Alibaba Cloud E-MapReduce

Product Introduction

Issue: 20190515

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning informatio n, supplementary instructions, and other content that the user must understand.	Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list -- instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich` *{stand | slave}*

Contents

Legal disclaimer..... I
Generic conventions..... I
1 What is E-MapReduce?.....1
2 Benefits..... 3
3 Architecture..... 4
4 Scenarios.....5
5 Versioning..... 7

1 What is E-MapReduce?

Alibaba Cloud Elastic MapReduce (or E-MapReduce) is a big data processing solution that facilitates the processing and analysis of massive amounts of data.

Built on Alibaba Cloud Elastic Compute Service (ECS) and based on open-source Apache Hadoop and Apache Spark, E-MapReduce flexibly manages your data in a wide range of scenarios, such as trend analysis, data warehousing, and online and offline data processing. It also makes it easy for you to import and export data to and from other cloud storage systems and database systems, such as Alibaba Cloud OSS and Alibaba Cloud RDS.

Using E-MapReduce

In general, to use a distributed processing system such as Hadoop or Spark, follow these steps:

1. Evaluate the business characteristics.
2. Select a machine type.
3. Purchase a machine.
4. Prepare the hardware environment.
5. Install an operating system.
6. Deploy applications (such as Hadoop and Spark).
7. Start a cluster.
8. Write applications.
9. Run a job.
10. Obtain data or perform another operation.

Steps 1-7 are preliminary tasks and may take some time to complete. Steps 8-10, however, concern application logic. E-MapReduce provides an integrated set of cluster management tools, including those used to build, configure, run, and manage clusters, configure and run jobs, as well as select hosts, deploy environments, and monitor performance.

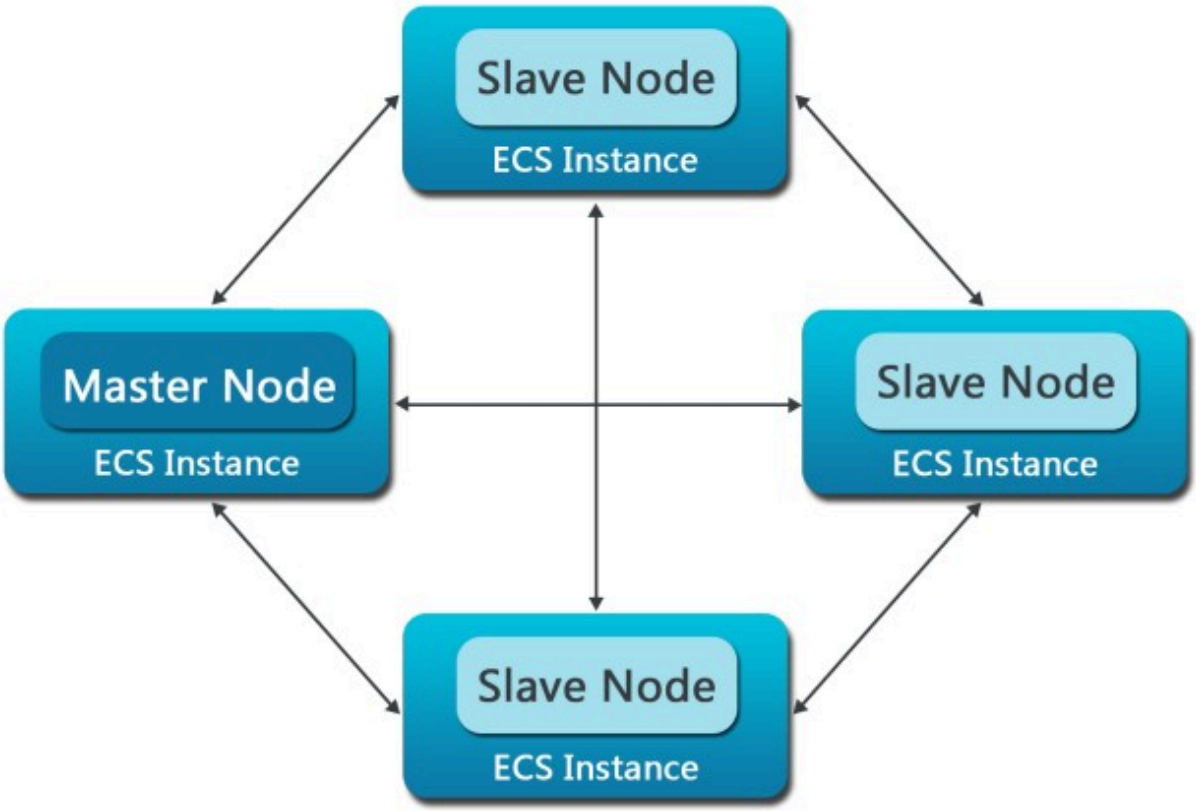
With E-MapReduce, processes such as procurement, preparation, operation, and maintenance are all managed, allowing you to focus on the processing logic of your applications. E-MapReduce also provides flexible combination modes, allowing you to select different cluster services according to your needs. For example, if you want

