



このコンテンツは公開から3年以上経過しており内容が古い可能性があります
最新情報については[サービス別資料](#)もしくはサービスのドキュメントをご確認ください

[AWS Black Belt Online Seminar]

Amazon EMR

サービスカットシリーズ

Solutions Architect

半場 光晴

2019/10/23

AWS 公式 Webinar

<https://amzn.to/JPWebinar>



過去資料

<https://amzn.to/JPArchive>



自己紹介



- 名前：
半場 光晴（はんば みつはる）
- 所属：
アマゾンウェブサービスジャパン
株式会社 技術統括本部
ソリューションアーキテクト
- 好きな AWS サービス：
AWS サポート、Amazon EMR、
Amazon Elasticsearch Service、
Amazon Kinesis

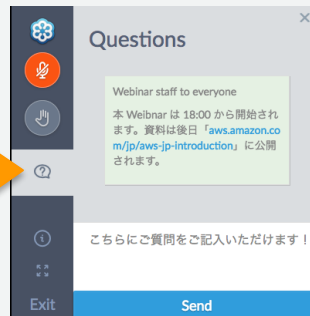
AWS Black Belt Online Seminar とは

「サービス別」「ソリューション別」「業種別」のそれぞれのテーマに分かれて、アマゾンウェブサービスジャパン株式会社が主催するオンラインセミナーシリーズです。

質問を投げることができます！

- 書き込んだ質問は、主催者にしか見えません
- 今後のロードマップに関するご質問は
お答えできませんのでご了承下さい

- ① 吹き出しをクリック
- ② 質問を入力
- ③ Sendをクリック



Twitter ハッシュタグは以下をご利用ください
#awsblackbelt

内容についての注意点

- 本資料では2019年10月22日時点のサービス内容および価格についてご説明しています。最新の情報はAWS公式ウェブサイト (<http://aws.amazon.com>) にてご確認ください。
- 資料作成には十分注意しておりますが、資料内の価格とAWS公式ウェブサイト記載の価格に相違があった場合、AWS公式ウェブサイトの価格を優先とさせていただきます。
- 価格は税抜表記となっております。日本居住者のお客様が東京リージョンを使用する場合、別途消費税をご請求させていただきます。
- AWS does not offer binding price quotes. AWS pricing is publicly available and is subject to change in accordance with the AWS Customer Agreement available at <http://aws.amazon.com/agreement/>. Any pricing information included in this document is provided only as an estimate of usage charges for AWS services based on certain information that you have provided. Monthly charges will be based on your actual use of AWS services, and may vary from the estimates provided.

本日のアジェンダ

- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

What's New in Amazon EMR

<https://aws.amazon.com/blogs/big-data/run-spark-applications-with-docker-using-amazon-emr-6-0-0-beta/>



Services

Resource Groups



Tokyo

Support

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

What's new

Sep 5, 2019 : Simplify your Spark application dependency management with Docker and Hadoop 3 with EMR 6.0.0 (Beta).

Hadoop 3 [Beta]

EMR 6.0.0 (Beta) allows users to define application and library dependencies using Docker images from [Docker Hub](#) and [Amazon Elastic Container Registry \(ECR\)](#) using Spark 2.4.3 and Hadoop 3.1.0. Using Hadoop 3, Docker, and EMR, Spark users no longer have to install library dependencies on individual cluster hosts, and application dependencies can now be scoped to individual Spark applications. You can use Docker images to package your own library dependencies, and can even run containers with different versions of R and Python on the same cluster. The EMR 6.0.0 (Beta) release is available in the US East (N. Virginia), and US West (Oregon) regions. [Learn more](#)

Aug 19, 2019 : Introducing Block public access configuration to secure EMR clusters from unintentional network exposure.

パブリックアクセス禁止

Amazon EMR has introduced a new account level configuration called Block public access that allows administrators to enforce common public network access rules for EMR clusters. You can enable this configuration and prevent your account users from launching clusters with security group rules that open ports for inbound traffic from IPv4 0.0.0.0/0 or IPv6 :::0. You can configure exceptions in Block Public access configuration to allow public access on a port or range of ports before you launch EMR clusters. [Learn more](#)

Aug 6, 2019 : Introducing EBS encryption for encrypting EBS volumes attached to EMR cluster

EBS 暗号化が簡単に

With EMR release 5.24.0 and later, you can enable EBS encryption for encrypting data on EBS volumes attached to EMR cluster. This feature is available as part of the security configuration settings. [Learn more](#)

Apr 30, 2019 : Support for multiple master nodes to enable high availability for EMR applications

マルチマスター

With EMR release 5.23.0 and later, you can launch an EMR cluster with three master nodes and enable high Availability of applications running on EMR clusters. [Learn more](#)

Apr 30, 2019 : Support for reconfiguring applications on running EMR Clusters

クラスター設定の、後から上書き

With Amazon EMR version 5.21.0 and later, you can re-configure cluster configurations and specify additional configuration classifications for each instance group in a running cluster. [Learn more](#)

Mar 11, 2019 : Support for Flink 1.7.0 on Amazon EMR release 5.21.0

You can now use Apache Flink 1.7.0 on Amazon EMR release 5.21.0. Flink 1.7.0 has many new features such as the ability to change the state schema of the streaming application which allows you to change the features captured by the application without restarting it, support for a new connector to write to Kafka 2.0, and an exactly-once S3 file sink for writes to Amazon S3. Flink 1.7.0 also features several streaming SQL improvements such as support for temporal tables that tracks the history of data changes and several new built-in SQL functions to handle complex data types. [Learn more](#)

Nov 19, 2018 : Introducing EMR Notebooks, a managed analytics environment based on Jupyter Notebooks.

ノートブック

EMR Notebooks allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analysis using EMR clusters. EMR Notebooks is pre-configured for Spark, supporting kernels for languages such as PySpark, Spark SQL, Spark R, and Scala, as well as open-source library packages found in Anaconda. Integrated Spark monitoring capabilities allow you to monitor job progress and debug code within the notebook editor. There is no additional cost for using EMR Notebooks. You only pay for the EMR cluster attached to the notebook. EMR Notebooks is available in the US East (N. Virginia and Ohio), US West (N. California and Oregon), Canada (Central), EU (Frankfurt, Ireland, and London), and Asia Pacific (Mumbai, Seoul, Singapore, Sydney, and Tokyo) regions. [Learn more](#)



本日のアジェンダ

- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

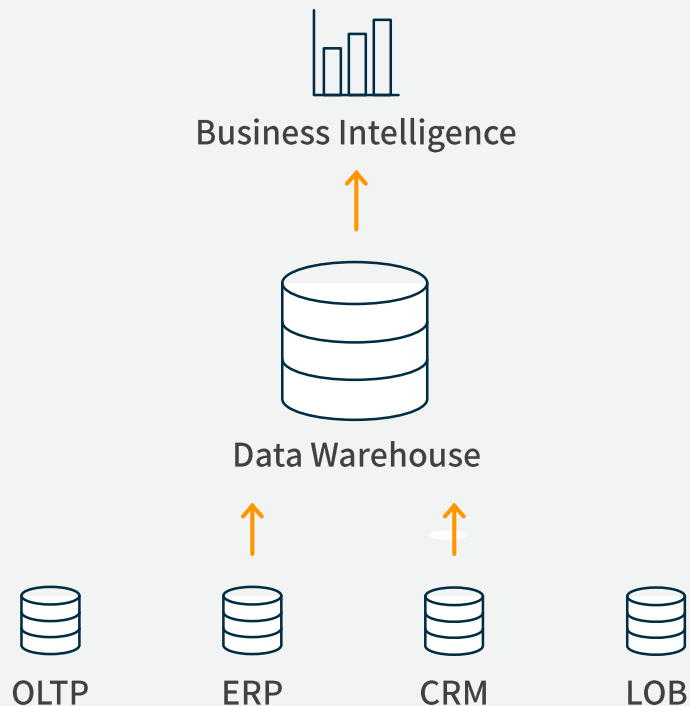
データを差別化要因にするために、求められること

新たな分析のタイプ



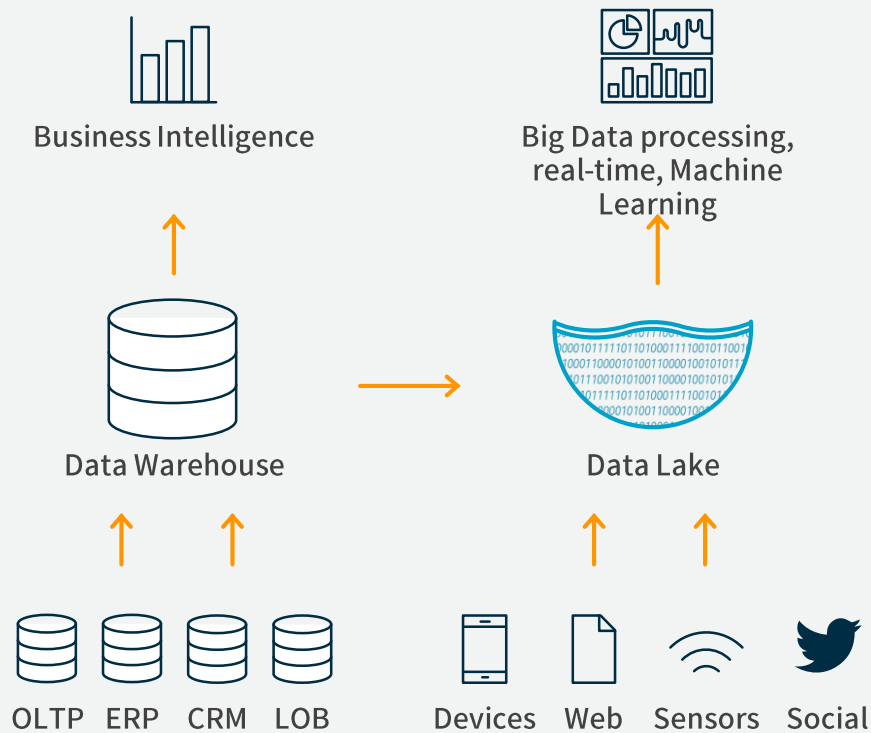
- 新たな非リレーショナルデータをPB-EBスケールでリアルタイムにキャプチャして保存する
- バッチレポートにとどまらず、リアルタイム、予測、音声、画像認識を組み込む新しいタイプの分析をする必要がある
- 安全かつ管理された方法でデータへのアクセスを民主化する

従来までの分析アプローチによくある姿



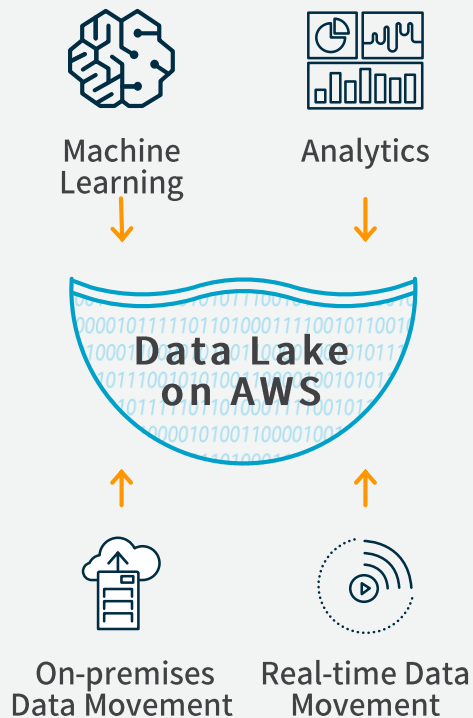
- リレーショナルデータ
- TBs-PBs スケール
- データロード前のスキーマ定義
- 運用レポート作成とアドホック操作
- 大きな初期設備投資額に加え、TBあたり年間およそ10万-50万ドルのランニングコスト

従来のアプローチを拡張するデータレイク



- リレーショナルデータに加えて、非リレーショナルデータ
- TBs-EBs スケール
- 多様な分析エンジン
- 低コストのストレージと分析

データレイクと AWS の分析系サービス



 オープンかつ包括的

 安全

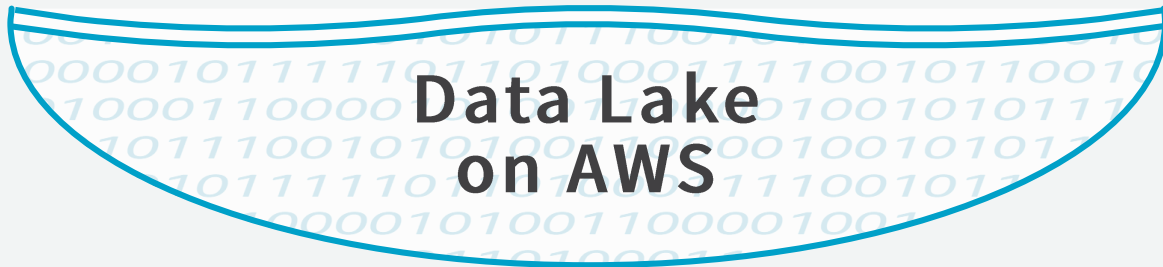
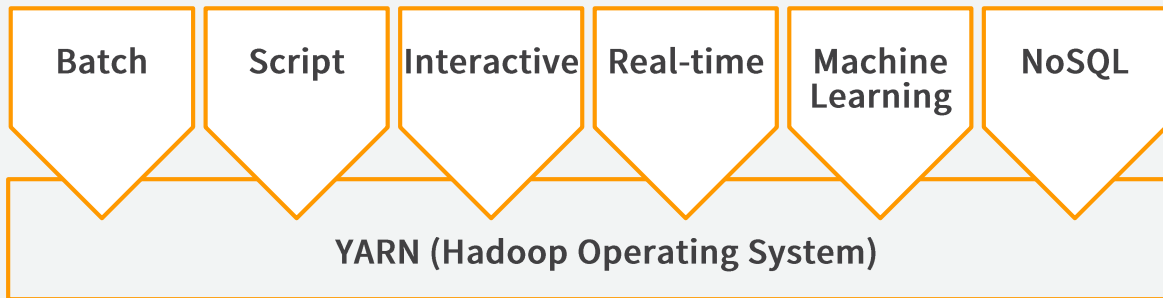
 拡張性と耐久性

 経済性

Apache Hadoop とデータレイク



Apacheは、Apache Software Foundationの登録商標または商標です



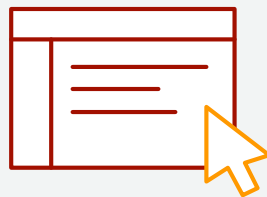
- 分散処理
- 多様な分析
 - Batch/Script (Hive/Pig)
 - Interactive (Spark, Presto)
 - Real-time (Spark)
 - Machine Learning (Spark)
 - NoSQL (HBase)
- 幅広いユースケース
 - Log and clickstream analysis
 - Machine Learning
 - Real-time analytics
 - Large-scale analytics
 - Genomics
 - ETL

