Alibaba Cloud

E-MapReduce

Product Introduction

C-J Alibaba Cloud

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 because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should 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 this document based on the "status quo", "being defective", and "existing functions" of its products and services. 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 take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, 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 contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, 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 this document 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 directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
⚠ Danger	A danger notice 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.
🔔 Warning	A warning notice 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 restart an instance.
🔊 Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	🔊 **Notice:** If the weight is set to 0, the server no longer receives new requests.
❓ Note	A note indicates supplemental instructions, best practices, tips, and other content.	❓ **Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
Courier font	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.What is E-MapReduce?

Alibaba Cloud E-MapReduce (EMR) is a big data processing solution provided by Alibaba Cloud.

Introduction

EMR is built on top of Alibaba Cloud Elastic Compute Service (ECS) and developed based on open source Apache Hadoop and Apache Spark. It allows you to conveniently use peripheral systems in the Hadoop and Spark ecosystems to analyze and process data. EMR can also read data from or write data to other Alibaba Cloud storage systems and database systems, such as Object Storage Service (OSS) and ApsaraDB RDS.
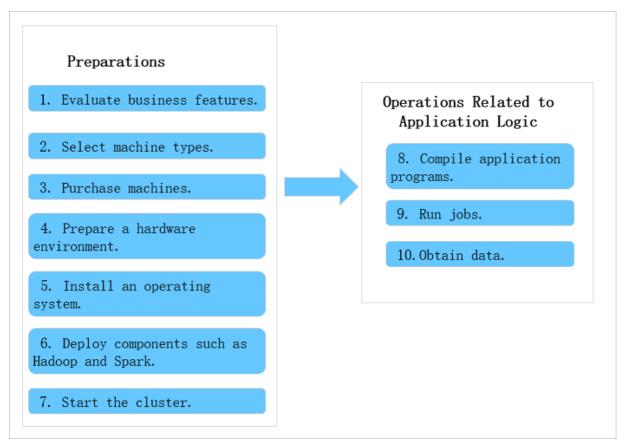
SmartData is a storage service for the EMR Jindo engine. SmartData provides centralized storage, optimized caching, and accelerated cache computing for EMR computing engines. It extends the storage features.

> ⑦ Note
> - For more information about Apache Hadoop, visit the Apache Hadoop official website.
> - For more information about Apache Spark, visit the Apache Spark official website.
> - For more information about Apache Hive, visit the Apache Hive official website.
> - For more information about Apache HBase, visit the Apache HBase official website.
> - For more information about SmartData, see SmartData.

Advantages over open source big data ecosystems

If you use an open source distributed processing system, such as Hadoop or Spark, to process data without using EMR, you must perform all the steps in the following figure.
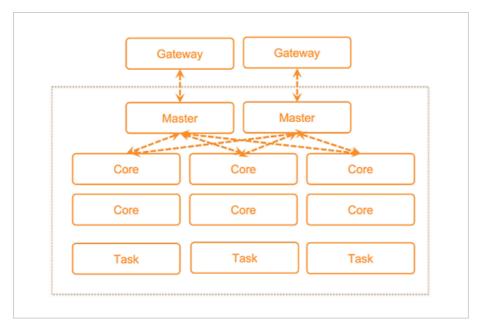
In this procedure, only the last three steps are related to your application logic. The first seven steps are all preparations, which are complex and time-consuming. EMR integrates all the required cluster management tools to provide the following features: host selection, environment deployment, cluster building, cluster configuration, cluster running, job configuration, job running, cluster management, and performance monitoring. This frees you from all the tedious procurement, preparation, and O&M work required to build clusters. You need only to focus on the processing logic of your applications.

EMR also offers different combinations of cluster services to meet your business requirements. For example, to perform daily data measurement and batch computing, you need only to run the Hadoop service for EMR. If you also want to perform stream computing and real-time computing, you can add the Spark service.

Composition of EMR clusters

Clusters are the core user-oriented component of EMR. An EMR cluster is a Hadoop, Flink, Druid, or ZooKeeper cluster that is deployed on one or more ECS instances. For example, a Hadoop cluster consists of some daemon processes, such as NameNode, DataNode, ResourceManager, and NodeManager. These daemon processes run on the ECS instances of the cluster. Each ECS instance corresponds to a node.

The following figure shows a Hadoop cluster that consists of master nodes, core nodes, and task nodes and is associated with a gateway cluster.

- Master nodes: run the following services and processes: NameNode of HDFS, JournalNode of HDFS, ZooKeeper, ResourceManager of YARN, and HMaster of HBase. You can determine whether to create a high-availability (HA) cluster based on the use scenario of the cluster. We recommend that you create a non-HA cluster in a test environment and create an HA cluster in a production environment. In an HA cluster, you can deploy two or three master nodes. If you deploy two master nodes, JournalNode of HDFS and ZooKeeper are deployed on the emr-worker-1 node. We recommend that you deploy three master nodes when you create an HA cluster in a production environment.

- Core nodes: run the DataNode process of HDFS and the NodeManager process of YARN to support data storage and computing. Core nodes do not support auto scaling.

- Task nodes: run the NodeManager process of YARN to support computing. You can flexibly add or remove task nodes based on your business requirements.

- Gateway cluster: is configured with Hadoop client files. You can use a gateway cluster to submit jobs to the Hadoop cluster associated with this gateway cluster. This way, you do not need to log on to the Hadoop cluster, which avoids security issues and environment isolation issues. You must create a Hadoop cluster first, and then create a gateway cluster and associate it with the Hadoop cluster.

2.Benefits

Alibaba Cloud E-MapReduce (EMR) provides you with convenient and controllable open source big data services at the enterprise level. It allows you to easily deploy open source big data services, such as Hadoop, Spark, Flink, Kafka, and HBase.

Stable and reliable open source components

- Open source components are used. Each version of EMR provides the latest versions of open source components. For more information about the mappings between EMR versions and open source component versions, see Overview.

- EMR is fully adapted to open source components and has eliminated the version compatibility issues among various open source components.

- EMR provides an enhanced deployment environment in Alibaba Cloud for open source components. This ensures a much higher performance than that in the open source community.

Cost effectiveness

- Compared with traditional HDFS-based clusters that use fixed configurations, EMR clusters use auto scaling and tiered storage mechanisms, which help reduce costs by more than 50%.

- You can create preemptible instances. Compared with the pay-as-you-go billing method, preemptible instances help reduce fees by 50% to 80%. For more information, see Overview.

Ease of use

- You can create or scale out a cluster within minutes. You do not need to manually deploy or start services.

- EMR provides a comprehensive monitoring and alerting system, which covers the hardware and Hadoop service of clusters. You can configure alert templates.

Scalability

- Compute-storage separation: Computing and storage are decoupled to support the elastic use of resources.

- Custom cluster environment: You can use bootstrap actions and cluster scripts to flexibly configure a cluster environment and deploy third-party optimization or cluster management tools in EMR. For more information, see Manage bootstrap actions and Cluster scripts.

- Self-managed maintenance: You can log on to the master node of a cluster, view the logs and deployment environment of the cluster, and optimize the configurations. For more information, see Common file paths.

- Auto scaling: EMR can automatically scale in or out clusters based on your business requirements.
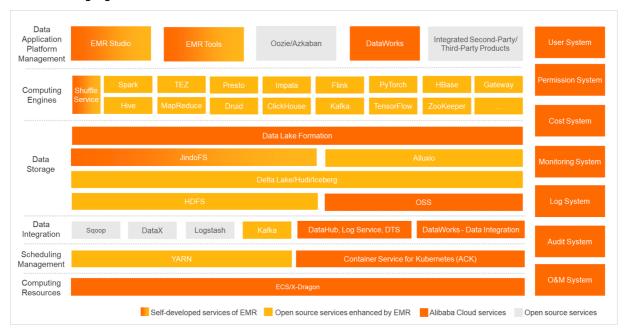
Deep integration

- You can deploy EMR clusters based on Alibaba Cloud Elastic Compute Service (ECS) and Container Service for Kubernetes (ACK). Various ECS instance types are supported. You can determine the instance types to use based on your business requirements. For more information, see ECS instances.

- EMR is integrated into DataWorks. You can use EMR as a job computing and data storage engine in DataWorks.

- Data Lake Formation (DLF) is integrated into EMR. In data lake scenarios, EMR allows you to manage metadata for multiple engines in a centralized manner.

3.Architecture

This topic describes the architecture of E-MapReduce (EMR).

The following figure shows the architecture of EMR.

EMR consists of four types of services:

- Open source services

  Apache big data services, such as Hadoop, Hive, and HBase, are integrated into EMR. The versions of the open source services are updated with EMR versions. For more information, see release notes in Overview.

  > 🔊 **Notice**   You are not allowed to update the version of a service in an existing EMR cluster.

- Open source services enhanced by EMR

  EMR enhances the performance and features of some open source services. Examples:

  ○ Spark Streaming SQL is added to Spark. This significantly improves the performance of Spark. For more information, see Common keywords.

  ○ The Z-ordering and Data Skipping features are added to Delta Lake. For more information, see Overview.

- Self-developed services of EMR

  EMR provides the following self-developed services, which ensure that open source components and services can better run on the Alibaba Cloud infrastructure:

  ○ Shuffle Service is an extended component of EMR. It is used to optimize the shuffle operations of computing engines. For more information, see ESS overview.

  ○ SmartData optimizes storage, caching, and computing for various EMR computing engines in a centralized manner and extends storage features. For more information, see SmartData.

- Alibaba Cloud services

EMR connects to both open source big data ecosystems and the Alibaba Cloud ecosystem. You can deploy EMR clusters on Alibaba Cloud Elastic Compute Service (ECS) instances or Container Service for Kubernetes (ACK) clusters and store data in Alibaba Cloud Object Storage Service (OSS). You can learn and use Machine Learning Platform for AI (PAI) in an EMR Data Science cluster. EMR is integrated into DataWorks, and you can use EMR as a job computing engine or data storage engine in DataWorks.

4.Scenarios
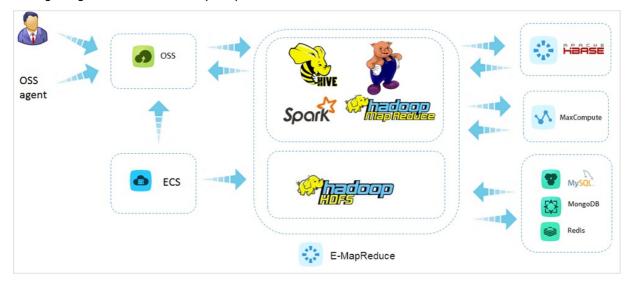
E-MapReduce (EMR) clusters are suitable for all the scenarios supported by the Hadoop ecosystem and Spark.

EMR is a cluster service based on Hadoop and Spark. You can use the Alibaba Cloud Elastic Compute Service (ECS) instances on which EMR clusters are deployed as your dedicated physical machines. The following examples show the classic scenarios of EMR:

- Batch processing of data
- Ad hoc queries for data analysis
- Online analysis of large amounts of data
- Processing of streaming data
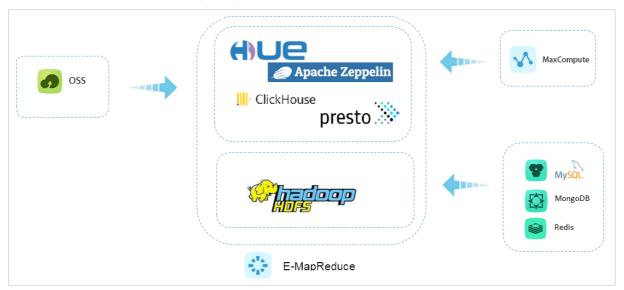
Batch processing of data

You can synchronize a large number of logs to the data nodes of EMR. Then, you can use tools such as Hue and a mainstream computing framework such as Hive, Spark, or Presto to get a quick insight into the data. You can also use tools such as Sqoop to load data from ApsaraDB RDS or other storage engines to EMR. Then, you can analyze the data and synchronize the results to ApsaraDB RDS or other storage engines. This feature helps implement data visualization.
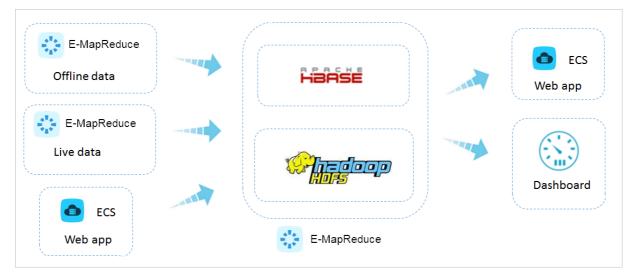
Ad hoc queries for data analysis

EMR allows you to directly import massive data to or use an external table to import massive data to online analytical processing (OLAP) analysis engines, such as ClickHouse, Presto, and Impala, for efficient, real-time, and flexible data analysis. This feature is suitable for business scenarios such as user persona analysis, audience selection, BI reporting, and business analysis.

Online analysis of large amounts of data
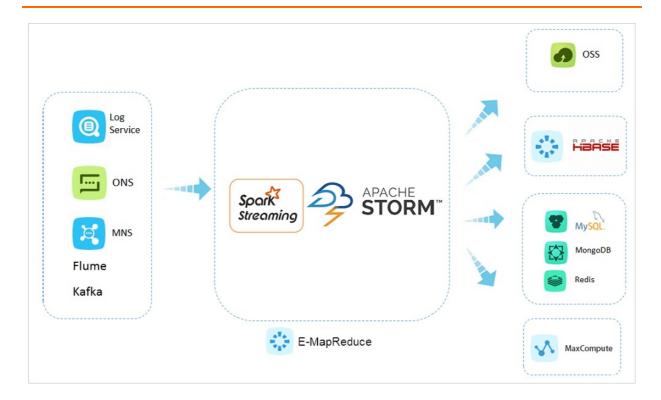
EMR analyzes petabytes of structured, semi-structured, or unstructured data generated by web apps and mobile apps. This allows web apps or data visualization services to visualize data in real time based on the analysis results obtained from EMR.

Processing of streaming data

EMR allows you to use and process real-time streaming data from services such as Log Service, Message Queue, Message Service (MNS), and Apache Kafka based on Spark Streaming and Storm.

EMR analyzes streaming data in fault-tolerant mode and writes analysis results to Object Storage Service (OSS) or HDFS.

5.Limits

When you use E-MapReduce (EMR), cluster instability may occur or clusters may become unavailable due to unexpected operations. Take note of the information in this topic to avoid these issues. This topic describes the limits of EMR.

> 🔊 **Notice**   When you use EMR, you must perform all operations in the EMR console. We recommend that you do not perform operations in the Elastic Compute Service (ECS) console. This may cause cluster instability or abnormalities. Take note of the information in this topic. In the case that you perform the operations that are prohibited, you will bear the consequences and responsibilities.

High-risk operations (prohibited)

Operation	Possible result	Suggestion
Delete or modify the hosts file that is stored in the etc/ directory.	You cannot find the services that run on the nodes of your cluster, which causes service exceptions.	Add information to the hosts file.
Modify parameters in component configuration files in the ECS console.	After specific services are restarted, the settings of the parameters that are modified are overwritten.	Modify the parameters in the EMR console.
Redeploy ECS instances in the ECS console.	The EMR service is affected.	submit a ticket.
Attach disks to the nodes of your EMR cluster in the ECS console.	The disks are unavailable because EMR cannot recognize and initialize the disks.	submit a ticket to replace disks if necessary.
Detach disks from the nodes of your EMR cluster in the ECS console.	This may cause data loss because EMR is unaware of the disk detaching operation.	
Remove core nodes in the ECS console.	This causes data loss, and execution failures of jobs on the removed nodes.	None

Operation	Possible result	Suggestion
Remove master nodes in the ECS console.	• For a high availability (HA) cluster, if you remove master nodes, the switchover of HDFS NameNode HA, YARN ResourceManager, or HBase HMaster fails. In this case, you must purchase a new EMR cluster to migrate data or tasks. • For a non-HA cluster, if you remove the master node, the cluster becomes unavailable, and you cannot migrate data or tasks.	None
Remove task nodes in the ECS console.	The jobs that you run on the removed nodes fail.	None
Stop the MySQL service of the master node. (Type is set to Built-in MySQL when you create an EMR cluster.)	The MySQL service deployed on the emr-header-1 node is associated with Hive MetaStore, Oozie, and Ranger. If you stop the MySQL service, the associated components cannot access the specific database.	None
Change the password of the root user that is used to access the MySQL service deployed on the emr-header-1 node. (Type is set to Built-in MySQL when you create an EMR cluster.)	The associated component such as Hue or Ranger fails.	None
Modify the security group to which ECS instances belong when an EMR cluster is running.	• The network connection between nodes is abnormal. • Components become unavailable.	None

Operation	Possible result	Suggestion
Change the billing method in the ECS console.	After you change the billing method, you can no longer change the new billing method back to the original one.	Change the billing method in the EMR console. For more information, see Switch from pay-as-you-go to subscription. 📢 **Notice** You cannot change the billing method from subscription to pay-as-you-go in the EMR console. If you want to change the billing method from subscription to pay-as-you-go, .