to receive daily statistics or perform simple batch operations, you can choose to only run Hadoop services in E-MapReduce. If you then want to implement stream-oriented and real-time computing at a later stage, you can add in Spark.

Structure of E-MapReduce

Clusters are the core component of E-MapReduce. An E-MapReduce cluster is essentially a Spark or Hadoop cluster that consists of multiple Alibaba Cloud ECS instances. For example, in Hadoop, the daemons that typically run on each ECS instance (such as NameNode, DataNode, ResourceManager, and NodeManager) form a Hadoop cluster. The nodes that run NameNode and ResourceManager are known as master nodes, while those that run DataNode and NodeManager are called slave nodes.

The following figure shows an E-MapReduce cluster that consists of one master node and three slave nodes:



2 Benefits

E-MapReduce has some practical strength over the self-built clusters. For example, it provides some convenient and controllable means to manage its clusters.

In addition, it also has the following strengths:

- Easy to use

  The user can select the required ECS types and disks and select the required software for automatic deployment.

  The user can apply for cluster resources at the corresponding position according to the geographical location where the user or the data source is located. Alibaba Cloud ECS currently supports the following regions: China (Hangzhou), China (Shanghai), China (Qingdao), China (Beijing), China (Zhangjiakou), China (Hohhot), China (Shenzhen), Singapore, Hong Kong, Australia (Sydney), Malaysia (Kuala Lumpur), Indonesia (Jakarta), Japan (Tokyo), Germany (Frankfurt), US (Silicon Valley), US (Virginia), India, and UAE (Dubai). E-MapReduce supports all of the regions supported by Alibaba Cloud ECS.
- Low price