Amazon EMR

大幅なコスト節減を可能にする、クラウドを利用したマネージドな Hadoop と Spark



高い品質



簡単



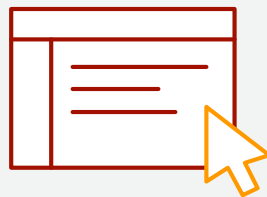
低コスト

Amazon EMR

大幅なコスト節減を可能にする、クラウドを利用したマネージドな Hadoop と Spark



高い品質



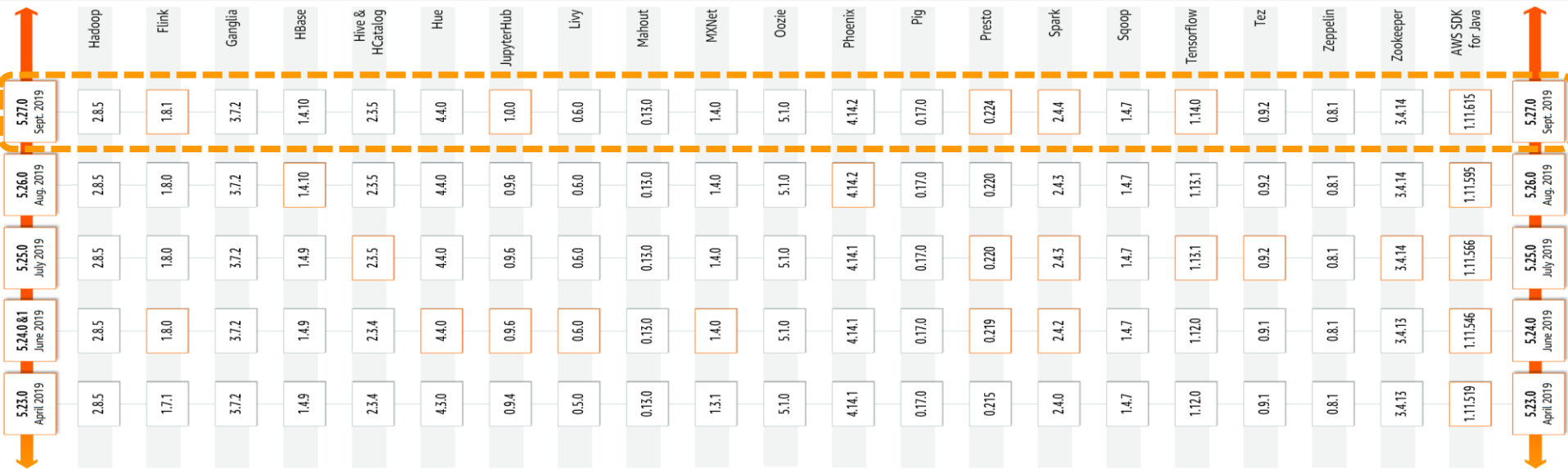
簡単



低コスト

高い品質の Hadoop と Spark

最新の Hadoop および Spark エコシステムのリリースをデプロイ

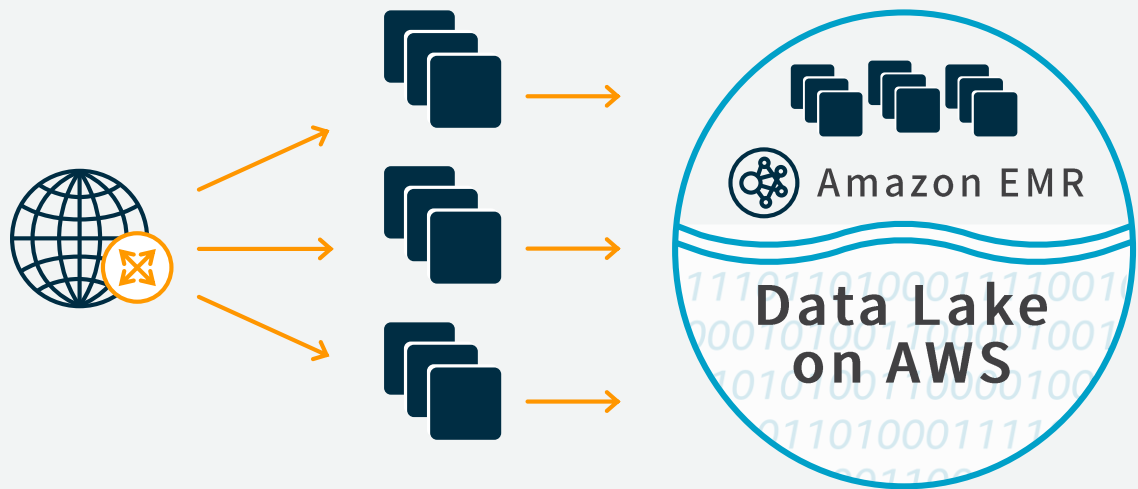


- 20のオープンソースプロジェクト: Apache Hadoop, Spark, HBase, Presto など

- リリースからおおよそ30日以内に、最新のオープンソースフレームワークへ更新を継続

高い品質の Hadoop と Spark

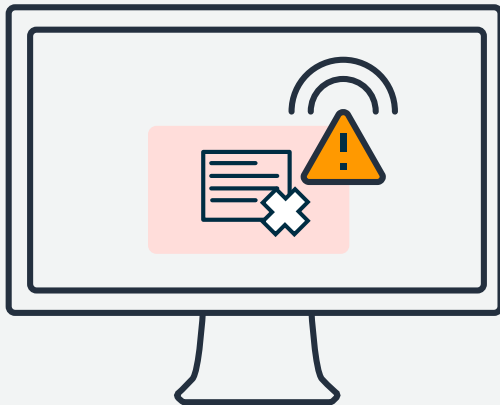
任意のサイズに拡張可能



- コンピューティング（EMR）とストレージ（S3）を個別にスケールリング
- PBからEBまで、あらゆる量のデータを保存、そして、処理
- 1、100、はたまた数千ノードのクラスターをプロビジョニング
- オートスケールリング

高い品質の Hadoop と Spark

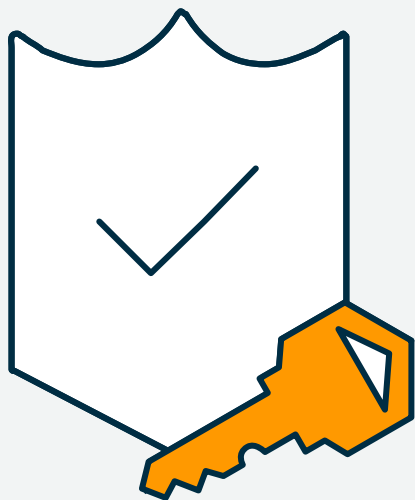
高い可用性と耐久性



- S3 は、99.9999999% の耐久性を提供するように設計されている
- EMR は、クラスターを監視し、パフォーマンスの低いノードと障害が発生したノードを置き換え、サービスを再起動する
- Amazon CloudWatch を使用してクラスターを監視する
- EMRには、ジョブ履歴を表示してログを閲覧できる組み込みコンソールがある
- EMR は、データ永続性のために、（S3利用に加えて）クラスター内に HDFS も備えている

高い品質の Hadoop と Spark

高い安全性



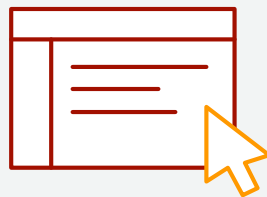
- 保存時および転送中のデータの暗号化
- Amazon Macie によるMLによるセキュリティ
- Amazon VPC を使用したネットワーク分離
- IAM ポリシーによるアクセスと権限の制御
- AWS CloudTrail を使用したアクティビティの記録と監査
- Kerberos サポートによるマイクロソフト AD との統合

Amazon EMR

大幅なコスト節減を可能にする、クラウドを利用したマネージドな Hadoop と Spark



高い品質



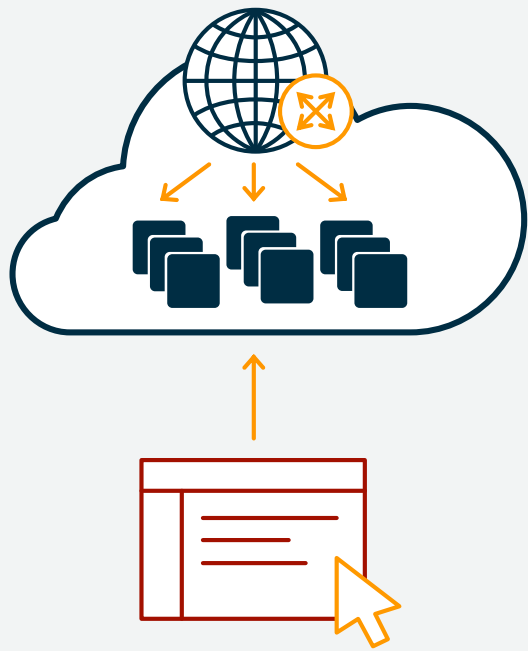
簡単



低コスト

簡単

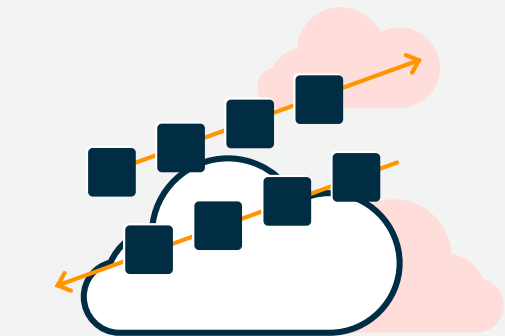
わずか数分でクラスターの起動・利用開始



- Hadoop/Sparkクラスターを数分で起動
- Hadoopのインストールやメンテナンス不要
- クラスターのチューニングと構成を自動実行
- リリースからおよそ 30 日以内に最新の Hadoop バージョンを利用可能

簡単

自動、そして、伸縮自在の、スケーリング



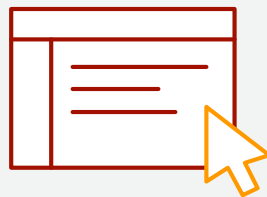
- スケーリングポリシーに基づいてクラスタを自動的に拡張
- 処理完了時にクラスタをシャットダウン可能
- 一時的なクラスターと長時間稼働クラスターの両方に最適化
- 手動による介入は不要

Amazon EMR

大幅なコスト節減を可能にする、クラウドを利用したマネージドな Hadoop と Spark



高い品質



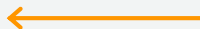
簡単



低コスト

低コスト

リザーブドインスタンスとスポットの活用による、75-90%のコスト節約



- リザーブドインスタンスで最大 75% の節約
- スポットで最大 90% 節約
 - オンデマンド料金のほんの一部の支払い
 - 入札価格が市場価格を超えた場合にリソースを取得
 - スポットとオンデマンドの組み合わせでインスタンスタイプのリストからプロビジョニング
 - 容量/価格に基づいて最適な AZ で起動
 - スポットブロックのサポート

低コスト

低い TCO (Total Cost of Ownership)

オンプレ

Support Costs

Server Costs

Hardware—Server, Rack, Chassis, PDUs, Tor Switches (+Maintenance)

Software—OS, Virtualization Licenses (+Maintenance)

Network Costs

Network Hardware—LAN Switches, Load Balancer Bandwidth costs

Software—Network Monitoring

IT Labor Costs

Server admin, virtualization admin, storage admin, network admin, support team

Extras

Project planning, advisors, legal, contractors, managed services, training, cost of capital

EMR

Subscription Fee Support Costs

- Hadoop クラスターの管理とサポートにかかる管理時間の短縮
- 前払いコストなし: ハードウェアの取得、設置
- オペレーションコストの節約: データ・センターのスペース、電力、冷却
- ビジネス価値: 遅延コスト、リスクプレミアム、競争力、ガバナンスなど

多くの Hadoop および Spark プロジェクトを支えている EMR



Amazon EMR な This is My Architecture

from scratch: Resource Manager Controls Task Distribution with Multiple Amazon EMR Clusters

<https://youtu.be/nM-AkqNh7Yo>

The video player shows a presentation slide with two men, Hiroaki Idobata and Mitsuharu Hamba, standing in front of a screen. The screen displays a diagram of an AWS architecture. The diagram includes icons for S3, SQS, EC2, CloudWatch, and EMR. The text 'This is My Architecture' is visible on the screen. The video player interface shows the title 'Today's guest is Mr. Hiroaki Idobata,' and the video progress bar is at 0:15 / 5:51.

Hiroaki Idobata
CTO
from scratch

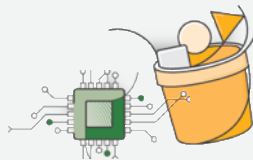
Mitsuharu Hamba
Solutions Architect
AWS

Today's guest is Mr. Hiroaki Idobata,

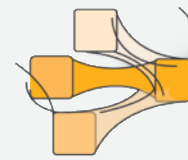
Why Amazon EMR?



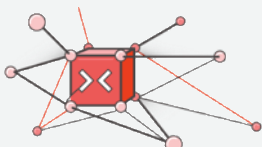
自動化



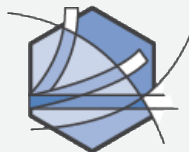
疎結合



弾力性



統合

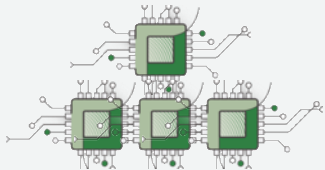


継続的な更新

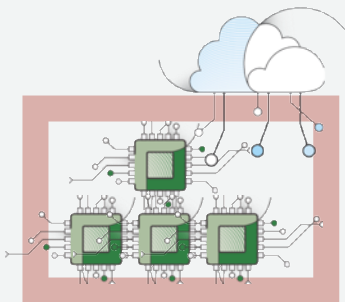


低コスト

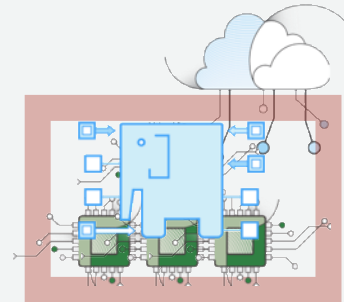
Why Amazon EMR? : 自動化



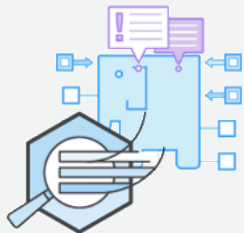
EC2 のプロビジョニング



クラスターのセットアップ



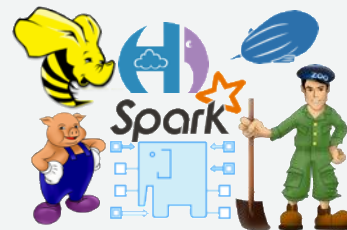
Hadoop の設定



監視と
失敗のハンドリング



ジョブの送信

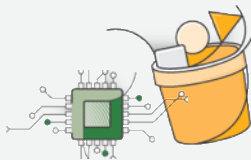


アプリケーションの
インストール

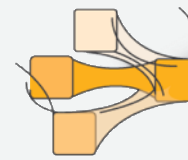
Why Amazon EMR?



