

Alibaba Cloud E-MapReduce

プロダクト紹介

Document Version20190604

目次

1 E-MapReduce の概要.....	1
2 利点.....	3
3 アーキテクチャ.....	4
4 シナリオ.....	5
5 バージョン管理.....	7

1 E-MapReduce の概要

Alibaba Cloud Elastic MapReduce (または E-MapReduce) は、大量のデータの処理と分析を容易にするビッグデータ処理ソリューションです。

E-MapReduce は、ECS (Alibaba Cloud Elastic Compute Service) 上に構築され、オープンソースの Apache Hadoop と Apache Spark に基づいて、トレンド分析、データウェアハウス、オンラインおよびオフラインのデータ処理など、幅広いシナリオでデータを柔軟に管理します。また、Alibaba Cloud OSS や Alibaba Cloud RDS など、他のクラウドストレージシステムやデータベースシステムとの間の、データのインポートやエクスポートが容易です。

E-MapReduce の使用

通常、Hadoop や Spark などの分散処理システムを使用するには、以下の手順に従います。

1. ビジネス特性を評価します。
2. 機種を選択します。
3. マシンを購入します。
4. ハードウェア環境を準備します。
5. オペレーティングシステムをインストールします。
6. アプリケーション (Hadoop や Spark など) をデプロイします。
7. クラスターを起動します。
8. アプリケーションを記述します。
9. ジョブを実行します。
10. データを取得します。または別の操作を実行します。

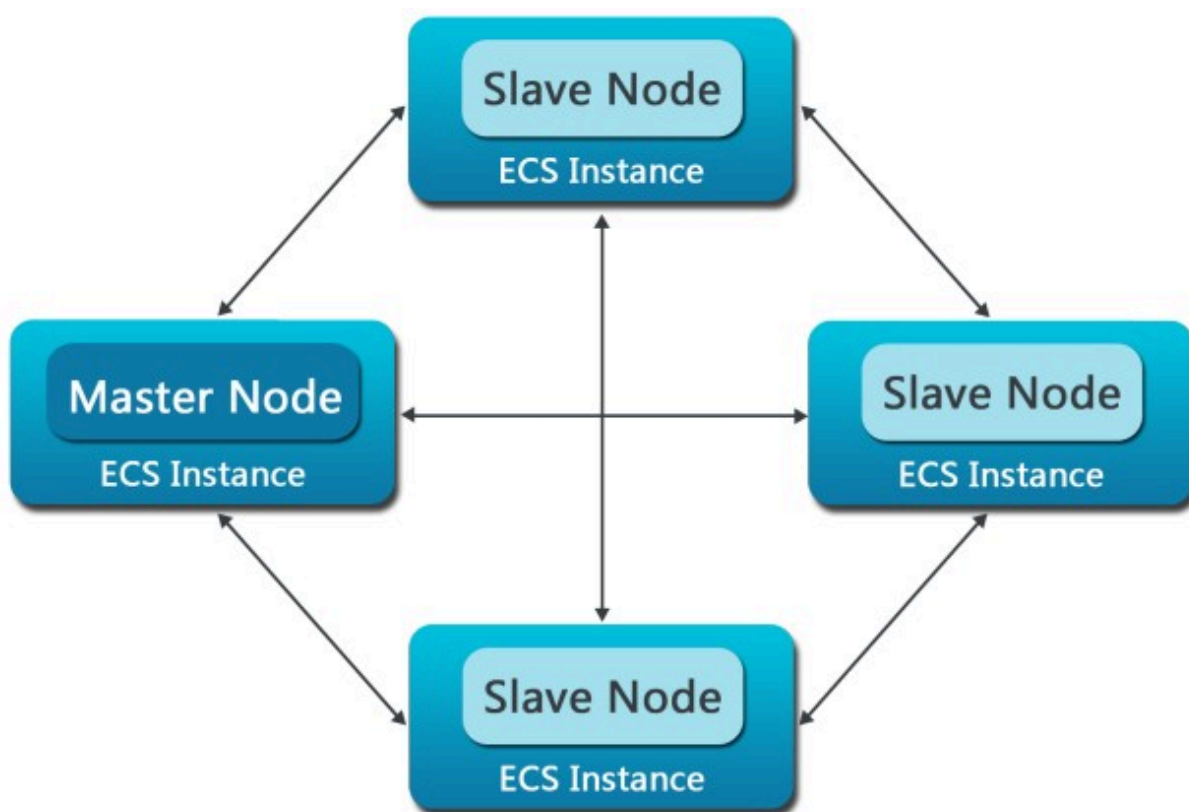
手順 1 ～ 7 は準備作業であり、完了するまでに時間がかかる可能性があります。一方、手順 8 ～ 10 はアプリケーションロジックに関するものです。E-MapReduce では、クラスタの構築、構成、実行および管理、ジョブの構成および実行、ホストの選択、環境のデプロイ、およびパフォーマンスの監視に使用されるものを含む、統合された一連のクラスタ管理ツールが提供されます。