FAQ

Problem description	Solution
Insufficient disk capacity	Increase the capacity of a single disk or add core nodes in the EMR console. EMR clusters do not support the addition of disks.
Excess disk capacity	Purchase a new cluster and release the original one. For more information, see Create a cluster. EMR clusters do not support scale-down of disk capacity.
Insufficient computing capabilities	Add task nodes in the EMR console. For more information, see Scale out a cluster.
Excess computing capabilities	Resolve this issue based on the billing method of your cluster. • For a pay-as-you-go cluster, remove one or more task nodes from the cluster in the EMR console. • For a subscription cluster, stop the NodeManager of YARN on a specific task node, change the billing method of the ECS instance that serves as the task node to pay-as-you-go in the ECS console, and then release the instance.
Outdated component versions	Purchase a cluster of a later version. You are not allowed to update a specific component of an existing cluster. For more information, see Create a cluster.
Conversion from a non-HA cluster to an HA cluster	Non-HA clusters cannot be converted to HA clusters. We recommend that you purchase an HA cluster.
Deployment of third-party software or services on EMR	We recommend that you use bootstrap actions to install third-party software or third-party services when you create a cluster. If you manually install third-party software or third-party services after you create a cluster, you must reinstall the software or services when you add nodes.

6.Technical support

Alibaba Cloud E-MapReduce (EMR) is a solution provided by Alibaba Cloud. EMR is built on top of Alibaba Cloud Elastic Compute Service (ECS) or Container Service for Kubernetes (ACK) in semi-hosted mode. You are granted all management and operation permissions on EMR clusters that you created. This topic describes the EMR technical support for you from the following aspects:

- Support scope
- Seek help from Alibaba Cloud technical support
- SLA

Support scope

- Services supported
  - The EMR technical support team provides support for the use of the EMR console, provides related consultation services, and handles stability issues and product defects related to the console. The EMR console excludes components that are deployed in EMR clusters. You can perform the following operations in the EMR console:
    - Purchase an EMR cluster.
    - Create an EMR cluster.
    - Scale an EMR cluster.
    - Add services to an EMR cluster.
    - Configure auto scaling rules.
    - Use the monitoring and alerting feature.
    - Develop data.

  - The EMR technical support team provides support for the use of self-developed components in EMR, provides related consultation services, and handles stability issues and product defects related to the self-developed components. EMR provides the following self-developed components:
    - SmartData
    - Bigboot
    - ESS
    - DLF-Auth

  - The EMR technical support team assists in troubleshooting issues that occur on open source components in EMR and provides solutions in an iterative schedule based on the product plan. EMR supports the following open source components:
    - Hadoop components, including Hadoop Distributed File System (HDFS), YARN, and MapReduce
    - Components for data integration, such as Flume and Sqoop
    - Components for data storage, such as HDFS, HBase, Kudu, Iceberg, Delta Lake, Hudi, and Kafka
    - Compute engine components, such as Spark, Hive, Tez, Druid, Flink, and Phoenix
    - Online analytical processing (OLAP) components, such as Presto, ClickHouse, and Impala
    - Other components, such as ZooKeeper, Knox, Ganglia, OpenLDAP, Hue, Zeppelin, Livy, Superset, Ranger, Alluxio, and Oozie

- The EMR technical support team assists in troubleshooting issues that occur on other Alibaba Cloud services on which EMR depends. The service teams of these Alibaba Cloud services are responsible for the handling of service defects and provide solutions to meet service requirements. EMR depends on the following Alibaba Cloud services:
    - ECS or ACK
    - Object Storage Service (OSS)
    - CloudMonitor

- Services not supported

    - The EMR technical support team provides reliable O&M features such as monitoring and alerting and parameter configuration. However, the EMR technical support team does not perform O&M operations for cluster or component management. The O&M operations include cluster capacity planning, component performance optimization, and specific O&M operations on a cluster, such as service restart and parameter modification.

    - The EMR technical support team does not provide technical consultation on the development of business applications on top of EMR components or troubleshoot related issues. The development covers the following processes: business implementation, code development, job performance tuning, and business migration.

    - The EMR technical support team does not troubleshoot issues in a single job if no obvious exceptions in cluster components occur or no service defects exist.

    - The EMR technical support team does not provide support for non-standard capabilities of EMR, such as upgrade of the cluster version or component configurations, scale-in of core nodes, and reduction of disk capacity.

    - The EMR technical support team does not troubleshoot the issues that occur on third-party components.

    - The EMR technical support team does not troubleshoot the issue of cluster instability or cluster unavailability due to unexpected operations. For more information about unexpected operations, see Limits.

Seek help from Alibaba Cloud technical support

If you have questions, we recommend that you submit a ticket on the Alibaba Cloud official website to seek help from Alibaba Cloud technical support

1. Log on to the ticket system.

2. Click the ⌄ icon below **More Product And Service Category**.

3. Click **E-MapReduce** in the **More** section.

4. On the **New Ticket** page, configure the parameters.

5. Click **Submit**.

You can scan the following QR code to join the DingTalk group of EMR. In this group, you can discuss issues about EMR with other users.

> ⑦ **Note**   This DingTalk group is used only for communication between users. If you have any questions, submit a ticket.

SLA

For more information about SLA, see Service Level Comparison.

7.Release version
7.1. Overview

This topic describes the following information about the released versions of E-MapReduce (EMR): version format, version upgrade records, and release notes.

Version format

EMR versions follow the V**a.b.c** format, where:

- a indicates that major changes are made in a new version.
- b indicates that changes are made on some components in a new version.
- c indicates that several bugs are fixed in a new version and the version is forward compatible.

The following example describes the version number changes in a version upgrade:

- Upgrade from version 1.0.0 to 2.0.0: The value of a changes, which indicates that major changes are made in the new version. You must conduct a test to verify that all existing jobs can run properly after the upgrade.
- Upgrade from version 1.0.0 to 1.1.0: The value of b changes, which indicates that minor changes are made on some components in the new version. Most features remain unchanged after the upgrade. You must conduct a test to verify that all existing jobs can run properly after the upgrade.
- Upgrade from version 1.0.0 to 1.0.1: The value of c changes, which indicates that the new version is backward compatible with the previous version.

The following content describes the bundled software, cluster creation, and cluster upgrade of each EMR version:

- Bundled software: The software and software version are fixed in each EMR version. We recommend that you do not change the software version.
- Cluster creation: After you select an EMR version and create a cluster of this version, the cluster version is not automatically upgraded. Subsequent upgrades to the image used by this version do not affect the existing clusters. The new version of the image is used to create new clusters.
- Cluster upgrade: When you upgrade the version of your cluster, for example, from version 1.0.X to 1.1.X, you must conduct a test to verify that the existing jobs in the cluster can run properly in the new software environment.

Version upgrade records (EMR V5.X series)

- Release notes for EMR V5.6.X
- Release notes for EMR V5.5.X
- Release notes for EMR V5.4.X
- Release notes for EMR V5.3.X
- Release notes for EMR V5.2.X

Version upgrade records (EMR V4.X series)

- Release notes for EMR V4.10.X
- Release notes for EMR V4.9.X
- Release notes for EMR V4.8.X

- Release notes for EMR V4.6.X
- Release notes for EMR V4.5.X
- Release notes for EMR V4.4.X
- Release notes for EMR V4.3.X

Version upgrade records (EMR V3.X series)

- Release notes for EMR V3.40.X
- Release notes for EMR V3.39.X
- Release notes for EMR V3.38.X
- Release notes for EMR V3.37.X
- Release notes for EMR V3.36.X
- Release notes for EMR V3.35.X
- Release notes for EMR V3.34.X
- Release notes for EMR V3.33.X
- Release notes of EMR V3.32.X
- Release notes of EMR V3.30.X
- Release notes of EMR V3.29.X
- Release notes of EMR V3.28.X
- Release notes of EMR V3.27.X
- Release notes of EMR V3.26.X
- Release notes of EMR V3.25.X
- Release notes of EMR V3.24.X
- Release notes of EMR V3.23.X
- Release notes of EMR V3.22.X
- Release notes of versions earlier than E-MapReduce V3.22.X

EMR V4.3.0 to EMR V4.6.0

Version	EMR V4.6.0	EMR V4.5.0	EMR V4.4.1	EMR V4.3.0
Release date	2021.1	2020.12	2020.8	2020.5
Hadoop	3.2.1	3.1.3	3.1.3	3.1.3
HDFS	3.2.1	3.1.3	3.1.3	3.1.3
YARN	3.2.1	3.1.3	3.1.3	3.1.3
Hive	3.1.2	3.1.2	3.1.2	3.1.1
Spark	2.4.7	2.4.5	2.4.5	2.4.5
Knox	1.1.0	1.1.0	1.1.0	1.1.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2

Version	EMR V4.6.0	EMR V4.5.0	EMR V4.4.1	EMR V4.3.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
SmartData	3.2.0	3.1.0	2.7.301	2.6.7
Bigboot	3.2.0	3.1.0	2.7.301	2.6.7
ESS	1.0.0	1.0.0	N/A	N/A
Hudi	0.6.0	0.6.0	N/A	N/A
OpenLDAP	2.4.44	2.4.44	2.4.44	2.4.44
Hue	4.4.0	4.4.0	4.4.0	4.4.0
HBase	2.1.9	2.1.9	2.1.9	2.1.9
ZooKeeper	3.5.6	3.5.6	3.5.6	3.5.6
Presto	338	338	331	331
Impala	3.4.0	3.4.0	3.4.0	3.2.0
Zeppelin	0.9.0	0.8.2	0.8.1	0.8.1
Pig	N/A	0.14.0	0.14.0	0.14.0
Flume	1.9.0	1.9.0	1.9.0	1.9.0
Livy	0.6.0	0.6.0	0.6.0	0.6.0
Superset	0.36.0	0.36.0	0.36.0	0.28.1
Ranger	2.1.0	2.1.0	2.0.0	2.0.0
Flink	1.10-vvr-1.0.2	1.10-vvr-1.0.2	1.10-vvr-1.0.2	1.10-vvr-1.0.2
Storm	1.2.2	1.2.2	1.2.2	1.2.2
Kudu	1.11.1	1.11.1	1.11.1	1.11.1
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Kafka	2.4.1	2.4.1	2.4.1	2.4.1
Kafka-Manager	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16

EMR V3.28.2 to EMR V3.33.0

Version	EMR V3.33.0	EMR V3.32.0	EMR V3.30.0	EMR V3.29.0	EMR V3.28.2
Release date	2021.1	2020.11	2020.6	2020.6	2020.5
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
Knox	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0
Spark	2.4.7	2.4.5	2.4.5	2.4.5	2.4.5
Hive	2.3.7	2.3.5	2.3.5	2.3.5	2.3.5
Tez	0.9.2	0.9.2	0.9.2	0.9.2	0.9.2
Pig	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
YARN	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
HDFS	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
Flink	1.11-vvr-2.0.1	1.11-vvr-2.0.1	1.11-vvr-2.0.1	1.10-vvr-1.0.4	1.10-vvr-1.0.4
Druid	0.20.0	0.18.1	0.18.0	0.18.0	0.17.1
HBase	1.4.9	1.4.9	1.4.9	1.4.9	1.4.9
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1	4.14.1
ZooKeeper	3.5.6	3.5.6	3.5.6	3.5.6	3.5.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0	0.6.0
Presto	338	338	338	338	331
Storm	1.2.2	1.2.2	1.2.2	1.2.2	1.2.2
Impala	2.12.2	2.12.2	2.12.2	2.12.2	2.12.2
Flume	1.9.0	1.9.0	1.9.0	1.9.0	1.9.0
Hue	4.4.0	4.4.0	4.4.0	4.4.0	4.4.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0	5.1.0
Zeppelin	0.9.0	0.8.1	0.8.1	0.8.1	0.8.1
Ranger	1.2.0	1.2.0	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
SmartData	3.2.0	3.1.0	3.0.0	2.7.301	2.7.2

Version	EMR V3.33.0	EMR V3.32.0	EMR V3.30.0	EMR V3.29.0	EMR V3.28.2
Kafka	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1
Kafka-Manager	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16
Superset	0.36.0	0.36.0	0.36.0	0.35.2	0.35.2
Bigboot	3.2.0	3.1.0	3.0.0	2.7.301	2.7.2
OpenLDAP	2.4.44	2.4.44	2.4.44	2.4.44	2.4.44
Kudu	1.10.0	1.10.0	1.10.0	1.10.0	1.10.0
Flink-Vvp	1.11-2.2.2	1.11-2.2.2	1.11-2.2.2	1.11-2.2.2	1.11-2.2.2
PAI-Alink	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0
Alluxio	2.4.0	2.4.0	N/A	N/A	N/A
Hudi	0.6.0	0.6.0	N/A	N/A	N/A

EMR V3.25.1 to EMR V3.28.1

Version	EMR V3.28.1	EMR V3.28.0	EMR V3.27.2	EMR V3.26.3	EMR V3.25.1
Release date	2020.6	2020.6	2020.5	2020.4	2020.1
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
Knox	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0
Spark	2.4.5	2.4.5	2.4.3	2.4.3	2.4.3
Hive	2.3.5	2.3.5	2.3.5	2.3.5	2.3.5
Tez	0.9.2	0.9.2	0.9.2	0.9.2	0.9.2
Pig	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
YARN	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
HDFS	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
Flink	1.10-vvr-1.0.2	1.10-vvr-1.0.2	1.9.1	1.9.1	1.9.1
Druid	0.18.0	0.18.0	0.17.1	0.16.0	0.16.0
HBase	1.4.9	1.4.9	1.4.9	1.4.9	1.4.9

Version	EMR V3.28.1	EMR V3.28.0	EMR V3.27.2	EMR V3.26.3	EMR V3.25.1
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1	4.14.1
ZooKeeper	3.5.6	3.5.6	3.5.6	3.5.6	3.5.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0	0.6.0
Presto	331	331	331	310	310
Storm	1.2.2	1.2.2	1.2.2	1.2.2	1.2.2
Impala	2.12.2	2.12.2	2.12.2	2.12.2	2.12.2
Flume	1.9.0	1.9.0	1.9.0	1.9.0	1.9.0
Hue	4.4.0	4.4.0	4.4.0	4.4.0	4.4.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0	5.1.0
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1	0.8.1
Ranger	1.2.0	1.2.0	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
SmartData	2.7.1	2.7.0	2.6.5	2.6.3	2.2.3
Kafka	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1
Kafka-Manager	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16
Superset	0.35.2	0.35.2	0.35.2	0.28.1	0.28.1
Bigboot	2.7.1	2.7.0	2.6.5	2.6.3	2.2.3
OpenLDAP	2.4.44	2.4.44	2.4.44	2.4.44	2.4.44
Kudu	1.10.0	1.10.0	1.10.0	1.10.0	1.10.0
Analytics Zoo	N/A	N/A	N/A	0.5.0	0.5.0
Tensorflow on Spark	N/A	N/A	N/A	1.0.1	1.0.1

EMR V3.22.1 to EMR V3.25.0

Version	EMR V3.25.0	EMR V3.24.3	EMR V3.24.0	EMR V3.23.0	EMR V3.22.1
Release date	2020.1	2019.11	2019.11	2019.9	2019.7
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5

Version	EMR V3.25.0	EMR V3.24.3	EMR V3.24.0	EMR V3.23.0	EMR V3.22.1
Knox	1.1.0	1.1.0	1.1.0	1.1.0	1.1.0
Spark	2.4.3	2.4.3	2.3.5	2.4.3	2.4.3
Hive	2.3.5	2.3.5	2.3.5	2.3.5	2.3.5
Tez	0.9.2	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
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
YARN	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
HDFS	2.8.5	2.8.5	2.8.5	2.8.5	2.8.5
Flink	1.9.1	1.9.1	1.9.1	1.8.2	1.7.2
Druid	0.16.0	0.16.0	0.16.0	0.15.1	0.14.2
HBase	1.4.9	1.4.9	1.4.9	1.4.9	1.4.9
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1	4.14.1
ZooKeeper	3.5.6	3.5.5	3.5.5	3.5.5	3.5.5
Livy	0.6.0	0.6.0	0.6.0	0.6.0	0.6.0
Presto	310	0.228	0.228	0.221	0.221
Storm	1.2.2	1.2.2	1.2.2	1.2.2	1.2.2
Impala	2.12.2	2.12.2	2.12.2	2.12.2	2.12.2
Flume	1.9.0	1.9.0	1.9.0	1.9.0	1.9.0
Hue	4.4.0	4.4.0	4.4.0	4.4.0	4.4.0
Oozie	5.1.1	5.1.0	5.1.0	5.1.0	5.1.0
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1	0.8.1
Ranger	1.2.0	1.2.0	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
Tensorflow on Spark	1.0.1	1.0.1	1.0.1	1.0.0	1.0.0
Analytics Zoo	0.5.0	0.5.0	N/A	0.5.0	0.5.0
Kafka	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1

