

Alibaba Cloud E-MapReduce

Product Introduction

Issue: 20181127

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 information, supplementary instructions, and other content that the user must understand.	 Note: 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 <code>cd /d C:/windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid Instance_ID</code>
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>
{ } or {a b}	It indicates that it is a required value, and only one item can be selected.	<code>swich {stand slave}</code>

Contents

Legal disclaimer	I
Generic conventions	I
1 What is EMR	1
2 Benefits	3
3 Architecture	4
4 Scenarios	5
5 Versioning	7
6 Release notes	8

1 What is EMR

Alibaba Cloud Elastic MapReduce (E-MapReduce) is a system solution for big data processing that runs on the Alibaba Cloud platform. E-MapReduce is built on Alibaba Cloud Elastic Compute Service (ECS) and is based on open-source Apache Hadoop and Apache Spark. It facilitates the use of other peripheral systems (for example, Apache Hive, Apache Pig, and HBase) in the Hadoop and Spark ecosystems to analyze and process data. You can also easily import data to and export data from other cloud data storage systems and database systems, such as Alibaba Cloud OSS and Alibaba Cloud RDS.

Use of E-MapReduce

In general, to use distributed processing systems, such as Hadoop and Spark, the following actions are recommended:

1. Evaluate business characteristics.
2. Select a machine type.
3. Purchase a machine.
4. Prepare 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 and so on.

Steps 8-10 relate to the application logic of users. Steps 1-7 are early preparations and tend to be difficult and cumbersome. E-MapReduce provides an integrated solution of cluster management tools, such as host selection, environment deployment, cluster building, cluster configuration, cluster running, job configuration, job running, cluster management, and performance monitoring.

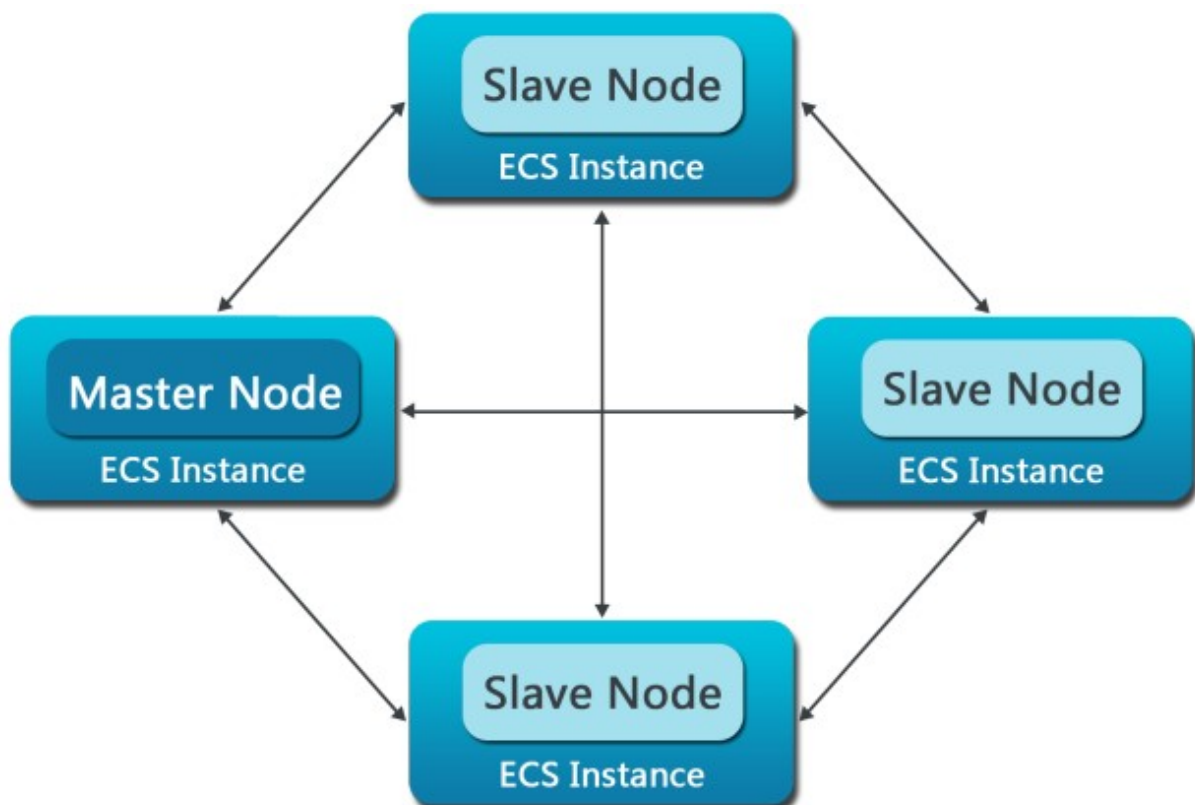
With E-MapReduce, processes such as procurement, preparation, operation, and maintenance are managed, allowing you to focus on the processing logics of your applications. E-MapReduce also provides flexible combination modes, allowing you to select different cluster services according to your needs. For example, if you want to implement daily statistics and simple batch operations, you can choose to run only Hadoop services in E-MapReduce; if you still want to