アプリケーションの処理ロジックに集中するには、E-MapReduce を使用して、調達、準備、運用、保守などのプロセスをすべて管理します。E-MapReduce には柔軟な組み合わせモードもあります。E-MapReduce では、必要に応じてさまざまなクラスタサービスを選択します。たとえば、毎日の統計を受信したり、単純なバッチ操作を実行する場合は、E-MapReduce で Hadoop サービスのみの実行を選択します。ストリーム指向コンピューティングおよびリアルタイムコンピューティングを後の段階で実装する場合は、Spark に追加します。

E-MapReduce の構造

クラスターは E-MapReduce のコアコンポーネントです。E-MapReduce クラスターは、基本的には複数の Alibaba Cloud ECS インスタンスからなる Spark または Hadoop クラスターです。たとえば、Hadoop では、通常は各 ECS インスタンス (namenode、datanode、resourcemanager、nodemanager など) で実行されるデーモンによって Hadoop クラスターが形成されます。Namenode と Resourcemanager を実行するノードはマスターノードと呼ばれ、Datanode と Nodemanager を実行するノードはスレーブノードと呼ばれます。

以下の図は、1つのマスターノードと3つのスレーブノードで構成される E-MapReduce クラスターを示しています。



2 利点

E-MapReduce には、自作クラスターに比べて多くの利点があります。たとえば、便利で制御可能な、クラスターの管理方法が数多くあります。

その他の利点は以下のとおりです。

- ・ ユーザビリティ

- 自動デプロイに必要なソフトウェアを選択すると同様に、必要な ECS モデルとディスクの選択は容易です。
- クラスターは要求に応じて数分で作成され、オフラインジョブの完了後にリリースされます。
- 地理的な場所またはデータソースの場所に応じて、クラスターリソースを申請します。

Alibaba Cloud ECS では、現在以下のリージョンがサポートされています。中国 (杭州)、中国 (上海)、中国 (青島)、中国 (北京)、中国 (深セン)、シンガポール、中国 (香港)、米国 (バージニア)、および米国 (シリコンバレー)。E-MapReduce では、中国 (北京)、中国 (杭州)、中国 (上海)、および中国 (深セン) がサポートされていますが、まもなく Alibaba Cloud ECS によってすべてのリージョンがカバーされる予定です。