Version	EMR V3.25.0	EMR V3.24.3	EMR V3.24.0	EMR V3.23.0	EMR V3.22.1
Kafka-Manager	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16	1.3.3.16
Superset	0.28.1	0.28.1	0.28.1	0.28.1	0.28.1
Bigboot	2.2.3	2.2.2	2.2.1	2.0.2	2.0.1
SmartData	2.2.3	2.2.2	2.2.1	2.0.2	2.0.1
OpenLDAP	2.4.44	2.4.44	2.4.44	2.4.44	2.4.44
Kudu	1.10.0	1.10.0	1.10.0	1.10.0	1.10.0

Version upgrade records (earlier than EMR V3.22.X)

- **EMR V3.19.0 to EMR V3.21.0**

Version	EMR V3.19.0	EMR V3.19.1	EMR V3.20.0	EMR V3.21.0
Release date	2019.3	2019.4	2019.5	2019.6
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Knox	1.1.0	1.1.0	1.1.0	1.1.0
ApacheDS	2.0.0	2.0.0	2.0.0	2.0.0
Spark	2.4.1	2.4.1	2.4.2	2.4.3
Hive	3.1.1	3.1.1	3.1.1	3.1.1
Tez	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
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
YARN	2.8.5	2.8.5	2.8.5	2.8.5
HDFS	2.8.5	2.8.5	2.8.5	2.8.5
Flink	1.7.2	1.7.2	1.7.2	1.7.2
Druid	0.13.0	0.13.0	0.13.0	0.14.2
HBase	1.4.9	1.4.9	1.4.9	1.4.9
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13
Livy	0.6.0	0.6.0	0.6.0	0.6.0

Version	EMR V3.19.0	EMR V3.19.1	EMR V3.20.0	EMR V3.21.0
Presto	0.213	0.213	0.213	0.213
Storm	1.2.2	1.2.2	1.2.2	1.2.2
Impala	2.12.2	2.12.2	2.12.2	2.12.2
Flume	1.8.0	1.8.0	1.8.0	1.8.0
Hue	4.1.0	4.1.0	4.1.0	4.4.0
Oozie	4.2.0	4.2.0	5.1.0	5.1.0
Zeppelin	0.8.0	0.8.0	0.8.1	0.8.1
Ranger	1.2.0	1.2.0	1.2.0	1.2.0
Ganglia	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
Tensorflow	N/A	N/A	1.8.0	1.8.0
Kafka	2.11-1.1.1	1.1.1	2.11	1.1.1
Superset	0.28.1	0.28.1	0.28.1	0.28.1
Jupyter	N/A	N/A	4.4.0	4.4.0
Analytics Zoo	N/A	N/A	0.2.0	0.5.0
Bigboot	N/A	N/A	1.0.0	1.0.0
OpenLDAP	N/A	N/A	N/A	N/A
Kudu	N/A	N/A	N/A	N/A

- **EMR V3.15.0 to EMR V3.18.1**

Version	EMR V3.15.0	EMR V3.16.0	EMR V3.17.0	EMR V3.18.1
Release date	2018.11	2018.12	2019.1	2019.2
Hadoop	2.7.2	2.7.2-1.3.2	2.7.2	2.8.5
Knox	0.13.0	0.13.0	1.1.0	1.1.0
ApacheDS	2.0.0	2.0.0	2.0.0	2.0.0
Spark	2.3.2	2.3.2-1.0.1	2.3.2	2.3.2
Hive	2.3.3	2.3.3-1.0.3	2.3.3	3.1.1

Version	EMR V3.15.0	EMR V3.16.0	EMR V3.17.0	EMR V3.18.1
Tez	0.9.1	0.9.1-1.0.2	0.9.1	0.9.1
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	1.4.7	1.4.7-1.0.0	1.4.7	1.4.7
YARN	2.7.2	2.7.2	2.7.2	2.8.5
HDFS	2.7.2	2.7.2	2.7.2	2.8.5
Flink	1.4.0	1.6.2-1.0.0	1.6.2	1.6.2
Druid	0.12.3	0.12.3-1.0.1	0.12.3	0.13.0
HBase	1.1.1	1.1.1-1.0.2	1.1.1	1.4.9
Phoenix	4.10.0	4.10.0-1.0.0	4.10.0	4.14.1
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13
Livy	0.5.0	0.5.0	0.5.0	0.5.0
Presto	0.208	0.208	0.213	0.213
Storm	1.1.2	1.2.2	1.2.2	1.2.2
Impala	2.10.0	2.10.0-1.0.0	2.12.2	2.12.2
Flume	N/A	1.8.0	1.8.0	1.8.0
Hue	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
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
Ranger	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
OS	CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4
Tensorflow	1.8.0	1.8.0	1.8.0	N/A
Kafka	2.11-1.0.1	2.11-1.1.0	2.11-1.1.1	2.11-1.1.1
Superset	0.27.0	0.28.1	0.28.1	0.28.1
Jupyter	4.4.0	4.4.0	4.4.0	N/A
Analytics Zoo	0.2.0	0.2.0	0.2.0	N/A

- **EMR V3.11.0 to EMR V3.14.0**

Version	EMR V3.11.0	EMR V3.12.0	EMR V3.13.0	EMR V3.14.0
Release date	2018.6	2018.7	2018.8	2018.10
Hadoop	2.7.2-emr-1.2.14	2.7.2-emr-1.2.14	2.7.2	2.7.2
Knox	0.13.0	0.13.0	0.13.0	0.13.0
ApacheDS	2.0.0	2.0.0	2.0.0	2.0.0
Spark	2.2.1	2.3.1	2.3.1	2.3.1
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Tez	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
Sqoop	1.4.6	1.4.7	1.4.7	1.4.7
YARN	2.7.2	2.7.2	2.7.2	2.7.2
HDFS	2.7.2	2.7.2	2.7.2	2.7.2
Flink	1.4.0	1.4.0	1.4.0	1.4.0
Druid	0.11.0	0.12.0	0.12.2	0.12.3
HBase	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
ZooKeeper	3.4.11	3.4.12	3.4.12	3.4.13
Presto	0.188	0.188	0.208	0.208
Storm	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
Hue	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
Zeppelin	0.7.3	0.7.3	0.8.0	0.8.0
Ranger	0.7.3	1.0.0	1.0.0	1.0.0
Ganglia	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

Version	EMR V3.11.0	EMR V3.12.0	EMR V3.13.0	EMR V3.14.0
Tensorflow	N/A	N/A	1.8.0	1.8.0
Kafka	N/A	N/A	2.11-1.0.1	2.11-1.0.1
Superset	N/A	N/A	0.25.6	0.25.6
Jupyter	N/A	N/A	N/A	N/A
Analytics Zoo	N/A	N/A	N/A	N/A

- **EMR V3.7.1 to EMR V3.10.1**

Version	EMR V3.7.1	EMR V3.8.1	EMR V3.9.1	EMR V3.10.1
Release date	2018.1	2018.1	2018.2	2018.4
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
Knox	0.13.0	0.13.0	0.13.0	0.13.0
ApacheDS	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
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Tez	0.8.4	0.8.4	0.8.4	0.9.1
Pig	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
YARN	2.7.2	2.7.2	2.7.2	2.7.2
HDFS	2.7.2	2.7.2	2.7.2	2.7.2
Flink	N/A	1.4.0	1.4.0	1.4.0
Druid	N/A	N/A	0.11.0	0.11.0
HBase	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
ZooKeeper	3.4.11	3.4.11	3.4.11	3.4.11
Presto	0.188	0.188	0.188	0.188
Storm	1.0.1	1.0.1	1.0.1	1.1.2
Impala	2.10.0	2.10.0	2.10.0	2.10.0

Version	EMR V3.7.1	EMR V3.8.1	EMR V3.9.1	EMR V3.10.1
Hue	3.12.0	3.12.0	3.12.0	4.1.0
Oozie	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
Ganglia	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

- **EMR V2.9.2 to EMR V2.11.0**

Version	EMR V2.9.2	EMR V2.10.0	EMR V2.11.0
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

- **EMR V1.0.0 to EMR V1.3.0**

Version	EMR V1.0.0	EMR V1.1.0	EMR V1.2.0	EMR V1.3.0
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	N/A	N/A	N/A	1.4.6
Hue	N/A	N/A	N/A	3.9.0
Zeppelin	N/A	N/A	N/A	0.5.6
HBase	N/A	N/A	1.1.1	1.1.1
Phoenix	N/A	N/A	N/A	N/A
ZooKeeper	N/A	N/A	3.4.6	3.4.6
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

> ? **Note**   About Hadoop versions:
>
> To ensure compatibility with Alibaba Cloud Object Storage Service (OSS), the emr-core component is deployed based on an open source Hadoop version. Existing APIs remain unchanged. The version number of the emr-core component is suffixed to the end of the Hadoop version.

7.2. Release notes (EMR V5.X series)
7.2.1. Release notes for EMR V5.6.X

This topic describes the release notes for E-MapReduce (EMR) V5.6.X, including the release date, updates, and release version information.

Release date

April 21, 2022 for EMR V5.6.0

Updates

Service	Description
JindoData	The JindoData service is added. The service version is 4.3.0.
JindoSDK	JindoSDK is updated to 4.3.0.

Service	Description

Spark	Spark is updated to 3.2.1.
Hive	The following issue is fixed: After speculative execution is enabled for Hive on Tez, both the original task and the speculative task are committed.
Presto	The following issue is fixed: After a Hadoop cluster is initialized and the Presto service is added to the cluster, the Presto service fails to start.
DeltaLake	Subqueries can be nested in DML statements.
Hudi	Hudi is updated to 0.10.1.
Iceberg	Iceberg is updated to 0.13.1.
YARN	A feature that allows you to run an ApplicationMaster only on a core node is added.
HBase	The issue of the HBase 2.3.4 kernel is fixed.
ZooKeeper	The configuration of Java Virtual Machine (JVM) parameters is optimized.
Impala	Impala is adapted to JindoSDK 4.3.0.
Sqoop	The PostgreSQL version is updated.
Zeppelin	The issue that the Java Database Connectivity (JDBC) interpreter fails to start is fixed.
Ranger	Delta tables and Hudi tables are supported by the Spark plugin of Ranger 1.2.0.
Flume	Flume is adapted to JindoSDK 4.3.0.
Oozie	Log4j is updated to 2.17.2.
DLF-Auth	DLF-Auth is updated to 2.0.0.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1

Service	Version
Hive	3.1.2
Spark	3.2.1
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
JindoData	4.3.0
DLF-Auth	2.0.0
Iceberg	0.13.1
Hudi	0.10.1
DeltaLake	1.1.0
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.3.0
HBase	2.3.4
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.1
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
RSS	1.0.0
Alluxio	2.5.0
Kudu	1.14.0

Service	Version
Oozie	5.2.1

Service	Version
RSS	1.0.0

7.2.2. Release notes for EMR V5.5.X

This topic describes the release notes for E-MapReduce (EMR) V5.5.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V5.5.1	March 25, 2022
EMR V5.5.0	February 15, 2022

Updates

EMR V5.5.1

> ⑦ **Note** Only OLAP clusters support this version. You can create an OLAP cluster in the new EMR console.

Service	Description
ClickHouse	The default values of some parameters are changed.
StarRocks	StarRocks is updated to 2.1.1.

EMR V5.5.0

Service	Description
SmartData	The service is no longer used.
Bigboot	
RSS	• EMR Remote Shuffle Service (ESS) is upgraded to Remote Shuffle Service (RSS). For more information, see RSS. • The service functionality and stability are enhanced.

Service	Description
JindoSDK	- The architecture of SmartData is upgraded to JindoData. - EMR is integrated with JindoSDK for JindoData 4.0.0 for the first time. JindoData connects to Alibaba Cloud Object Storage Service (OSS) and the Alibaba Cloud OSS-HDFS service.
Spark	- An IF expression can be used in the COUNT DISTINCT function, and the CASE WHEN syntax for the COUNT DISTINCT function is optimized. To use this feature, set spark.sql.optimizer.rewriteConditionalDistinctAggregates to true. - Fallback from Shuffle Hash Join to Sort Merge Join is supported. To use this feature, set spark.sql.join.preferSortMergeJoin to false and set spark.sql.join.enableShuffledHashJoinFallback to true. - Automatic merging of small files in non-dynamic partitions is supported. The data of merged files is written to tables. To use this feature, set spark.sql.adaptive.merge.output.small.files.enabled to true. - The concurrency in scenarios in which the GROUPING SETS clause or the DISTINCT function is used can be automatically adjusted. To use this feature, set spark.sql.execution.optimizeExpand to true. - Hive on Spark is optimized. - The syntax of the time travel feature is supported. - Spark is adapted to JindoSDK.
Tez	Tez is adapted to JindoSDK.
Hive	- The issue about batch deletion that occurs on Hive Jindo is fixed. - The out of memory (OOM) issue that occurs on HiveServer2 is fixed. - Hive on Spark is optimized. - Hive is adapted to JindoSDK.

Service	Description
Presto	- Presto is updated to 358. - By default, the MySQL, Iceberg, Hudi, Phoenix, Kudu, and Delta connector types are added. The default configurations of these connector types are updated. - Data lake analysis is supported. - User-defined functions (UDFs) can be dynamically loaded. - Presto is adapted to JindoSDK.
Delta Lake	- Version update - Delta Lake is updated to 1.1.0 and compatible with Spark 3.2.0. - All features that are released for commercial use are migrated to Delta Lake 1.1.0. - Metadata management - The feature of synchronizing metadata changes to metastores is optimized. - The statistics on table data are automatically reported to metastores. - SQL - The syntax of the time travel feature is supported. - The DROP PARTITION SQL syntax is supported. - SQL statements can be executed to overwrite data in dynamic partitions. - The ADD COLUMN statement can be used to add columns to specified locations (FIRST and AFTER). - Enhanced table management capabilities - The file size can be dynamically adjusted based on the table size. By default, this feature is enabled. - The auto-vacuum feature is supported and enabled by default. Concurrent vacuum operations are supported. - The logic of automatic compaction is optimized. By default, the automatic compaction feature is disabled. - The Z-ordering syntax is added. Z-ordering-based data processing is accelerated.
Hudi	- Hudi is updated to 0.10.0. - Spark 3.2.0 is supported. - JindoFS in block storage mode is supported.
HDFS	HDFS is adapted to JindoSDK.

Service	Description
YARN	• YARN is adapted to RSS memory configurations. • YARN is adapted to JindoSDK.
Flume	Flume is adapted to JindoSDK.
Impala	Impala is adapted to JindoSDK.
Ranger	• Spark 3.2.0 is supported. • Presto 358 is supported.
HBase	• The issue about default parameter settings is fixed. • The issue about the date format of garbage collection (GC) logs is fixed.
ClickHouse	• The transaction feature is added. For more information, see Use a transaction. • The Hadoop Distributed File System (HDFS) and OSS disk types are added to support the separation of hot and cold data. For more information, see Separate hot and cold data by using HDFS and Separate hot and cold data by using OSS. • When you use an engine of the Replicated*MergeTree type, the following disk types are supported: zero copy provided by Amazon S3, OSS, and HDFS. • The logic of processing data when the ClickHouse component stops working is optimized.
Iceberg	• Iceberg is updated to 0.13.0. • Presto 358 is supported.
DLF-Auth	• Spark 3.2.0 is supported. • Presto 358 is supported.

Release version information

> ⑦ **Note**　To view the information about an OLAP cluster, you must log on to the new EMR console.

○ Hadoop cluster ○ ClickHouse cluster ○ Shuffle Service cluster ○ OLAP cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1

Service	Version
Hive	3.1.2
Spark	3.2.0
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
DLF-Auth	1.0.4
Iceberg	0.13.0
Hudi	0.10.0
Delta Lake	1.1.0
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.0.0
HBase	2.3.4
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.2
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
RSS	1.0.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
ClickHouse	21.3.13.9

Service	Version
RSS	1.0.0

Service	Version
ClickHouse	21.3.13.9.2.9
StarRocks	2.1.1
ZooKeeper	3.6.3

7.2.3. Release notes for EMR V5.4.X

This topic describes the release notes for E-MapReduce (EMR) V5.4.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V5.4.3	December 2021
EMR V5.4.2	December 2021
EMR V5.4.1	November 2021
EMR V5.4.0	October 2021

Updates

- EMR V5.4.3

  The Log4j security vulnerability of all related components is fixed. For more information, see Vulnerability announcement | RCE vulnerability in Apache Log4j 2.