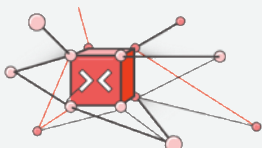
自動化



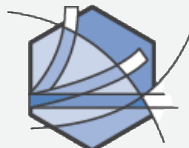
疎結合



弾力性



統合

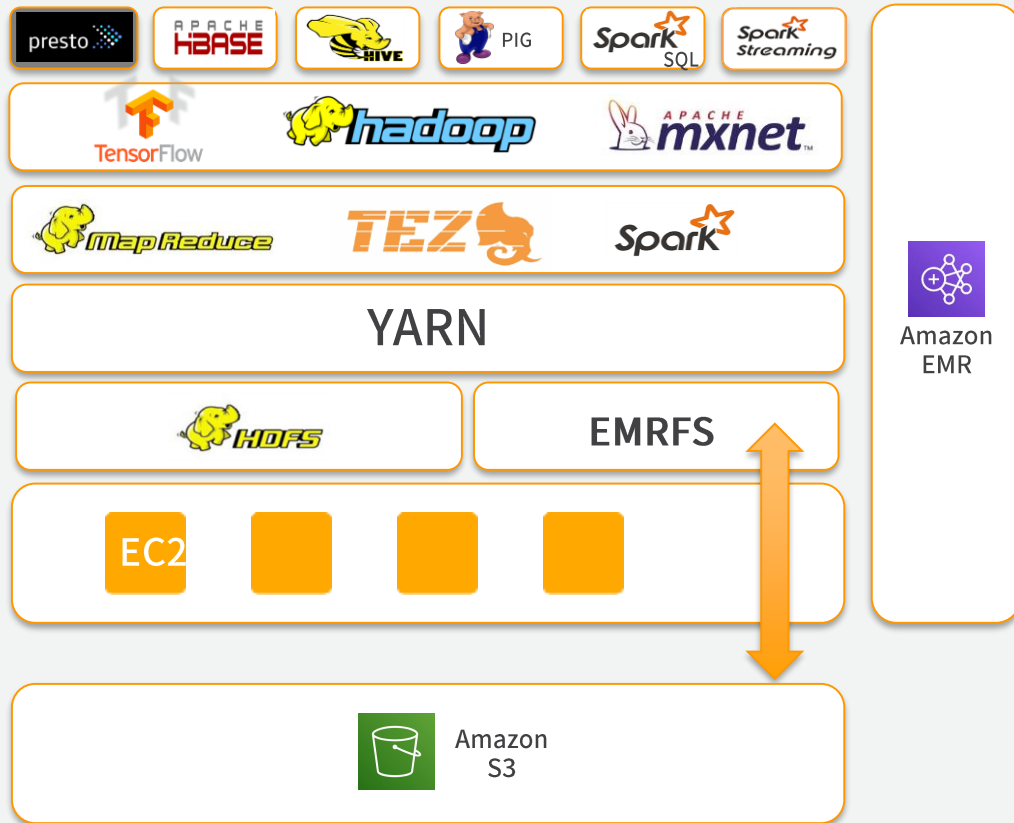


継続的な更新

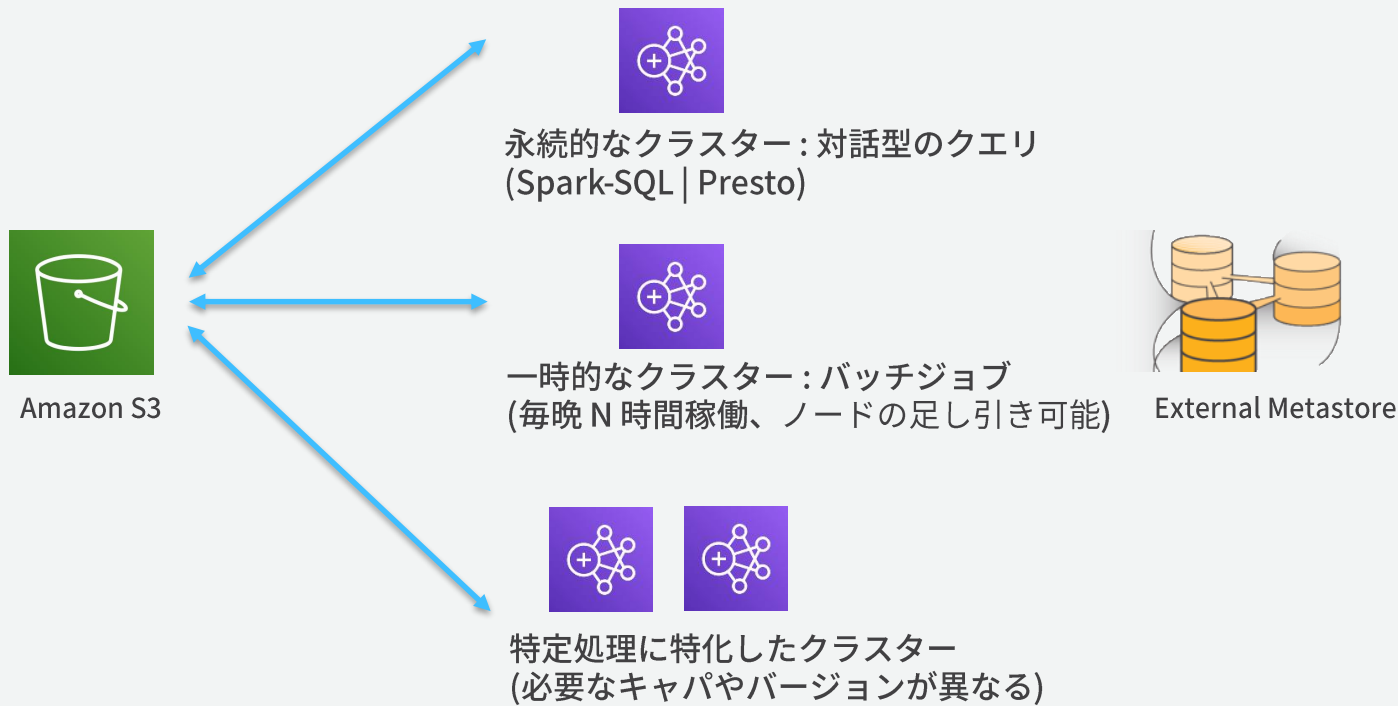


低コスト

Why Amazon EMR? : コンピューティングとストレージの分離



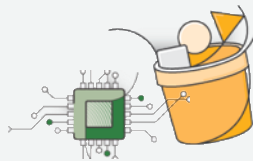
Why Amazon EMR? : コンピューティングとストレージの分離



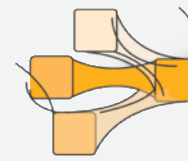
Why Amazon EMR?



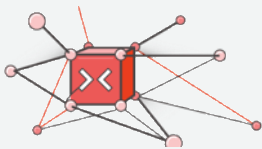
自動化



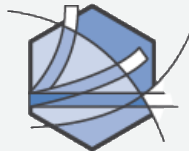
疎結合



弾力性



統合

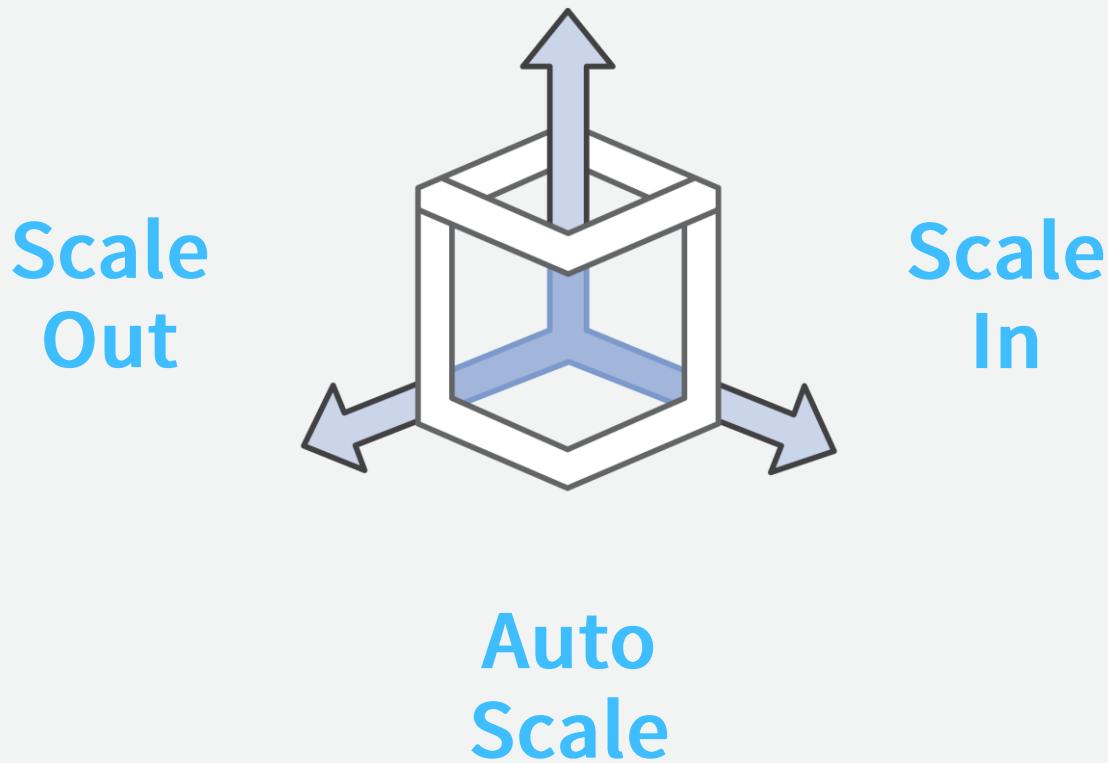


継続的な更新



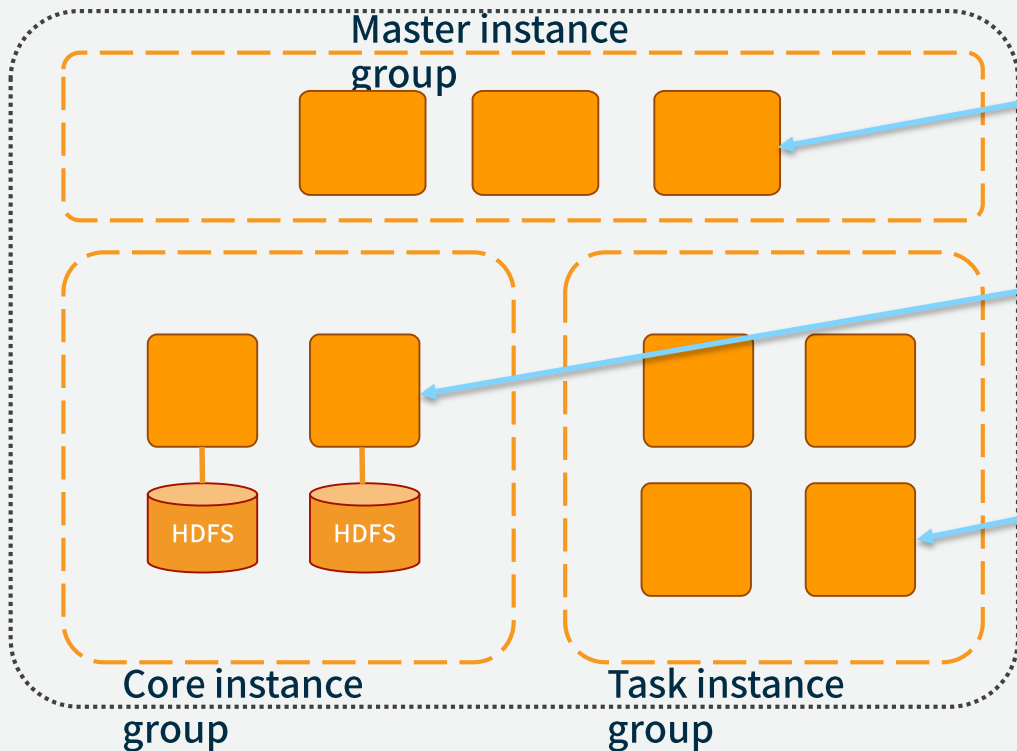
低コスト

Why Amazon EMR? 弾力性



EMR ノードの弾力性

EMR cluster



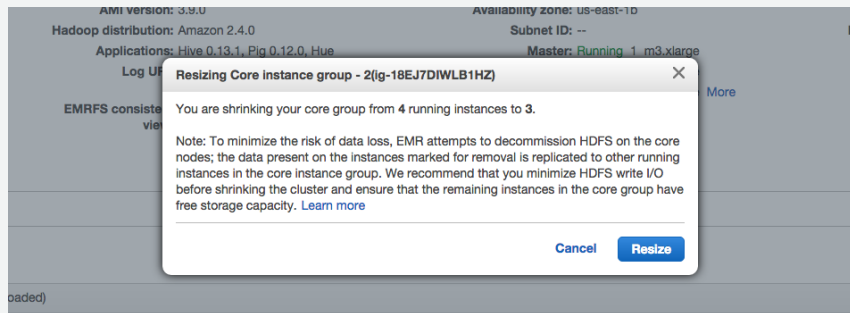
マスターノードは稼働し
続けねばならない

コアノードは Graceful に
追加および削除が可能

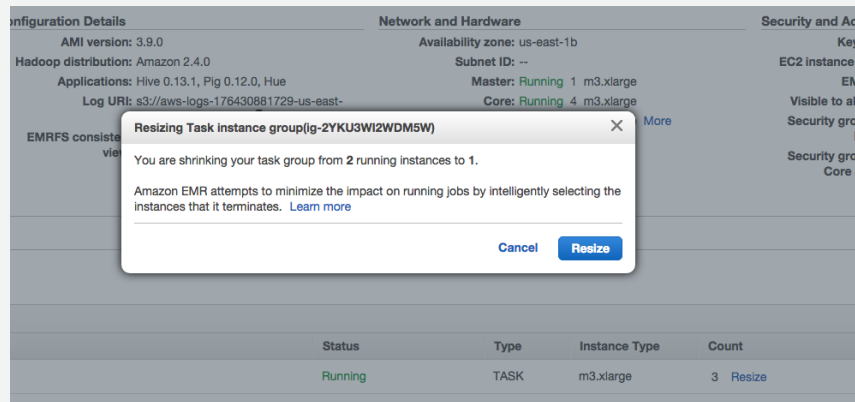
タスクノードのロストは、
クラスターに対する問題
なし

配慮されたスケールイン: Core ノードと Task ノード

Core



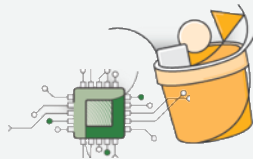
Task



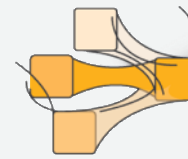
Why Amazon EMR?