implement stream-oriented and real-time computing, you can add Spark on the basis of the Hadoop service.

Composition of E-MapReduce

The core component directly oriented to an E-MapReduce user is the cluster. An E-MapReduce cluster is a Spark and Hadoop cluster consisting of multiple ECS Alibaba Cloud instances. For example, in Hadoop, generally some daemon processes running on each ECS instance (such as namenode, datanode, resourcemanager, and nodemanager) make up a Hadoop cluster. The nodes running namenode and resourcemanager are known as master nodes, while those running datanode and nodemanager are called slave nodes.

For example, the following figure shows an E-MapReduce cluster consisting of one master node and three slave nodes:



2 Benefits

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

- Usability

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

Users can apply for cluster resources at the corresponding position according to the geographical location where users or the data source are located. Now, Alibaba Cloud ECS supports regions, including China East 1, China East 2, China North 1, China North 2, China South 1, Singapore, Hong Kong, US East 1 and US West 1. E-MapReduce supports regions including China North 2, China East 1, China East 2 and China South 1, and later it will extend to all 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

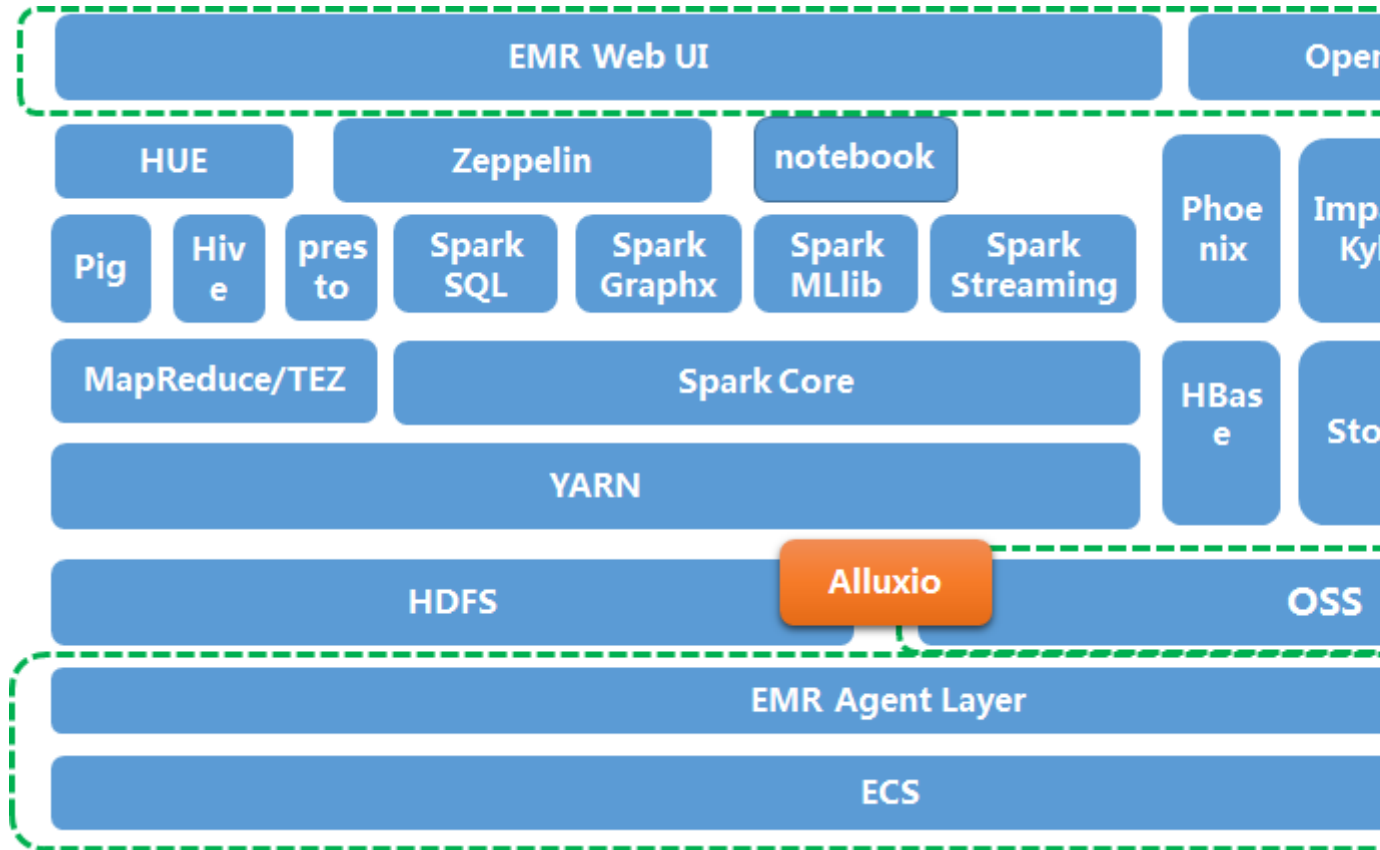
E-MapReduce can be subject to deep integration with other Alibaba Cloud products, so that they can be used as the input source or output destination of Hadoop or Spark computing engine in E-MapReduce.

- 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.

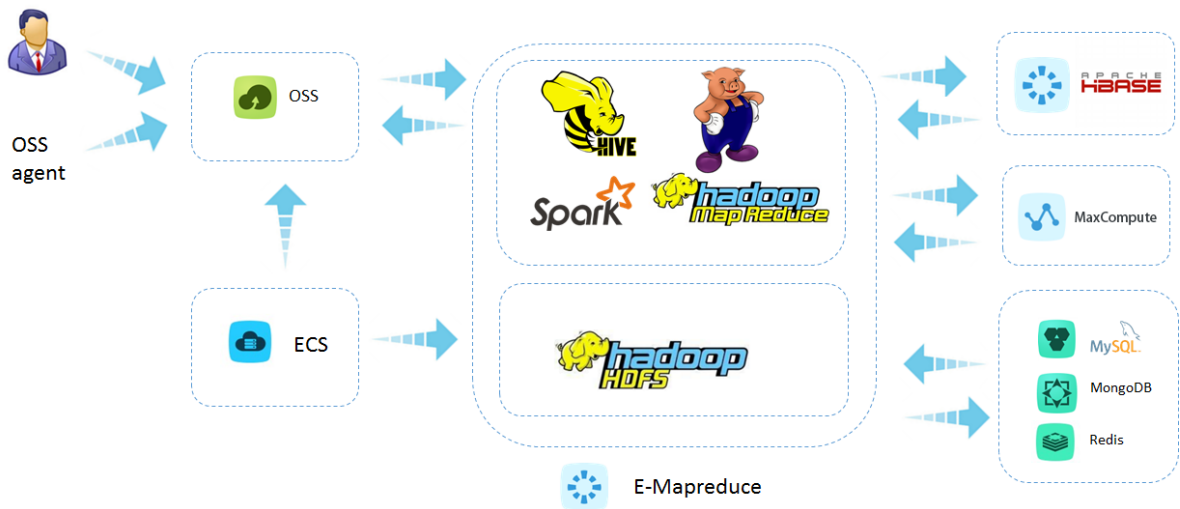


In the preceding figure, an E-MapReduce cluster is set based on the Hadoop ecological environment. It allows seamless data exchange with cloud services, such as Alibaba Cloud Object Storage Service (OSS) and ApsaraDB (RDS). This exchange allows users to share and transfer data between multiple systems and meet access needs for different types of businesses.

