



Amazon EMR 클러스터의 하드웨어 선택

AWS 권장 가이드



AWS 권장 가이드: Amazon EMR 클러스터의 하드웨어 선택

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon의 상표 및 트레이드 드레스는 Amazon 외 제품 또는 서비스와 함께, Amazon 브랜드 이미지를 떨어뜨리거나 고객에게 혼동을 일으킬 수 있는 방식으로 사용할 수 없습니다. Amazon이 소유하지 않은 기타 모든 상표는 Amazon과 제휴 관계이거나 관련이 있거나 후원 관계와 관계없이 해당 소유자의 자산입니다.

Table of Contents

소개	1
개요	1
클러스터 구성	2
선택 및 배포	3
선택 지침	3
인스턴스 선택	4
용량 추정	5
프라이머리 노드	5
코어 및 태스크 노드	5
코어 노드	5
태스크 노드	6
스토리지 선택	7
하둡 분산 파일 시스템	7
Amazon EMR에서 HDFS를 사용해야 하는 경우	7
EMR 파일 시스템	7
자동 규모 조정	8
Amazon EMR Managed Scaling	8
사용자 지정 자동 조정	8
조정 팁	9
모범 사례	10
인스턴스 추가	10
인스턴스 그룹	10
스팟 인스턴스	10
리소스	12
문서 기록	13
.....	xiv

Amazon EMR 클러스터의 하드웨어 선택

Sayde Aguilar, Amiin Samatar, Diego Valencia, Amazon Web Services(AWS)

2023년 8월([문서 기록](#))

Amazon EMR은 빅 데이터 처리를 위한 도구입니다. 오픈 소스 소프트웨어, 특히 Apache Spark 및 Apache Hudi와 같은 Apache 도구를 사용합니다. 또한 저렴한 pay-as-you-go제 모델을 구성하고 사용하기 위한 몇 가지 옵션을 제공합니다.

이 가이드에서는 이러한 탄력성을 기반으로 Amazon EMR 클러스터를 설계하는 방법을 설명하고 하드웨어를 선택할 때 따라야 할 모범 사례를 제공합니다.

개요

Amazon EMR은 방대한 양의 데이터를 처리하기 위한 프레임워크인 Apache Hadoop MapReduce를 사용하여 구축되었습니다. Hadoop MapReduce는 병렬 로직을 사용하여 분산 클러스터의 데이터를 동시에 처리합니다. 즉, 모든 프로세스에 자체 프로세서가 있습니다. Amazon EMR은 Amazon Elastic Compute Cloud(Amazon EC2)에 구조화된 가상 서버의 하둡 클러스터를 사용합니다. 즉, 모든 병렬 프로세스는 Amazon Web Services()에서 실행되는 독립 실행형 컴퓨터에서 이루어집니다AWS.

하둡 클러스터는 병렬 또는 분산 환경을 사용하여 대량의 비정형 데이터를 처리하는 데 사용되는 특정 유형의 컴퓨팅 클러스터입니다. 하둡 클러스터의 주요 특성은 확장성이 뛰어나고 데이터 처리를 가속화하도록 구성할 수 있다는 것입니다. 노드를 추가하거나 제거하여 처리량을 늘리거나 줄이면 확장성에 도달합니다. 하둡 클러스터에서는 각 데이터가 클러스터 노드 간에 복사되므로 노드에 장애가 발생할 경우 손실되는 데이터가 거의 없습니다.

Amazon EMR에서 탄력성은 동적 크기 조정 기능을 나타냅니다. 클러스터를 자동으로 확장하고 필요한 사항을 변경할 수 있습니다. 초기 하드웨어 설계에 의존할 필요가 없습니다.

이 가이드에서는 이러한 탄력성을 기반으로 Amazon EMR 클러스터를 설계하는 방법을 설명하고 하드웨어를 선택할 때 따라야 할 모범 사례를 제공합니다.

Amazon EMR 클러스터 구성

Amazon EMR을 사용하면 빅 데이터 프레임워크로 데이터를 처리하고 분석하는 클러스터를 설정하는데 몇 분 밖에 걸리지 않습니다. 그러나 클러스터를 구현하기 전에 스토리지를 준비하고, 작업 한도를 정의하고, 워크로드의 가장 효율적인 배포를 결정해야 합니다.

