://eu-west-1.elasticmapreduce/bootstrap-actions/configure-hadoop

Optional Arguments

--site-key-value io.file.buffer.size=65536

Add another Bootstrap Action

< Back

Continue

* Required field

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=NewJobFlowWizard

nazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings

Create a New Job Flow

Cancel

DEFINE JOB FLOW ✓ | SPECIFY PARAMETERS ✓ | CONFIGURE EC2 INSTANCES ✓ | BOOTSTRAP ACTIONS ✓ | REVIEW

Please review the details of your job flow and click "Create Job Flow" when you are ready to launch your Hadoop Cluster.

Job Flow Name:	My Job Flow
Type:	Word Count (Streaming)

Edit Job Flow Definition

Input Location:	s3n://eu-west-1.elasticmapreduce/samples/wordcount/input
Output Location:	s3n://mza/wordcount/output/2011-05-12
Mapper:	s3n://eu-west-1.elasticmapreduce/samples/wordcount/wordSplitter.py
Reducer:	aggregate
Extra Args:	

Edit Job Flow Parameters

Number of Instances:	10
Type of Instance:	m1.small
Amazon EC2 Key Pair:	aws
Amazon S3 Log Path:	
Enable Hadoop Debugging:	No

Edit EC2 Configs and Advanced Options

Bootstrap Actions:	1 Bootstrap Actions created for this Job Flow

Edit Bootstrap Actions

‹ Back

Create Job Flow

Note: Once you click "Create Job Flow," instances will be launched and you will be charged accordingly.

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

1 to 1 of 1 Job Flows

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
My Job Flow	STARTING	2011-05-12 13:40 GMT+0	0 hours 0 minutes	0

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: -

Description | Steps | Bootstrap Actions | Instance Groups

Name:	My Job Flow	**Creation Date:**	2011-05-12 13:40 GMT+0100
Start Date:	-	**End Date:**	-
Availability Zone:	-	**Instance Count:**	10
Master Instance Type:	m1.small	**Slave Instance Type:**	m1.small
Key Name:	aws	**Log URI:**	-
Master Public DNS Name:	-	**Hadoop Version:**	0.20

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | **Welcome, Matt Wood** | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

1 to 1 of 1 Job Flows

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
My Job Flow	STARTING	2011-05-12 13:40 GMT+0	0 hours 0 minutes	0

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: -

Description | **Steps** | Bootstrap Actions | Instance Groups

Step Name	State	Start Date	End Date	Jar	Main Class	Args
Streaming Job	PENDING	-	-	/home/hadoop/contrib/streaming/hadoop-streaming.jar	-	-input s3n://eu-west-1.elasticmapreduce/samples/wordcount/input -output s3n://mza/wordcount/output/2011-05-12 -mapper s3n://eu-west-1.elasticmapreduce/samples/wordcount/wordSplitter.py -reducer aggregate

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

1 to 1 of 1 Job Flows

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
My Job Flow	STARTING	2011-05-12 13:40 GMT+0	0 hours 0 minutes	0

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: -

Description | Steps | **Bootstrap Actions** | Instance Groups

Action Name	Path	Args
Configure Hadoop	s3n://eu-west-1.elasticmapreduce/bootstrap-actions/configure-hadoop	--site-key-value io.file.buffer.size=65536

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All | 1 to 1 of 1 Job Flows

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
My Job Flow	STARTING	2011-05-12 13:40 GMT+0	0 hours 0 minutes	0

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: -

Description | Steps | Bootstrap Actions | Instance Groups

Instance Group Id	Role	Instance Type	State	Running Count	Request Count	Creation DateTime	Last State Change Reason
ig-24XVTCKNOBXNY	MASTER	m1.small	PROVISIONING	0	1	2011-05-12 13:40 GMT+0100	-
ig-X0QNDMX7MACL	CORE	m1.small	PROVISIONING	0	9	2011-05-12 13:40 GMT+0100	-

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

1 to 1 of 1 Job Flows

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
☑	My Job Flow	COMPLETED	2011-05-12 13:40 GMT+0	0 hours 4 minutes	10

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: Steps completed

Description | Steps | Bootstrap Actions | Instance Groups

Name:	My Job Flow	**Creation Date:**	2011-05-12 13:40 GMT+0100
Start Date:	2011-05-12 13:44 GMT+0100	**End Date:**	2011-05-12 13:47 GMT+0100
Availability Zone:	eu-west-1a	**Instance Count:**	10
Master Instance Type:	m1.small	**Slave Instance Type:**	m1.small
Key Name:	aws	**Log URI:**	-
Master Public DNS Name: ec2-46-137-22-212.eu-west-1.compute.amazonaws.com		**Hadoop Version:**	0.20

AWS Management Console

https://console.aws.amazon.com/elasticmapreduce/home?region=eu-west-1#s=Home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Settings | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Your Elastic MapReduce Job Flows

Region: EU West | Create New Job Flow | Terminate | Debug | Show/Hide | Refresh | Help

Viewing: All

1 to 1 of 1 Job Flows

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
☑	My Job Flow	COMPLETED	2011-05-12 13:40 GMT+0	0 hours 4 minutes	10