Service	Description
Presto	The Log4j security vulnerability of the Elasticsearch connector is fixed.

Service	Description
Data Lake Formation (DLF) Metastore	○ By default, the logging feature for metastores is disabled. In earlier versions, the feature is enabled by default. ○ The error caused by an excessively long getTableStats URI for a metastore is fixed.
Delta Lake	The issue that schema changes fail to be synchronized to a metastore is fixed.
Sqoop	The issue that precision loss for the DECIMAL data type occurs when you use Sqoop to import data to HCatalog tables is fixed.

- EMR V5.4.2

Service	Description
SmartData	○ SmartData is updated to 3.8.0. For more information, see SmartData 3.8.X overview. ○ Authentication and authorization based on Kerberos and Ranger can be used to manage permissions on data in OSS.

- EMR V5.4.1

Service	Description
SmartData	SmartData is updated to 3.7.3. For more information, see SmartData 3.7.X overview.
Oozie	The issue that Jetty Server of Oozie fails to start due to JAR package conflicts in high availability (HA) scenarios is fixed.
Impala	The issue that the `no such method error` message appears when you query data in DLF metadata tables is fixed.
DLF-Auth	DLF-Auth is updated to 1.0.1.

- EMR V5.4.0

Service	Description
SmartData	SmartData is updated to 3.7.2. For more information, see SmartData 3.7.X overview.

Service	Description
Spark	- Spark is updated to 3.1.2. - In Spark 3.x, the Distinct computing performance is optimized for Spark SQL. The optimization feature is triggered if an aggregation operator contains multiple count(distinct case ... when ...) methods. - The array-index out of bounds error that is returned when some required statistics for Adaptive Query Execution (AQE) are missing is fixed. - Errors related to AQE and data caching in specific scenarios are fixed.
Hive	In JindoFS in block storage mode, the metadata of multiple Hive tables can be optimized at the same time. By default, this feature is disabled.
Presto	Storage handlers can be used to query data in Delta tables.
Delta Lake	- Delta Lake is updated to 1.0.0. - The same Delta Lake connectors are used in Hive 2 and Hive 3. - The error that is returned when you use Delta Lake connectors to query data from multi-level partitioned tables is fixed. - The SQL syntax that supports various features, such as Data Skipping, Optimize, and Z-ordering, can be used. - Metadata can be synchronized to a metastore.
Hudi	- Hudi is updated to 0.9.0. - The issue about the compatibility of sql.extension between Delta Lake and Hudi is fixed. ⓘ Note Spark 3.1.2 is supported.
HDFS	By default, the reserved space of NameNode adaptively increases. This way, NameNode enters the Safe mode in a timely manner when the disk space is insufficient.
Storm	The service is no longer used.
Zeppelin	Zeppelin is updated to 0.10.0.
Hue	- The issue that YARN Job Browser sometimes cannot present or terminate jobs is fixed. - YARN Job Browser is accessible by default. - The Presto protocol is supported by default.

Service	Description
Druid	The following issue is fixed: After a server is unexpectedly shut down, the related node fails to restart because a PID file is not deleted.
ClickHouse	- Some default configurations are updated. - Clusters can be scaled out. - The MetaChecker feature is supported. - Object Storage Service (OSS) table engines and OSS table functions can be used to read data.
Iceberg	- Iceberg is updated to 0.12.0-1.0.1. - Errors related to the Hive runtime dependency are fixed.
Knox	The issue that the first access to the Spark UI fails is fixed.
DLF-Auth	The service is added. The permissions of using Hive or Spark to access DLF can be configured. The supported version is 1.0.0.

Release version information

○ Hadoop cluster ○ Kafka cluster ○ ClickHouse cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	3.1.2
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	EMR V5.4.0: 3.7.2
Bigboot	EMR V5.4.1: 3.7.3 EMR V5.4.2: 3.8.0
Iceberg	0.12.0
DLF-Auth	1.0.0

Service	Version
Hudi	0.9.0
Delta Lake	1.0.0
OpenLDAP	2.4.44
Hue	4.9.0
HBase	2.3.4
ZooKeeper	3.6.3
Presto	338
Impala	3.4.0
Zeppelin	0.10.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
ESS	1.2.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
ClickHouse	21.3.13.9

7.2.4. Release notes for EMR V5.3.X

This topic describes the release notes for E-MapReduce (EMR) V5.3.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V5.3.1	September 2021
EMR V5.3.0	August 2021

Updates

- EMR V5.3.1

Service	Description
SmartData	SmartData is updated to 3.7.1.
Hue	The issue that Impala cannot be used in a high-security cluster is fixed.
Kudu	Kerberos is supported.
HBase	○ The time required to restart HBase in a high-security cluster is reduced. ○ The issue that Spark 3.1.1 cannot be integrated with HBase is fixed. ○ The Graceful Stop process is optimized.

- EMR V5.3.0

Service	Description
SmartData	SmartData is updated to 3.7.0.
Spark	The incompatibility issue between Spark and Delta Lake is fixed.

Service	Description
Hive	In JindoFS in block storage mode, the metadata of multiple Hive tables can be optimized at the same time. By default, this feature is disabled.
Delta Lake	- The partitioning feature of Delta Lake is supported. - The incompatibility issue between the `DESC DETAIL` command and Spark 3.1.1 is fixed.
YARN	- Information about app IDs, CPU utilization, and memory usage is added to the REST APIs of containers for nodes. - The issue that the Application Master (AM) logs of an automatically released node cannot be viewed is fixed. - The issue that a cluster cannot be accessed due to historical state store data is fixed. - The data of a node that is automatically released based on the decommissioning logic of auto scaling can be deleted. - The Graceful Decommission logic of auto scaling is optimized. The node on which NodeManager runs is marked deprecated only after the NodeManager process is complete.
ZooKeeper	ZooKeeper is updated to 3.6.3.
Flink	- The SmartData service is added. - Object Storage Service (OSS) can be accessed in password-free mode when you log on to a DataFlow-Flink cluster in SSH mode and submit a job in the cluster.
Impala	The issue that the LIST operation is repeatedly performed on directly deleted OSS partition directories is fixed.
Hue	The UI display exception that occurs when you use Hue together with Oozie is fixed.
Kudu	Kudu is updated to 1.14.0.
ClickHouse	ClickHouse is updated to 21.3.13.9.
Iceberg	The Iceberg service is added. Iceberg 0.12.0 is used.

Release version information

○ Hadoop cluster ○ ClickHouse cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	3.1.1
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	EMR V5.3.0: 3.7.0
Bigboot	EMR V5.3.1: 3.7.1
Iceberg	0.12.0
Hudi	0.8.0
Delta Lake	0.8.0
OpenLDAP	2.4.44
Hue	4.9.0
HBase	2.3.4
ZooKeeper	3.6.3
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
Storm	1.2.2
ESS	1.2.0

Service	Version
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
ClickHouse	21.3.13.9

Service	Version
ESS	1.2.0

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

7.2.5. Release notes for EMR V5.2.X

E-MapReduce (EMR) V5.2.X is the first stable version of the EMR V5.X series. This topic describes the release notes for EMR V5.2.X, including the release date, updates, and release version information.

Release date

July 16, 2021 for EMR V5.2.1

Updates

Service	Description

Service	Description
SmartData	SmartData is updated to 3.6.1. For more information, see SmartData 3.6.X overview.
Hive	- The issue that the output of the `show create table` command based on Data Lake Formation (DLF) metadata is inaccurate is fixed. - The default parameters of Hive are optimized to improve the performance of Hive jobs. - In the EMR console, the parameter names on the **hive-env** tab of the **Configure** tab for the Hive service are changed to uppercase. This facilitates the use of the parameters. - The issue that user-defined functions (UDFs) cause HiveServer2 memory leak is fixed. - The error message that is reported because of the incompatibility between the file system and Hive metastore when you write data to a Hive table is optimized.
HDFS	The data compression algorithm Zstandard is supported.
Delta Lake	- Delta Lake is updated to 0.8.0. - Spark 3 is supported.
Flink	Flink is updated to 1.12-vvr-3.0.2.
Hudi	- Hudi is updated to 0.8.0. - Hudi can be integrated with Spark SQL.
Spark	⚲ **Notice** In EMR V5.2.1, Spark (3.1.1) and Kudu (1.11.1) are incompatible with each other. - Delta Lake and Hudi are supported. - Remote Shuffle Service is supported. - Livy is supported. - In the EMR console, the parameter names on the **spark-defaults** tab of the **Configure** tab for the Spark service are optimized. - The cost-based optimization (CBO), dynamic partition pruning, and Z-order features are optimized. The performance of these features is 50% higher than in Spark 3. - Log Service, DataHub, and Message Queue for Apache RocketMQ can be used as data sources.
Tez	The default parameters of Tez are optimized to improve the performance of Tez jobs.

Service	Description
Ranger	• The warning error contained in logs about starting Spark in Ranger is fixed. • The issue that user information fails to be automatically synchronized after Ranger is connected to a Lightweight Directory Access Protocol (LDAP) server is fixed.
Knox	• Knox is adapted to Kudu. • Knox is adapted to HBase.
Kafka	• The Cruise Control component can be used to provide the balance feature for Kafka clusters. • Disks for Kafka clusters are hot-swappable. You can replace a damaged disk without the need to stop the Kafka broker of your cluster. • The default values of some parameters are changed.
Phoenix	The issue that no JDBC driver is found when Hive or Spark SQL is used to access Phoenix tables is fixed.
EMR Remote Shuffle Service (ESS)	Spark 3 is supported.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	3.1.1
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.6.1
Bigboot	3.6.1
Hudi	0.8.0

Service	Version
OpenLDAP	2.4.44
Hue	4.9.0
HBase	2.3.4
ZooKeeper	3.6.2
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
Storm	1.2.2
ESS	1.0.0
Alluxio	2.5.0
Kudu	1.11.1
Oozie	5.1.0

Service	Version
ESS	1.0.0

Service	Version
ZooKeeper	3.6.2
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
knox	1.1.0

Service	Version
Ranger	2.1.0

7.3. Release notes (EMR V4.X series)

7.3.1. Release notes for EMR V4.10.X

This topic describes the release notes for E-MapReduce (EMR) V4.10.X, including the release date and updates.

Release date

March 23, 2022 for EMR V4.10.0

Updates

Service	Description
SmartData	The service is no longer used.
Bigboot	
JindoSDK	• The architecture of SmartData is upgraded to JindoData. • EMR is integrated with JindoSDK for JindoData 4.0.0 for the first time. JindoData connects to Alibaba Cloud Object Storage Service (OSS) and the Alibaba Cloud OSS-HDFS service.

Service	Description
Spark	- Spark is updated to 2.4.8. - The issue that adaptive execution does not take effect in some scenarios is fixed. - The issue that statistical aggregate functions are used in different manners in Spark and Hive is fixed. - The issue that Spark cannot read valid data of the `CHAR` type from a Hive ORC table is fixed. - The default configurations of Thrift Server are optimized. - In the EMR console, the parameter names on the **spark-defaults** tab of the Configure tab for the Spark service are optimized. - Hive on Spark is optimized. - The array-index out of bounds error that is returned when some required statistics for Adaptive Query Execution (AQE) are missing is fixed. - Errors related to AQE and data caching in specific scenarios are fixed. - Log4j Metrics Appender is removed because the configuration is invalid. - The null pointer exception that occurs when SparkContext is started is fixed. - The data compression algorithm Zstandard is supported.
Hive	- The issue about HiveServer2 memory leaks caused by user-defined functions (UDFs) is fixed. - The issue that the output of the Show Create Table command based on Data Lake Formation (DLF) metadata is inaccurate is fixed. - The default parameters of Hive are optimized to improve the performance of Hive jobs. - In the EMR console, the parameter names on the **hive-env** tab of the Configure tab for the Hive service are changed to uppercase. This facilitates the use of the parameters. - The error message that is reported because of the incompatibility between the file system and Hive metastore when you write data to a Hive table is optimized. - In JindoFS in block storage mode, the metadata of multiple Hive tables can be optimized at the same time. By default, this feature is disabled.
Ranger	- The warning error contained in logs about starting Spark in Ranger is fixed. - The issue that user information fails to be automatically synchronized after Ranger is connected to a Lightweight Directory Access Protocol (LDAP) server is fixed.

Service	Description
HDFS	- The data compression algorithm Zstandard is supported. - By default, the reserved space of NameNode adaptively increases. This way, NameNode enters the Safe mode at the earliest opportunity when the disk space is insufficient.
YARN	- Information about app IDs, CPU utilization, and memory usage is added to the RESTful APIs of containers for nodes. - The issue that the Application Master (AM) logs of an automatically released node cannot be viewed is fixed. - The issue that a cluster cannot be accessed due to historical state store data is fixed. - The data of a node that is automatically released based on the decommissioning logic of auto scaling can be deleted. - The Graceful Decommission logic of auto scaling is optimized. The node on which NodeManager runs is marked deprecated only after the NodeManager process is complete.
Knox	- Knox is adapted to Kudu. - Knox is adapted to HBase. - The issue that the first access to the Spark UI fails is fixed.
Tez	The default parameters of Tez are optimized to improve the performance of Tez jobs.
Sqoop	The issue that precision loss for the DECIMAL data type occurs when you use Sqoop to import data to HCatalog tables is fixed.

Service	Description
Delta Lake	- Metadata management - The built-in catalog of Spark, instead of an API operation that is called by using the Hive CLI, is used to synchronize metadata and partition information. - The statistics on table data are automatically reported to metastores. - SQL - The syntax of the time travel feature is supported. - The DROP PARTITION SQL syntax is supported. - The ADD COLUMN statement can be used to add columns to specified locations (FIRST and AFTER). - Enhanced table management capabilities - The file size can be dynamically adjusted based on the table size. By default, this feature is enabled. - The auto-vacuum feature is supported and enabled by default. Concurrent vacuum operations are supported. - The logic of automatic compaction is optimized. By default, the automatic compaction feature is disabled. - The Z-ordering syntax is added. Z-ordering-based data processing is accelerated.
Hudi	- Hudi is updated to 0.10.0. - The issue about the compatibility of sql.extension between Delta Lake and Hudi is fixed.
Iceberg	The service is added. The supported version is 0.13.0.
Hue	- The issue that garbled characters are displayed when Hue is used to query historical records is fixed. - The UI display exception that occurs when you use Hue together with Oozie is fixed. - The issue that YARN Job Browser sometimes cannot present or terminate jobs is fixed. - YARN Job Browser is accessible by default. - The Presto protocol is supported by default.
DLF-Auth	The service is added. The supported version is 1.0.4.

Service	Description
HBase	- The time required to restart HBase in a high-security cluster is reduced. - The issue that Spark 3.1.1 cannot be integrated with HBase is fixed. - The Graceful Stop process is optimized.
ZooKeeper	ZooKeeper is updated to 3.6.3.
Presto	- Presto is updated to 358. - User-defined functions (UDFs) can be dynamically loaded. For more information, see Dynamically load UDFs. - Data lake analysis is supported.
Impala	- The issue that the LIST operation is repeatedly performed on directly deleted OSS partition directories is fixed. - The following issue is fixed: The `no such method error` message appears when you query data in DLF metadata tables.
Zeppelin	Zeppelin is updated to 0.10.0.
Oozie	The issue that Jetty Server of Oozie fails to start due to JAR package conflicts in high availability (HA) scenarios is fixed.

Release version information

○ Hadoop cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	2.4.8
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
DLF-Auth	1.0.4
Iceberg	0.13.0

Service	Version
Hudi	0.10.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.0.0
HBase	2.3.4
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.2
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
RSS	1.0.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

7.3.2. Release notes for EMR V4.9.X

This topic describes the release notes for E-MapReduce (EMR) V4.9.X, including the release date and updates.