- ・ 費用対効果

- 基本的なインスタンスの管理にかかるコストが削減されます。
- 使用するインスタンス分の料金を、請求に応じて支払います。

- ・ 統合のしやすさ

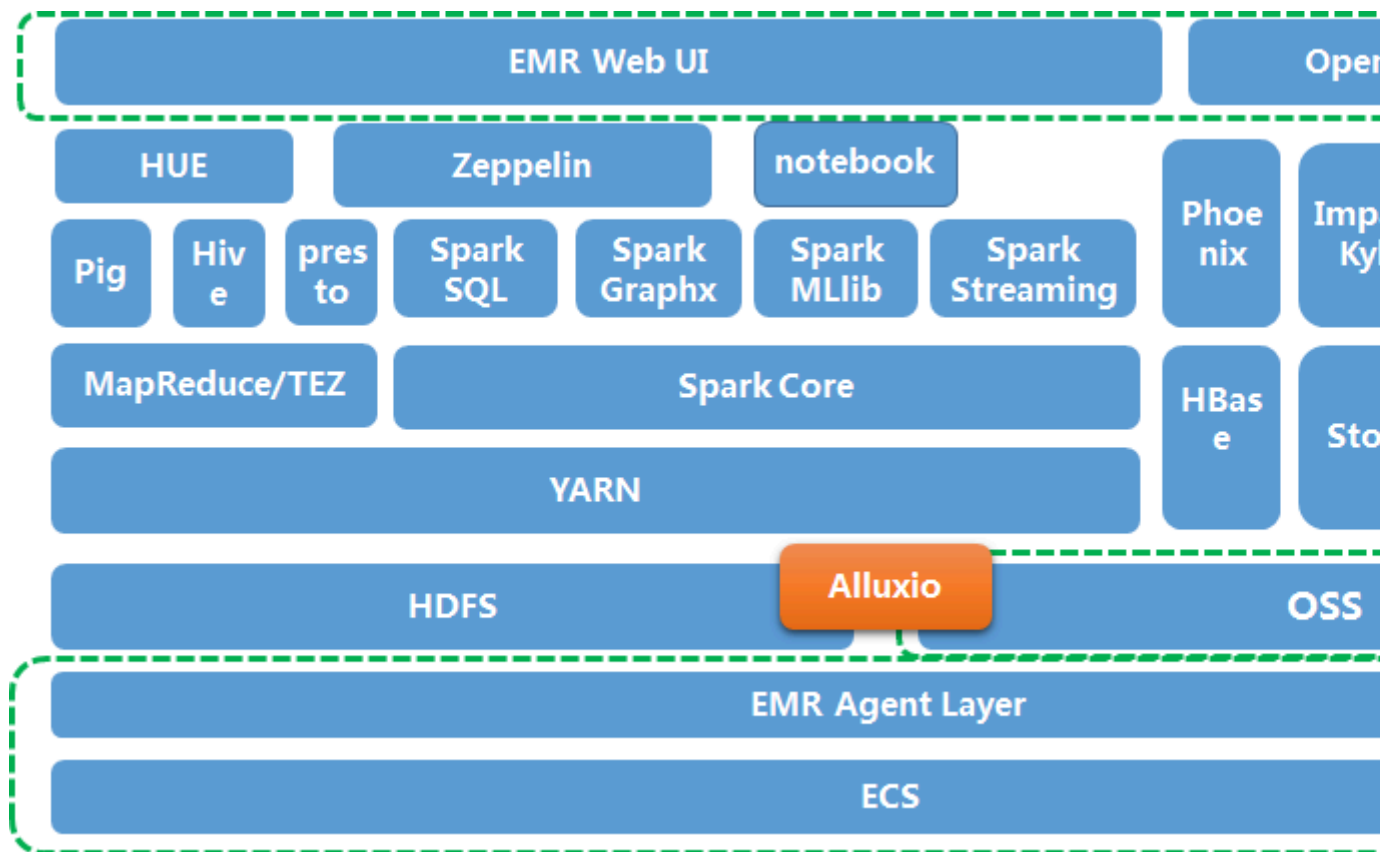
E-MapReduce は他の Alibaba Cloud プロダクトとシームレスに統合され、Hadoop/Spark コンピューティングエンジンの入力ソースまたは出力先として使用されます。