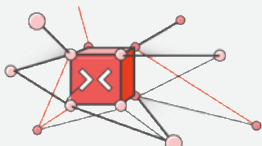
自動化



疎結合



弾力性



統合

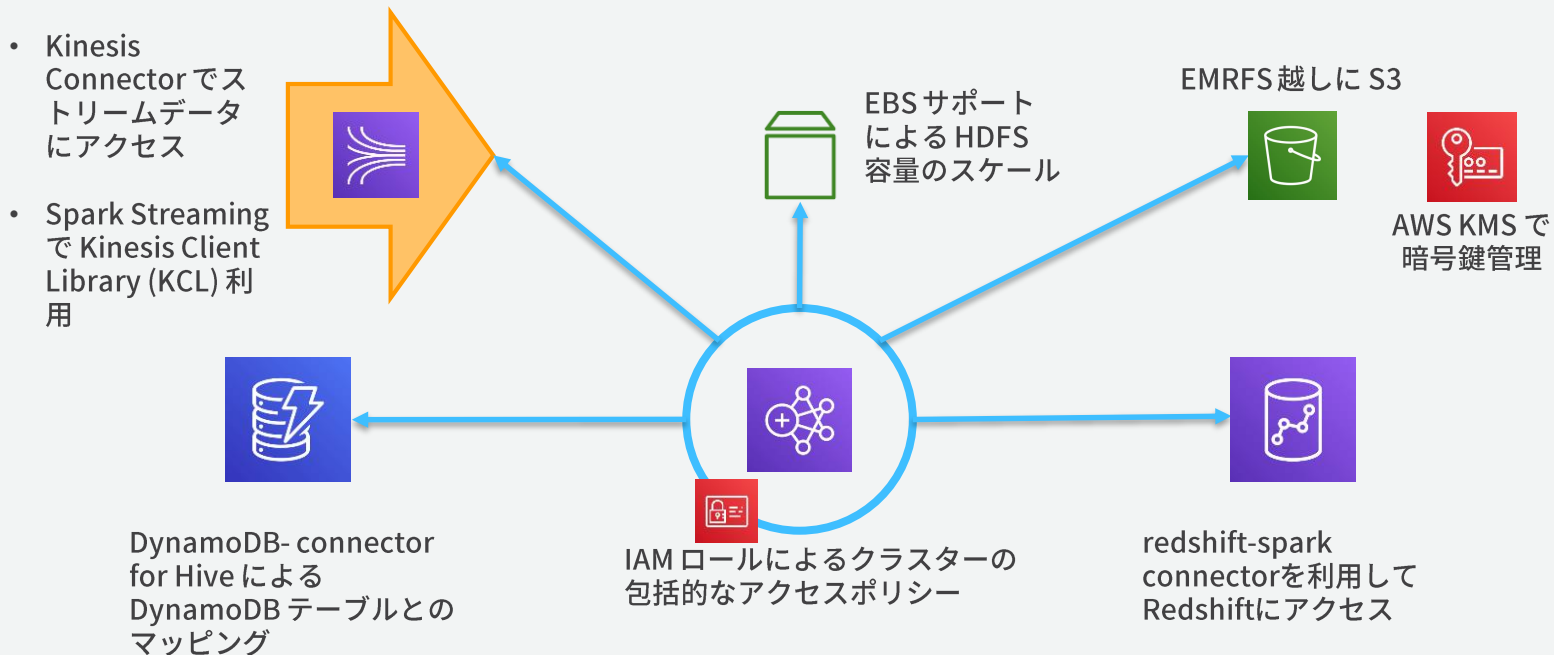


継続的な更新



低コスト

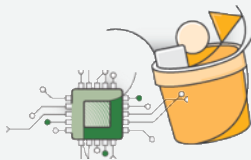
Why Amazon EMR? : AWS サービスとの統合



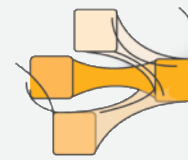
Why Amazon EMR?



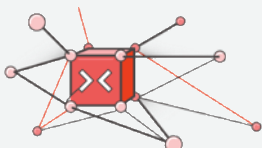
自動化



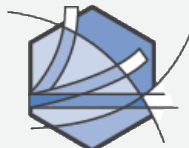
疎結合



弾力性



統合



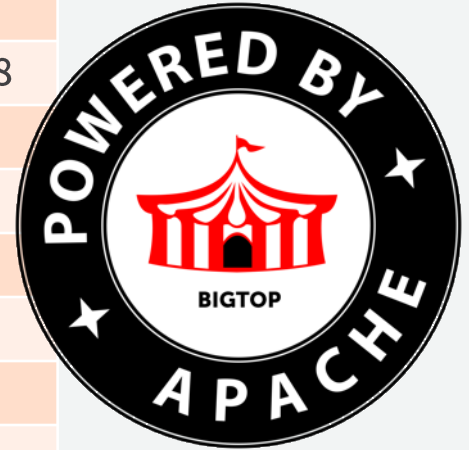
継続的な更新



低コスト

Why Amazon EMR? : 継続的な更新

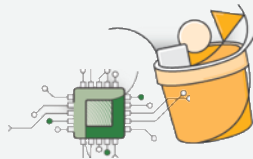
Application	Open source release	EMR release
Spark 2.4.4	Sep 01, 2019	Sep 23, 2019
Spark 2.4.3	May 08, 2019	July 17, 2019
Spark 2.4.2	April 23, 2019	June 26, 2019
Spark 2.4.0	November 2, 2018	December 18, 2018
...		
Spark 2.3.0	February 28, 2018	April 10, 2018
Spark 2.2.0	July 11, 2017	August 10, 2017
Spark 2.1.0	December 28, 2016	January 26, 2017
Spark 2.0	July 26, 2016	August 2, 2016
...		



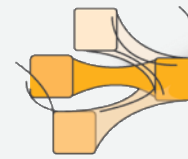
Why Amazon EMR?



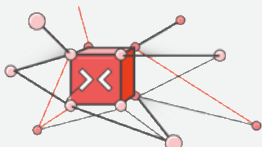
自動化



疎結合



弾力性



統合

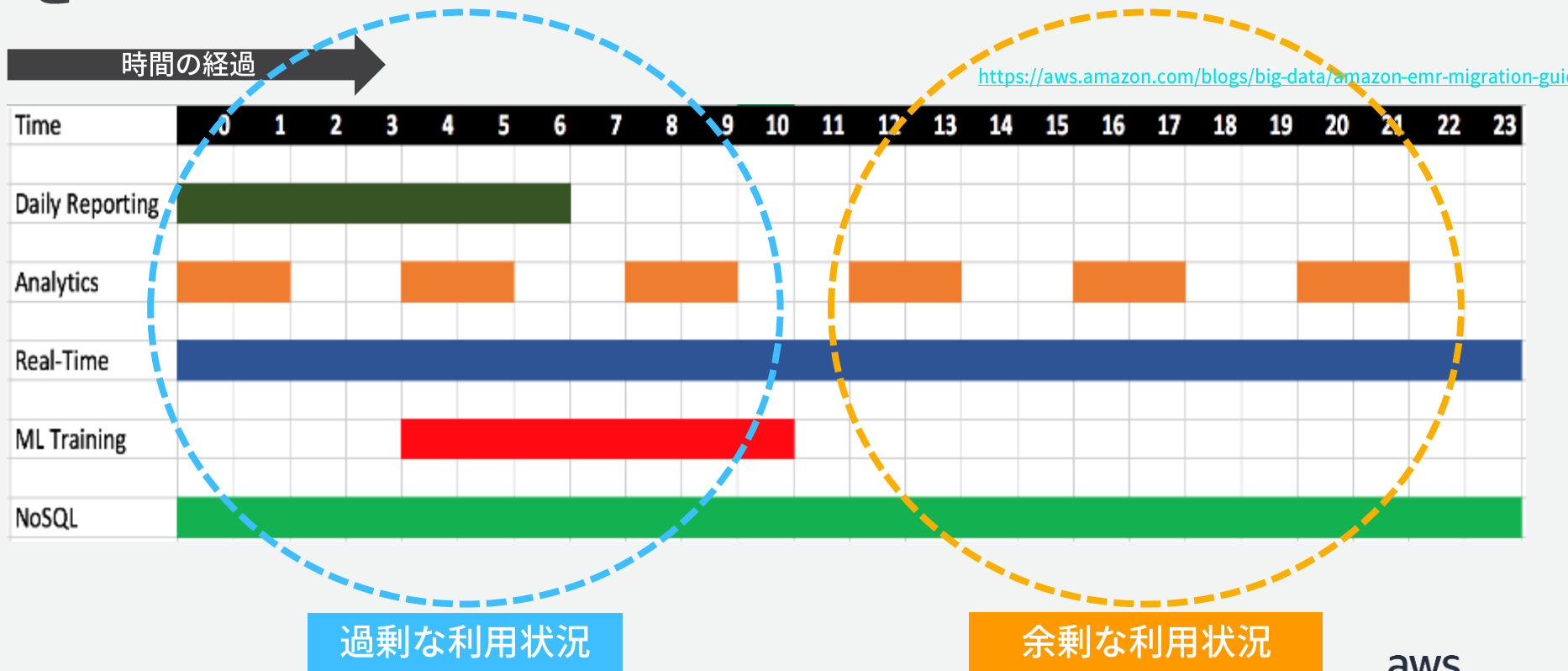


継続的な更新



低コスト

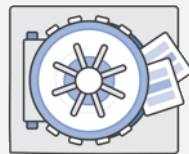
例えば、 オンプレでのジョブの実行状況をプールのレーンのように並べてみると



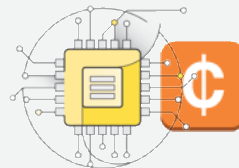
Why Amazon EMR? : 低コスト



一時的なクラスター



リザーブド
インスタンス



スポット
と
インスタンスフリート

スポットのマーケット特性について、おさらい

M5	1a	1c	1d	On Demand
24XL	\$1.712	\$1.677	\$1.677	\$5.952
12XL	\$0.838	\$0.839	\$0.838	\$2.976
4XL	\$0.320	\$0.286	\$0.279	\$0.992
2XL	\$0.152	\$0.147	\$0.140	\$0.496
XL	\$0.07	\$0.07	\$0.07	\$0.248

インスタンスファミリー

インスタンスサイズ

アベイラビリティゾーン (AZ)

リージョン

それぞれ個別にスポットマーケットがある

Spot Instance Advisor の活用

<https://aws.amazon.com/ec2/spot/instance-advisor/>

Spot Instance Advisor

Region: OS:

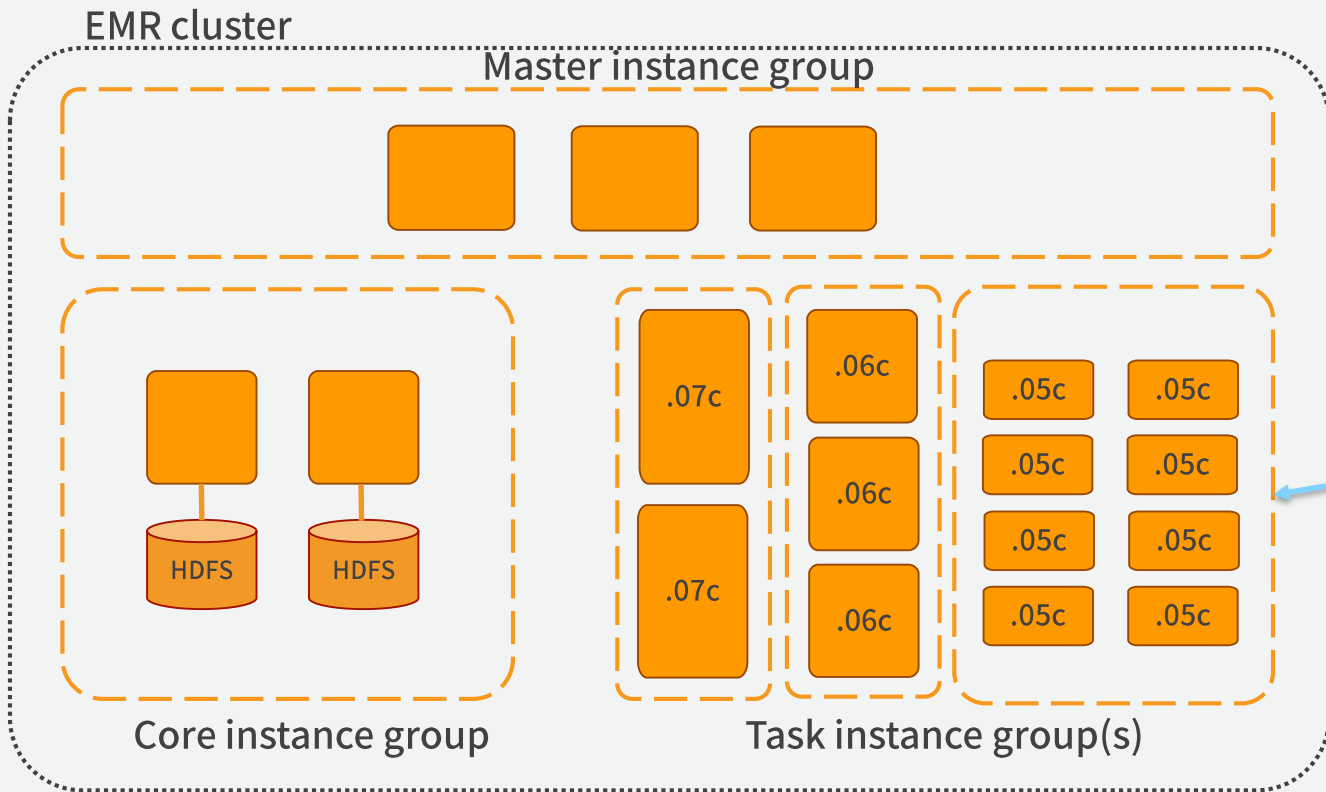
Instance type filter:

vCPU (min): Memory GiB (min):