Release date

April 21, 2021 for EMR V4.9.0

Updates

Service	Description

Service	Description
SmartData	SmartData is updated to 3.5.0. For more information, see SmartData 3.5.X.
Spark	- The issue that adaptive execution does not take effect in some scenarios is fixed. - The issue that statistical aggregate functions are used in different manners in Spark and Hive is fixed. - The issue that Spark cannot read valid data of the CHAR type from a Hive ORC table is fixed.
HDFS	The SM4 encryption algorithm is supported.
Hue	Hue is updated to 4.9.0.
Alluxio	Alluxio is updated to 2.5.0.
Livy	Livy is updated to 0.7.1.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	2.4.7
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.5.0
Bigboot	3.5.0
Hudi	0.6.0
OpenLDAP	2.4.44
Hue	4.9.0

Service	Version
HBase	2.3.4
ZooKeeper	3.6.2
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	2.1.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2
Alluxio	2.5.0
ESS	1.0.0
Kudu	1.11.1
Oozie	5.1.0

Service	Version
ESS	1.0.0

ZooKeeper	3.6.2
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

7.3.3. Release notes for EMR V4.8.X

This topic describes the release notes for E-MapReduce (EMR) V4.8.X, including the release date and updates.

Release date

March 15, 2021 for EMR V4.8.0

Updates

Service	Description
SmartData	SmartData is updated to 3.4.0. For more information, see SmartData 3.4.X.
Spark	- Some default configurations are optimized. - Performance is optimized. Window-based top-k queries can be pushed down. - The capability of reading data from and writing data to Hive tables in the CSV or JSON format is enhanced. - All the column names of a table can be omitted in the ANALYZE statement. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication. - Spark Beeline is easier to use.
Hive	- Some default configurations are optimized. - Performance is optimized. The cost-based optimization (CBO) feature is enhanced. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
YARN	The risk caused by unauthorized access from a Hadoop cluster to the YARN web UI is fixed. If you access the YARN web UI by using SSH Tunnel, you no longer need to explicitly specify user.name in the URL.
Tez	Some default configurations are optimized.

Service	Description
Ranger	- The issue caused by filter pushdown in Spark is fixed. - The issue that prevents Presto from being enabled after you disable Presto in Ranger is fixed. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
Hue	LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
Impala	- Impala is updated to 3.4.0. - Shiro is updated to 1.7.0. - Metadata stored in Alibaba Cloud Data Lake Formation (DLF) is supported. - Data in the Delta format can be queried. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication. - The exception that occurs when you use the INSERT OVERWRITE statement to overwrite data stored in OSS is fixed.
Hudi	- SQL statements can be executed to query data in Hudi tables. - The issue that causes the query results on some data to be inaccurate is fixed. - Partition pruning is supported if you query data in a Copy on Write table of Hudi by using Spark. - The bucket-based index mechanism is supported to improve write performance.

Service	Description
Delta Lake	- The issue that prevents metadata from being synchronized to a Hive metastore based on existing Delta tables is fixed. - The issue that prevents the MERGE statement from parsing asterisks (`*`) in data is fixed. - The issue that causes an error to be reported when you convert data in the Parquet format into a Delta table and create table metadata is fixed. - The issue that causes the OPTIMIZE command to fail if no files need to be compacted is fixed. - A subquery can be used as the source in the MERGE statement. - Data can be cached if you use Presto to query data in a Delta table. This improves query efficiency. - Impala can be used to query data in Delta tables.
ESS	- Exceptions in the shuffle read stage, such as ClosedChannelException, IndexOutOfBoundsException, and excessive off-heap memory usage, are fixed. - The issue that causes NullPointerException (NPE) to be reported after metric monitoring is enabled is fixed.
HAS	The issue that prevents the admin.keytab file from being initiated again after a HAS installation error is reported is fixed.
Presto	LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
HBase	- HBase is updated to 2.2.6. - Access control based on Ranger is no longer supported.
Sqoop	Files in the Parquet format can be imported to OSS.
Superset	- The issue that prevents the admin user from logging on to the web UI is fixed. - Datasets are compatible with Druid clusters. - Spark SQL datasets are no longer supported.

Service	Description
Knox	- Access to Presto by using Knox is supported. - The issue that causes the Druid web UI to be inaccessible is fixed. - The limit that you can access the Ranger web UI based on HTTP only by using Knox in high security mode is removed.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	2.4.7
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.4.0
Bigboot	3.4.0
Hudi	0.6.0
OpenLDAP	2.4.44
Hue	4.4.0
HBase	2.3.4
ZooKeeper	3.5.6
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0

Service	Version
Livy	0.6.0
Superset	0.36.0
Ranger	2.1.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2
Alluxio	2.4.1
ESS	1.0.0
Kudu	1.11.1
Oozie	5.1.0

Service	Version
ESS	1.0.0

ZooKeeper	3.5.6
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

7.3.4. Release notes for EMR V4.6.X

This topic describes the release notes for E-MapReduce (EMR) V4.6.X, including the release date and updates.

Release date

January 15, 2021 for EMR V4.6.0

Updates

Service	Description
SmartData	SmartData is updated to 3.2.0. For more information, see SmartData 3.2.X.
Spark	- Spark is updated to 2.4.7. - jQuery is updated to 3.5.1. - Spark is compatible with Hive to automatically update table and partition sizes. - Spark metadata and job running information can be sent to DataWorks.
Hive	- Metadata from Alibaba Cloud Data Lake Formation (DLF) in an HCatalog table is supported. - Hive metadata and job running information can be sent to DataWorks.
Metastore	- The statistics feature for Hive is added. - Metadata from Alibaba Cloud DLF in an HCatalog table is supported. - Methods to obtain an STS token are optimized.
HDFS	- jQuery is updated to 3.5.1. - HDFS is updated to 3.2.1.
YARN	- YARN is updated to 3.2.1. - jQuery is updated to 3.5.1. - Configurations of Fair Scheduler are adjusted. - Timeline Server is optimized.
Zeppelin	Zeppelin is updated to 0.9.0.
OpenLDAP	- The audit feature is added. - By default, the SSL port 10636 is enabled. - Presto can be enabled with one click.
Hue	Presto is supported.
EMRHook	- The EMRHook service is added. - hive-hook is used to send Hive metadata and job running information to DataWorks. - spark-hook is used to send Spark metadata and job running information to DataWorks.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.2.1
YARN	3.2.1
Hive	3.1.2
Spark	2.4.7
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.2.0
Bigboot	3.2.0
Hudi	0.6.0
OpenLDAP	2.4.44
Hue	4.4.0
HBase	2.1.9
ZooKeeper	3.5.6
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.6.0
Superset	0.36.0
Ranger	2.1.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2

Service	Version
ESS	1.0.0
Kudu	1.11.1
Oozie	5.1.0

Service	Version
ESS	1.0.0

ZooKeeper	3.5.6
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

7.3.5. Release notes for EMR V4.5.X

This topic describes the release notes for E-MapReduce (EMR) V4.5.X, including the release date and updates.

EMR V4.5.1

- Release date

  December 13, 2020

- Updates
  - The issue that occurs when you query partitioned tables by using Hive or Presto is fixed.
  - EMR V4.5.1 is available only in the China (Hangzhou), China (Shanghai), and China (Beijing) regions.

EMR V4.5.0

- Release date

  December 7, 2020

- New features

Service	Description

Service	Description
EMR Remote Shuffle Service (ESS)	ESS 1.0.0 is supported. For more information, see ESS.
Hudi	Hudi 0.6.0 is supported.
Delta Lake	Delta Lake 0.6.1 is supported.

- Updates

Service	Description
Ranger	○ Ranger is updated to 2.1.0. ○ Ownership-related permissions are supported.
Presto	○ Presto is updated to 338. ○ Metadata stored in Alibaba Cloud Data Lake Formation (DLF) is supported.
Zeppelin	Zeppelin is updated to 0.8.2.
SmartData	SmartData is updated to 3.1.0. For more information, see SmartData 3.1.x.
Bigboot	Bigboot is updated to 3.1.0.
Hive	○ Metadata stored in Alibaba Cloud DLF is supported. ○ Ownership-related permissions of Ranger are supported.
Spark	Metadata stored in Alibaba Cloud DLF is supported.
DLF Metastore	○ The issue that Presto cannot be started in a high-security cluster is fixed. ○ Hive 3 and metadata cache are supported. ○ The issue that occurs when you query data by using Hive or Presto is fixed.
Impala	Custom configuration of parameters on the *catalogd.flgs*, *impalad.flgs*, and *statestored.flgs* tabs is supported in the EMR console.
Tez	Vulnerabilities related to autoDeploy on the web UI of Tez are fixed.

Service	Description
OpenLDAP	A rule that determines whether port 10389 is in the waiting state is added.
Hue	Security vulnerabilities of the MySQL backend are fixed.
Kerberos	○ Apache Kerby is updated to 2.0.1. ○ The issue that the kadmin principal of an external Kerberos cluster cannot be customized is fixed.
Sqoop	○ File formats such as Parquet, Avro, and ORC are supported. ○ Metadata stored in Alibaba Cloud DLF is supported.

Release version information

○ Hadoop cluster ○ Shuffle Service cluster ○ Kafka cluster

Service	Version
HDFS	3.1.3
YARN	3.1.3
Hive	3.1.2
Spark	2.4.5
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.1.0
Bigboot	3.1.0
Hudi	0.6.0
OpenLDAP	2.4.44
Hue	4.4.0
HBase	2.1.9

Service	Version
ZooKeeper	3.5.6
Presto	338
Impala	3.4.0
Zeppelin	0.8.2
Pig	0.14.0
Flume	1.9.0
Livy	0.6.0
Superset	0.36.0
Ranger	2.1.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2
ESS	1.0.0
Kudu	1.11.1
Oozie	5.1.0

Service	Version
ESS	1.0.0

ZooKeeper	3.5.6
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.1.0

7.3.6. Release notes for EMR V4.4.X

This topic describes the release notes for E-MapReduce (EMR) V4.4.X, including the release date and updates.

Release date

September 15, 2020 for EMR V4.4.1

Updates

Service	Description
YARN	- The *hadoop/tools/lib* directory is deleted from the value of the *yarn.application.classpath* parameter. - Default parameter settings for MapReduce jobs are optimized.
Hive	Default parameter settings are optimized.
Tez	
Ranger	- Impala-based access control is supported. - The jackson-databind version is updated.
Impala	- Integration with Ranger is supported. - Shiro is updated to 1.6.0.
SmartData	SmartData and Bigboot are updated to 2.7.301.
Bigboot	
Knox	- The web UI of Tez can be independently viewed. Knox is compatible with Tez on the web UI of YARN. - Shiro is updated to 1.6.0.
EMRDOCTOR	The following issue is fixed: Job information is not collected if a time-based configuration file is empty.
Ganglia	The detection feature of the service RPC port for Hadoop Distributed File System (HDFS) is enabled.
Oozie	- The issue that the web UI cannot be accessed is fixed. - The jackson-databind version is updated.
ZooKeeper	Internal IP addresses can be bound to Elastic Compute Service (ECS) instances to start service ports.
Superset	The startup script is repaired.

Service	Description
Livy	The versions of jackson-databind and Fastjson are updated.
Zepplin	The versions of jackson-databind and Shiro are updated.
HAS	The versions of jackson-databind and Fastjson are updated.
Flume	The Fastjson version is updated.

Release version information

○ Hadoop cluster ○ Kafka cluster

Service	Version
HDFS	3.1.3
YARN	3.1.3
Hive	3.1.2
Spark	2.4.5
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	2.7.301
Bigboot	2.7.301
OpenLDAP	2.4.44
Hue	4.4.0
HBase	2.1.9
ZooKeeper	3.5.6
Presto	331
Impala	3.4.0
Zeppelin	0.8.1
Pig	0.14.0

Service	Version
Flume	1.9.0
Livy	0.6.0
Superset	0.36.0
Ranger	2.0.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2
Kudu	1.11.1
Oozie	5.1.0

ZooKeeper	3.5.6
Ganglia	3.7.2
Kafka	2.4.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	2.0.0

7.3.7. Release notes for EMR V4.3.X

This topic describes the release notes for E-MapReduce (EMR) V4.3.X, including the release date and updates.

Release date

May 20, 2020 for EMR V4.3.0

Updates

Service	Description

Service	Description
Ranger	- Custom deployment of Hadoop Distributed File System (HDFS), Hive, and Spark plug-ins is supported. You can enable plug-ins on required service nodes. - The RangerUserSync and RangerAdmin components can be configured in the EMR console.
Presto	The Kudu client is updated.
Spark	- Spark is updated to 2.4.5. - The associated Delta Lake is updated to 0.6.0. - The issue that PySpark cannot properly run after Ranger Hive is enabled is fixed.
HDFS	- The issue that the HDFS_NAMENODE_OPTS parameter does not take effect is fixed. - Custom deployment is supported.
YARN	Custom deployment is supported.
Hive	Custom deployment is supported.
Knox	Information on the web UI of HDFS NameNode in Hadoop 3.X can be viewed.
Zeppelin	The issue that a *zepping.keytab* file fails to be generated is fixed.
Kafka	Kafka is updated to 2.4.1.
Kudu	Kudu is updated to 1.11.1.
Impala	Issues related to HAProxy are fixed.
Livy	Issues related to xmllint are fixed.
Hue	- Hue can be deployed on gateway clusters. - Multiple Hue instances can be enabled on a single node.

Release version information

○ Hadoop cluster ○ Kafka cluster

Service	Version
HDFS	3.1.3

Service	Version
YARN	3.1.3
Hive	3.1.1
Spark	2.4.5
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	2.6.7
Bigboot	2.6.7
OpenLDAP	2.4.44
Hue	4.4.0
HBase	2.1.9
ZooKeeper	3.5.6
Presto	331
Impala	3.2.0
Zeppelin	0.8.1
Pig	0.14.0
Flume	1.9.0
Livy	0.6.0
Superset	0.28.1
Ranger	2.0.0
Flink	1.10-vvr-1.0.2
Storm	1.2.2
Kudu	1.11.1
Oozie	5.1.0
	ZooKeeper	3.5.6

Kafka	Ganglia	3.7.2
	Kafka	2.4.1
	Kafka-Manager	1.3.3.16
	OpenLDAP	2.4.44
	Knox	1.1.0
	Ranger	2.0.0

7.4. Release notes (EMR V3.X series)

7.4.1. Release notes for EMR V3.40.X

This topic describes the release notes for E-MapReduce (EMR) V3.40.X, including the release date, updates, and release version information.

Release date

April 21, 2022 for EMR V3.40.0

Updates

Service	Description
JindoData	The JindoData service is added. The service version is 4.3.0.
JindoSDK	JindoSDK is updated to 4.3.0.
Spark	Spark is updated to 3.2.1.
Hive	• The following issue is fixed: After speculative execution is enabled for Hive on Tez, both the original task and the speculative task are committed. • The following issue is fixed: User-defined functions (UDFs) can be called only after you reload the functions.
Presto	The following issue is fixed: After a Hadoop cluster is initialized and the Presto service is added to the cluster, the Presto service fails to start.
DeltaLake	The incompatibility issue between Delta Lake and Streaming SQL is fixed.
Hudi	Hudi is updated to 0.10.1.
Iceberg	Iceberg is updated to 0.13.1.

Service	Description
YARN	- A feature that allows you to run an ApplicationMaster only on a core node is added. - The issue that the value of the mareduce.map.java.opts parameter does not include taihaodoctor is fixed.
ZooKeeper	The configuration of Java Virtual Machine (JVM) parameters is optimized.
Flink	Flink, Impala, Flume, and Druid are adapted to JindoSDK 4.3.0.
Impala	
Flume	
Druid	
Sqoop	The PostgreSQL version is updated.
Zeppelin	The issue that the Java Database Connectivity (JDBC) interpreter fails to start is fixed.
Ranger	Hudi tables are supported by the Spark plugin of Ranger 1.2.0.
Oozie	Log4j is updated to 2.17.2.
HBase	The issue that the RegionServer of HBase 1.4.9 fails to start is fixed.
DLF-Auth	DLF-Auth is updated to 2.0.0.

Release version information

◎ Hadoop cluster ◎ Druid cluster ◎ Presto cluster

Service	Version
HDFS	2.8.5
YARN	2.8.5
Hive	2.3.9
Spark	2.4.8
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7

Service	Version
JindoData	4.3.0
Iceberg	0.13.1
DLF-Auth	2.0.0
Hudi	0.10.1
DeltaLake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.3.0
Spark	3.2.1
HBase	1.4.9
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.1
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	1.2.0
Phoenix	4.14.1
RSS	1.0.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
HDFS	2.8.5
Druid	0.20.1

Service	Version
ZooKeeper	3.6.3
Knox	1.1.0
Ganglia	3.7.2
JindoData	4.3.0
OpenLDAP	2.4.44
JindoSDK	4.3.0
YARN	2.8.5
Superset	0.36.0

Component	Version
Knox	1.1.0
Presto	358
Ganglia	3.7.2
JindoData	4.3.0
Iceberg	0.13.1
Hudi	0.10.1
DeltaLake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.3.0
Alluxio	2.5.0

7.4.2. Release notes for EMR V3.39.X

This topic describes the release notes for E-MapReduce (EMR) V3.39.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V3.39.2	March 25, 2022

Version	Date
EMR V3.39.1	February 15, 2022

Updates

EMR V3.39.2

> ⑦ **Note** Only OLAP and Dataflow clusters support this version. You can create an OLAP or Dataflow cluster in the new EMR console.