Amazon EMR 구성의 경우 이 안내서에서는 다음 주제를 다룹니다.

- [Amazon EMR 클러스터 선택 및 배포](#)
- [Amazon EMR 클러스터 용량 추정](#)
- [올바른 스토리지 선택](#)
- [자동 조정 접근 방식 선택](#)
- [구성 모범 사례](#)

Amazon EMR 클러스터 선택 및 배포

노드 유형을 식별하고 구성합니다. Amazon EMR 클러스터를 정의할 때 해당 하드웨어를 이해하는 것이 중요합니다. 어떻게 작동하나요? 어떻게 구성되나요? 이러한 질문에 대한 답변은 세 부분으로 구성됩니다.

- 노드 유형
- 각 노드가 수행하는 함수
- 각 노드에 가장 효율적인 EC2 인스턴스의 유형

처음에는 기본 노드가 일반 리소스 관리를 담당합니다. 분산 애플리케이션의 기본 구성 요소를 실행합니다. 예를 들어 Hadoop Distributed File System(HDFS) NameNode 서비스를 실행하고, 클러스터에서 수행할 작업을 추적하고, 시스템 상태를 모니터링합니다.

또한 Amazon EMR에는 코어 노드와 태스크 노드가 있습니다. 코어 노드는 프라이머리 노드에서 관리합니다. 코어 노드는 작업 노드를 실행하고 클러스터의 HDFS에 데이터를 저장하는 역할을 합니다. 작업 노드는 클러스터로 들어오는 작업을 관리하는 역할을 합니다. 작업 노드는 데이터를 저장하지 않습니다. (작업 노드는 필수가 아닙니다.)

Amazon EMR 클러스터를 구성하고 배포할 때 중요한 고려 사항은 클러스터 노드를 나타내는 EC2 인스턴스를 올바르게 선택하는 것입니다. 클러스터에 대한 인스턴스 그룹 구성 또는 인스턴스 플릿 구성을 사용하는지 여부에 따라 EC2 인스턴스를 클러스터에 추가하는 방법에는 여러 가지가 있습니다. 지원되는 인스턴스 유형에 대한 자세한 내용은 [AWS 설명서를](#) 참조하세요.

다음 지침은 대부분의 [Amazon EMR 클러스터](#)에 적용됩니다. [클러스터 구성 모범 사례](#)를 검토할 수도 있습니다.

인스턴스 선택 지침

일반적으로 Amazon EMR 구현에 선호되는 인스턴스는 실행 중인 작업에 따라 달라집니다. 다음 질문을 고려하세요.

- 작업 메모리가 집약적입니까?
- 작업 CPU가 집약적입니까?
- 대용량 스토리지가 필요합니까?
- 작업에 GPU 용량이 필요합니까?

이러한 질문은 필요한 인스턴스 유형과 필요한 실제 특성을 이해하는 데 도움이 됩니다. 동시에 처리할 작업 수와 작업을 처리해야 하는 속도를 결정합니다. Amazon EMR 사용량은 시간당 단위로 청구되므로 이는 중요합니다. 클러스터를 켜면 전체 시간에 대한 요금이 부과됩니다.

서로 다른 AWS 리전에서 실행되는 각 인스턴스의 비용을 확인할 수 있습니다. 리전 간 가격을 비교하려면 [AWS 요금 계산기](#)를 사용하고 위치에 따라 값을 변경할 수 있습니다.

EC2 인스턴스 선택

이전 질문에 답했다면 해당 요구 사항에 따라 인스턴스를 선택해야 합니다. 처리 작업 요구 사항을 이해한 후 필요한 특성에 따라 인스턴스 유형을 결정합니다.