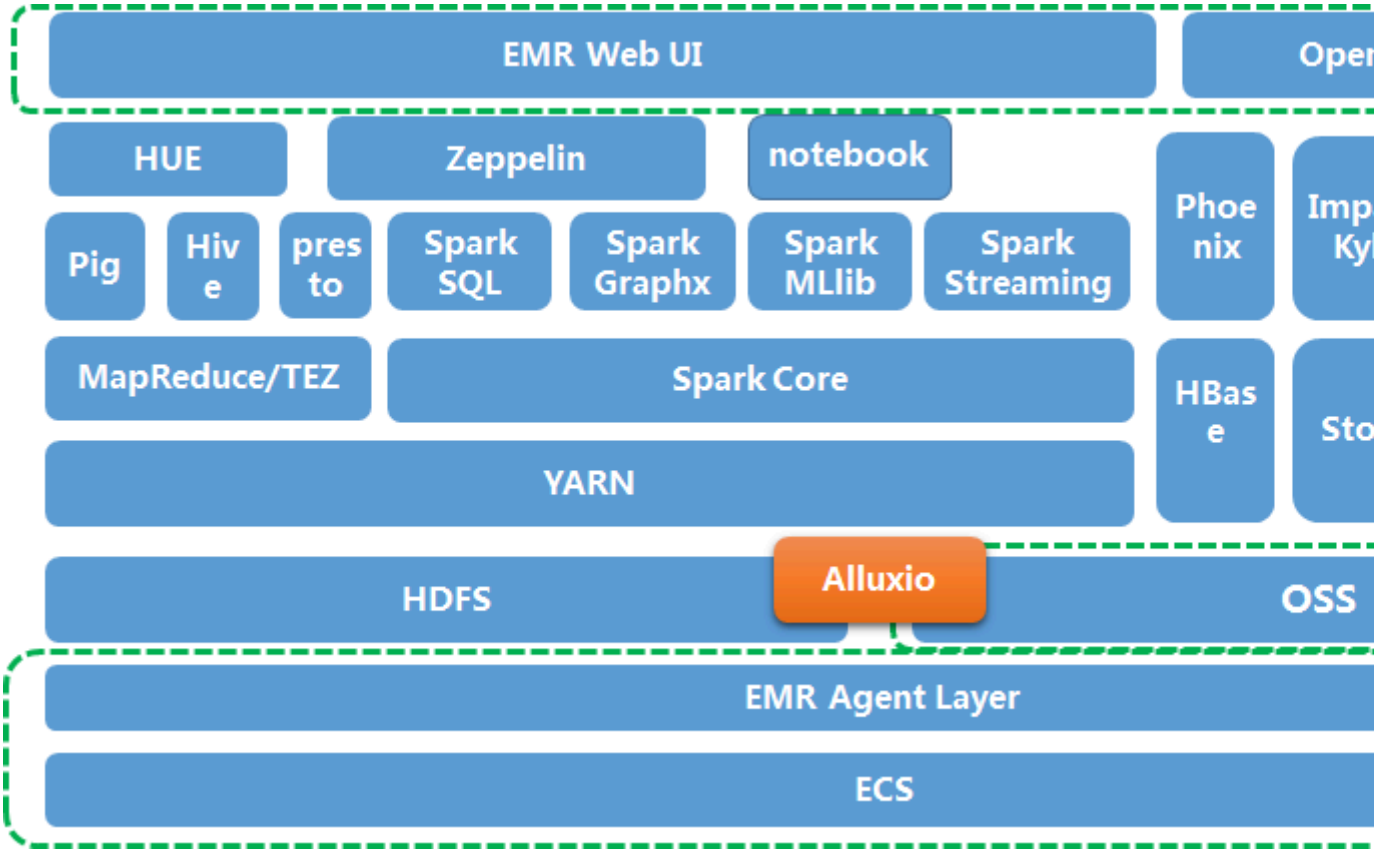
  The user can create a cluster as needed, that is, it can release the cluster after an offline task running is completed and add a node dynamically when needed.
- Deep integration

  E-MapReduce is deeply integrated with other Alibaba Cloud products (such as OSS, MNS, RDS, and MaxCompute) and can be used as the input source or output destination of the Hadoop/Spark computing engine.
- Security

  E-MapReduce integrates Alibaba Cloud RAM resource permission management system, so that it can isolate the service permissions through the primary account or sub-accounts.

3 Architecture

The product architecture of E-MapReduce is detailed in the following figure.