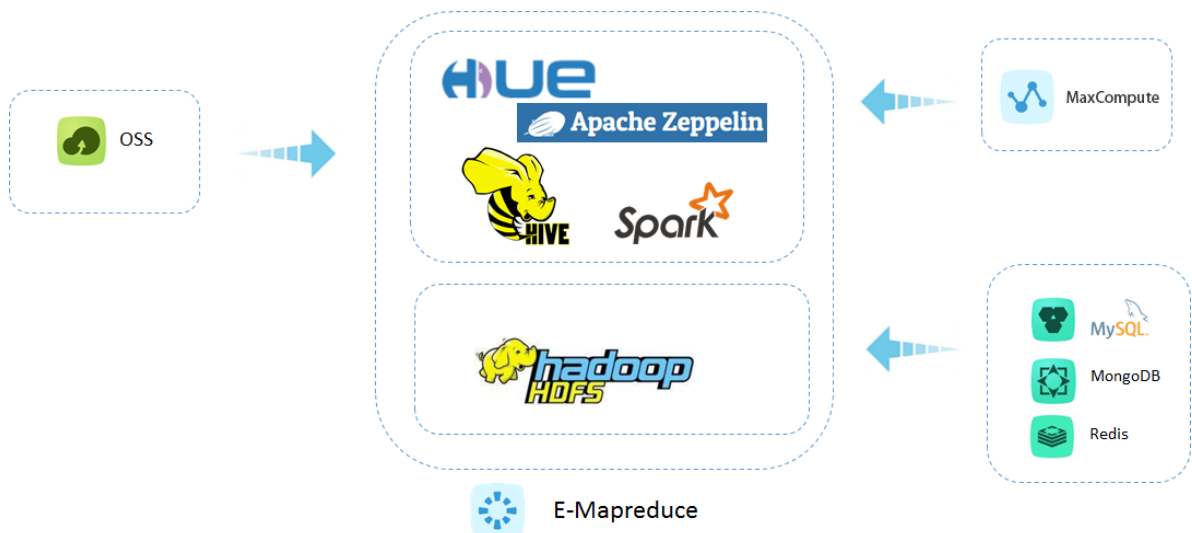
4 Scenarios

E-MapReduce clusters apply to a wide variety of application scenarios. E-MapReduce supports all scenarios that Hadoop ecosystem and Spark can support. This is because E-MapReduce is essentially taken as cluster services of Hadoop and Spark, allowing you to regard the Alibaba Cloud's ECS host used by E-MapReduce as your exclusive physical host. The following figures detail some classic application scenarios of E-MapReduce.