- 범용 인스턴스가 필요한 경우 M6g, T4g 또는 M5 인스턴스를 선택합니다.
- 컴퓨팅 최적화 인스턴스가 필요한 경우 C6g 또는 C5 인스턴스를 선택합니다.
- 메모리 최적화 인스턴스가 필요한 경우 R6g, X1, R5 또는 z1d 인스턴스를 선택합니다.
- 스토리지를 최적화해야 하는 경우 높은 I/O 성능을 제공하는 I3 인스턴스를 선택합니다.
- GPU와 같은 가속 컴퓨팅이 필요한 경우 P3, G4 또는 Inf1 인스턴스를 선택합니다. 이러한 인스턴스 유형은 다른 프로세스 중에서도 기계 학습 및 유체 역학에 높은 성능을 제공합니다.

인스턴스 유형과 기능을 이해하는 또 다른 방법은 각 인스턴스 유형에 대한 [기본 메모리](#)를 분석하는 것입니다. 이 지표는 MapReduce 작업의 성능을 조정하고 개선하는 데 도움이 됩니다. 자세한 내용은 [하둡 데몬 구성 설정](#)을 참조하세요.

필요한 인스턴스 유형을 알고 있으면 클러스터 용량을 계획할 수 있습니다.

Amazon EMR 클러스터 용량 추정

Amazon EMR은 크기 조정이 가능한 플랫폼이지만 클러스터 크기를 적절하게 조정하는 것이 중요합니다. 크기를 적절하게 조정하면 클러스터 크기가 작을 경우 발생하는 느린 속도, 클러스터 크기가 너무 클 경우 발생하는 높은 비용을 방지할 수 있습니다. 이러한 문제를 예상하기 위해 워크로드에 필요한 노드의 수와 유형을 계산할 수 있습니다.

프라이머리 노드

이 노드 유형은 데이터 및 프로세스의 배포를 조정하는 역할을 합니다. 앞서 언급한 것처럼 프라이머리 노드의 컴퓨팅 요구 사항은 낮습니다. 단일 프라이머리 노드를 사용하여 Amazon EMR 클러스터를 관리할 수 있습니다. 하지만 단일 장애 지점이 없도록 최대 3개의 프라이머리 노드를 사용할 수 있습니다. 하나의 프라이머리 노드에 장애가 발생하면 Amazon EMR은 다른 두 프라이머리 노드 중 하나로 장애 조치합니다.

코어 및 태스크 노드

코어 노드와 태스크 노드의 차이점은 태스크 노드는 데이터를 저장하지 않고 병렬 컴퓨팅 태스크를 수행할 능력만 제공한다는 것입니다.

코어 노드와 태스크 노드의 수를 계산하려면 데이터 크기와 대략적인 메모리 사용량을 알아야 합니다.

코어 노드

코어 노드는 태스크를 실행하여 데이터를 처리하고 Hadoop 분산 파일 시스템(HDFS)에 데이터를 저장하는 역할을 합니다. 코어 노드의 용량을 계산하려면 코어 노드 수를 정의하고, 노드 수에 각 노드의 Amazon Elastic Block Store(Amazon EBS) 스토리지를 곱합니다.

예를 들어, 1TiB의 데이터를 처리하는 코어 노드 10개를 정의하고, Amazon EBS 스토리지가 64GiB인 m5.xlarge 인스턴스 유형을 사용하는 경우 $10 \text{ nodes} \times 64 \text{ GiB}$, 즉 640GiB의 용량을 갖게 됩니다. HDFS [복제 계수](#) 3을 기준으로 노드에 데이터 크기가 세 번 복제되므로 1TiB의 데이터에는 3TiB의 용량이 필요합니다. 이 예제에서는 640GiB뿐이므로 용량이 3TiB가 될 때까지 노드 수를 늘리거나 인스턴스 유형을 변경해야 합니다.

m5.4xlarge 인스턴스 유형의 스토리지는 256GiB입니다. m5.4xlarge 인스턴스 유형으로 변경하고 12개의 인스턴스를 선택하면 충분한 용량이 제공됩니다.