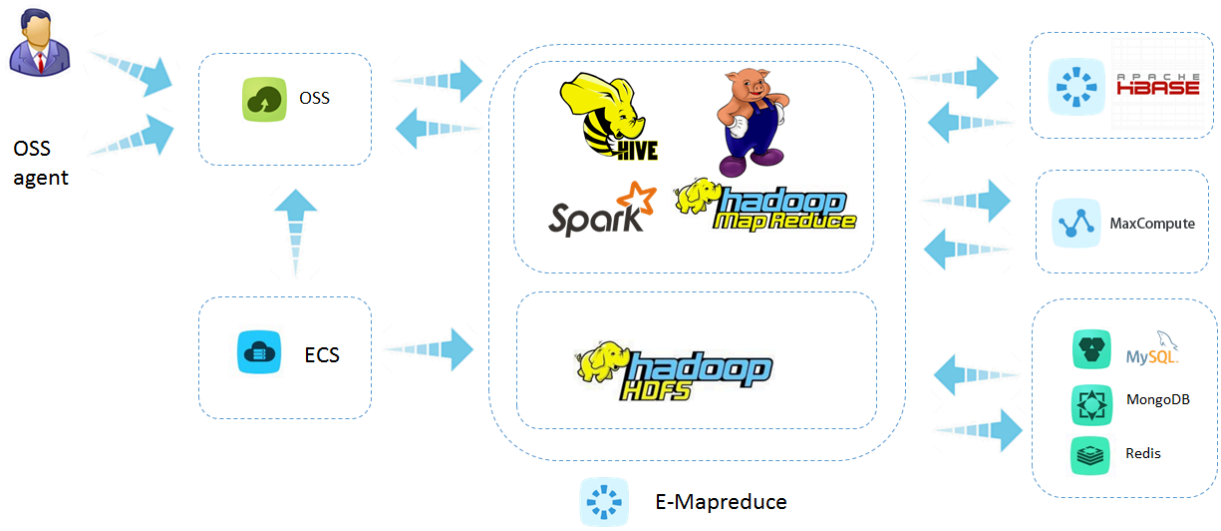
As shown in the figure, an E-MapReduce cluster is built based on the Hadoop ecosystem. This allows data to be exchanged seamlessly with cloud services, such as Alibaba Cloud Object Storage Service (OSS) and ApsaraDB (RDS), and enables you to share and transfer data between multiple systems. This meets the access needs of different types of businesses.

4 Scenarios

E-MapReduce clusters can be used in various scenarios, as E-MapReduce supports all scenarios that Apache Hadoop and Spark support.

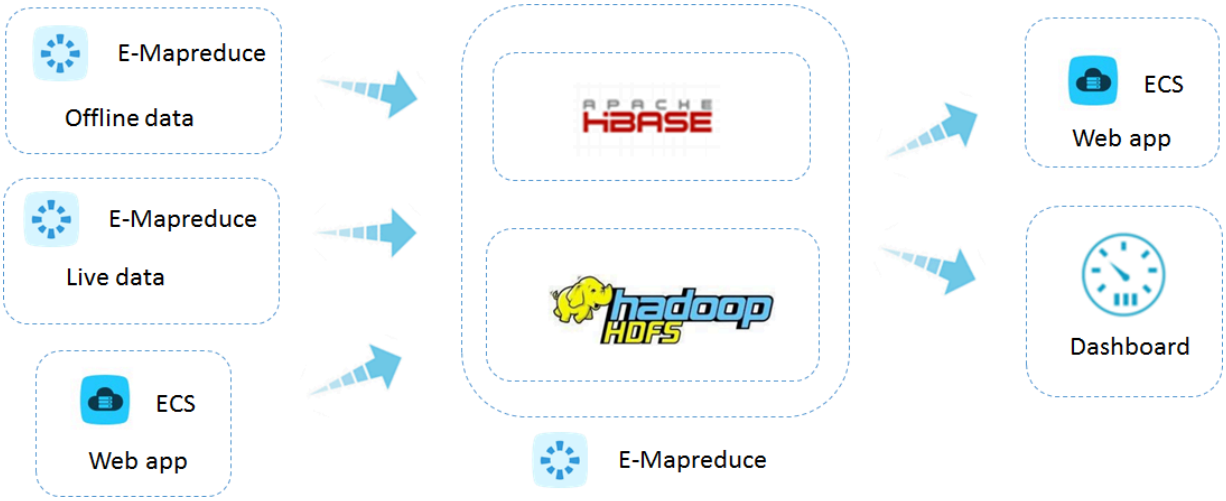
Because E-MapReduce is based on Hadoop and Spark clusters, you can use your Alibaba Cloud ECS host as your own physical host. The following figures detail some typical application scenarios of E-MapReduce.