- ・ セキュリティ

E-MapReduce は Alibaba Cloud RAM を統合し、プライマリーまたはサブアカウントを使用してサービス権限を分離します。

3 アーキテクチャ

次の図で、E-MapReduce のプロダクトアーキテクチャを詳しく説明します。



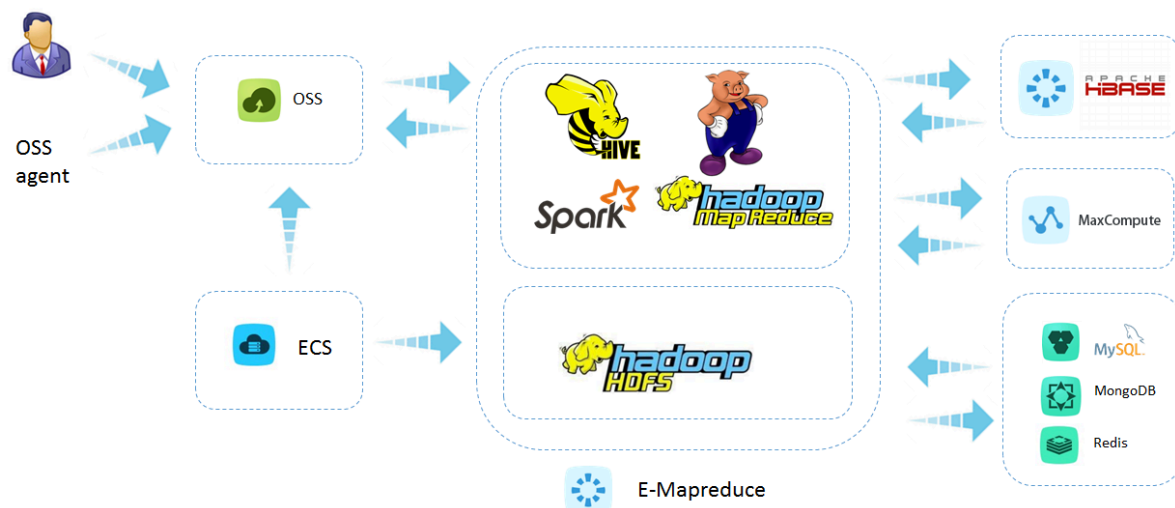
図のように、E-MapReduce クラスターは Hadoop エコシステムに基づいて構築されています。これにより、Alibaba Cloud OSS (オブジェクトストレージサービス) および ApsaraDB (RDS) などのクラウドサービスとシームレスにデータ交換され、複数のシステム間でデータが共有され、転送されます。これにより、さまざまなタイプの業種のアクセスニーズが満たされます。

4 シナリオ

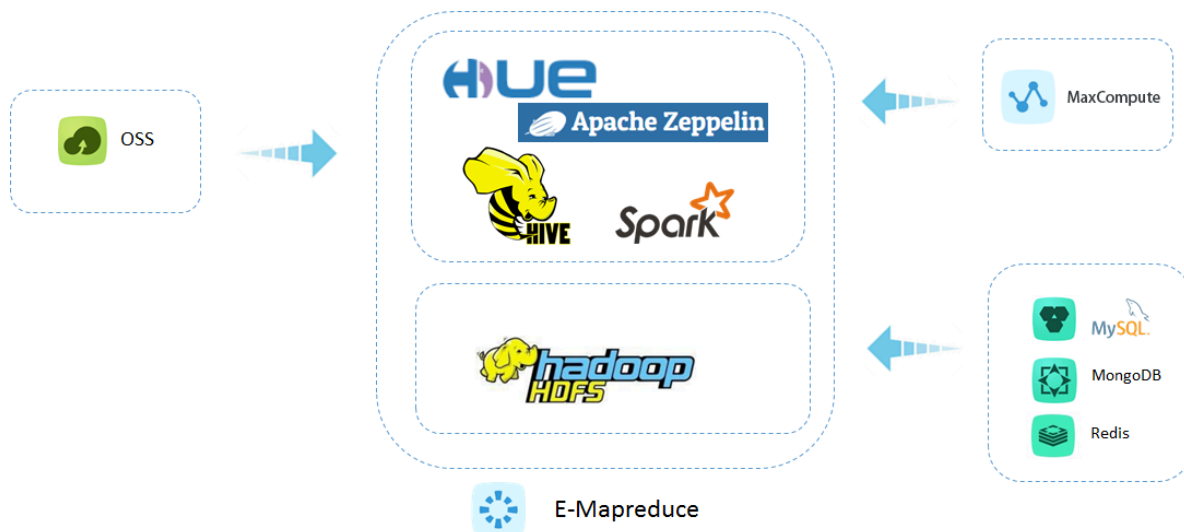
E-MapReduce クラスターはさまざまなシナリオで使用されます。E-MapReduce では、Apache Hadoop および Spark でサポートされているすべてのシナリオがサポートされます。

E-MapReduce は Hadoop および Spark クラスターに基づいているため、Alibaba Cloud ECS は独自の物理ホストと同じように使用されます。以下の図では、E-MapReduce のいくつかの典型的なアプリケーションシナリオの詳細が示されます。