Instance types supported by EMR

Instance Type	vCPU	Memory GiB	Savings over On-Demand [†]	Frequency of interruption ▾
i3.metal	72	512	70%	<5% □□□□□
r4.16xlarge	64	488	78%	<5% □□□□□
c5n.xlarge	4	10.5	76%	<5% □□□□□
m5d.24xlarge	96	384	76%	<5% □□□□□
m1.large	2	7.5	90%	<5% □□□□□
i3.xlarge	4	30.5	70%	<5% □□□□□
m3.xlarge	4	15	84%	<5% □□□□□
c3.8xlarge	32	60	76%	<5% □□□□□
m5a.24xlarge	96	384	69%	<5% □□□□□
r3.4xlarge	16	122	82%	<5% □□□□□

複数の Task グループ割り当て



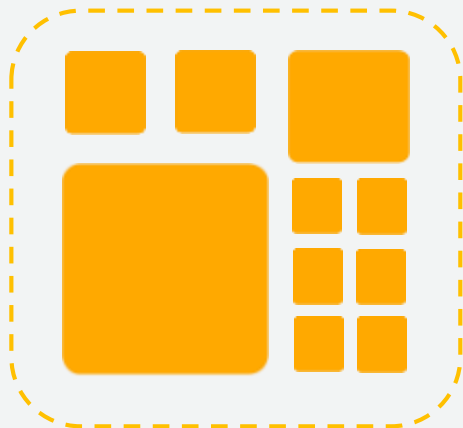
グループ毎に、異なる
インスタンスタイプや
スポット入札額を指定
できる

インスタンスフリートを利用したスポットのさらなる活用

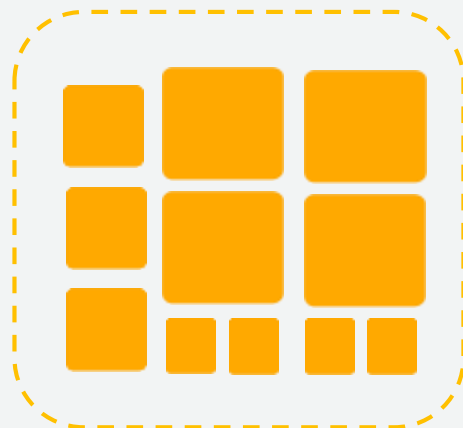
Master Node



Core Instance Fleet



Task Instance Fleet

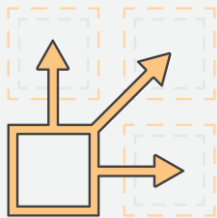


- 指定したインスタンスタイプのリストから、スポットとオンデマンドをプロビジョニング
- 容量/価格に基づいて最適なアベイラビリティゾーンで起動
- スポットブロックのサポート

本日のアジェンダ

- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

What is Amazon EMR



簡単

わずか数分でクラスターを起動



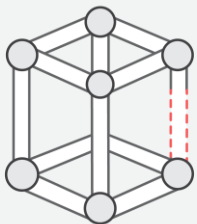
低コスト

秒単位の課金



豊富なオープンソース

最新バージョンのソフトウェア



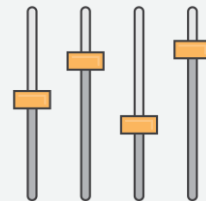
マネージド

監視に費やす労力を節減



安全

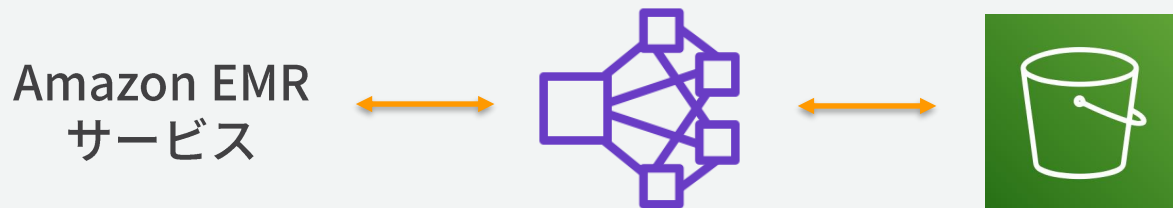
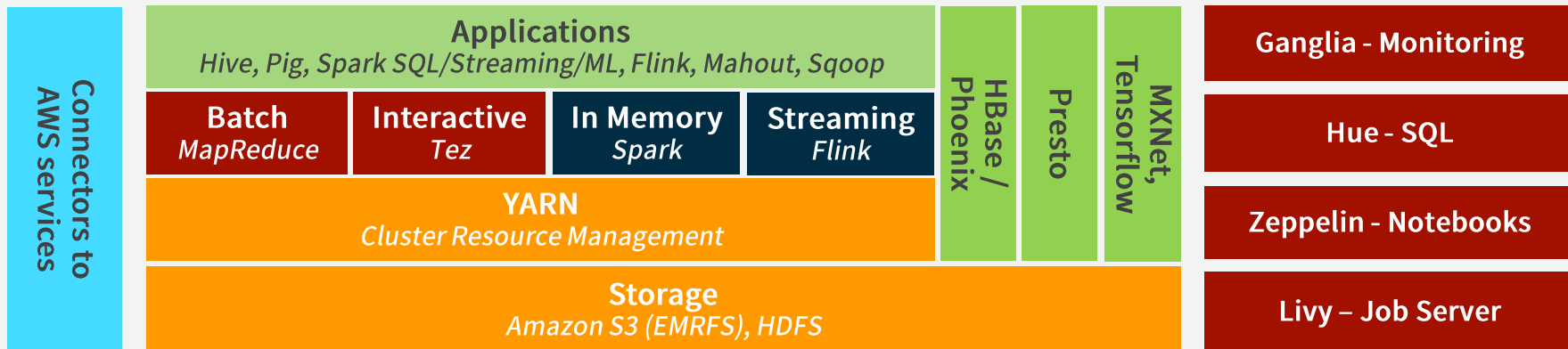
簡単にオプション設定可能



柔軟

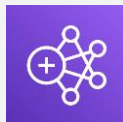
自在なカスタムと制御

オープンソースのアプリケーションたち



ジョブ送信の選択肢

Spark アプリケーション
の送信



Amazon EMR
Step API

AWS Lambda で、EMR Step
APIをコール、または、クラ
スター内の Spark を直接
コールして、アプリケー
ションを送信する

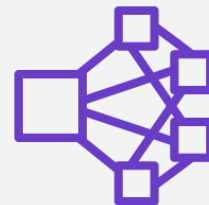
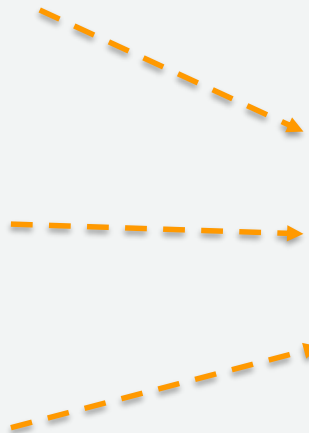


AWS Lambda

ジョブ送信のスケジューリ
ングや、複雑なワークフ
ローを定義するために、パイ
プラインを構成する



Airflow、Luigi、Digdag、その
他任意のスケジューラー on EC2

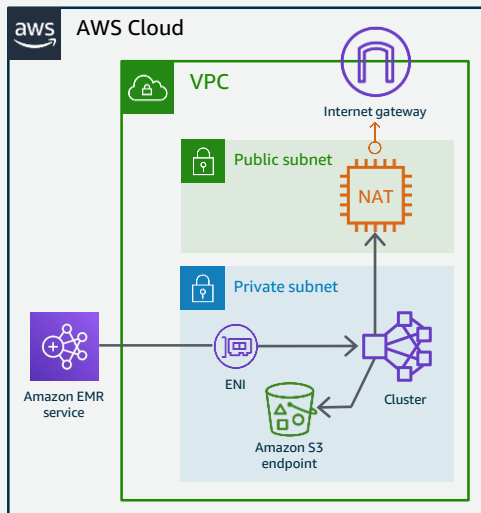


Amazon EMR

クラスター内の
Oozie を利用し
てジョブの DAG
を構成

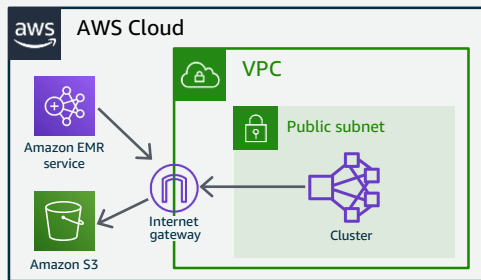
ネットワーク : VPC オプション

Private Subnet 配置



- S3 への接続に Amazon S3 エンドポイントを使用する
- 他のサービスまたはインターネットへの接続に、マネージド NAT を使用する
- セキュリティグループを使用してトラフィックを制御する
 - ElasticMapReduce-Master-Private
 - ElasticMapReduce-Slave-Private
 - ElasticMapReduce-ServiceAccess

Public Subnet 配置



- **Block Public Access** を有効にすると、すべてのパブリックアドレスからのインバウンドトラフィックを許可するルールがクラスターのセキュリティグループにある場合、EMR クラスターの起動を防止する

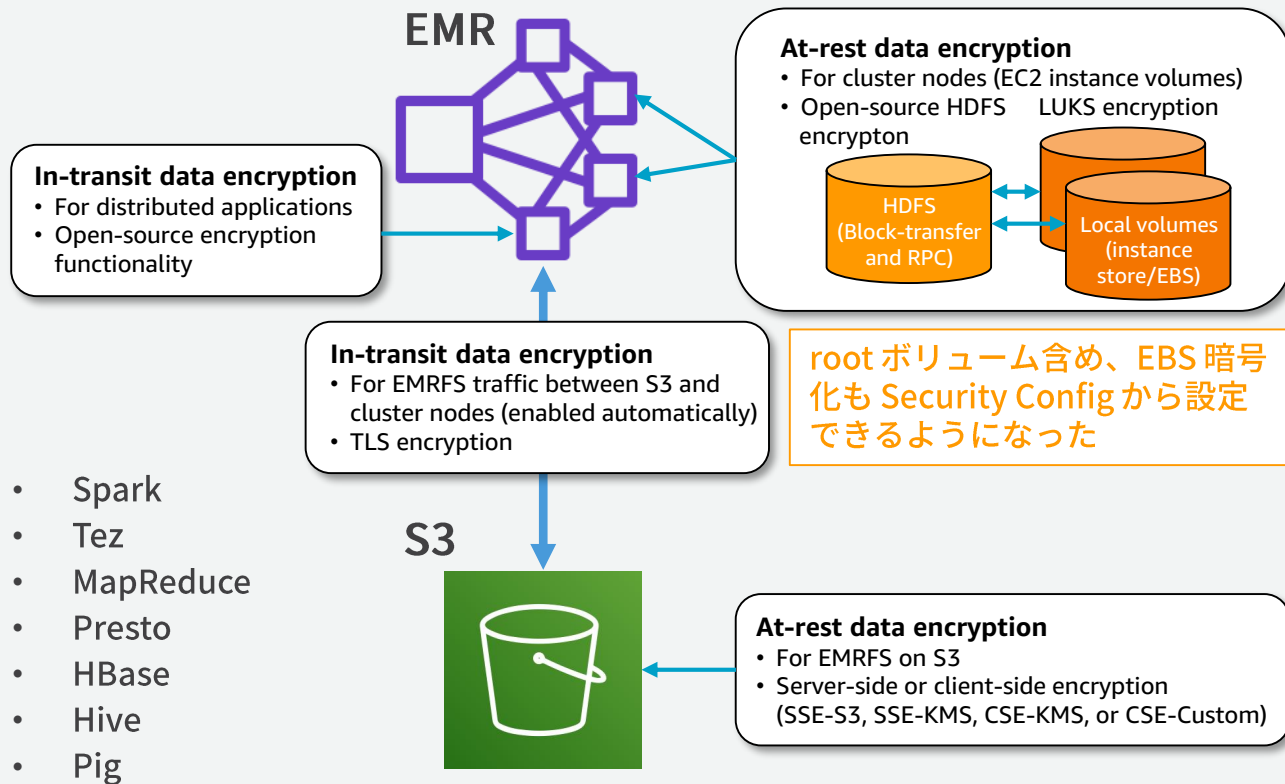
アクセス制御 : IAM Users and Roles

- Amazon EMR サービスにアクセスするための IAM ポリシー (IAM ユーザーや Federated ユーザー向け)
 - AmazonElasticMapReduceFullAccess
 - AmazonElasticMapReduceReadOnlyAccess
- Amazon EMR クラスターのための IAM ポリシー
 - サービスロール (AmazonElasticMapReduceRole) : EC2 インスタンスの作成など関連する AWS サービスにアクセスしてクラスターをプロビジョニングする、Amazon EMR サービスに許可されるアクション
 - インスタンスプロファイル (AmazonElasticMapReduceforEC2Role) : クラスターからの EMRFS 越しの Amazon S3 へのアクセスなど、Amazon EMR で実行されるアプリケーションに許可されるアクション
 - オートスケーリングロール (AmazonElasticMapReduceforAutoScalingRole) : クラスターのオートスケーリングがトリガーされた際に EC2 インスタンスの起動や終了を実施する、Amazon EMR サービスに許可されるアクション