Offline data processing

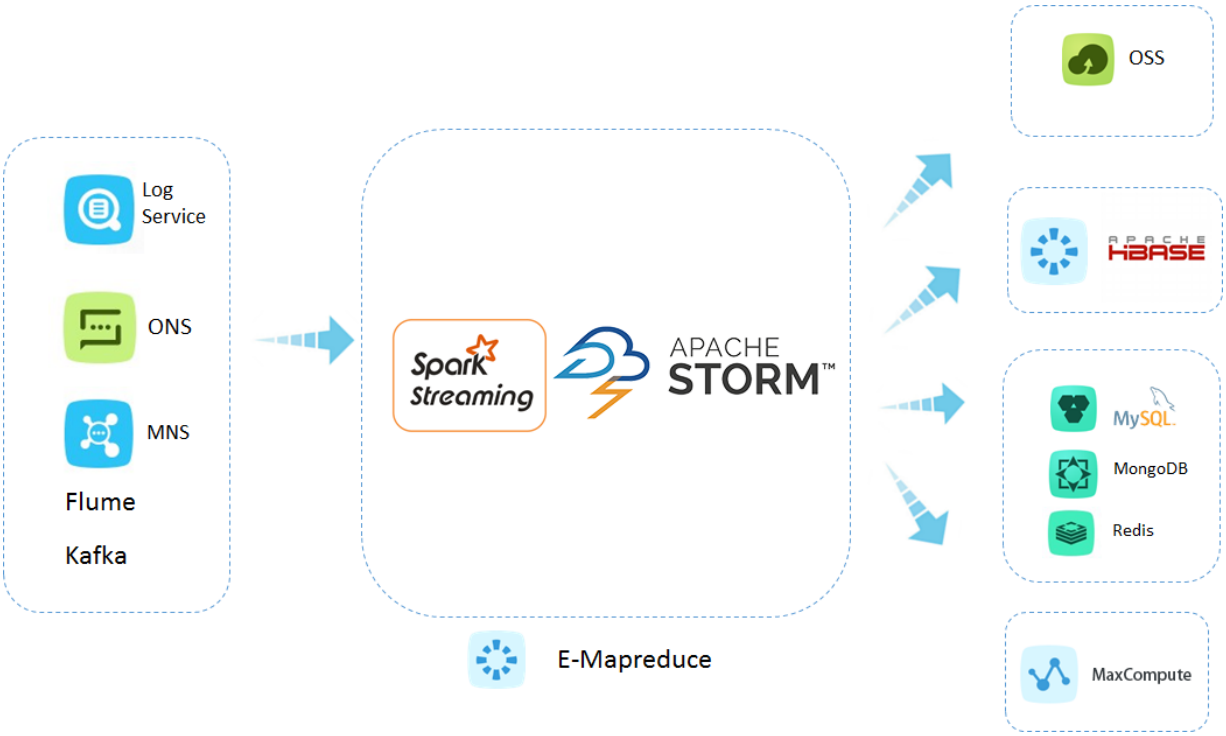


Ad-hoc data analysis

Online services for massive amounts of data



Stream data processing



5 Versioning

- E-MapReduce follows an X.Y.Z format in its software version numbers:
  - X indicates that major changes have been made in the version.
  - Y indicates that minor changes have been made to some components in the version.
  - Z indicates that while bugs have been fixed in the version, it is still compatible with previous versions.