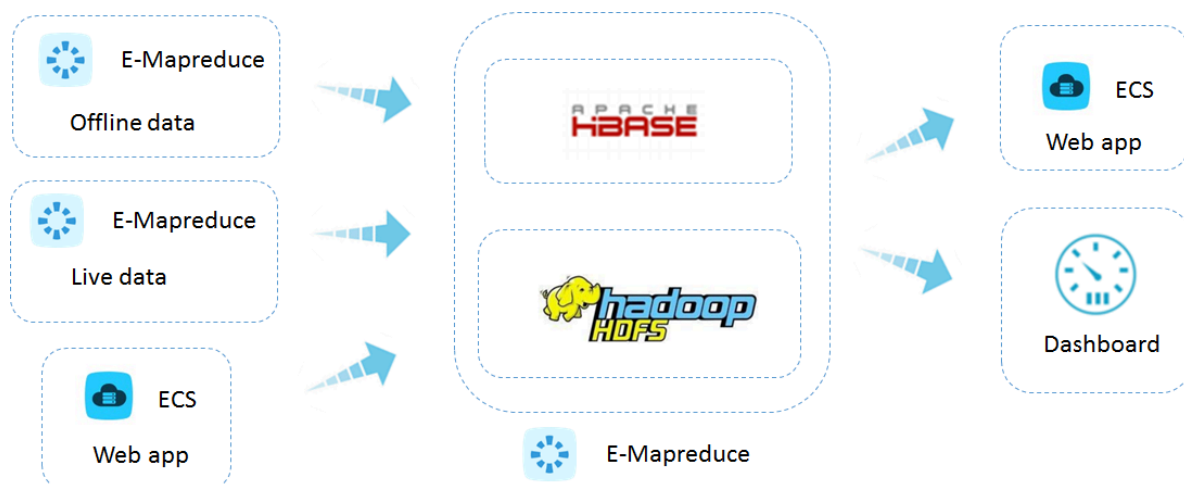
・ オフラインデータ処理



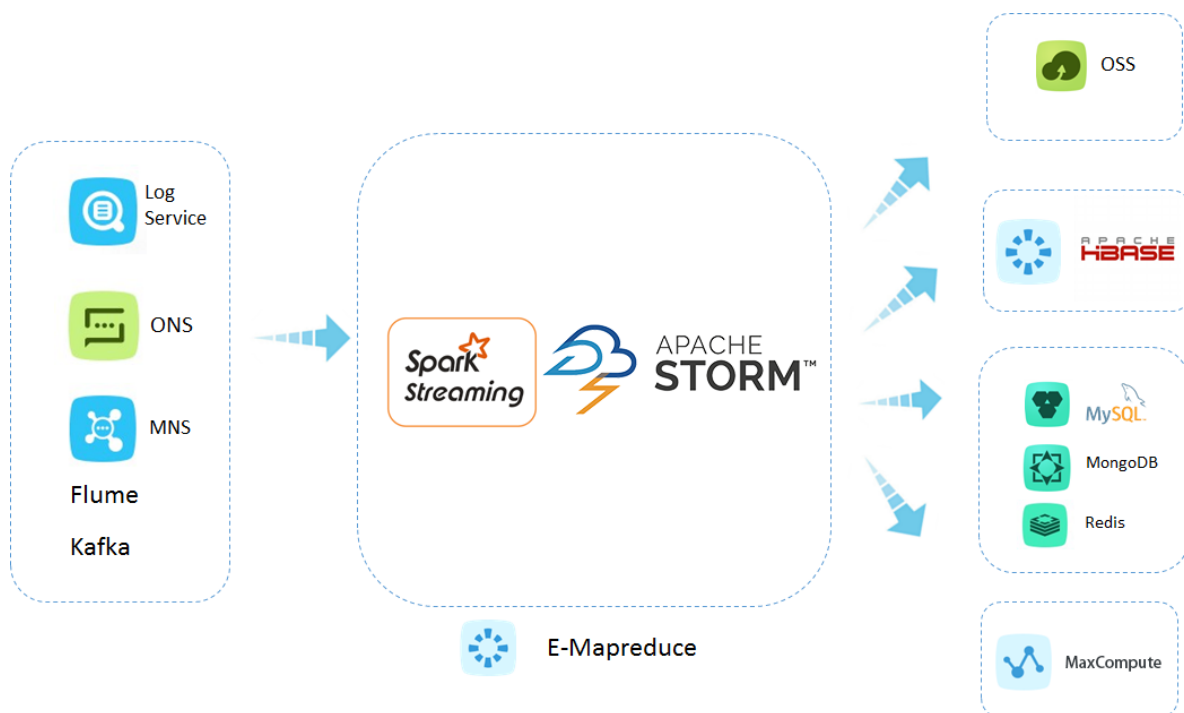
・ アドホックデータ分析



・ 大量のデータのオンラインサービス



・ ストリームデータ処理



5 バージョン管理

- ・ E-MapReduce では、ソフトウェアのバージョン番号は X.Y.Z の形式になります。
 - X は、このバージョンで大幅な変更があったことを示します。
 - Y は、このバージョンで一部のコンポーネントに若干の変更があったことを示します。
 - Z は、このバージョンではバグが修正されているが、以前のバージョンと互換性が保たれていることを示します。

たとえば、バージョン 1.0.0 から 2.0.0 へアップグレードする場合は、多くの変更が加えられます。したがって、バージョンアップの際は、以前のジョブがすべて正常に実行されるか確認することを推奨します。バージョン 1.0.0 から 1.1.0 へのアップデートでは、通常はコンポーネントがアップグレードされます。ジョブが正常に実行されるか確認するために、同様のテストを行うことを推奨します。バージョン 1.0.0 から 1.0.1 へのアップデートは、通常は一種のメンテナンスとして実施され、以前のバージョンと完全な互換性があります。

- ・ E-MapReduce の各リリースでは、ソフトウェアとソフトウェアのバージョンが修正されます。複数のソフトウェアバージョンを選択することはできません。また、ソフトウェアバージョンを手動で変更することは推奨されません。
- ・ E-MapReduce のバージョンを選択してクラスターで使用する場合は、使用中の E-MapReduce は自動的にアップグレードされません。このバージョンに対応するイメージがアップグレードされた場合は、すでに作成したクラスターは影響を受けません。新しいクラスターのみが新しいイメージを使用します。
- ・ クラスターのバージョンをアップグレードするときは (たとえば、1.0.x から 1.1.x へのアップグレードなど)、新しいソフトウェア環境でジョブが正常に実行されるかを確認するために、テストを実施することを推奨します。

リリースノート

- ・ 3.x

バージョン	EMR-3.7.1	EMR-3.8.1	EMR-3.9.1	EMR-3.10.1	EMR-3.11.0	EMR-3.12.0	EMR-3.13.0	EMR-3.14.0	EMR-3.15.0	EMR-3.16.0	EMR-3.17.0	EMR-3.18.1	EMR-3.19.0	EMR-3.19.1	EMR-3.20.0