容易な End-to-End のセキュリティ設定

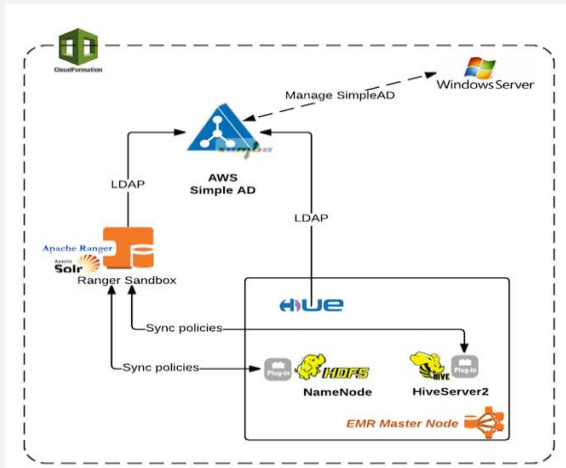
- 保管中のデータの暗号化
 - すべての Amazon S3 暗号化機能をサポート
 - ローカルディスクと HDFS 暗号化を設定可能
- 転送中のデータの暗号化
 - VPC プライベートサブネットでの EMR クラスター実行
 - Hadoop、MapReduce、Spark 用の暗号化されたノード間通信
 - SSL を介した他のサービスへのデータ転送
- AWS IAM との統合
 - IAM ロール、バケットポリシーと ACL、タグベースのアクセス許可のサポート
- ネイティブな Hadoop エコシステム機能による認証認可
- コンプライアンスおよび監査
 - SOC 1/2/3、PCI-DSS、FedRAMP、HIPAA、などに適合
 - CloudTrail で、すべての API 呼び出しをログに記録
 - S3 データのオブジェクトアクセスロギング

セキュリティ：暗号化



<https://aws.amazon.com/blogs/big-data/secure-amazon-emr-with-encryption/>

セキュリティ: 認証と認可

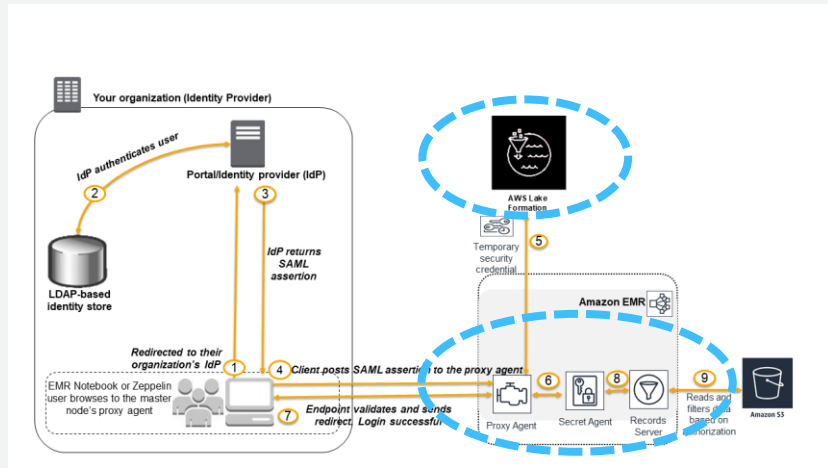


Apache

Ranger

Hive、HBase、YARN、およびHDFS用のプラグイン

- Hive の行レベルの認証
 - データマスキングを使用
- 組み込み検索による充実の監査機能
- エッジノードで Ranger を実行する



AWS Lake Formation との統合 [Beta]

- Spark、Zeppelin、EMR ノートブックをサポート
- AD FS などの既存の SAML ベースの IdP を使用した、企業ですでに利用されている ID による管理
- AWS Glue データカタログをメタデータストアとして使用
- EMR ノートブックまたは Apache Zeppelin を使用して、AWS Glue と Lake Formation によって管理されるデータにアクセス
- Lake Formation のアクセス許可に従い、AWS Glue データカタログのデータベース、テーブル、列にアクセス

Kerberos による認証



<https://aws.amazon.com/blogs/big-data/use-kerberos-authentication-to-integrate-amazon-emr-with-microsoft-active-directory/>

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



EMRFS きめ細かな (fine-grained) 認可

例

User: aduser
Group: analyst

IAM role: analytics_prod



例

User: aduser2
Group: dev

IAM role: analytics_dev



IAM ロールを、
ユーザー、グループ、または S3 プレフィックスにマッピング可能

<https://aws.amazon.com/blogs/big-data/securely-analyze-data-from-another-aws-account-with-emrfs/>

セキュリティ：統制と監査

- EMR API には AWS CloudTrail
- EMR クラスターからの S3 アクセスには S3 アクセスログ
- YARN とアプリケーションのログ
 - EMR の ロギング、デバッグ、アプリケーション履歴を活用
- Apache Ranger をアプリケーションレベルの監査用 UI に利用

監視

- Amazon S3
 - バケットアクセスログ
 - CloudTrail データイベント
 - CloudWatch メトリックス
- Amazon EMR
 - さまざまなログファイルを 5 分間隔で Amazon S3 にアーカイブする
 - ロギング、デバッグ、アプリケーション履歴の活用
 - ログファイルは、クラスタの終了後も利用可能
- CloudWatch メトリックス
 - クラスタ、ノード、YARN、ストレージ（S3 と HDFS）、メモリー、HBase
 - 5分ごとに更新
- Ganglia

AWS Glue データカタログを共通のメタデータストアとして適用可能

The screenshot shows the AWS Glue console interface. On the left is a navigation menu with options like 'Data catalog', 'Databases', 'Tables', 'Connections', 'Crawlers', 'Classifiers', 'ETL', 'Jobs', 'Triggers', 'Dev endpoints', and 'Tutorials'. The main area displays the configuration for a crawler job named '2015'. The configuration includes details such as 'Name', 'Description', 'Database', 'Classification', 'Location', 'Connection', 'Deprecated', 'Last updated', 'Input format', 'Output format', and 'Serde serialization lib'. Below this, 'Serde parameters' are listed, including 'paths' and 'sizeKey'. 'Table properties' are also shown, such as 'CrawlerSchemaSerializer/Version', 'recordCount', 'averageRecordSize', 'CrawlerSchemaDeserializer/Version', 'compressionType', and 'typeOfData'. At the bottom, a 'Schema' table is displayed with columns for 'Column name', 'Data type', and 'Key'. The schema table lists 8 columns: 'id' (string), 'type' (string), 'actor' (struct), 'repo' (struct), 'payload' (struct), 'public' (boolean), 'created_at' (string), and 'org' (struct). A large orange arrow points from the 'payload' and 'org' rows of the schema table to a detailed view window.

Column name	Data type	Key
1	id	string
2	type	string
3	actor	struct
4	repo	struct
5	payload	struct
6	public	boolean
7	created_at	string
8	org	struct

- Spark、Hive、および Prestoをサポート
- スキーマとパーティションの自動生成
- テーブル更新をマネージドで提供

The screenshot shows a dialog box titled 'payload schema details'. It displays a hierarchical JSON schema for the 'payload' column. The schema is a STRUCT containing several fields: 'ref:STRING', 'ref_type:STRING', 'master_branch:STRING', 'description:STRING', 'pusher_type:STRING', 'push_id:INT', 'size:INT', 'distinct_size:INT', 'head:STRING', 'before:STRING', 'commits:ARRAY', 'action:STRING', 'release:STRUCT', 'uri:STRING', 'assets_uri:STRING', and 'upload_uri:STRING'. The 'release' field is expanded to show its sub-structure. A 'Close' button is located at the bottom of the dialog.

本日のアジェンダ

- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

EMR クラスター作成手段の選択肢



AWS
マネージメント
コンソール



AWS CLI



AWS SDKs



Compute

- EC2
- Lightsail [↗](#)
- ECR
- ECS
- EKS
- Lambda
- Batch
- Elastic Beanstalk
- Serverless Application Repository



Storage

- S3
- EFS
- FSx
- S3 Glacier
- Storage Gateway
- AWS Backup



Database

- RDS
- DynamoDB



Robotics

- AWS RoboMaker



Blockchain

- Amazon Managed Blockchain



Satellite

- Ground Station



Management & Governance

- AWS Organizations
- CloudWatch
- AWS Auto Scaling
- CloudFormation
- CloudTrail
- Config
- OpsWorks
- Service Catalog
- Systems Manager
- Trusted Advisor
- Managed Services



Analytics

- Athena
- EMR
- CloudSearch
- Elasticsearch Service
- Kinesis
- QuickSight [↗](#)
- Data Pipeline
- AWS Glue
- MSK



Security, Identity, & Compliance

- IAM
- Resource Access Manager
- Cognito
- Secrets Manager
- GuardDuty
- Inspector
- Amazon Macie [↗](#)
- AWS Single Sign-On
- Certificate Manager
- Key Management Service
- CloudHSM



Business Applications

- Alexa for Business
- Amazon Chime [↗](#)
- WorkMail



End User Computing

- WorkSpaces
- AppStream 2.0
- WorkDocs
- WorkLink



Internet Of Things

- IoT Core
- Amazon FreeRTOS
- IoT 1-Click
- IoT Analytics
- IoT Device Defender
- IoT Device Management
- IoT Events
- IoT Greengrass
- IoT SiteWise
- IoT Things Graph

クラスターの作成 : マネージメントコンソール

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

Create cluster

View details

Clone

Terminate

Filter:

All clusters



Filter clusters ...

No clusters found



Name

ID

Status

Creation

Create Cluster - Quick Options

[Go to advanced options](#)

General Configuration

Cluster name

Logging ⓘ

S3 folder 📁

Launch mode Cluster ⓘ Step execution ⓘ

Software configuration

Release ⓘ

Applications

- Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.5, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- HBase: HBase 1.4.10 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.5, Hue 4.4.0, Phoenix 4.14.2, and ZooKeeper 3.4.14
- Presto: Presto 0.224 with Hadoop 2.8.5 HDFS and Hive 2.3.5 Metastore
- Spark: Spark 2.4.4 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.1

Use AWS Glue Data Catalog for table metadata ⓘ

Hardware configuration

Instance type ⓘ The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#) 📄

Number of instances (1 master and 2 core nodes)

Security and access

EC2 key pair ⓘ [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](#) ⓘ

詳細
オプション

クイックに
作成

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

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input type="checkbox"/> Presto 0.224 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input type="checkbox"/> Oozie 5.1.0 |
| <input type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.14.0 |

Multi-master support

- Enable multi-master support ?

AWS Glue Data Catalog settings (optional)

- Use for Hive table metadata ?

Edit software settings ?

- Enter configuration Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Add steps (optional) ?

Step type ?

- Auto-terminate cluster after the last step is completed

Cancel

Next

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release  

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input type="checkbox"/> Presto 0.224 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input type="checkbox"/> Oozie 5.1.0 |
| <input type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.14.0 |

Multi-master support

Enable multi-master support 

AWS Glue Data Catalog settings (optional)

Use for Hive table metadata 

Edit software settings

Enter configuration Load JSON from S3

Add steps (optional)

Step type 

Auto-terminate cluster after the last step is completed

Cancel

Next

EMR マルチマスター

- EMR **マルチマスター**を選択すると、3つのマスターノードを持つクラスターが作成される
- **マルチマスター**機能により、HBase、YARN リソースマネージャー、HDFS ネームノード、Spark、Hive、および Ganglia の高可用性をサポートする
- **マルチマスター**機能により、プライマリマスターノードに障害が発生した場合、Amazon EMR は自動的にスタンバイマスターノードにフェイルオーバーする

Create Cluster - Advanced Options

[Go to quick options](#)

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release

- | | | |
|--|--|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input checked="" type="checkbox"/> Presto 0.224 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input type="checkbox"/> Oozie 5.1.0 |
| <input checked="" type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.14.0 |

Multi-master support

Enable multi-master support

AWS Glue Data Catalog settings (optional)

- Use for Hive table metadata
- Use for Presto table metadata
- Use for Spark table metadata

Edit software settings

Enter configuration Load JSON from S3

Add steps (optional)

Step type

Auto-terminate cluster after the last step is completed

Glue データカタログの設定

- Glue データカタログを Hive、Spark、および Presto でサポートされている外部メタストアとして指定できる
- 永続的なメタストア、または、異なるクラスター、サービス、アプリケーション、さらに、AWS アカウント、によって共有されるメタストアが必要な場合に、非常に便利なオプション

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release  

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input type="checkbox"/> Presto 0.224 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input type="checkbox"/> Oozie 5.1.0 |
| <input type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.14.0 |

Multi-master support

Enable multi-master support 

AWS Glue Data Catalog settings (optional)

Use for Hive table metadata 

Edit software settings

Enter configuration Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

Add steps (optional)

Step type 

Auto-terminate cluster after the last step is completed

Cancel

Next

ソフトウェア設定の編集

- Hadoop アプリケーションのデフォルト設定を上書きできる
- S3 に配置したオブジェクトから、またはインラインの JSON ファイル
- キャパシティスケジューラー、core-site、hadoop-env、hadoop-log4j、hdfs、httpfs-env、https-site、maapred-env、pred-site、yarn-env、yarn-site、hive-ec-log4j、hive-log4j、hive-site、pig-properties、pig-log4j、などなど
- さらに、クラスター作成後も、ソフトウェア設定を再構成し、実行中のクラスター内の各インスタンスグループに設定を追加・更新できるようになった

<https://aws.amazon.com/blogs/big-data/modifying-your-cluster-on-the-fly-with-amazon-emr-reconfiguration/>

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Software Configuration

Release  

- | | | |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input type="checkbox"/> Zeppelin 0.8.1 | <input type="checkbox"/> Livy 0.6.0 |
| <input type="checkbox"/> JupyterHub 1.0.0 | <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Flink 1.8.1 |
| <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 1.4.10 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 2.3.5 | <input type="checkbox"/> Presto 0.224 | <input type="checkbox"/> ZooKeeper 3.4.14 |
| <input type="checkbox"/> MXNet 1.4.0 | <input type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> Mahout 0.13.0 |
| <input checked="" type="checkbox"/> Hue 4.4.0 | <input type="checkbox"/> Phoenix 4.14.2 | <input type="checkbox"/> Oozie 5.1.0 |
| <input type="checkbox"/> Spark 2.4.4 | <input type="checkbox"/> HCatalog 2.3.5 | <input type="checkbox"/> TensorFlow 1.14.0 |

Multi-master support

Enable multi-master support 

AWS Glue Data Catalog settings (optional)

Use for Hive table metadata 

Edit software settings

Enter configuration Load JSON from S3

Add steps (optional)

Step type 

Auto-terminate cluster after the last step is completed

Cancel

Next

ステップの追加

- ステップは、クラスターに送信するジョブの単位
- 直列に実行される
- 完了時にクラスターを自動終了するように選択も可能
- ストリーミング、Hive、Pig、Spark、およびカスタム JAR の事前設定オプションがある

Add steps (optional) ⓘ

Step type Auto-t

- ✓ Select a step
- Streaming program
- Hive program
- Pig program
- Spark application
- Custom JAR

Configure

ハードウェアの設定

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration ⓘ

If you need more than 20 EC2 instances, see [this topic](#) ⓘ.