$12 \text{ instances} \times 256 \text{ GiB of storage} = 3072 \text{ GiB} = 3 \text{ TiB available}$

태스크 노드

태스크 노드는 태스크만 실행합니다. 데이터를 저장하지는 않습니다. 태스크 노드 수를 계산하려면 메모리 사용량을 추정해야 합니다. 이 용량은 코어와 태스크 노드로 분할할 수 있습니다. 필요한 태스크 노드 수를 계산하려면 이전 단계에서 계산한 코어 노드가 제공하는 메모리를 메모리 사용량에서 빼면 됩니다.

메모리 범위를 넓히려면 필요한 메모리에 3을 곱하는 것이 좋습니다.

각각 20GiB의 프로세스 28개가 있다고 가정해 보겠습니다.

$3 \times 28 \text{ processes} \times 20 \text{ GiB of memory} = 1680 \text{ GiB of memory}$

이 예제에서 코어 노드에는 64GiB 메모리(m5.4xlarge 인스턴스)가 있습니다. 코어 노드가 제공하는 메모리는 $64 \text{ GiB} \times 12 \text{ nodes} = 768 \text{ GiB of memory}$ 로, 이 예제에서는 충분하지 않습니다.

부족한 용량을 확보하려면 필요한 총 메모리에서 코어 노드 메모리를 뺍니다.

$1680 \text{ GiB} - 768 \text{ GiB core node memory} = 912 \text{ GiB memory shortage.}$

태스크 노드는 나머지 912GiB 메모리를 제공할 수 있습니다. 이 예제에서 태스크 노드의 메모리는 32GiB(m5.2xlarge 인스턴스)입니다. 필요한 태스크 노드 수를 구하려면 부족한 메모리를 인스턴스 유형 메모리로 나눕니다.

$912 \text{ GiB} / 32 \text{ GiB} = 28.5 \text{ task nodes}$

태스크 노드는 소수가 될 수 없으므로 태스크 노드를 29개로 올림해야 합니다.

올바른 스토리지 선택

클러스터의 [인스턴스 유형 및 용량](#)을 구성합니다. 코어 및 태스크 노드에는 처리 및 컴퓨팅 성능이 필요하지만 코어 노드만 데이터를 저장합니다. 코어 노드에 가장 비용 효율적인 스토리지 유형을 선택합니다.

대량의 데이터를 처리하기 위해 Amazon EMR을 사용하는 경우 Amazon S3에서 데이터를 이동하는 몇 가지 옵션이 있습니다. 가장 좋은 옵션은 워크로드에 따라 다릅니다. 다음 섹션에서는 어떤 스토리지 유형이 적합한지 결정할 때 고려해야 할 몇 가지 핵심 사항을 제공합니다.

하둡 분산 파일 시스템

Hadoop 분산 파일 시스템(HDFS)은 Hadoop을 위한 확장 가능하고 이동 가능한 분산 파일 시스템입니다. HDFS의 이점은 클러스터를 관리하는 하둡 클러스터 노드와 개별 단계를 관리하는 하둡 클러스터 노드 간의 데이터 인식입니다.

Amazon EMR에서 HDFS를 사용해야 하는 경우

동일한 데이터 세트 또는 디스크 I/O 집약적 워크로드에 반복 읽기가 있는 경우 중간 결과를 캐싱하는데 HDFS를 사용하고 데이터 처리를 위한 핫 스토리지로 사용할 수 있습니다. HDFS는 임시적이므로 인스턴스가 종료될 때 회수됩니다.

EMR 파일 시스템

EMR 파일 시스템(EMRFS)은 Amazon EMR 클러스터가 Amazon EMR에서 Amazon S3로 직접 일반 파일을 읽고 쓰는 데 일반적으로 사용하는 HDFS의 구현입니다.

각 실행에서 데이터 세트를 한 번 읽을 때 EMRFS를 사용할 수 있습니다. EMRFS는 스토리지를 컴퓨팅에서 분리하므로 데이터를 저장하기 위해 특별히 코어 노드를 프로비저닝할 필요가 없으며 HDFS에서 데이터 복제 비용을 지불할 필요가 없습니다. 이로 인해 비용이 절감되고 여러 클러스터에 대한 데이터의 가용성이 제공됩니다. 또한 클러스터를 종료한 후 데이터를 유지하는 이점도 있습니다.

