Amazon Elastic MapReduce

Elastic MapReduce

- **Amazon Elastic MapReduce Overview**
- FAQs
- Pricing

Developer Resources

- AWS Management Console
- Documentation
- Release Notes
- Sample Code & Libraries
- Developer Tools
- Articles & Tutorials
- Community Forum

Featured Case Studies

Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Using Amazon Elastic MapReduce, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics research. Amazon Elastic MapReduce lets you focus on crunching or analyzing your data without having to worry about time-consuming set-up, management or tuning of Hadoop clusters or the compute capacity upon which they sit.

What's New

KARMASPHERE

Karmasphere Analyst is a visual, desktop workspace for analyzing data on Amazon Elastic MapReduce. It provides graphical tools to perform SQL-based querying of structured and unstructured data and visualize the results. Karmasphere Analyst is available with hourly pricing and no upfront fees or long-term commitments.

› Read more

Using Spot Instances

Watch this video to learn how to save money by using EC2 Spot Instances with Amazon Elastic MapReduce through the AWS Management Console.

Need Help?

Ask on the Elastic MapReduce forum

This page contains the following categories of information. Click to jump down:

Amazon Elastic MapReduce Functionality

Amazon Elastic MapReduce automatically spins up a Hadoop implementation of the MapReduce framework on Amazon EC2 instances, sub-dividing the data in a job flow into smaller chunks so that they can be processed (the " "f) ll l d ll b h d d h f l l (h " d "

To use Amazon Elastic MapReduce, you simply:

Develop your data processing application. Amazon Elastic MapReduce enables job flows to be developed in SQL-like languages, such as Hive and Pig, making it easy to write data analytical scripts without in-depth knowledge of the MapReduce development paradigm. If desired, more sophisticated applications can be authored in your choice of Cascading, Java, Ruby, Perl, Python, PHP, R, or C++. There are several code samples and tutorials available in the Getting Started Guide that will help you get up and running quickly.

Upload your data and your processing application into Amazon S3. Amazon S3 provides reliable, scalable, easy-to-use storage for your input and output data.

Log in to the AWS Management Console to start an Amazon Elastic MapReduce "job flow." Simply choose the number and type of Amazon EC2 instances you want, specify the location of your data and/or application on Amazon S3, and then click the "Create Job Flow" button. Alternatively you can start a job flow by specifying the same information mentioned above via our Command Line Tools or APIs. For more sophisticated workloads you can choose to install additional software or alter configuration of your Amazon EC2 instances using Bootstrap Actions.

Monitor the progress of your job flow(s) directly from the AWS Management Console, Command Line Tools or APIs. And, after the job flow is done, retrieve the output from Amazon S3. You can optionally track progress and identify issues in steps, jobs, tasks, or task attempts of your job flows directly from the job flow debug window in the AWS Management Console. Amazon Elastic MapReduce uses Amazon SimpleDB to store job flow state information.

Pay only for the resources that you actually consume. Amazon Elastic MapReduce monitors your job flow, and unless you specify otherwise, shuts down your Amazon EC2 instances after the job completes.

Service Highlights

Elastic — Amazon Elastic MapReduce enables you to use as many or as few compute instances running Hadoop as you want. You can commission one, hundreds, or even thousands of instances to process gigabytes, terabytes, or even petabytes of data. You can modify the number of instances while your job flow is running and you can run as many job flows concurrently as you wish. You can instantly spin up large Hadoop job flows which will start processing within minutes, not hours or days. When your job flow completes, unless you specify otherwise, the service automatically tears down your instances.

Easy to use — You don't need to worry about setting up, running, or tuning the performance of Hadoop clusters; instead, you can concentrate on data analysis. We provide easy-to-use tools and sample data processing applications that let you get up and running without writing a single line of code. Once you start a job flow, Amazon Elastic MapReduce handles Amazon EC2 instance provisioning, security settings, Hadoop configuration and set-up, log collection, health monitoring, and other hardware-related complexities such as automatically removing faulty instances from your running job flow.

Reliable — Amazon Elastic MapReduce is built on Amazon's highly reliable infrastructure, and has tuned Hadoop's performance specifically for Amazon's infrastructure environment. The service also monitors your job flow execution—retrying failed tasks, shutting down problematic instances, and provisioning new nodes to replace those that fail.

Seamlessly integrated with other AWS services — Amazon Elastic MapReduce is designed to integrate easily with other AWS services such as Amazon S3 and EC2, providing the infrastructure for data processing applications. The service runs job flows in Amazon EC2 and stores input and output data in Amazon S3.

Secure — Amazon Elastic MapReduce automatically configures Amazon EC2 firewall settings that control network access to and between instances that run your job flows.