  If you upgrade from version 1.0.0 to version 2.0.0, for example, a number of major changes will have been made. We therefore recommend verifying that all previous jobs can run normally. An update from version 1.0.0 to 1.1.0, however, is typically made to upgrade a component. We recommend performing a similar test to verify that jobs are running normally. An update from version 1.0.0 to version 1.0.1 is typically to perform some kind of maintenance, and it remains fully compatible with previous versions.
- In each release of E-MapReduce, the software and software version are fixed. You cannot select multiple software versions, and changing the software version manually is not recommended.
- If you select a version of E-MapReduce and use it in a cluster, it is not upgraded automatically. If the images corresponding to this version are upgraded, the cluster you created is not impacted. Only new clusters will use the new images.
- When you upgrade the version of your cluster (for example, from 1.0.x to 1.1.x), we recommend performing a test to verify that your jobs are running normally in the new software environment.

Release notes

- 3.x

版本	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
发布时间	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
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	2.7.2-1.3.2	2.7.2	2.8.5	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	1.1.0	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	2.3.2-1.0.1	2.3.2	2.3.2	2.4.1	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	2.3.3-1.0.3	2.3.3	3.1.1	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	0.9.1-1.0.2	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	1.4.7-1.0.0	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	2.8.5	2.8.5	2.8.5	2.8.5

版本	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
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	2.8.5	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
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

版本	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
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
Tensorflow							1.8.0	1.8.0	1.8.0	1.8.0	1.8.0				1.8.0
Kafka							2.11-1.0.1	2.11-1.0.1	2.11-1.0.1	2.11-1.1.0	2.11-1.1.1	2.11-1.1.1	2.11-1.1.1	1.1.1	2.11
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

Version	EMR-2.9.2	EMR-2.10.0	EMR-2.11.0
Release Date	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

Version	EMR-2.9.2	EMR-2.10.0	EMR-2.11.0
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
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

Version	EMR-1.0.0	EMR-1.1.0	EMR-1.2.0	EMR-1.3.0
Release Date	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 version description:

  To support Alibaba Cloud OSS, the emr-core component is added based on the version of Hadoop, without any changes made to the original interface. The version of this component is added after the Hadoop version.

- EMR 3.1.x
  - OS is upgraded to CentOS7.2.
  - Spark is upgraded to 2.1.1.
  - emr-core is upgraded to 1.2.6.
  - Fixes bugs in AK-free OSS operations.
- EMR 3.0.x
  - EMR 3.0.2:
    - emr-core is upgraded to 1.2.5.
    - AK-free OSS supports more regions.
    - Adjusts the AK replacement strategy.
    - Fixes bugs in Hive and Hadoop.
  - EMR 3.0.1:
    - Supports interactive and unified table management.
    - Uses an external unified database to save Hive metadata.
    - Optimizes the read and write performance of OSS.
    - All clusters that use external Hive metadata share the same meta information.
    - Spark is upgraded to 2.0.2.
    - emr-core is upgraded to 1.2.4.
  - EMR 3.0.0:
    - Uses 3.0.1, which is fully compatible with 3.0.0.

- EMR 2.4.x
  - EMR 2.4.2:
    - emr-core is upgraded to 1.2.5.
    - AK-free OSS supports more regions.
    - Adjusts the AK replacement strategy.
    - Fixes bugs in Hive and Hadoop.
  - EMR 2.4.1:
    - Supports interactive and unified table management.
    - Uses an external unified database to save Hive metadata.
    - Optimizes the read and write performance of OSS.
    - All clusters that use external Hive metadata share the same meta information.
    - emr-core is upgraded to 1.2.4.
  - EMR 2.4.0: Uses 2.4.1, which is fully compatible with 2.4.0.
- EMR 2.3.x
  - EMR 2.3.1: Supports an interactive workbench where each cluster uses an independent internal database to save Hive metadata.
  - EMR 2.3.0: Under maintenance. Supports an interactive workbench, but does not support table management. Using this version is not recommended.
- EMR 2.2.0

  Deprecated version.
- EMR 2.1.x
  - EMR 2.1.1: Under maintenance. Supports an interactive workbench, but does not support table management. Using this version is not recommended.
  - EMR 2.1.0: Deprecated version.
- EMR 2.0.x

  Deprecated version.
- EMR 1.xDeprecated version.