1 Job Flow selected

Job Flow: j-3KCBEUQQ4GS5H

Last State Change Reason: Steps completed

Description | Steps | Bootstrap Actions | **Instance Groups**

Instance Group Id	Role	Instance Type	State	Running Count	Request Count	Creation DateTime	Last State Change Reason
ig-24XVTCKNOBXNY	MASTER	m1.small	ENDED	0	1	2011-05-12 13:40 GMT+0100	Job flow terminated
ig-X0QNDMX7MACL	CORE	m1.small	ENDED	0	9	2011-05-12 13:40 GMT+0100	Job flow terminated

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

AWS Management Console

https://console.aws.amazon.com/s3/home

aws.amazon.com | AWS | Products | Developers | Community | Support | Account | Welcome, Matt Wood | Sign Out

AWS Elastic Beanstalk | Amazon S3 | Amazon EC2 | Amazon VPC | Amazon CloudWatch | Amazon Elastic MapReduce | Amazon CloudFront | AWS CloudFormation | Amazon RDS | Amazon SNS | AWS IAM

Buckets

Create Bucket | Actions

- aws-slides
- aws-wp-assets
- cf-templates-xthrnrzfznsj-ap-n
- cf-templates-xthrnrzfznsj-eu-w
- cf-templates-xthrnrzfznsj-us-e
- cf-templates-xthrnrzfznsj-us-w
- elasticbeanstalk-us-east-1-545
- hpc-video
- my-website
- **mza**

Objects and Folders

Upload | Create Folder | Actions | Refresh | Properties | Transfers | Help

mza > wordcount > output > 2011-05-12

Name	Size	Last Modified
part-00000	19 KB	Thu May 12 12:46:24 GMT+100 2011
part-00001	20.8 KB	Thu May 12 12:46:23 GMT+100 2011
part-00002	19.8 KB	Thu May 12 12:46:27 GMT+100 2011
part-00003	19.7 KB	Thu May 12 12:46:24 GMT+100 2011
part-00004	20.1 KB	Thu May 12 12:46:23 GMT+100 2011
part-00005	18.8 KB	Thu May 12 12:46:23 GMT+100 2011
part-00006	19.4 KB	Thu May 12 12:46:27 GMT+100 2011
part-00007	18.9 KB	Thu May 12 12:46:25 GMT+100 2011
part-00008	19.1 KB	Thu May 12 12:46:24 GMT+100 2011
part-00009	19.2 KB	Thu May 12 12:46:32 GMT+100 2011
part-00010	19 KB	Thu May 12 12:46:32 GMT+100 2011
part-00011	19.3 KB	Thu May 12 12:46:33 GMT+100 2011
part-00012	19.8 KB	Thu May 12 12:46:34 GMT+100 2011
part-00013	19.6 KB	Thu May 12 12:46:33 GMT+100 2011
part-00014	19.9 KB	Thu May 12 12:46:33 GMT+100 2011

part-00000

```
abbreviated 27
abdellah    9
abderrazzak 3
abdication  6
abducted    3
abena   3
abu 83
acceded 24
accession   73
accorded    6
accumulate  3
acevedo 3
act 144
actual  24
adamkus 9
adams   6
add 19
additionally    21
address 668
adil    12
adiyaman    3
admonition  12
adolf   3
adrift  21
adulyadej   3
advantages  3
advisory    55
aer 3
aerosol 1
afar    24
afrikaans   6
```

Line: 1 Column: 1 Plain Text Tab Size: 4

Performance

Performance

Compute performance

Cluster Compute

Intel Xeon E5-2670

10 gig E non-blocking network

60.5 Gb

Placement groupings

amazon web services

Cluster Compute

Intel Xeon E5-2670

10 gig E non-blocking network

60.5 Gb

Placement groupings

+ GPU enabled instances

amazon web services

Performance

Compute performance

IO performance

Performance

Compute performance

NoSQL

Unstructured data storage.

amazon web services™

DynamoDB

Predictable, consistent performance

Unlimited storage

Single digit millisecond latencies

No schema for unstructured data

Backed on solid state drives

...and SSDs for all.

New Hi1 storage instances.

amazon web services

hi1.4xlarge

2 x 1Tb SSDs

10 GigE network

HVM: 90k IOPS read, 9k to 75k write

PV: 120k IOPS read, 10k to 85k write

"The hi1.4xlarge configuration is about **half the system cost** for the **same throughput**."