Inexpensive — Amazon Elastic MapReduce passes on to you the financial benefits of Amazon's scale. You pay a very low rate for the compute capacity you actually consume. Amazon Elastic MapReduce is optimized to save you money by monitoring progress of your job flows and turning off resources when a job flow is completed.

On-Demand Instances — On-Demand Instances let you pay for compute capacity by the hour 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. On-Demand Instances also remove the need to buy "safety net" capacity to handle periodic traffic spikes.

Reserved Instances — Reserved Instances give you the option to make a low, one-time payment for each instance you want to reserve and in turn receive a significant discount on the hourly usage charge for that instance. After the one-time payment for an instance, that instance is reserved for you, and you have no further obligation; you may choose to run that instance for the discounted usage rate for the duration of your term, or when you do not use the instance, you will not pay usage charges on it.

Spot Instances — Spot Instances allow you to bid on unused Amazon EC2 capacity and run those instances for as long as your bid exceeds the current Spot Price. The Spot Price changes periodically based on supply and demand and customers whose bids meet or exceed it gain access to the available

Multiple Locations — Amazon Elastic MapReduce uses geographically dispersed EC2 infrastructure and is currently available in the US East (Northern Virginia), US West (Oregon), US West (Northern California), EU (Ireland), APAC (Singapore), and APAC (Tokyo) Regions.

Third Party Tools — Amazon Elastic MapReduce is supported by Karmasphere Studio, a plug-in for the Eclipse IDE that provides a familiar graphical environment for managing the complete Hadoop development lifecycle. Please visit the Elastic MapReduce with Karamasphere Analytics detail page to learn more.

↑ Top

Instance Types

To use Amazon Elastic MapReduce, you need to first select the type and quantity of Amazon EC2 instances you want. Amazon Elastic MapReduce works with any Amazon EC2 Linux/Unix instance type. It supports both On-Demand and Reserved instances; if you have Reserved Instances they will be used first by your job flows. Note that only the instance types listed below are currently supported by Amazon Elastic MapReduce.

Standard Amazon EC2 Instances

Instances of this family are well suited for most applications.

Small Instance (Default) 1.7 GB of memory, 1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit), 160 GB of instance storage, 32-bit platform

Large Instance 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of instance storage, 64-bit platform

Extra Large Instance 15 GB of memory, 8 EC2 Compute Units (4 virtual cores with 2 EC2 Compute Units each), 1690 GB of instance storage, 64-bit platform

High-Memory Amazon EC2 Instances

Instances of this family offer large memory sizes for high throughput applications, including database and memory caching applications.

High-Memory Extra Large Instance 17.1 GB memory, 6.5 ECU (2 virtual cores with 3.25 EC2 Compute Units each), 420 GB of local instance storage, 64-bit platform

High-Memory Double Extra Large Instance 34.2 GB of memory, 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each), 850 GB of instance storage, 64-bit platform

High-Memory Quadruple Extra Large Instance 68.4 GB of memory, 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each), 1690 GB of instance storage, 64-bit platform

High-CPU Amazon EC2 Instances

Instances of this family have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications.

High-CPU Medium Instance 1.7 GB of memory, 5 EC2 Compute Units (2 virtual cores with 2.5 EC2 Compute Units each), 350 GB of instance storage, 32-bit platform

High-CPU Extra Large Instance 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of instance storage, 64-bit platform

High Performance Computing Amazon EC2 Instances

Instances of this family combine large memory sizes and high CPU resources with 10 Gbps networking. They are well-suited for high performance, I/O intensive applications, such as mapping genomes for scientific research, simulating aerospace and automotive designs for engineering activities, and mining data for business intelligence.

Cluster Compute Quadruple Extra Large 23 GB memory, 33.5 EC2 Compute Units, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet

Cluster GPU Quadruple Extra Large 22 GB memory, 33.5 EC2 Compute Units, 2 x NVIDIA Tesla “Fermi” M2050 GPUs, 1690 GB of local instance storage, 64-bit platform, 10 Gigabit Ethernet

EC2 Compute Unit (ECU) – One EC2 Compute Unit (ECU) provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Pricing

Amazon Elastic MapReduce currently is available in the US, EU and APAC Regions. Pay only for what you use – there is no minimum fee. Amazon Elastic MapReduce pricing is in addition to normal Amazon EC2 and Amazon S3 pricing.