リリース日	2018.1	2018.1	2018.2	2018.4	2018.6	2018.7	2018.8	2018.10	2018.11	2018.12	2019.1	2019.2	2019.3	2019.4	2019.5

バージョン	EMR-3.7.1	EMR-3.8.1	EMR-3.9.1	EMR-3.10.1	EMR-3.11.0	EMR-3.12.0	EMR-3.13.0	EMR-3.14.0	EMR-3.15.0	EMR-3.16.0	EMR-3.17.0	EMR-3.18.1	EMR-3.19.0	EMR-3.19.1	EMR-3.20.0
Hadoop	2.7.2-emr-1.2.10	2.7.2-emr-1.2.12	2.7.2-emr-1.2.13	2.7.2-emr-1.2.14	2.7.2-emr-1.2.14	2.7.2-emr-1.2.14	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.8.5	2.8.5	2.8.5
Knox	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	0.13.0	1.1.0	1.1.0	1.1.0	1.1.0
Apache DS	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0
Spark	2.2.1	2.2.1	2.2.1	2.2.1	2.2.1	2.3.1	2.3.1	2.3.1	2.3.2	2.3.2	2.3.2	2.3.2	2.3.2	2.4.1	2.4.2
Hive	2.3.2	2.3.2	2.3.2	2.3.2	2.3.3	2.3.3	2.3.3	2.3.3	2.3.3	2.3.3	2.3.3	3.1.3	3.1.1	3.1.1	3.1.1
Tez	0.8.4	0.8.4	0.8.4	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1	0.9.1
Pig	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6	1.4.6	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
YARN	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.8.5	2.8.5	2.8.5
HDFS	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.7.2	2.8.5	2.8.5	2.8.5
Flink		1.4.0	1.4.0	1.4.0	1.4.0	1.4.0	1.4.0	1.4.0	1.4.0	1.6.2-1.0.0	1.6.2	1.6.2	1.7.2	1.7.2	1.7.2