Service	Description
Flink	• The application performance management (APM) dashboard is optimized. Metrics such as sourceIdleTime are added. • Alerting by CloudMonitor is supported.
Kafka	• SSL and Simple Authentication and Security Layer (SASL) configurations are supported. • The default values of some parameters are changed.
ClickHouse	The default values of some parameters are changed.

EMR V3.39.1

Service	Description
SmartData	The service is no longer used.
Bigboot	
RSS	• EMR Remote Shuffle Service (ESS) is upgraded to Remote Shuffle Service (RSS). For more information, see RSS. • The service functionality and stability are enhanced.
JindoSDK	• The architecture of SmartData is upgraded to JindoData. • EMR is integrated with JindoSDK for JindoData 4.0.0 for the first time. JindoData connects to Alibaba Cloud Object Storage Service (OSS) and the Alibaba Cloud OSS-HDFS service.
Spark	• Hive on Spark is optimized. • Spark is adapted to JindoSDK.
Tez	Tez is adapted to JindoSDK.
Hive	Hive is adapted to JindoSDK.

Service	Description
Presto	- User-defined functions (UDFs) can be dynamically loaded. - The for ... as of syntax can be used in the time travel feature to query data in a Delta Lake table. - An independent Delta Lake catalog is added. Presto provides default configurations for a Delta connector and supports the Optimize, Z-ordering, and Data Skipping features based on the independent Delta Lake catalog. - The issue that data in Merge on Read tables of Hudi cannot be queried by using a Hudi connector is fixed. You cannot use a Hive connector to query Merge on Read tables of Hudi. - Presto is adapted to JindoSDK.
Delta Lake	- Metadata management - The built-in catalog of Spark, instead of an API operation that is called by using the Hive CLI, is used to synchronize metadata and partition information. - The statistics on table data are automatically reported to metastores. - SQL - The syntax of the time travel feature is supported. - The DROP PARTITION SQL syntax is supported. - The ADD COLUMN statement can be used to add columns to specified locations (FIRST and AFTER). - Enhanced table management capabilities - The file size can be dynamically adjusted based on the table size. By default, this feature is enabled. - The auto-vacuum feature is supported and enabled by default. Concurrent vacuum operations are supported. - The logic of automatic compaction is optimized. By default, the automatic compaction feature is disabled. - The Z-ordering syntax is added. Z-ordering-based data processing is accelerated.
Hudi	Hudi is updated to 0.10.0.
HDFS	HDFS is adapted to JindoSDK.
YARN	YARN is adapted to JindoSDK.
Flume	Flume is adapted to JindoSDK.

Service	Description
Flink	• By default, the *lib* directory of Flink is uploaded to your HDFS cluster. This way, you can configure the yarn.provided.lib.dirs parameter to use the directory. • Flink is adapted to JindoSDK.
Impala	Impala is adapted to JindoSDK.
Ranger	• The issue that Spark History Server fails to start is fixed. • Ranger is adapted to JindoSDK.
HBase	• The issue about default parameter settings is fixed. • The issue about the date format of garbage collection (GC) logs is fixed. • The restart issue that occurs if an IP address is configured for RegionServer is fixed.
Druid	Druid is adapted to JindoSDK.
ClickHouse	• The transaction feature is added. For more information, see Use a transaction. • The logic of processing data when the ClickHouse component stops working is optimized.
Iceberg	• Iceberg is updated to 0.13.0. • Default configuration items are hidden to improve user experience.
DLF-Auth	The issue that Spark History Server fails to start is fixed.
StarRocks	The service is added in the new EMR console. The 2.0.1 version of this service is released.

Release version information

> ⑦ **Note** To view the information about an OLAP cluster, you must log on to the new EMR console.

◉ Hadoop cluster ◉ Druid cluster ◉ Dataflow cluster ◉ ClickHouse cluster ◉ Presto cluster ◉ OLAP cluster

Service	Version
HDFS	2.8.5
YARN	2.8.5

Service	Version
Hive	2.3.9
Spark	2.4.8
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
Iceberg	0.13.0
DLF-Auth	1.0.4
Hudi	0.10.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.0.0
Spark	3.2.0
HBase	1.4.9
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.2
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	1.2.0
Phoenix	4.14.1
RSS	1.0.0
Alluxio	2.5.0

Service	Version
Kudu	1.14.0
Oozie	5.2.1

Service	Version
HDFS	2.8.5
Druid	0.20.1
ZooKeeper	3.6.3
Knox	1.1.0
Ganglia	3.7.2
OpenLDAP	2.4.44
JindoSDK	4.0.0
YARN	2.8.5
Superset	0.36.0

Component	Version
HDFS	2.8.5
YARN	2.8.5
ZooKeeper	3.6.3
Knox	1.1.0
Flink	1.13-vvr-4.0.11
OpenLDAP	2.4.44
JindoSDK	4.0.0
Ganglia	3.7.2
Kafka	1.1.1
Kafka-Manager	1.3.3.16

Service	Version
ZooKeeper	3.6.3

Service	Version
Ganglia	3.7.2
ClickHouse	20.8.12.2

Component	Version
Knox	1.1.0
Presto	358
Ganglia	3.7.2
Iceberg	0.13.0
Hudi	0.10.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
JindoSDK	4.0.0
Alluxio	2.5.0

Service	Version
ClickHouse	20.8.12.2.2.17
StarRocks	2.0.1
ZooKeeper	3.6.3

7.4.3. Release notes for EMR V3.38.X

This topic describes the release notes for E-MapReduce (EMR) V3.38.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V3.38.3	December 2021
EMR V3.38.2	December 2021
EMR V3.38.1	November 2021

Version	Date
EMR V3.38.0	October 2021

Updates

- EMR V3.38.3

  The Log4j security vulnerability of all related components is fixed. For more information, see
  Vulnerability announcement | RCE vulnerability in Apache Log4j 2.

Service	Description
Presto	- The error that is reported when Presto in a high availability cluster is used to query data in a Hudi table is fixed. - The Log4j security vulnerability of the Elasticsearch connector is fixed.
DLF Metastore	- By default, the logging feature for metastores is disabled. In earlier versions, the feature is enabled by default. - The error caused by an excessively long getTableStats URI for a metastore is fixed.
Delta Lake	The issue that schema changes fail to be synchronized to a metastore is fixed.
Flink	- VVR is updated to 4.0.11. This version supports the following features: - The commercial features of Flink Change Data Capture (CDC) are released: - Schema evolution is supported. - Flink SQL semantics can be used to synchronize an entire database at a time. - GeminiStateBackend can be used to store state data in Object Storage Service (OSS). - A Hudi connector of Enterprise Edition is provided, and a built-in Data Lake Formation (DLF) catalog is used in the connector to manage metadata.
Sqoop	The following issue is fixed: Precision loss for the DECIMAL data type occurs when you use Sqoop to import data to HCatalog tables.

- EMR V3.38.2

Service	Description

Service	Description
SmartData	○ SmartData is updated to 3.8.0. For more information, see SmartData 3.8.X overview. ○ Authentication and authorization based on Kerberos and Ranger can be used to manage permissions on data in OSS.

- EMR V3.38.1

Service	Description
SmartData	SmartData is updated to 3.7.3. For more information, see SmartData 3.7.X overview.
Spark	○ Log4j Metrics Appender is removed because the configuration is invalid. ○ The null pointer exception that occurs when SparkContext is started is fixed.
Presto	○ The following issue is fixed: You must configure host parameters for the Presto service before you can use Presto in a high-availability Hadoop cluster to query data from a Hive table. ○ The following issue is fixed: Presto cannot be started by default when the memory is small. ○ The issue that modifications to *worker-jvm* do not take effect is fixed. ○ Ranger is supported.
Impala	The following issue is fixed: The `no such method error` message appears when you query data in DLF metadata tables.
Ranger	○ Presto is supported. ○ The exceptions that occur when you use Ranger to configure the permissions of using Spark to insert data into ORC or Parquet tables are fixed. ○ The following issue is fixed: Hive role permissions configured in Ranger do not take effect after Kerberos is enabled.
DLF-Auth	○ DLF-Auth is updated to 1.0.1. ○ The permissions of using Presto to access DLF can be configured. ○ The issue that data cannot be cached for RAM users is fixed.

- EMR V3.38.0

Service	Description
SmartData	SmartData is updated to 3.7.2. For more information, see SmartData 3.7.X overview.
Spark	○ Spark is updated to 2.4.8. ○ Both of Spark 2.4.8 and Spark 3.1.2 are supported. ⑦ **Note** Delta and Remote Shuffle Service are not supported in Spark 3. ○ In Spark 3.x, the Distinct computing performance is optimized for Spark SQL. The optimization feature is triggered if an aggregation operator contains multiple count(distinct case ... when ...) methods. ○ The array-index out of bounds error that is returned when some required statistics for Adaptive Query Execution (AQE) are missing is fixed. ○ Errors related to AQE and data caching in specific scenarios are fixed.
Hive	Hive is updated to 2.3.9.
Presto	○ An independent Presto cluster can be created. ○ Presto is updated to 358. ◁ **Notice** This version does not support Ranger. ○ Connectors such as Hudi and MySQL connectors are supported by default, and the default configurations are updated. ○ Auto scaling is supported by Presto clusters. ⑦ **Note** To use the auto scaling feature, . ○ Data lake analysis is supported.
Delta Lake	○ The same Delta Lake connectors are used in Hive 2 and Hive 3. ○ The error that is returned when you use Delta Lake connectors to query data from multi-level partitioned tables is fixed.
Hudi	○ Hudi is updated to 0.9.0. ○ The issue about the compatibility of sql.extension between Delta Lake and Hudi is fixed.

Service	Description
HDFS	By default, the reserved space of NameNode adaptively increases. This way, NameNode enters the Safe mode in a timely manner when the disk space is insufficient.
Flink	○ Flink is updated to 1.13-vvr-4.0.10, which corresponds to Apache Flink 1.13.1. ○ Commercial connectors, such as a Hologres connector, are added to Flink. ○ Metric reporters are added to report metrics on the APM dashboard. ○ A SchemaRegistry-based Kafka catalog is added to the Kafka connector. This way, you can read data from or write data to existing Kafka topics without the need to execute DDL statements.
Storm	The service is no longer used.
Zeppelin	Zeppelin is updated to 0.10.0.
Ranger	If Presto 358 is used, Ranger cannot be used to configure permissions related to Presto.
Hue	○ The issue that YARN Job Browser sometimes cannot present or terminate jobs is fixed. ○ YARN Job Browser is accessible by default. ○ The Presto protocol is supported by default.
Druid	The following issue is fixed: After a server is unexpectedly shut down, the related node fails to restart because a PID file is not deleted.
ClickHouse	○ Some default configurations are updated. ○ Clusters can be scaled out. ○ The MetaChecker feature is supported. ○ OSS table engines and OSS table functions can be used to read data. ○ Table-level custom ZooKeeper addresses are supported.
Iceberg	The service is added. The supported versions range from 0.12.0 to 1.0.1.
Knox	The issue that the first access to the Spark UI fails is fixed.
DLF-Auth	The service is added. The permissions of using Hive or Spark to access DLF can be configured. The supported version is 1.0.0.

Service	Description
ESS	ESS is updated to 1.2.0.

Release version information

○ Hadoop cluster ○ Druid cluster ○ Dataflow cluster ○ ClickHouse cluster ○ Presto cluster

Service	Version
HDFS	2.8.5
YARN	2.8.5
Hive	2.3.9
Spark	2.4.8
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	EMR V3.38.0: 3.7.2
Bigboot	EMR V3.38.1: 3.7.3 EMR V3.38.2: 3.8.0
Iceberg	0.12.0
DLF-Auth	1.0.0
Hudi	0.9.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
Spark	3.1.2
HBase	1.4.9
ZooKeeper	3.6.3
Presto	358
Impala	3.4.0
Zeppelin	0.10.0

Service	Version
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	1.2.0
Phoenix	4.14.1
ESS	1.2.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
HDFS	2.8.5
Druid	0.20.1
ZooKeeper	3.6.3
Knox	1.1.0
Ganglia	3.7.2
SmartData	3.7.2
Bigboot	
OpenLDAP	2.4.44
YARN	2.8.5
Superset	0.36.0

Cluster mode	Component	Version
	HDFS	2.8.5
	YARN	2.8.5
	ZooKeeper	3.6.3
	Knox	1.1.0
Flink		

	Flink	EMR V3.38.0: 1.13-vvr-4.0.10 EMR V3.38.3: 1.13-vvr-4.0.11
	SmartData	EMR V3.38.0: 3.7.2 EMR V3.38.3: 3.8.0
	Bigboot	
	OpenLDAP	2.4.44
Kafka	ZooKeeper	3.6.3
	Ganglia	3.7.2
	Kafka	1.1.1
	Kafka-Manager	1.3.3.16
	OpenLDAP	2.4.44
	Knox	1.1.0
	Ranger	1.2.0

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
ClickHouse	20.8.12.2

Component	Version
Knox	1.1.0
Presto	358
Ganglia	3.7.2
SmartData	EMR V3.38.0: 3.7.2 EMR V3.38.1: 3.7.3 EMR V3.38.2: 3.8.0
Bigboot	
Hudi	0.9.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
Alluxio	2.5.0

7.4.4. Release notes for EMR V3.37.X

This topic describes the release notes for E-MapReduce (EMR) V3.37.X, including the release dates, updates, and release version information.

Release dates

Version	Date
EMR V3.37.1	September 2021
EMR V3.37.0	August 2021

Updates

- EMR V3.37.1

Service	Description
SmartData	SmartData is updated to 3.7.1.
Hue	The issue that Impala cannot be used in a high-security cluster is fixed.
Kudu	Kerberos is supported.

- EMR V3.37.0

Service	Description
SmartData	SmartData is updated to 3.7.0.
Spark	The incompatibility issue between Spark and Delta Lake is fixed.
Delta Lake	- The connectors for Delta Lake are updated. This way, you can use a storage handler to create tables and query data. - The exception that occurs when you perform an INSERT OVERWRITE operation on a partitioned table is fixed. - The exception that occurs when you perform an OPTIMIZE operation to write virtual fields to a file in the G-SCD scenario is fixed.

Service	Description
YARN	- Information about app IDs, CPU utilization, and memory usage is added to the REST APIs of containers for nodes. - The issue that the Application Master (AM) logs of an automatically released node cannot be viewed is fixed. - The data of a node that is automatically released based on the decommissioning logic of auto scaling can be deleted. - The Graceful Decommission logic of auto scaling is optimized. The node on which NodeManager runs is marked deprecated only after the NodeManager process is complete.
ZooKeeper	ZooKeeper is updated to 3.6.3.
Flink	- The SmartData service is added. - Object Storage Service (OSS) can be accessed in password-free mode when you log on to a DataFlow-Flink cluster in SSH mode and submit a job in the cluster.
Impala	The issue that the LIST operation is repeatedly performed on directly deleted OSS directories is fixed.
Hue	The UI display exception that occurs when you use Hue together with Oozie is fixed.
Kudu	Kudu is updated to 1.14.0.
ClickHouse	Some default configurations are updated.

Release version information

⊙ Hadoop cluster ⊙ Druid cluster ⊙ Flink cluster ⊙ Kafka cluster ⊙ ClickHouse cluster

Service	Version
HDFS	2.8.5
YARN	2.8.5
Hive	2.3.8
Spark	2.4.7
Knox	1.1.0
Tez	0.9.2

Service	Version
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	EMR V3.37.0: 3.7.0
Bigboot	EMR V3.37.1: 3.7.1
Hudi	0.8.0
Delta Lake	0.6.1
OpenLDAP	2.4.44
Hue	4.9.0
HBase	1.4.9
ZooKeeper	3.6.3
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	1.2.0
Storm	1.2.2
Phoenix	4.14.1
ESS	1.0.0
Alluxio	2.5.0
Kudu	1.14.0
Oozie	5.2.1

Service	Version
HDFS	2.8.5
Druid	0.20.1

Service	Version
ZooKeeper	3.6.3
Knox	1.1.0
Ganglia	3.7.2
SmartData	EMR V3.37.0: 3.7.0 EMR V3.37.1: 3.7.1
Bigboot	
OpenLDAP	2.4.44
YARN	2.8.5
Superset	0.36.0

Service	Version
HDFS	2.8.5
YARN	2.8.5
ZooKeeper	3.6.3
Knox	1.1.0
Flink	1.12-vvr-3.0.2
SmartData	EMR V3.37.0: 3.7.0 EMR V3.37.1: 3.7.1
Bigboot	
OpenLDAP	2.4.44

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
Kafka	1.1.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	1.2.0

Service	Version
ZooKeeper	3.6.3
Ganglia	3.7.2
ClickHouse	20.8.12.2

7.4.5. Release notes for EMR V3.36.X

This topic describes the release notes for E-MapReduce (EMR) V3.36.X, including the release date, updates, and release version information.