Instance group configuration

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

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

Network vpc-2d04ce49 (172.31.0.0/16) (default) | default ⓘ [Create a VPC](#) ⓘ ⓘ

EC2 Subnet subnet-87b244f1 | Default in us-west-2a ⓘ

Root device EBS volume size 10 GiB ⓘ

Choose the instance type, number of instances, and a purchasing option. You can choose to use On-Demand Instances, Spot Instances, or both. The instance type and purchasing option apply to all EC2 instances in each instance group, and you can only specify these options for an instance group when you create it. [Learn more about instance purchasing options](#) ⓘ

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	1 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ⓘ	Not available for Master ⓘ
Core Core - 2 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	2 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ⓘ	Not enabled ⓘ
Task ✕ Task - 3 ⓘ	m5.xlarge ⓘ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB ⓘ ⓘ Add configuration settings ⓘ	0 Instances	<input checked="" type="radio"/> On-demand ⓘ <input type="radio"/> Spot ⓘ Use on-demand as max price ⓘ	Not enabled ⓘ

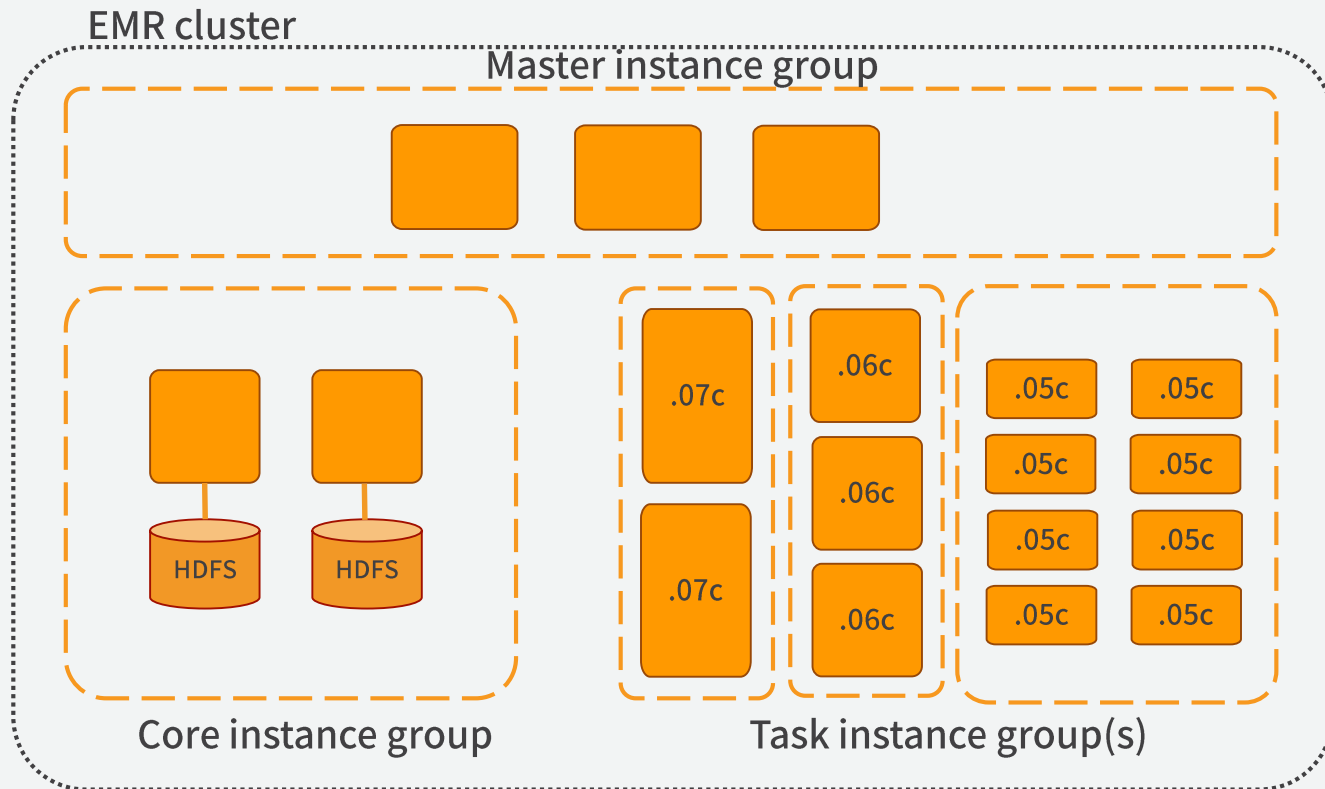
[+ Add task instance group](#)

Cancel

Previous

Next

ハードウェアの設定：複数の Task グループ



ハードウェア設定 : オートスケーリング

group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not available for Master
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled

[+ Add task instance group](#)

オートスケーリング

Auto Scaling rules



Maximum instances:

Minimum instances:

Scale out

Default-scale-out-1: Add instance if YARNMemoryAvailablePercentage is less than for five-minute period with a cooldown of seconds

Default-scale-out-2: Add instance if ContainerPendingRatio is greater than for five-minute period with a cooldown of seconds

[+ Add rule](#)

Scale in

Default-scale-in: Terminate instance if YARNMemoryAvailablePercentage is greater than for five-minute period with a cooldown of seconds

[+ Add rule](#)

Done

ハードウェア設定：スポットインスタンスの利用

group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not available for Master
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	2 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not enabled

+ Add task instance group

ハードウェア設定：スポットインスタンスの利用

group when you create it. [Learn more about instance purchasing options](#)

Node type	Instance type	Instance count	Purchasing option	Auto Scaling
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot Use on-demand as max price	Not available for Master
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	2 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Use on-demand as max price	Not enabled
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	<input type="radio"/> On-demand <input checked="" type="radio"/> Spot Use on-demand as max price	Not enabled

+ Add task instance group

Current spot price

Availability zone	Price
us-west-2a	\$0.084
us-west-2b	\$0.081
us-west-2c	\$0.077
us-west-2d	\$0.069 lowest

現在の
Spot 落札価格

インスタンスフリート：もうひとつのハードウェア設定方法

- スポットとオンデマンドを組み合わせ、指定したインスタンスタイプのリストからプロビジョニングする
- 容量/価格に基づいて、最適なアベイラビリティゾーンで起動する
- スポットブロックもサポート

Instance group configuration



Uniform instance groups

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



Instance fleets

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

Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration ?

If you need more than 20 EC2 instances, [see this topic](#).

Instance group configuration **Uniform instance groups**
Specify a single instance type and purchasing option for each node type.

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

Network vpc-2d04ce49 (172.31.0.0/16) (default) | default Create a VPC ?

EC2 Subnet **Public**

Root device EBS volume size GiB ?

For each fleet, specify up to five instance types. For core and task fleets, enter target capacities for on-demand and spot instances. Amazon EMR launches instances from among the types you specify to fulfill the targets. For the master fleet, the target is always one. For each instance type, choose a maximum spot price. The advanced Spot options for each fleet determine Spot provisioning behavior.

Node type	Fleet instance types	Target capacity	Advanced Spot options	Auto Scaling
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Add / remove instance types to fleet	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot <i>The master fleet consists of one EC2 instance</i>		Not available for instance fleets
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Each instance counts as 4 units Add / remove instance types to fleet	<input type="text" value="0"/> On-demand units <input type="text" value="0"/> Spot units 0 Total units	Defined duration Not set Provisioning timeout Terminate cluster after 60 min. of Spot unavailability	Not available for instance fleets
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Each instance counts as 4 units Add / remove instance types to fleet	<input type="text" value="0"/> On-demand units <input type="text" value="0"/> Spot units 0 Total units	Defined duration Not set Provisioning timeout Terminate cluster after 60 min. of Spot unavailability	Not available for instance fleets

ネットワーク
VPC と (複数の) Subnet

ノードタイプ、
インスタンスの
タイプと数



Step 1: Software and Steps

Step 2: Hardware

Step 3: General Cluster Settings

Step 4: Security

Hardware Configuration ?

If you need more than 20 EC2 instances, [see this topic](#).

- Instance group configuration**
- Uniform instance groups**
Specify a single instance type and purchasing option for each node type.
 - Instance fleets**
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#)

Network vpc-2d04ce49 (172.31.0.0/16) (default) | default [Create a VPC](#)

EC2 Subnet Public
 subnet-87b244f1 | Default in us-west-2a
 subnet-ca8f74e1 | Default in us-west-2d
 subnet-df7d7e86 | Default in us-west-2c

Size 10 GiB

For core and task fleets, enter target capacities for on-demand and spot instances. Amazon EMR fulfills the target capacity for each instance type you specify to fulfill the targets. For the master fleet, the target is always one. For each instance type, the target capacity and the advanced Spot options for each fleet determine Spot provisioning behavior. [Learn more](#)

Node type	Fleet instance types	Target capacity	Advanced Spot options	Auto Scaling
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Add / remove instance types to fleet	<input checked="" type="radio"/> On-demand <input type="radio"/> Spot <i>The master fleet consists of one EC2 instance</i>		Not available for instance fleets
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Each instance counts as 4 units Add / remove instance types to fleet	<input type="checkbox"/> On-demand units <input type="checkbox"/> Spot units 0 Total units	Defined duration Not set Provisioning timeout Terminate cluster after 60 min. of Spot unavailability	Not available for instance fleets
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Maximum Spot price: % On-Demand 100 Each instance counts as 4 units Add / remove instance types to fleet	<input type="checkbox"/> On-demand units <input type="checkbox"/> Spot units 0 Total units	Defined duration Not set Provisioning timeout Terminate cluster after 60 min. of Spot unavailability	Not available for instance fleets

インスタンスタイプごとの
ユニット数

フリートごとの
合計ユニット数

General Options and Tags

General Options

Cluster name

Logging 

S3 folder 

Debugging 

Termination protection 

Tags

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

全般的なオプション: ロギング

- デフォルトで、マスターノードに書き込まれるログ: /mnt/var/log
 - ステップのログ
 - Hadoop および YARN コンポーネントのログ
 - ブートストラップアクションのログ
 - インスタンス状態のログ
- ロギングがオンになっている場合、ログは S3 にも書き込まれる
 - 5 分間隔で書き込む
 - コンソールではデフォルトで ON
 - CLI ではデフォルトで OFF

全般的なオプション: デバッグ

- クラスタでデバッグを有効にすると、Amazon EMR はログファイルを Amazon S3 にアーカイブするのに加えて、それらのファイルのインデックスを作成する
- 有効にした後、コンソールを使用して、クラスタのステップ、ジョブ、タスク、およびタスク試行ログを参照できる
- デバッグ用のログも、5 分間隔で S3 にプッシュされる

全般的なオプション: 追加オプション

Additional Options

EMRFS consistent view ⓘ

EMRFS metadata store

EmrFSMetadata ⓘ

Number of retries


5 ⓘ

Retry period (in seconds)

10 ⓘ

Custom AMI ID ⓘ

▼ Bootstrap Actions

Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. [Learn more](#) 

Add bootstrap action ⓘ

追加オプション: EMRFS Consistent View

- S3 は結果整合性を持つ、そこで、EMRFS Consistent View
 - DynamoDB をファイルレジストリとして使用する
 - EMRFS によって書き込まれた、または EMRFS と同期された Amazon S3 オブジェクトについて、EMR クラスターがリストと書き込み後の読み取りの一貫性をチェックできるようにする
- 以下の項目を調整できる
 - 矛盾を検出した後、EMFRS が S3 を呼び出す回数
 - 最初の再試行までの時間間隔
 - その後の再試行では、指数 (Exponential) バックオフが使用される

追加オプション: カスタム AMI

- メリット

- ブートストラップアクションを使用する代わりに、アプリケーションをプリインストールするなどカスタマイズを事前に行うことで、クラスタの開始時間を短縮できる
- 予期しないブートストラップアクションでのエラーを防止する
- ~~Amazon EBS ルートボリューム暗号化のサポート~~
 - root ボリューム含め、EBS の暗号化も Security Configuration から設定できるようになった

- 必須要件

- アマゾン Linux AMI である (Amazon Linux 2 AMI はサポートされていない)
- HVM および EBS-Backed の AMI である
- 64 ビット AMI である
- アプリケーションと同じ名前のユーザーを持つことはできない (例: hadoop、hdfs、yarn、spark など)

<https://aws.amazon.com/blogs/big-data/create-custom-amis-and-push-updates-to-a-running-amazon-emr-cluster-using-amazon-ec2-systems-manager/>

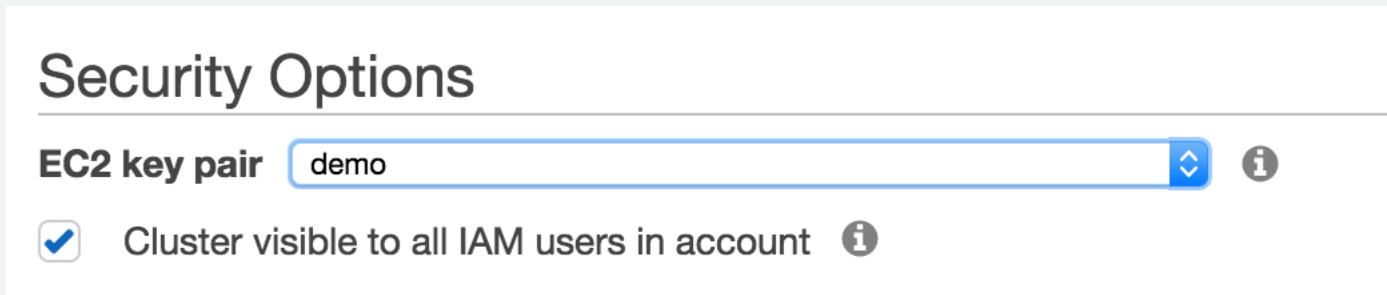
追加オプション: ブートストラップアクション

- Hadoop が各ノードで開始する前に実行されるスクリプト
- 通常、追加のソフトウェアのインストールに使用
- 最大16のブートストラップアクションを実行可能

- 条件指定で実行
 - instance.json または job-flow.json に定義されたインスタンス固有の値に対する条件
 - 例: IsMaster=true

- 柔軟なカスタム
 - 任意のカスタムスクリプトを実行する
 - 例: S3から各ノードにファイルをコピーする

セキュリティオプション - EC2 キーペアと可視性



- EC2 キーペア
 - マスターノードに SSH 接続できるようにするには、キーペアをアタッチする必要がある
- クラスタ表示
 - OFF にすると、クラスター作成者（と root ユーザー）のみが、CLI とコンソールでクラスタを表示できるようになる

セキュリティオプション：権限

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](#) ⓘ

Auto Scaling role [EMR_AutoScaling_DefaultRole](#) ⓘ

- EMR はロールを使用して AWS リソースにアクセスする
 - EMR ロール
 - EMR が EC2 などのリソースにアクセスできるようにする
 - EC2 インスタンスプロファイル
 - クラスター内の EC2 インスタンスが S3 などのリソースにアクセスすることを許可する
 - オートスケーリングロール
 - オートスケーリングによるインスタンスの追加と終了を許可する
- デフォルトのロールを使用するか、独自のロールを選択する



セキュリティオプション: 暗号化とセキュリティグループ

▼ Encryption Options

Security configuration ⓘ

▼ EC2 Security Groups

An EC2 security group acts as a virtual firewall for your cluster nodes to control inbound and outbound traffic. There are two types of security groups you can configure, [EMR managed security groups](#) and [additional security groups](#). EMR will [automatically update](#) the rules in the EMR managed security groups in order to launch a cluster. [Learn more](#).