バージョン	EMR-3.7.1	EMR-3.8.1	EMR-3.9.1	EMR-3.10.1	EMR-3.11.0	EMR-3.12.0	EMR-3.13.0	EMR-3.14.0	EMR-3.15.0	EMR-3.16.0	EMR-3.17.0	EMR-3.18.1	EMR-3.19.0	EMR-3.19.1	EMR-3.20.0
Druid			0.11.0	0.11.0	0.11.0	0.12.0	0.12.2	0.12.3	0.12.3	0.12.3-1.0.1	0.12.3	0.13.0	0.13.0	0.13.0	0.13.0
HBase	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1-1.0.2	1.1.1	1.4.9	1.4.9	1.4.9	1.4.9
Phoenix	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0	4.10.0-1.0.0	4.10.0	4.14.1	4.14.1	4.14.1	4.14.1
ZooKeeper	3.4.11	3.4.11	3.4.11	3.4.11	3.4.11	3.4.12	3.4.12	3.4.13	3.4.13	3.4.13	3.4.13	3.4.13	3.4.13	3.4.13	3.4.13
Presto	0.188	0.188	0.188	0.188	0.188	0.188	0.208	0.208	0.208	0.208	0.213	0.213	0.213	0.213	0.213
Storm	1.0.1	1.0.1	1.0.1	1.1.2	1.1.2	1.1.2	1.1.2	1.1.2	1.1.2	1.2.2	1.2.2	1.2.2	1.2.2	1.2.2	1.2.2
Impala	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0	2.10.0-1.0.0	2.12.2	2.12.2	2.12.2	2.12.2	2.12.2
Hue	3.12.0	3.12.0	3.12.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0	4.1.0
Oozie	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	5.1.0
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.1	0.7.3	0.7.3	0.8.0	0.8.0	0.8.0	0.8.0	0.8.0	0.8.0	0.8.0	0.8.0	0.8.1
Ranger			0.7.1	0.7.1	0.7.3	1.0.0	1.0.0	1.0.0	1.0.0	1.0.0	1.0.0	1.0.0	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2
OS	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4

バージョン	EMR-3.7.1	EMR-3.8.1	EMR-3.9.1	EMR-3.10.1	EMR-3.11.0	EMR-3.12.0	EMR-3.13.0	EMR-3.14.0	EMR-3.15.0	EMR-3.16.0	EMR-3.17.0	EMR-3.18.1	EMR-3.19.0	EMR-3.19.1	EMR-3.20.0
Tensorflow							1.8.0	1.8.0	1.8.0	1.8.0	1.8.0				1.8.0
Kafka							2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0	2.11.1.0
Superset							0.25.6	0.25.6	0.27.0	0.28.1	0.28.1	0.28.1	0.28.1	0.28.1	0.28.1
Jupyter								4.4.0	4.4.0	4.4.0					4.4.0
Analytics Zoo								0.2.0	0.2.0	0.2.0					0.2.0

・ 2.x

バージョン	EMR-2.9.2	EMR-2.10.0	EMR-2.11.0
リリース日	2018.2	2018.4	2018.7
Hadoop	2.7.2-emr-1.2.12	2.7.2-emr-1.2.12	2.7.2-emr-1.2.12
Spark	1.6.3	1.6.3	1.6.3
Hive	2.3.2	2.3.2	2.3.3
Tez	0.8.4	0.9.1	0.9.1
Pig	0.14.0	0.14.0	0.14.0
Sqoop	1.4.6	1.4.6	1.4.6
Hue	3.12.0	4.1.0	4.1.0
Zeppelin	0.7.1	0.7.1	0.7.3
HBase	1.1.1	1.1.1	1.1.1
Phoenix	4.10.0	4.10.0	4.10.0
Storm	1.0.1	1.1.2	1.1.2
Presto	0.188	0.188	0.188.0