Release date

July 16, 2021 for EMR V3.36.1

Updates

Service	Description
SmartData	SmartData is updated to 3.6.1. For more information, see SmartData 3.6.X overview.
Hive	• Hive is updated to 2.3.8. • The issue that the output of the `show create table` command based on Data Lake Formation (DLF) metadata is inaccurate is fixed. • The default parameters of Hive are optimized to improve the performance of Hive jobs. • In the EMR console, the parameter names on the **hive-env** tab of the **Configure** tab for the Hive service are changed to uppercase. This facilitates the use of the parameters. • The error message that is reported because of the incompatibility between the file system and Hive metastore when you write data to a Hive table is optimized.
HDFS	The data compression algorithm Zstandard is supported.
Flink	Flink is updated to 1.12-vvr-3.0.2. ⑦ **Note**　Flink is removed from Hadoop clusters.
Hudi	• Hudi is updated to 0.8.0. • Hudi can be integrated with Spark SQL.

Service	Description
Spark	- In the EMR console, the parameter names on the **spark-defaults** tab of the **Configure** tab for the Spark service are optimized. - The performance of log collection is optimized. - The data compression algorithm Zstandard is supported.
Impala	The core dump error reported when you use HDFS is fixed.
Tez	The default parameters of Tez are optimized to improve the performance of Tez jobs.
Knox	- Knox is adapted to Kudu. - Knox is adapted to Impala. - Knox is adapted to HBase.
Phoenix	The issue that no JDBC driver is found when Hive or Spark SQL is used to access Phoenix tables is fixed.
ClickHouse	The application performance management (APM) feature is available. This feature is used to support monitoring and alerting.

Release version information

⦿ Hadoop cluster ○ Druid cluster ○ Flink cluster ○ Kafka cluster ○ ClickHouse cluster

Service	Version
HDFS	2.8.5
YARN	2.8.5
Hive	2.3.8
Spark	2.4.7
Knox	1.1.0
Tez	0.9.2
Ganglia	3.7.2
Sqoop	1.4.7
SmartData	3.6.1
Bigboot	3.6.1
Hudi	0.8.0

Service	Version
OpenLDAP	2.4.44
Hue	4.9.0
HBase	1.4.9
ZooKeeper	3.6.2
Presto	338
Impala	3.4.0
Zeppelin	0.9.0
Flume	1.9.0
Livy	0.7.1
Superset	0.36.0
Ranger	1.2.0
Storm	1.2.2
Phoenix	4.14.1
ESS	1.0.0
Alluxio	2.5.0
Kudu	1.10.0
Oozie	5.2.1

Service	Version
HDFS	2.8.5
Druid	0.20.1
ZooKeeper	3.6.2
Knox	1.1.0
Ganglia	3.7.2
SmartData	3.6.1
Bigboot	3.6.1
OpenLDAP	2.4.44

Service	Version
YARN	2.8.5
Superset	0.36.0

Service	Version
HDFS	2.8.5
YARN	2.8.5
ZooKeeper	3.6.2
Knox	1.1.0
Flink	1.12-vvr-3.0.2
OpenLDAP	2.4.44

Service	Version
ZooKeeper	3.6.2
Ganglia	3.7.2
Kafka	1.1.1
Kafka-Manager	1.3.3.16
OpenLDAP	2.4.44
Knox	1.1.0
Ranger	1.2.0

Service	Version
ZooKeeper	3.6.2
Ganglia	3.7.2
ClickHouse	20.8.12.2

7.4.6. Release notes for EMR V3.35.X

This topic describes the release notes for E-MapReduce (EMR) V3.35.X, including the release date and updates.

Release date

April 21, 2021 for EMR V3.35.0

Updates

Component	Description
SmartData	SmartData is updated to 3.5.0. For more information, see SmartData 3.5.X.
Spark	• The issue that adaptive execution does not take effect in some scenarios is fixed. • The issue that statistical aggregate functions are used in different manners in Spark and Hive is fixed. • The issue that Spark cannot read valid data of the CHAR type from a Hive ORC table is fixed.
HDFS	The SM4 encryption algorithm is supported.
Hue	Hue is updated to 4.9.0.
Alluxio	Alluxio is updated to 2.5.0.
Druid	• Druid is updated to 0.20.1. • Security is enhanced.
Livy	Livy is updated to 0.7.1.

7.4.7. Release notes for EMR V3.34.X

This topic describes the release notes for E-MapReduce (EMR) V3.34.X, including the release date and updates.

Release date

March 15, 2021 for EMR V3.34.0

Updates

Component	Description
SmartData	SmartData is updated to 3.4.0. For more information, see SmartData 3.4.X.

Component	Description
Spark	- Some default configurations are optimized. - Performance is optimized. Window-based top-k queries can be pushed down. - The capability of reading data from and writing data to Hive tables in the CSV or JSON format is enhanced. - All the column names of a table can be omitted in the ANALYZE statement. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication. - Spark Beeline is easier to use.
Hive	- Some default configurations are optimized. - Performance is optimized. The cost-based optimization (CBO) feature is enhanced. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication. - Calcite is updated to 1.12.0. - The **hive.security.authorization.sqlstd.confwhit elist.append** parameter is added.
Presto	LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
YARN	The risk caused by unauthorized access from a Hadoop cluster to the YARN web UI is fixed. If you access the YARN web UI by using SSH Tunnel, you no longer need to explicitly specify user.name in the URL.
ZooKeeper	ZooKeeper is updated to 3.6.2.
Flink	The *config.sh* file is updated during initialization to fix the issue related to the HADOOP_CLASSPATH parameter.

Component	Description
Impala	- Impala is updated to 3.4.0. - Shiro is updated to 1.7.0. - Data Lake Formation (DLF) metadata is supported. - Data in the Delta format can be queried. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
Tez	Some default configurations are optimized.
HAS	The issue that prevents the admin.keytab file from being initiated again if a HAS installation error is reported is fixed.
Ranger	- The issue caused by filter pushdown in Spark is fixed. - The issue that prevents Presto from being enabled after you disable Presto in Ranger is fixed. - LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
Knox	The link to access Druid 0.20.0 by using Knox is fixed.
Hue	LDAP authentication can be enabled or disabled with a click. For more information, see Manage LDAP authentication.
Hudi	- SQL statements can be executed to query data in Hudi tables. - The issue that causes the query results on some data to be inaccurate is fixed. - Partition pruning is supported if you query data in a Copy on Write table of Hudi by using Spark. - The bucket-based index mechanism is supported to improve write performance.

Component	Description
Delta Lake	• The issue that prevents metadata from being synchronized to a Hive metastore based on existing Delta tables is fixed. • The issue that prevents the MERGE statement from parsing asterisks (`*`) in data is fixed. • The issue that causes an error to be reported when you convert data in the Parquet format into a Delta table and create table metadata is fixed. • The issue that causes the OPTIMIZE command to fail if no files need to be compacted is fixed. • A subquery can be used as the source in the MERGE statement. • Data can be cached if you use Presto to query data in a Delta table. This improves query efficiency. • Impala can be used to query data in Delta tables.
Superset	• The issue that prevents the admin user from logging on to the web UI is fixed. • Datasets are compatible with Druid clusters. • Spark SQL datasets are no longer supported.
Sqoop	Files in the Parquet format can be imported to OSS.
Alluxio	Alluxio is updated to 2.4.1.
Phoenix	Field settings are supported by Hive on Phoenix.
Pig	Pig is removed.

7.4.8. Release notes for EMR V3.33.X

This topic describes the release notes for E-MapReduce (EMR) V3.33.X, including the release date and updates.

Release date

January 15, 2021 for EMR V3.33.0

Updates

Component	Description
SmartData	SmartData is updated to 3.2.0. For more information, see SmartData 3.2.X.

Component	Description
Spark	- Spark is updated to 2.4.7. - jQuery is updated to 3.5.1. - Spark is compatible with Hive to automatically update table and partition sizes. - Spark metadata and job running information can be sent to DataWorks.
Hive	- Hive is updated to 2.3.7. - Metadata from Alibaba Cloud Data Lake Formation (DLF) in an HCatalog table is supported. - Hive metadata and job running information can be sent to DataWorks.
Metastore	- The statistics feature for Hive is added. - Metadata from Alibaba Cloud DLF in an HCatalog table is supported. - Methods to obtain an STS token are optimized.
HDFS	jQuery is updated to 3.5.1.
YARN	- jQuery is updated to 3.5.1. - Configurations of Fair Scheduler are adjusted. - Timeline Server is optimized.
Zeppelin	Zeppelin is updated to 0.9.0.
Ranger	- Audit log configurations of Hive are added. - Configurations of Log4j Audit are added.
OpenLDAP	- The audit feature is added. - By default, the SSL port 10636 is enabled. - Presto can be enabled with one click.
Knox	- Vulnerabilities of the Spring framework are fixed. - The issues caused when you view the Executors page on the web UI of Spark are fixed. - The issues occurred on the job status page of the web UI of Oozie are fixed.
Hue	Presto is supported.
Druid	Druid is updated to 0.20.0.

Component	Description
EmrHook	- Software services are added. - hive-hook: sends Hive metadata and job running information to DataWorks. - spark-hook: sends Spark metadata and job running information to DataWorks.

7.4.9. Release notes of EMR V3.32.X

This topic describes the release notes of E-MapReduce (EMR) V3.32.X, including the release date and updates.

Release date

November 23, 2020 for EMR V3.32.0

Updates

Component	Description
SmartData	SmartData is updated to 3.1.0. For more information, see SmartData 3.1.x.
Alluxio	- Alluxio 2.4.0 is supported. - The default configurations can be adjusted based on the sizes of cluster nodes. - By default, EMR HDFS is used as UnderFS to provide out-of-box services. - Alluxio OSS UnderFS is enhanced to support the new features of different OSS versions. - Engines such as Hadoop, Hive, Spark, and Presto are supported.
Hudi	Hudi 0.6.0 is supported.
Spark	The data collection feature of JindoTable can be enabled or disabled.

Component	Description
Hive	• The connection leak issue of the HiveServer connection pool is fixed. • The data collection feature of JindoTable can be enabled or disabled. • The performance of `ADD COLUMN` is optimized. • The issue that causes data read from Hudi tables to be invalid is fixed. • The default configurations can be adjusted based on the sizes of cluster nodes.
HDFS	The number of supported snapshots is increased.
YARN	The default configurations can be adjusted based on the sizes of cluster nodes.
Tez	The default configurations can be adjusted based on the sizes of cluster nodes.
Sqoop	The issue that files in the Avro format cannot be imported is fixed.

7.4.10. Release notes of EMR V3.30.X

This topic describes the release notes of E-MapReduce (EMR) V3.30.X, including the release date and updates.

Release date

October 26, 2020 for EMR V3.30.0

Updates

Component	Description
SmartData	SmartData is updated to 3.0.0. For more information, see SmartData 3.0.x.
Spark	• Metadata from Alibaba Cloud Data Lake Formation (DLF) is supported. • Has dependencies are updated to 2.0.1. • Issues caused by backticks (`) in Streaming SQL are fixed. • The JAR package of Delta is removed. Delta is separately deployed. • Logs are stored in HDFS directories.

Component	Description
Hive	• Metadata from Alibaba Cloud DLF is supported. • The issue caused when you read an empty Delta table directory and write data into a dummy file is fixed. • Has dependencies are updated to 2.0.1.
Presto	• Metadata from Alibaba Cloud DLF is supported. • The limits on reading a Delta table are eliminated. • The issue of no Java Virtual Machine (JVM) configurations in high security mode is fixed. • Has dependencies are updated to 2.0.1.
HDFS	• The Hot-swap Disk Module mode is supported. • Has dependencies are updated to 2.0.1.
YARN	• Issues related to YARN RMZKStateStore are fixed. • Snappy files generated from Log Service are supported. • Directory configurations in Local mode of MapReduce are modified to resolve the problems that occur in directory permission checks. • The Hot-swap Disk Module mode is supported. • Logs are stored in HDFS directories. • Has dependencies are updated to 2.0.1.
Zookeeper	• Internal IP addresses can be bound to Elastic Compute Service (ECS) instances to start service ports. • Has dependencies are updated to 2.0.1.
Flink-Vvp	• Flink-VVP is updated to a version in the range of 1.11 to 2.2.2. • SQL and Autopilot features are supported. ⑦ **Note** Only Dataflow clusters support Flink-VVP.

Component	Description
Flink	- Data can be written into Object Storage Service (OSS) in cache mode. Exactly-once semantics are supported based on the Flink checkpoint and the data sources that support data retransmission. - Features of open source Flink 1.11.1 are supported. An SQL statement can insert data into different destination tables or partitions and generate multiple channels of output. - Has dependencies are updated to 2.0.1.
Impala	- Custom configuration of parameters on the *catalogd.flgs*, *impalad.flgs*, and *statestored.flgs* tabs is supported. - Shiro is updated to 1.6.0. - Has dependencies are updated to 2.0.1.
Tez	- Default settings of AM memory parameters are optimized. - Has dependencies are updated to 2.0.1.
HAS	Has dependencies are updated to 2.0.1.
Storm	
Zeppelin	
Ranger	
OpenLDAP	
Oozie	
Knox	
Kafka	
HUE	
HBase	
Druid	

7.4.11. Release notes of EMR V3.29.X

This topic describes the release notes of E-MapReduce (EMR) V3.29.X, including the release date and updates.

Release date

July 29, 2020 for EMR V3.29.0

Updates

Component	Description
Bigboot	• Bigboot is updated to 2.7.301. • JindoDistCp allows you to write data into OSS in IA or Archive mode. • Features of Jindo Fuse are enhanced to support multiple namespaces. • The metadata caching feature of JindoFS in cache mode is enhanced.
Spark	• Spark is updated to 2.4.5.2.0. • A third-party metastore is supported. • The datalake metastore-client parameter is added.
Hive	• Hive is updated to 2.3.5.6.0. • A third-party metastore is supported. • The datalake metastore-client parameter is added.
Presto	Presto is updated to 338.
Ranger	• The Ranger software package is updated to 1.2.0-1.5.0. • Presto 338 is supported. • The Description parameter is added to the configuration file.
HDFS	The space reserved for DataNode is configured as needed.
Knox	Knox is compatible with Impala, later Flink versions, and PAI.
Druid	Druid is updated to 0.18.1.
SmartData	SmartData is updated to 2.7.301.

7.4.12. Release notes of EMR V3.28.X

This topic describes the release notes of E-MapReduce (EMR) V3.28.X, including the release date, new features, and updates.

Release date

June 12, 2020 for EMR V3.28.0

New features

Component	Description
Bigboot	- JindoTable is released for data statistics based on the popularity of tables or partitions. - Complete storage policies in block storage mode are supported. Tiered storage policies, including the IA and Archive storage classes, are supported. - The data migration tool Jindo DistCp is used. - The performance of Jindo Fuse is improved and issues related to Jindo Fuse are fixed. - The integration of JFS Scheme on Jindo Job Committer and a Hive engine is optimized in cache mode. - The weight for reading OSS data in block storage mode is configured to alleviate the pressure on reading locally cached data. - JindoFS is divided into the following software modules: Bigboot (management layer), SmartData (distributed service), and JindoFS SDK. These modules are independently updated and maintained.

Updates

Component	Description
Flink	Open source Flink is upgraded to the enterprise-level Ververica Platform. Ververica Platform is customized based on open source Flink 1.10 and provides value-added features such as self-developed storage engine Gemini.
Bigboot	Bigboot is updated to 2.7.0.
Delta	- Delta is updated to 0.6.0. - Delta code is decoupled from Spark code.
Spark	- Spark is updated to 2.4.5. - Spark is compatible with the Streaming SQL scripts of DataFactory. - Delta 0.6.0 is supported.
Hive	Delta 0.6.0 is supported.

Component	Description
Ranger	• Custom deployment of HDFS, Hive, and Spark is supported. • You can configure parameters on the ranger-admin-site and ranger-ugsync-site tabs for the Ranger service in the EMR console.
HDFS	If no DataNode is available when you write data to HDFS, the relevant DataNode exception information (HDFS-9023) is displayed.
Hue	• Hue can be deployed on Gateway clusters. • Multiple Hue instances can be deployed on a single node.
DataFactory	Delta 0.6.0 is supported.
Druid	Druid is updated to 0.18.0.
Knox	• Knox is updated to a version that ranges from 1.0.7 to 1.1.0. • Information on the web UI of HBase can be viewed.