	Amazon EC2 Price	Amazon Elastic MapReduce Price
Standard On-Demand Instances		
Small (Default)	$0.085 per hour	$0.015 per hour
Large	$0.34 per hour	$0.06 per hour
Extra Large	$0.68 per hour	$0.12 per hour
Hi-Memory On-Demand Instances		
Extra Large	$0.50 per hour	$0.09 per hour
Double Extra Large	$1.00 per hour	$0.21 per hour
Quadruple Extra Large	$2.00 per hour	$0.42 per hour
Hi-CPU On-Demand Instances		
Medium	$0.17 per hour	$0.03 per hour
Extra Large	$0.68 per hour	$0.12 per hour
Cluster Compute On-Demand Instances		
Quadruple Extra Large	$1.30 per hour	$0.33 per hour
Cluster GPU On-Demand Instances		
Quadruple Extra Large	$2.10 per hour	$0.42 per hour

Amazon EC2, Amazon S3 and Amazon SimpleDB charges are billed separately. Pricing for Amazon Elastic MapReduce is per instance-hour consumed for each instance type, from the time job flow began processing until it is terminated. Each partial instance-hour consumed will be billed as a full hour. For additional details on Amazon EC2 Instance Types, Amazon EC2 Reserved Instances Pricing, Amazon S3 Pricing, or Amazon SimpleDB Pricing, follow the links below:

Amazon EC2 Instance Types

Amazon EC2 Reserved Instances Pricing

Amazon S3 Pricing

Amazon SimpleDB Pricing

↑ Top

Resources

Developer Resources

› view all

- AWS Management Console
- Sample Data-Processing Applications
- WSDL
- Release Notes
- Documentation
- Sample Code & Libraries
- Developer Tools
- Community Forum
- Articles & Tutorials

Additional Product Information

- FAQs
- Amazon Web Services Customer Agreement
- Service Health Dashboard

Related Services

- Amazon Elastic Compute Cloud
- Amazon S3
- AWS Import/Export

↑ Top

Detailed Description

Amazon Elastic MapReduce uses Apache Hadoop as its distributed processing engine. Hadoop is an open source Java software framework that supports data-intensive distributed applications running on large clusters of commodity hardware. Hadoop implements a computational model named “MapReduce,” in which the job is divided into many small fragments of work, each of which may be executed on any node in the cluster. This framework has been used by developers, enterprises, and startups and has proven to be a reliable software platform for processing up to petabytes of data on clusters of thousands of commodity machines.

Amazon Elastic MapReduce allows you to implement data processing applications in many languages including Java, Perl, Ruby, Python, PHP, R, or C++. You can test these applications on different instance types and job flow

Amazon S3, and then click the "Create Job Flow" button. Alternatively you can start a job flow by specifying the same information mentioned above via our Command Line Tools or APIs. Amazon Elastic MapReduce employs a simple web service interface that is easy to use and highly flexible:

RunJobFlow: Creates a job flow request, starts EC2 instances and begins processing.

DescribeJobFlows: Provides status of your job flow request(s).

AddJobFlowSteps: Adds additional step to an already running job flow.

TerminateJobFlows: Terminates running job flow and shutdowns all instances.

If you wish to run job flows with more than 20 instances, please complete the instance request form.

Paying for What You Use

You are only charged for the resources actually consumed. For example, let's say you launched 100 Amazon EC2 Standard Small instances for an Amazon Elastic MapReduce job flow, where the Amazon Elastic MapReduce cost is an incremental $0.015 per hour. The Amazon EC2 instances will begin booting immediately, but they won't necessarily all start at the same moment. Amazon Elastic MapReduce will track when each instance starts and will check it into the cluster so that it can accept processing tasks.

In the first 10 minutes after your launch request, Amazon Elastic MapReduce either starts your job flow (if all of your instances are available) or checks in as many instances as possible. Once the 10 minute mark has passed, Amazon Elastic MapReduce will start processing (and charging for) your job flow as soon as 90% of your requested instances are available. As the remaining 10% of your requested instances check in, Amazon Elastic MapReduce starts charging for those instances as well.

So, in the above example, if all 100 of your requested instances are available 10 minutes after you kick off a launch request, you'll be charged $1.50 per hour (100 * $0.015) for as long as the job flow takes to complete. If only 90 of your requested instances were available at the 10 minute mark, you'd be charged $1.35 per hour (90 * $0.015) for as long as this was the number of instances running your job flow. When the remaining 10 instances checked in, you'd be charged $1.50 per hour (100 * $0.015) for as long as the balance of the job flow takes to complete. Each job flow will run until one of the following occurs: you terminate the job flow with the *TerminateJobFlows* API call (or an equivalent tool), the job flow shuts itself down, or the job flow is terminated due to software or hardware failure. Partial instance hours consumed are billed as full hours.

↑ Top

Intended Usage and Restrictions

Your use of this service is subject to the Amazon Web Services Customer Agreement

↑ Top