자동 조정 접근 방식 선택

탄력성은 Amazon EMR 사용의 주요 이점 중 하나입니다. 리소스를 자동으로 조정하기 위한 두 가지 주요 옵션이 있습니다.

- Managed Scaling
- 사용자 지정 조정 정책

관리형 조정 또는 사용자 지정 자동 조정 정책을 사용하면 필요한 리소스만 사용하도록 노드를 스케일 인 및 스케일 아웃할 수 있습니다. 확장은 용량이 더 필요할 때 리소스를 더 추가하는 데 사용됩니다. 스케일 인하면 사용되지 않는 리소스를 제거하여 비용 효율성이 향상됩니다. Amazon EMR 서비스 내에서 Amazon CloudWatch 지표는 리소스를 모니터링하도록 활성화되어 클러스터를 확장할 수 있습니다. CloudWatch는 5분마다 데이터 포인트를 가져옵니다.

각 자동 조정 접근 방식에는 다양한 고려 사항이 있습니다.

Amazon EMR Managed Scaling

워크로드가 다음 기준을 충족하는 경우 [EMR 관리형 조정](#)을 사용합니다.

- 관리형 경험이 필요합니다.
- Amazon EMR 5.330 이상이 사용됩니다.
- 1분의 평가 빈도가 필요합니다.
- 이 솔루션은 인스턴스 플릿을 사용하여 1~5개의 인스턴스 옵션을 제공합니다.
- 애플리케이션은 Apache Spark, Apache Hive 또는 Apache Hadoop YARN을 기반으로 합니다.

사용자 지정 자동 조정

워크로드가 다음 기준을 충족하는 경우 [사용자 지정 자동 조정 정책](#)을 사용합니다.

- 조정을 위한 지표를 제어해야 합니다.
- Amazon EMR 4.0 이상이 사용됩니다.
- 높은 평가 빈도가 필요하지 않습니다.
- 연속 크기 조정 사이에 휴지 시간을 제어할 필요는 없습니다.

- 조정 시 추가하거나 제거할 인스턴스 수를 제어하는 것이 중요합니다.
- 솔루션에는 사용자 지정 조정 작업이 필요합니다. 예를 들어 5분 동안 하나 이상의 노드를 확장할 수 있습니다. 또는 휴지 기간을 조정할 수 있습니다.
- 인스턴스 그룹에서 다른 인스턴트 유형을 사용하는 데는 제한이 없습니다.

클러스터에 자동 조정을 추가할 때의 팁

- 처리할 데이터의 양을 알고 있어야 합니다. 데이터 크기가 가장 큰 사례를 사용하여 예측합니다.
- 클러스터의 [크기를 조정합니다](#).
- 필요에 맞는 [스토리지 유형](#)을 선택합니다.
- Amazon EMR 클러스터의 [지표](#)를 이해합니다.
- 클러스터 크기 조정에 [적합한 지표](#)를 결정하는 방법을 알아봅니다.
- 스팟 인스턴스, 균일한 인스턴스 그룹 또는 인스턴스 풀릿을 사용할지 여부를 결정합니다.
- 정보 및 제한 사항에 따라 선호하는 조정 접근 방식, Amazon EMR 관리형 조정 또는 사용자 지정 자동 조정 정책을 결정합니다.
- 관리형 조정 또는 사용자 지정 정책을 구성합니다.
- 사용자 지정 자동 조정 정책을 선택한 경우 Amazon EMR 지표를 모니터링하여 정책의 임계값을 조정합니다.

Amazon EMR 구성 모범 사례

Amazon EMR 클러스터를 구성할 때 인스턴스 추가, 인스턴스 그룹 작업 및 스팟 인스턴스 사용에 대한 다음 모범 사례를 사용합니다.

인스턴스 추가

EMR 클러스터를 구성할 때 중요한 고려 사항은 클러스터 노드를 나타내는 EC2 인스턴스를 올바르게 선택하는 것입니다. 클러스터가 실행 중일 때는 스팟 인스턴스를 온디맨드 인스턴스로 변경하는 등 인스턴스 유형을 변경할 수 없습니다. 기본 노드를 변경하려면 클러스터를 종료하고 새 노드를 생성해야 합니다. 따라서 가동 중지 시간을 최소화할 수 있도록 올바른 인스턴스 유형을 선택해야 합니다. 자세한 내용은 [클러스터 구성 지침 및 모범 사례](#)를 참조하세요.