Netflix

http://techblog.netflix.com/2012/07/benchmarking-high-performance-io-with.html

EBS

Elastic Block Store

amazon web services

Provisioned IOPS

Provision required IO performance

amazon web services

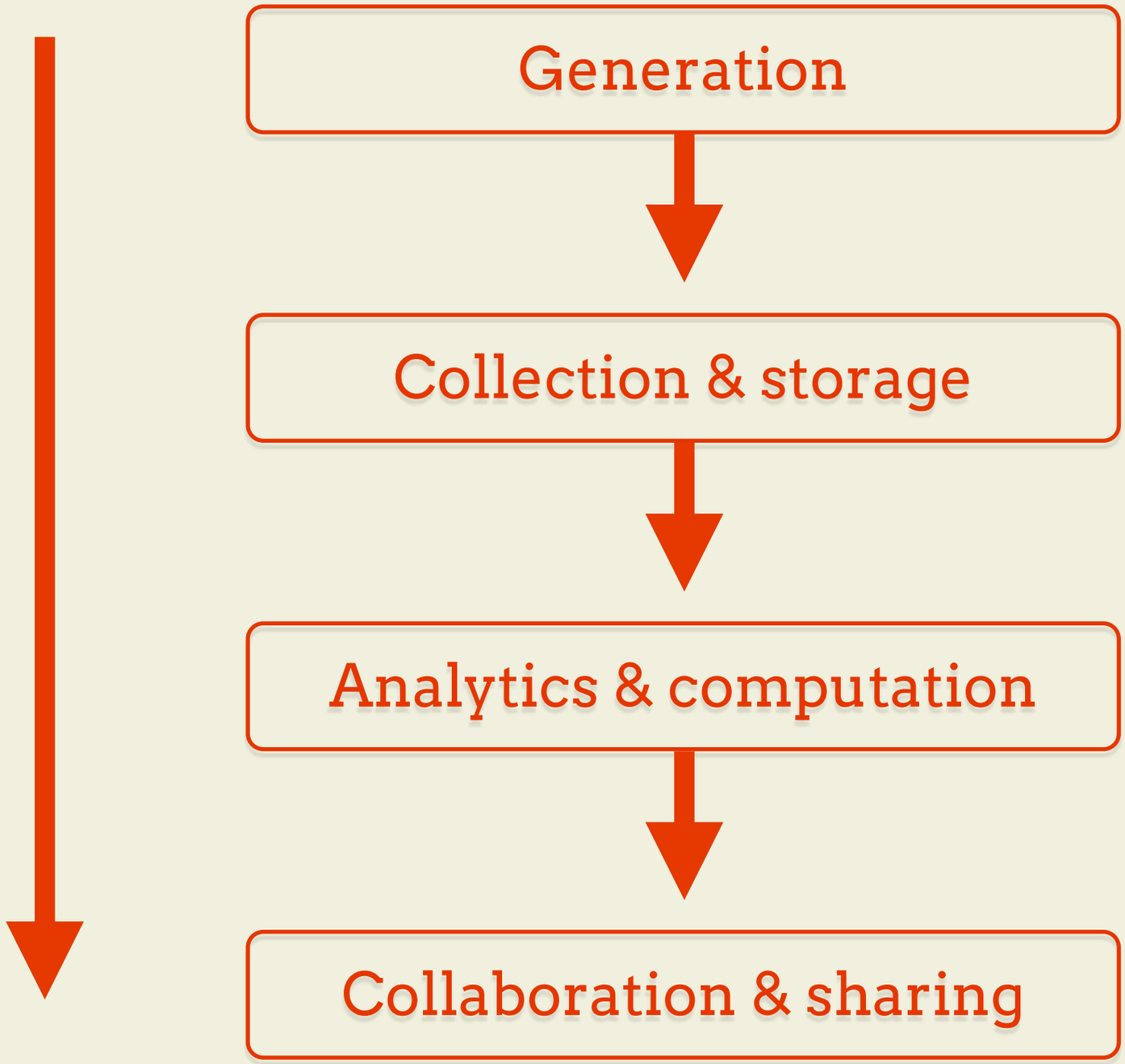
Provisioned IOPS

Provision required IO performance

+

EBS-optimized instances
with dedicated throughput

amazon web services



Performance + ease of use

3

PARTNER INTEGRATION

Extend platform with partners

Innovate on behalf of customers

Remove undifferentiated heavy lifting

amazon web services

MapR distribution for EMR

Rolled the Amazon Hadoop optimizations into MapR

Choice for EMR customers

Easy deployment for MapR customers

amazon web services

MapR distribution for EMR

Hadoop distribution

Integrated into EMR

NFS and ODBC drivers

High availability and cluster mirroring

Informatica on EMR

Enterprise data toolchain

"Swiss army knife" for data formats

Data integration

Available to all on EMR

amazon web services

AWS Marketplace

Karmasphere, Marketshare, Acunu Cassandra, Metamarkets, Aspera and more.

aws.amazon.com/marketplace

4

PARTNER SUCCESS STORIES

Razorfish

3.5 billion records
71MM unique cookies
1.7MM targeted ads
per day

3.5 billion records
71MM unique cookies
1.7MM targeted ads
per day

500% improvement in return on ad spend.

amazon web services™

Cycle Computing

+ Schrodinger

amazon web services

30k cores, $4200 an hour

(compared to $10+ million)

amazon web services™

Marketshare

+ Ticketmaster

Optimize live event pricing

amazon web services

Reduced developer infrastructure management time by **3 hours a day**

amazon web services™

Thank you!

amazon web services

Q & A

matthew@amazon.com

@mza on Twitter