バージョン	EMR-2.9.2	EMR-2.10.0	EMR-2.11.0
Impala	2.10.0	2.10.0	2.10.0
Zookeeper	3.4.6	3.4.11	3.4.11
Oozie	4.2.0	4.2.0	4.2.0
Ranger	0.7.1	0.7.1	0.7.3
Ganglia	3.7.2	3.7.2	3.7.2
OS	CentOS 7.4	CentOS 7.4	CentOS 7.4

・ 1.x

バージョン	EMR-1.0.0	EMR-1.1.0	EMR-1.2.0	EMR-1.3.0
リリース日	2015.11	2016.3	2016.4	2016.5
Hadoop	2.6.0	2.6.0	2.6.0	2.6.0-emr-1.1.1
Spark	1.4.1	1.6.0	1.6.1	1.6.1
Hive	1.0.1	1.0.1	2.0.0	2.0.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	-	-	-	1.4.6
Hue	-	-	-	3.9.0
Zeppelin	-	-	-	0.5.6
HBase	-	-	1.1.1	1.1.1
Phoenix	-	-	-	-
Zookeeper	-	-	3.4.6	3.4.6
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

・ Hadoop のバージョンの説明

Alibaba Cloud OSS をサポートするために、元のインターフェイスを変更せず、Hadoop のバージョンに基づいて emr-core コンポーネントが追加されます。このコンポーネントのバージョンは、Hadoop のバージョンの後に追加されます。

・ EMR 3.1.x

- OS は CentOS7.2 にアップグレードされます。
- Spark は 2.1.1 にアップグレードされます。
- emr-core は 1.2.6 にアップグレードされます。
- AK-free OSS 操作のバグが修正されます。

- ・ EMR 3.0.x
 - EMR 3.0.2
 - emr-core は 1.2.5 にアップグレードされます。
 - AK-free OSS でサポートされるリージョンが増えます。
 - AK 交換戦略が調整されます。
 - Hive と Hadoop のバグが修正されます。
 - EMR 3.0.1
 - インタラクティブかつ統合されたテーブル管理がサポートされます。
 - Hive メタデータの保存には、外部の統合データベースが使用されます。
 - OSS の読み書きのパフォーマンスが最適化されます。
 - 外部のHive メタデータを使用するすべてのクラスターでは、同じメタ情報が共有されます。
 - Spark は 2.0.2 にアップグレードされます。
 - emr-core は 1.2.4 にアップグレードされます。
 - EMR 3.0.0
 - 3.0.0 と完全な互換性がある 3.0.1 を使用します。
- ・ EMR 2.4.x
 - EMR 2.4.2
 - emr-core は 1.2.5 にアップグレードされます。
 - AK-free OSS でサポートされるリージョンが増えます。
 - AK 交換戦略が調整されます。
 - Hive と Hadoop のバグが修正されます。
 - EMR 2.4.1
 - インタラクティブかつ統合されたテーブル管理がサポートされます。
 - Hive メタデータの保存には、外部の統合データベースが使用されます。
 - OSS の読み書きのパフォーマンスが最適化されます。
 - 外部のHive メタデータを使用するすべてのクラスターでは、同じメタ情報が共有されます。
 - emr-core は 1.2.4 にアップグレードされます。
 - EMR 2.4.0 2.4.0 と完全な互換性がある 2.4.1 を使用します。

- ・ EMR 2.3.x
 - EMR 2.3.1 独立した内部データベースへ Hive メタデータを保存する各クラスターでは、インタラクティブな workbench がサポートされます。
 - EMR 2.3.0 メンテナンス中です。インタラクティブな workbench がサポートされますが、テーブル管理はサポートされません。このバージョンの使用は推奨されません。
- ・ EMR 2.2.0

非推奨のバージョンです。
- ・ EMR 2.1.x
 - EMR 2.1.1 メンテナンス中です。インタラクティブな workbench がサポートされますが、テーブル管理はサポートされません。このバージョンの使用は推奨されません。
 - EMR 2.1.0 非推奨のバージョンです。
- ・ EMR 2.0.x

非推奨のバージョンです。
- ・ EMR 1.x は非推奨のバージョンです。