클러스터에 대한 인스턴스 그룹 구성 또는 인스턴스 플릿 구성을 사용하는지 여부에 따라 클러스터에 EC2 인스턴스를 추가하는 방법에는 여러 가지가 있습니다.

- EC2 인스턴스 수동 추가
- 인스턴스 그룹에 작업을 수동으로 추가하여 인스턴스 자동 추가
- 자동 조정 설정

인스턴스 그룹

구성에 EC2 인스턴스를 추가할 때는 인스턴스 그룹을 사용하는 것이 좋습니다. 인스턴스를 수동으로 추가하는 경우 기존 코어 및 태스크 인스턴스 그룹에 동일한 유형의 인스턴스를 추가할 수 있습니다. 또한 다른 인스턴스 유형을 사용할 수 있는 태스크 인스턴스 그룹을 추가할 수 있습니다.

마지막으로 인스턴스 그룹에 대해 Amazon EMR에서 자동 조정을 설정합니다. 지정한 Amazon CloudWatch 지표의 값에 따라 인스턴스를 자동으로 추가 및 제거할 수 있습니다. 그렇지 않으면 인스턴스 플릿을 사용하는 경우 단일 태스크 인스턴스 플릿을 추가합니다. 그런 다음 기존 코어 및 태스크 인스턴스 플릿의 온디맨드 인스턴스 및 스팟 인스턴스의 목표 용량을 변경합니다.

스팟 인스턴스

작업 노드에서 스팟 인스턴스를 사용합니다. 작업 노드는 데이터를 처리하지만 Hadoop 분산 파일 시스템(HDFS)에 영구 데이터를 보관하지 않습니다. 스팟 가격이 최대 스팟 가격보다 높아져 작업 노드가 종료되면 데이터가 손실되지 않으며 클러스터에 미치는 영향이 최소화됩니다.

태스크 인스턴스 그룹을 스팟 인스턴스로 시작하면 Amazon EMR은 최대 스팟 가격을 사용하여 가능한 한 많은 태스크 노드를 프로비저닝합니다. 예를 들어 노드가 6개인 태스크 인스턴스 그룹을 요청할 수 있습니다. 스팟 인스턴스를 최대 스팟 가격 이하로 5개만 사용할 수 있는 경우 Amazon EMR은 노드가 5개인 인스턴스 그룹을 시작합니다. Amazon EMR은 나중에 가능하면 여섯 번째 노드를 추가합니다. 자세한 내용은 [클러스터 구성 지침 및 모범 사례를](#) 참조하세요.

리소스

- [Amazon EMR이란 무엇입니까?](#)
- [Amazon EMR 기능](#)
- [Amazon EMR에 대한 고려 사항 및 모범 사례](#)
- [Amazon EMR에 지원되는 인스턴스 유형](#)
- [클러스터 구성 지침 및 모범 사례](#)
- [노드 유형 이해: 프라이머리, 코어 및 태스크 노드](#)
- [작업 구성](#)
- [하둡 데몬 구성 설정](#)
- [관리 콘솔을 AWS 사용하여 관리형 조정 구성](#)
- [CloudWatch를 사용하여 지표 모니터링](#)
- [Amazon EMR Managed Scaling 소개 - 클러스터 크기를 자동으로 조정하여 비용 절감](#)

문서 이력

아래 표에 이 가이드의 주요 변경 사항이 설명되어 있습니다. 향후 업데이트에 대한 알림을 받으려면 [RSS 피드](#)를 구독하십시오.

변경 사항	설명	날짜
설명	용량 추정 예제 를 수정했습니다.	2023년 8월 10일
=	최초 게시	2021년 10월 29일

기계 번역으로 제공되는 번역입니다. 제공된 번역과 원본 영어의 내용이 상충하는 경우에는 영어 버전이 우선합니다.