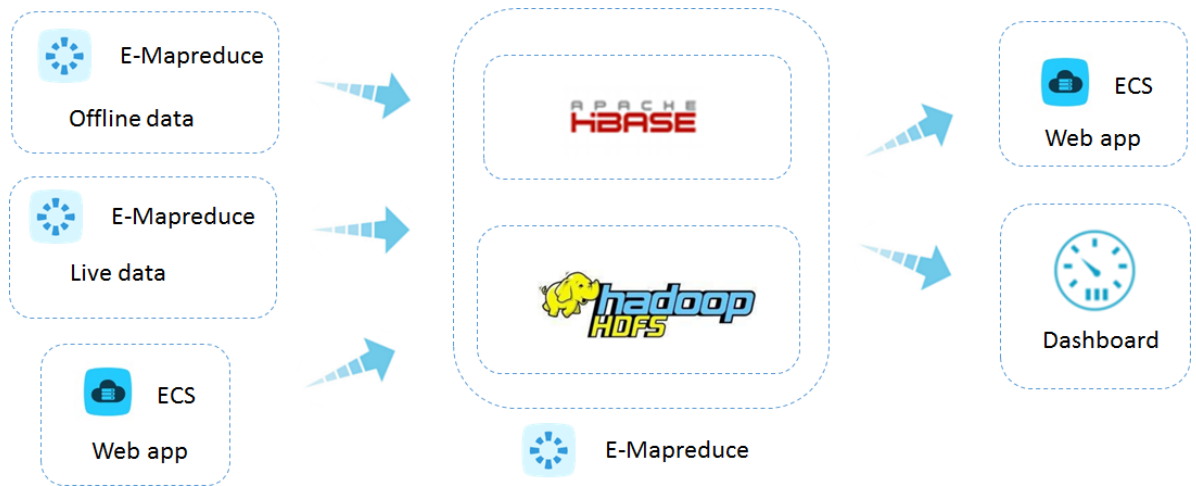
- **Offline data processing**



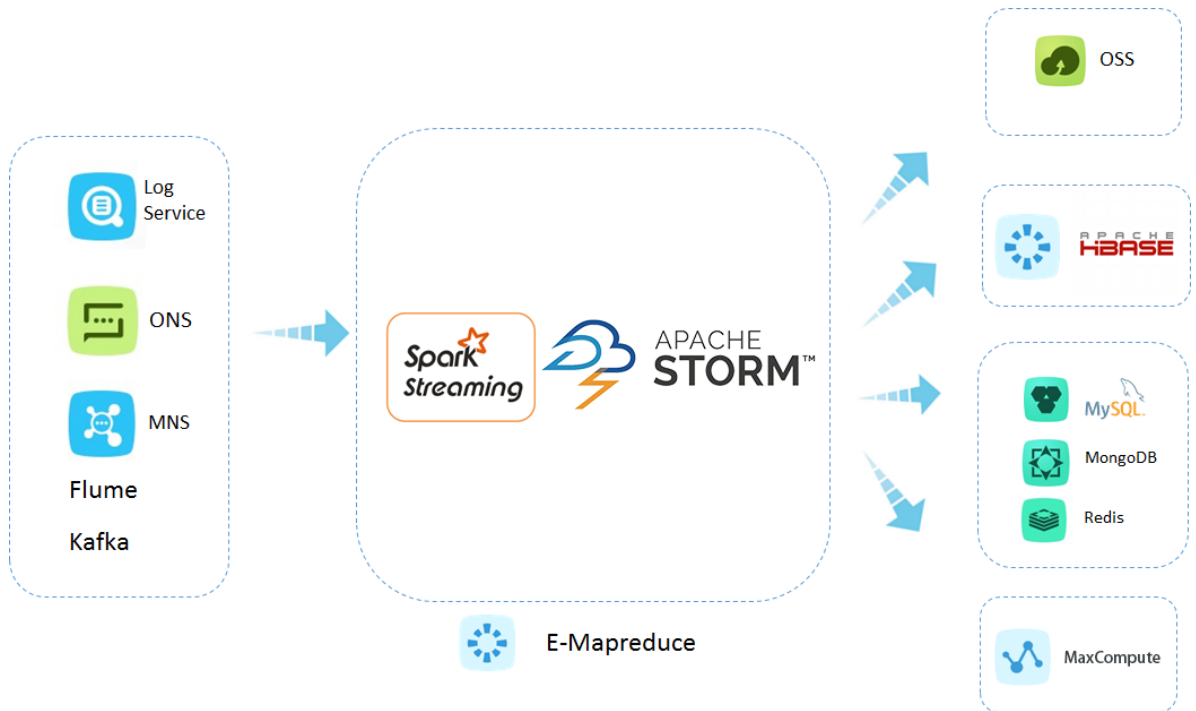
- **Ad-hoc data analysis queries**



- **Online massive data services**



• Stream data processing



5 Versioning

- E-MapReduce applies a version number rule in the a.b.c format:
 - a indicates major changes to the version.
 - b indicates moderate changes to some components in the version.
 - c indicates bug fixes in the version and can be compatible with previous versions.

For example, updates from 1.0.0 to 2.0.0 are major version changes. After a version upgrade, we recommend that you test to make sure all previous jobs can run normally. An update from 1.0.0 and 1.1.0 is a change generally conducted to upgrade a component version. We recommend that you perform a similar test to verify jobs run normally. An update from 1.0.0 and 1.0.1 is a c position change, and remains fully compatible with previous versions.