Type	EMR managed security groups EMR will automatically update the selected group	Additional security groups EMR will not modify the selected groups
Master	<input type="text" value="Default: sg-f712b29a (ElasticMapReduce-master)"/>	No security groups selected 
Core & Task	<input type="text" value="Default: sg-f112b29c (ElasticMapReduce-slave)"/>	No security groups selected 

[Create a security group](#)

クラスター作成の前に
作成しておく必要がある

Amazon EMR

Clusters

Security configurations

Block public access

VPC subnets

Events

Notebooks

Help

What's new

security configuration

Name

At-rest encryption

Enable and choose options for at-rest data encryption features in Amazon EMR, including Amazon S3 with EMRFS, local volumes attached to cluster instances, and block-transfer encryption for HDFS. [Learn more](#)

S3 encryption ⓘ

Encryption mode ⓘ

AWS KMS Key ⓘ

Local disk encryption ⓘ

Key provider type

AWS KMS Key ⓘ

In-transit encryption

Enable and choose options for open-source encryption features that apply to in-transit data for specific applications. Available encryption options may vary by Amazon EMR release. [Learn more](#)

TLS certificate provider

Certificate provider type ⓘ

S3 object ⓘ

Authentication

Kerberos

Enable Kerberos authentication for interactions between certain application components on your cluster using Kerberos principals. You can choose between having EMR install a KDC server on the master node of the cluster or you can share your own KDC details that EMR cluster can use. [Learn more](#)

Provider Cluster dedicated KDC External KDC

Ticket lifetime hours

Admin server

KDC server

Active Directory integration

内部か、外部か、
KDC の配置を指定する

IAM roles for EMRFS

Use IAM roles for EMRFS requests to Amazon S3

When an Amazon S3 request is made through EMRFS, each **Basis for access** is evaluated in order. EMRFS assumes the corresponding **IAM role** for the first match. Specify the cluster **Users** or **Groups**, or **S3 prefixes** as the **Basis for access**. If no **Basis for access** matches the request, EMRFS uses the cluster's EMR role for EC2. [Learn more](#)

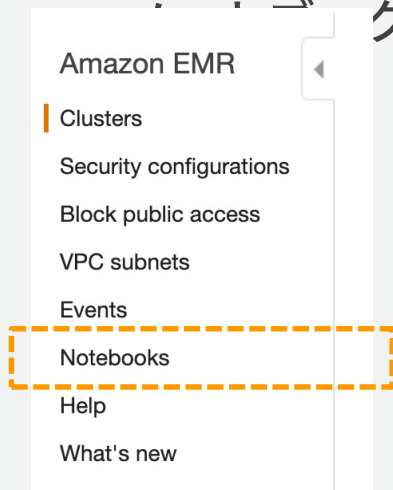
IAM role	Basis for access
<input type="text" value="Admin"/>	<input type="text" value="Users"/> <input type="text" value="Enter user names, separated by comma"/>

+ Add role mapping

EMR ノートブック

- Jupyterをベースとしており、データをインタラクティブに分析するためにある
- ノートブックを作成し、Hadoop、Spark、および Livy を実行している Amazon EMR クラスターにアタッチする

ノートブックへの記述内容は、クラスターとは別に Amazon S3 に保存される



A screenshot of the Amazon EMR console 'Notebooks' page. The page title is 'Notebooks'. Below the title is a description: '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](#)'. Below the description are buttons: 'Create notebook' (highlighted with a dashed orange box), 'View details', 'Open', 'Start', 'Stop', and 'Delete'. Below the buttons is a filter section: 'Filter: All notebooks x' and 'Showing 0 of 1 notebooks'. Below the filter is a table with columns: Name, Status, Cluster, Creation time (UTC-4), and Last modified. The table is currently empty, with 'Showing 0 of 1 notebooks' displayed below it.

<https://aws.amazon.com/blogs/big-data/emr-notebooks-a-managed-analytics-environment-based-on-jupyter-notebooks/>

<https://aws.amazon.com/blogs/big-data/install-python-libraries-on-a-running-cluster-with-emr-notebooks/>

クラスターの詳細

[Clone](#)[Terminate](#)[AWS CLI export](#)

Cluster: My cluster **Waiting** Cluster ready after last step completed.

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

Connections: [Enable Web Connection](#) – Spark History Server, Resource Manager ... (View All)

Master public DNS: ec2-3-113-248-228.ap-northeast-1.compute.amazonaws.com [SSH](#)

Tags: -- [View All / Edit](#)

Summary

ID: j-33UGWUVJGUQG7

Creation date: 2019-10-14 07:01 (UTC+9)

Elapsed time: 7 hours, 55 minutes

Auto-terminate: No

Termination protection: On [Change](#)

Configuration details

Release label: emr-5.27.0

Hadoop distribution: Amazon 2.8.5

Applications: Hive 2.3.5, Spark 2.4.4

Log URI: s3://aws-logs--ap-northeast-1/elasticmapreduce/ 

EMRFS consistent view: Disabled

Custom AMI ID: --

Network and hardware

Availability zone: ap-northeast-1d

Subnet ID: [subnet-0ce23027b751c53fc](#) 

Master: **Running** 1 units

Core: **Running** 4 (1 Requested)

Task: **Running** 4 (1 Requested)

Security and access

Key name: --

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for [sg-06e0d088240e79417](#) 

Master: (ElasticMapReduce-master)

Security groups for [sg-0ce79503cfce021b4](#) 

Core & Task: (ElasticMapReduce-slave)

本日のアジェンダ

- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

Amazon EMR と関連する AWS サービスの使い分け

□ Amazon Athena

- RDBなどS3 以外もデータソースにする必要がある、Athenaの同時実行クエリ数の制限を回避したい、いずれかに該当するなら、Presto on Amazon EMR

□ AWS Glue

- ETL 処理に Spot を使いたい、単一のノードのスペックを非常に高くしたい、ETL 処理に Spark 以外のアプリを使いたい、いずれかに該当するなら、Apache Spark on Amazon EMR

□ Amazon Kinesis Data Analytics for SQL / for Java App

- ストリーミングアプリを、より柔軟にカスタマイズしたいなら、Apache Spark on Amazon EMR

□ Amazon DynamoDB

Amazon EMR と関連する AWS サービスの使い分け

- ❑ つまり、こんな時は Amazon EMR
 - ❑ Hadoop と各種アプリをより柔軟にカスタマイズしたい
 - ❑ オンプレの Hadoop クラスターをシンプルに移行したい
 - ❑ マネージドサービスの制限を回避したい
 - ❑ できるだけ各種 OSS アプリのより新しいバージョンを使いたい
 - ❑ アプリケーションのバージョンを固定したい
 - ❑ 複数種類のアプリケーションを同一クラスター上に稼働させて連携させる必要がある

Amazon EMR の Apache Spark 性能向上の取り組み (一部抜粋)

EMRFS S3-optimized Committer

- Apache Parquet ファイルをEMRFS越しにS3に書く性能を向上

Dynamic Partition Pruning

- クエリが、あるテーブルに属する特定の Partition から読み書きする性能を向上

Flatten Scalar Subqueries

- 特定のテーブルの行に複数の条件を適用する必要があるクエリの性能を向上

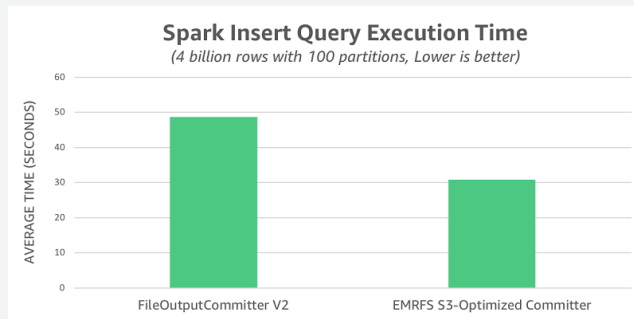
DISTINCT before INTERSECT

- INTERSECT を計算する前に、各コレクションの重複値を排除し、ホスト間でシャッフルされるデータの量を減らして性能を向上

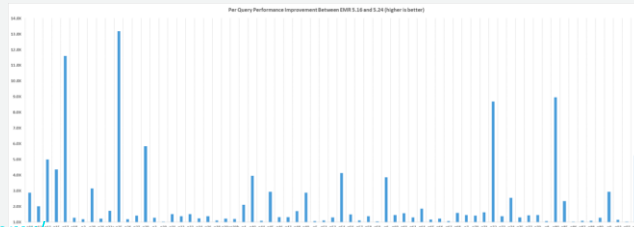
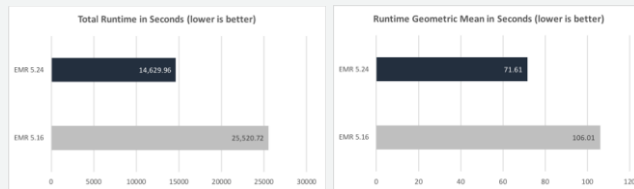
<https://aws.amazon.com/blogs/big-data/spark-enhancements-for-elasticity-and-resiliency-on-amazon-emr/>

<https://aws.amazon.com/blogs/big-data/improve-apache-spark-write-performance-on-apache-parquet-formats-with-the-emrfs-s3-optimized-committer/>

<https://aws.amazon.com/blogs/big-data/performance-updates-to-apache-spark-in-amazon-emr-5-24-up-to-13x-better-performance-compared-to-amazon-emr-5-16/>



TPC-DS benchmark



本日のアジェンダ

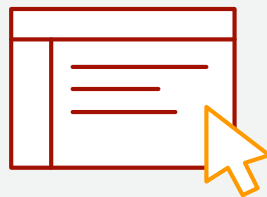
- Why Amazon EMR
- What is Amazon EMR
- How to Amazon EMR
- Amazon EMR and Friends
- まとめ

Amazon EMR

大幅なコスト節減を可能にする、クラウドを利用したマネージドな Hadoop と Spark



高い品質



簡単

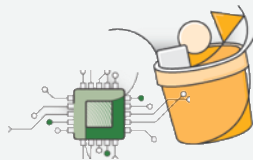


低コスト

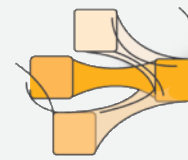
Why Amazon EMR?



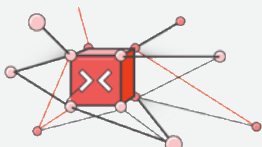
自動化



疎結合



弾力性



統合

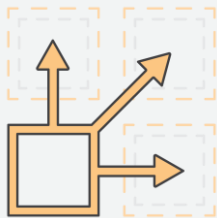


継続的な更新



低コスト

What is Amazon EMR



簡単

わずか数分でクラスターを起動



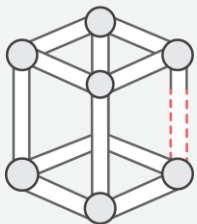
低コスト

秒単位の課金



豊富なオープンソース

最新バージョンのソフトウェア



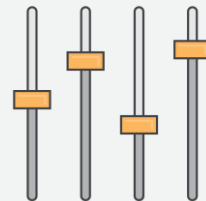
マネージド

監視に費やす労力を節減



安全

簡単にオプション設定可能



柔軟

自在なカスタムと制御

參考資料

- ❑ Amazon EMR documentation : <https://docs.aws.amazon.com/emr/>
- ❑ AWS Bigdata Blog : <https://aws.amazon.com/blogs/big-data/category/analytics/amazon-emr/>
- ❑ Amazon EMR Migration Guide :
https://d1.awsstatic.com/whitepapers/amazon_emr_migration_guide.pdf

AWS の日本語資料の場所「AWS 資料」で検索



The screenshot shows the AWS Japanese website header with the AWS logo, navigation links for '日本語' (Japanese), 'アカウント' (Account), and 'サポート' (Support), and a 'コンソールにサインイン' (Sign in to the console) button. Below the header is a navigation bar with links for '製品' (Products), 'ソリューション' (Solutions), '料金' (Pricing), 'ドキュメント' (Documentation), '学習' (Learning), 'パートナー' (Partners), 'AWS Marketplace', and 'その他' (Other). The main content area features the heading 'AWS クラウドサービス活用資料集トップ' (AWS Cloud Service Usage Resource Collection Top) and a paragraph of introductory text. At the bottom of the main content area are four buttons: 'AWS Webinar お申込' (AWS Webinar Registration), 'AWS 初心者向け' (AWS for Beginners), '業種・ソリューション別資料' (Resources by Industry/Solution), and 'サービス別資料' (Resources by Service).

aws

日本担当チームへお問い合わせ サポート 日本語 ▼ アカウント ▼ コンソールにサインイン

製品 ソリューション 料金 ドキュメント 学習 パートナー AWS Marketplace その他 🔍

AWS クラウドサービス活用資料集トップ

アマゾン ウェブ サービス (AWS) は安全なクラウドサービスプラットフォームで、ビジネスのスケールと成長をサポートする処理能力、データベースストレージ、およびその他多種多様な機能を提供します。お客様は必要なサービスを選択し、必要な分だけご利用いただけます。それらを活用するために役立つ日本語資料、動画コンテンツを多数ご提供しております。(本サイトは主に、AWS Webinar で使用した資料およびオンデマンドセミナー情報を掲載しています。)

AWS Webinar お申込 »

AWS 初心者向け »

業種・ソリューション別資料 »

サービス別資料 »

<https://amzn.to/JPArchive>

AWS Well-Architected 個別技術相談会

毎週”W-A個別技術相談会”を実施中

- AWSのソリューションアーキテクト(SA)に
対策などを相談することも可能

- 申込みはイベント告知サイトから
(<https://aws.amazon.com/jp/about-aws/events/>)

AWS イベント

で[検索]



AWS Well-Architected



aws

ご視聴ありがとうございました

AWS 公式 Webinar

<https://amzn.to/JPWebinar>



過去資料

<https://amzn.to/JPArchive>