7.4.13. Release notes of EMR V3.27.X

This topic describes the release notes of E-MapReduce (EMR) V3.27.X, including release dates, new features, and updates.

Release dates

Version	Release date
EMR V3.27.0	April 29, 2020
EMR V3.27.1	May 8, 2020
EMR V3.27.2	May 20, 2020

New features

Feature	Description

Feature	Description
Custom component deployment	Components on the master node can be customized. The following components are supported: • Hadoop • Spark • Hive • Zookeeper • Presto
Graceful release during auto scaling	Graceful release during auto scaling is supported. After you enable graceful release, the nodes that are removed during a scale-in are not immediately released. Instead, they are released after tasks are completed within the specified period of time.

Updates

Component	Description
Spark	• Partition fields of the date type are supported in the cube. • The stack depth in the spark-submit scripts is increased.
Delta	• DDL statements, including CREATE, SHOW, and DESCRIBE, are enhanced. • The OPTIMIZE statement for Z-order is supported.
Knox	• Information on the web UI of Druid can be viewed. • Deployment on multiple master nodes is supported.
Hive	• The magic committer in an HCatalog table is supported. • Some outdated default configurations are removed.
Bigboot	• Bigboot is updated to 2.6.3. • Deployment on multiple master nodes is supported.
SmartData	• SmartData is updated to 2.6.3. • Deployment on multiple master nodes is supported.

Component	Description
Ranger	- Solr is supported. - PrestoSQL 311 is supported.
Tez	The ScratchDir option can be used to specify a temporary directory in OSS.
Presto	Presto is updated to 331.
Druid	Druid is updated to 0.17.1.
Superset	Superset is updated to 0.35.2.
Sqoop	- The MySQL JDBC JAR package is updated to 5.1.48. - Custom encoding is supported by using `--mysql-charset` in MySQL direct export mode.

7.4.14. Release notes of EMR V3.26.X

This topic describes the release notes of E-MapReduce (EMR) V3.26.X, including release dates and updates.

Release dates

Version	Release date
EMR-3.26.3	April 16, 2020

Updates

Component	Description
Bigboot	- Bigboot is updated to 2.6.3. - Tablestore metadata and high availability for namespaces are supported.
SmartData	
Hive	The direct committer in an HCatalog table is supported.
YARN	JindoOssCommitter is used as the default committer.
HDFS	JindoFS is updated.
Spark	JindoOssCommitter is used as the default committer.

7.4.15. Release notes of EMR V3.25.X

This topic describes the release notes of E-MapReduce (EMR) V3.25.X, including the release date, new feature, and updates.

Release date

January 13, 2020 for EMR V3.25.0

New feature

Operations related to Ranger Presto are supported.

Updates

Component	Description
Ranger	• The database of Ranger Admin is initialized in a high-availability cluster. • The security issues that are caused when you run scripts on Ranger UserSync are fixed.
Spark	• Parameters related to Delta, such as `spark.sql .extensions` , can be configured in the EMR console.
Delta	• Data from a Delta table can be read by using Hive to avoid manual configuration of InputFormat. • The ALTER TABLE SET TBLPROPERTIES and UNSET TBLPROPERTIES statements are supported.
Hive	The issue that causes execution failures of MapReduce tasks in automatic local mode is fixed.
Presto	• Presto is updated to 310. • Joda-Time is updated to 2.10.5.
Tez	• Tez is updated to 0.9.2. • The issue that the application running progress cannot be properly displayed on the web UI of Tez is fixed. • The issue that you cannot view the application running history on the web UI of Tez is fixed.
Impala	The issue that an LZO table cannot be accessed by using Impala is fixed.
HDFS	JAR packages related to Mongo-Hadoop are removed.

Component	Description
ZooKeeper	ZooKeeper is updated to 3.5.6.
YARN	Information on the web UI of Tez can be viewed. You can set **yarn.resourcemanager.system-metrics-publisher.enabled** to true on the **yarn-site** tab.
Bigboot	• Bigboot is updated to 2.2.3.
SmartData	• The renaming operation is supported in OSS cache mode.
Knox	Dependencies are updated.
Oozie	Dependencies are updated.

7.4.16. Release notes of EMR V3.24.X

This topic describes the release notes of E-MapReduce (EMR) V3.24.X, including the release date, new features, and updates.

Release date

November 18, 2019 for EMR V3.24.0

New features

Component	Description
Delta	• The following SQL statements are supported: ALTER, CONVERT, CREATE, CTAS, DELETE, DESC, INSERT, MERGE, OPTIMIZE, UPDATE, and VACUUM. • The syntax of OPTIMIZE is optimized and built in Delta. • The Hive connector is supported. • The features of open source Delta are supported.
Grafana	Grafana 6.4.2 is supported in independent Flink clusters.
Prometheus	Prometheus 2.13.0 is supported in independent Flink clusters.
Alertmanager	Alertmanager 0.19.0 is supported in independent Flink clusters.

Component	Description
Tensorflow on Spark	- The TensorFlow framework can be deployed on Spark. TensorFlow and Spark are integrated to optimize task scheduling and data exchange, and support the entire deep learning process from the pre-processing of data to deep learning training. - Streaming tasks are supported.

Updates

Component	Description
SmartData	- The cache mode of JindoFS is optimized. The usage of OssFileSystem is integrated into the cache mode to cache data and metadata. Data caching and metadata caching are controlled by using separate switches and are disabled by default. - The read/write performance of the block storage mode and cache mode is optimized. - Hot data cached in local disks is measured more accurately and cleared in a timely manner to ensure that disk usage does not exceed the upper limit. - Both the block storage mode and cache mode are supported in gateway clusters. - The separated deployment of a single JindoFS cluster from multiple computing clusters is supported.
Spark	- Parameters related to Delta are supported. - The Spark plug-in can be configured in Ranger. - JindoCube is updated to 0.3.0.
Hive	- SQL statement compatibility can be checked. - Hive 2.3.5 and Hadoop 2.8.5 are released as a combination. - When Hive is restarted, the content in *hiveserver2-site.xml* is not synchronized to *hive-site.xml* in the spark-conf folder. - The MSCK command can be used to add incremental directories. - The bug triggered by the reuse of a Tez container in Hive is fixed. - The MSCK command can be used to optimize column directories.

Component	Description
Bigboot	Bigboot is updated to 2.2.1. Native code can be executed on all server models.
Ranger	- The deployment mode of the Spark plug-in is modified. - The bug that the emr-header-2 node of a high-availability cluster cannot obtain keytab files is fixed.
Kudu	The startup logic is repaired.
ZooKeeper	Four-letter-word commands can be configured. By default, all four-letter-word commands are enabled.
HDFS	HDFS is compatible with JindoFS.
YARN	- The default value of **yarn.scheduler.capacity.node-locality-delay** is changed to -1. - YARN is compatible with JindoFS.
Has	Has is interconnected with OpenLDAP.
OpenLDAP	OpenLDAP is interconnected with Has.
Presto	Presto is updated to 0.228.
Kafka	Bad disks (d1 instance family) are automatically removed.
Druid	Druid is updated to 0.16.0.
Flume	Flume is updated to 1.9.0.
Flink	- Flink is updated to 1.9.1. - Independent Flink clusters are supported. An IP address whitelist must be configured for an independent Flink cluster.

7.4.17. Release notes of EMR V3.23.X

This topic describes the release notes of E-MapReduce V3.23.X, including the release date and updates.

Release date

September 18, 2019

Updates

Component	Description
Druid	• Upgraded Druid to version 0.15.1. • Added the Router component. • Upgraded Fastjson to version 1.2.60.
Spark	• Updated the code for Spark SQL Thrift Server to fix the issue where IsolatedClassLoader cannot load classes in certain cases. • Refactored the code related to Spark transactions to improve stability. • Fixed the issue where optimized row columnar (ORC) files cannot be read or written after the built-in Hive is upgraded to version 2.3. • Supports the MERGE INTO syntax. • Supports the SCAN and STREAM syntaxes. • Supports exactly-once semantics (EOS) for the Structured Streaming Kafka sink. • Upgraded Delta to version 0.4.0.
Hive	• Removed Hive hooks configured in earlier versions of Hive. • Supports using multiple COUNT(DISTINCT) for hive.groupby.skew in data optimization. • Fixed the issue of data loss when joining tables with different bucket versions.
Flink	Upgraded Flink to version 1.8.2.
Bigboot	• Upgraded the HDFS FsImage analysis tool. • Updated the JAR package for Object Storage Service (OSS) to fix the issue where the non-Daemon thread keeps alive after the main thread is terminated.
Kafka	• Supports the DepolymentSet-Aware replication distribution strategy in Kafka. • Removed the Fastjson dependency.
HDFS	• Optimized the deployment logic of the JAR package for SmartData OSS. • Updated the JAR package for SmartData OSS.
Flume	Upgraded Fastjson to version 1.2.60.
Tensorflow on Spark	This is a new component.

Component	Description
Has	Upgraded Fastjson to version 1.2.60.
Livy	Upgraded Fastjson to version 1.2.60.

7.4.18. Release notes of EMR V3.22.X

This topic describes the release notes of E-MapReduce V3.22.X, including the release date, new features, and updates.

Release date

July 28, 2019

New features

Component	Description
Kudu	- The new component, Kudu, is added. Kudu fills in the gaps in the Hadoop ecosystem. Similar to HBase, you can use Kudu to quickly read, write, and modify data. In addition, you can use Kudu to query and analyze tremendous Parquet data or larger amount of data stored in HDFS. - Kudu provides the C++ API and Java API for secondary development. - Kudu integrates Impala, Spark, and Hive Metastore. - The Kudu component of E-MapReduce V3.22.0 is based on the open-source Apache Kudu 1.10.0.
OpenLDAP	- The new component, OpenLDAP, is added to replace the previous component ApacheDS. - High availability is provided.

Updates

Component	Description

Component	Description
JindoFS	- Provides two storage modes: - Block mode: JindoFS uses Object Storage Service (OSS) as the backend storage. In the block storage mode, JindoFS stores data as blocks in OSS and uses Namespace Service to maintain metadata. The block mode provides better performance than the cache mode when you read and write data or query metadata. The block mode supports multiple storage policies, including WARM, COLD, HOT, TEMP, and ALL_HDD. The WARM policy stores one local backup and one backup in OSS. The COLD policy stores only one backup in OSS. The HOT policy stores multiple local backups and one backup in OSS. The TEMP policy stores only one local backup. The ALL_HDD policy stores multiple local backups. The default storage policy is WARM. You can set different storage policies for directories as required. - Cache mode: This mode is compatible with OSS. In the cache mode, JindoFS stores data files as objects in OSS. When you access OSS objects stored in JindoFS, JindoFS can cache data and metadata of these OSS objects in the local cluster so that you can quickly access them next time. The cache mode provides multiple policies for you to synchronize metadata as required. - Offers an external client: - JindoFS provides an external client so that you can access JindoFS from the outside of an E-MapReduce cluster. You can use the external client to access namespaces in the block mode. However, you cannot use the external client to access the data cached by JindoFS in E-MapReduce clusters. In addition, the performance of accessing data through the external client is worse than that of data access within E-MapReduce clusters. - The cache mode is compatible with the existing OSS semantics. JindoFS accelerates data caching in E-MapReduce clusters. Therefore, you can use the OSS client to directly access JindoFS from the outside of E-MapReduce clusters. For example, you can use the OSS SDK or OssFileSystem of E-MapReduce to access JindoFS from the outside of E-MapReduce clusters. - Supports multiple ecosystem components: - Currently, JindoFS supports various computing engines in E-MapReduce, such as Spark, Flink, Hive, MapReduce, Impala, and Presto. - If you need to separate data computing from data storage, you can store data processing logs and storage logs in JindoFS, such as Spark event logs and YARN container logs. - JindoFS can be used as the backend storage of HBase to store HFile files. This enhances the storage capability of HBase.
OssFileSystem	- Added the feature of automatically detecting bad disks. This feature can resolve cache write failures caused by bad disks when you write data to OSS. - Completed the configurations of OssFileSystem.

Component	Description
Bigboot	- Upgraded Bigboot to version 2.0.0. - Supports multiple namespaces, OSS, multiple storage modes, and access from external clients. - Fixed the issue where the Bigboot monitor status is abnormal during server restart. - Improved the service specifications of Kudu. - Supports a validity check for the specification of each component.
Hadoop	- HDFS - Adapted to HDFS Federation. You can create an HDFS Federation cluster through custom configurations or API operations to avoid formatting namenodes more than one time. - Optimized the logic for detecting bad disks. If you use a local disk, the system detects the bad disk when the datanode blockreport is triggered by the dfsadmin command. - YARN Fixed the issue where the list of history MapReduce jobs is not updated when YARN container logs are stored in JindoFS or JindoFS-based OssFileSystem.

Component	Description

Spark	• Relational cache
	
	Supports using a relational cache to accelerate data queries through pre-computing. You can create a relational cache to pre-compute data. During a data query, Spark Optimizer automatically detects an appropriate relational cache, optimizes the SQL execution plan, and continues data computing based on the relational cache. This accelerates data queries. For example, you can use relational caches to implement multidimensional online analytical processing (MOLAP), generate data reports, create data dashboards, and synchronize data across clusters.
	
	  ◦ Supports using DDL to perform operations such as CACHE, UNCACHE, ALTER, and SHOW. A relational cache supports all data sources and data formats of Spark.
	  ◦ Supports updating caches automatically or by using the REFRESH command. Supports incremental caching based on the specified partitions.
	  ◦ Supports optimizing the SQL execution plan based on a relational cache.
	
	• Streaming SQL
	  ◦ Normalized the parameter settings of Stream Query Writer.
	  ◦ Optimized the schema compatibility check of Kafka data tables.
	  ◦ Supports automatically registering a schema with Schema Registry for a Kafka data table that does not have a schema.
	  ◦ Optimized log information recorded when a Kafka schema is incompatible.
	  ◦ Fixed the issue where the column name must be explicitly specified when the query result is written to a Kafka data table.
	  ◦ Removed the restriction that streaming SQL queries only support the Kafka and LogHub data sources.
	
	• Delta
	
	Added the Delta component. You can use Spark to create a Delta data source to perform streaming data writing, transactional reading and writing, data verification, and data backtracking. For more information, see Delta details.
	
	  ◦ You can call the DataFrame API to read data from or write data to Delta.
	  ◦ You can call the Structured Streaming API to read or write data by using Delta as the data source or sink.
	  ◦ You can call the Delta API to update, delete, merge, vacuum, and optimize data.
	  ◦ You can use SQL statements to create Delta tables, import data to Delta, and read data from Delta tables.
	
	• Others
	  ◦ Supports primary keys and foreign keys. This is a constraint feature.
	  ◦ Resolved JAR conflicts such as the servlet conflict.

Flink	Supports the log rollback feature of Log4j.
Kafka	• Supports the log rollback feature of Log4j. • Upgraded Fastjson.

Component	Description
Zeppelin	Upgraded the dependent commons-lang3 package to version 3.7 to fix the issue where PySpark cannot write data to OSS. For more information, see Spark 2.4 incompatibility with commons-lang3 in Zeppelin.
Ranger	Supports the SHOW GRANTS command.
Analytics-Zoo	Fixed the NumPy installation error.
Impala	Supports Apache Kudu 1.10.0.
Presto	Upgraded Presto to version 0.221.
ZooKeeper	Upgraded ZooKeeper to version 3.5.5.

7.4.19. Release notes of versions earlier than E-MapReduce V3.22.X

This topic describes the release notes of versions earlier than E-MapReduce V3.22.X.

E-MapReduce V3.1.1

- Upgraded the operating system to CentOS 7.2.
- Upgraded Spark to version 2.1.1.
- Upgraded emr-core to version 1.2.6.
- Fixed several bugs of AccessKey-free operations in Object Storage Service (OSS).

E-MapReduce V3.0.2

- Upgraded emr-core to version 1.2.5.
- Supports AccessKey-free OSS operations in more regions.
- Adjusted the AccessKey replacement strategy of Resource Access Management (RAM) roles.
- Fixed several bugs related to Hive and Hadoop.

E-MapReduce V3.0.1

- Supports table management in an interactive and unified manner, and stores Hive metadata in an external unified database. All clusters that use external Hive metadata share the same metadata.
- Upgraded emr-core to version 1.2.4 to improve the read and write performance of OSS.
- Upgraded Spark to version 2.0.2.

> ⑦ Note   E-MapReduce V3.0.1 is fully compatible with E-MapReduce V3.0.0.

E-MapReduce V3.0.0

Released the initial version of E-MapReduce.