- The software and software version bound on each E-MapReduce release version are fixed. E-MapReduce does not support selection from multiple different versions of software, and manual changes to the software version are not recommended.
- If a released version of E-MapReduce is selected, and is then created on a cluster, the version used by the cluster is not upgraded automatically. The images corresponding to the subsequent version do not affect the currently cluster created after upgrade as only new clusters use the new images.
- When you upgrade the version of a cluster (for example, from 1.0.x to 1.1.x), we recommend that you must test your jobs to make sure that they run normally in the new software environment.

For more information about version details of E-MapReduce, see [Release notes](#).

6 Release notes

3.x

Version	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
Release Date	2018.1	2018.1	2018.2	2018.4	2018.6	2018.7	2018.8	2018.10	2018.11
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
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
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
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
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
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
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
Druid			0.11.0	0.11.0	0.11.0	0.12.0	0.12.2	0.12.3	0.12.3
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
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
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
Presto	0.188	0.188	0.188	0.188	0.188	0.188	0.208	0.208	0.208
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
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
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
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
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
Ranger			0.7.1	0.7.1	0.7.3	1.0.0	1.0.0	1.0.0	1.0.0
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
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
Tensorflow							1.8.0	1.8.0	1.8.0

Version	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
Kafka							2.11-1.0.1	2.11-1.0.1	2.11-1.0.1
Superset							0.25.6	0.25.6	0.27.0
Jupyter									4.4.0
Analytics Zoo									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
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 provide support for Alibaba Cloud OSS, the emr-core component is added based on the version of open-source Hadoop without any changes made to the original interface. The version of this component will be added after the Hadoop version. The emr-core component version follows the hadoop version.

EMR 3.1.x

EMR 3.1.1: OS is upgraded to CentOS7.2. Spark is upgraded to 2.1.1. emr-core is upgraded to 1.2.6. Fixed the OSS AK-free operation bug.

EMR 3.0.x

- **EMR 3.0.2:** emr-core is upgraded to 1.2.5. AK-free OSS supports more regions. Adjusted the replacement strategy of AK, fixed Hive, Hadoop related bugs.
- **EMR 3.0.1:** Supports interactive and unified table management. Saves hive meta using external unified database. emr-core is upgraded to 1.2.4. Optimized the read and write performance of OSS. All the clusters that use external hive meta share the same meta information. Spark is upgraded to 2.0.2.
- **EMR 3.0.0:** Use the upgraded version 3.0.1. It is fully compatible with the earlier version 3.0.0.

EMR 2.5.x

EMR 2.5.1: OS is upgraded to CentOS7.2. emr-core is upgraded to version 1.2.6. Fixed OSS AK-free operational bugs.

EMR 2.4.x

- **EMR 2.4.2 :** emr-core is upgraded to 1.2.5. OSS AK-free operation supports more regions. Adjusted the replacement strategy of the role AK. Fixed Hive and Hadoop related bugs.
- **EMR 2.4.1:** Supports interactive and unified table management, and uses external unified databases to save hive meta. emr-core is upgraded to version 1.2.4. Optimized the read and write performance of OSS. All the clusters using external hive meta share the same meta information .
- **EMR 2.4.0:** Use the upgraded version 2.4.1. It is fully compatible with version 2.4.0.

EMR 2.3.x

- **EMR 2.3.1:** Supports an interactive workbench where each cluster can use an independent internal database to save hive meta.
- **EMR 2.3.0:** An earlier version that is under maintenance. Supports an interactive workbench, but doesn't support table management. It is not recommended to use.

EMR 2.2.0

An temporary version that is deprecated.

EMR 2.1.x

- **EMR 2.1.1:** An earlier version that is under maintenance. Supports an interactive workbench, but doesn't support table management. It is not recommended to use.
- **EMR 2.1.0:** An earlier version that is deprecated.

EMR 2.0.x

An earlier version that is deprecated.

EMR 1.x

An earlier version that has few features. It is deprecated for the moment.