Elastic MapReduce
Product Introduction

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

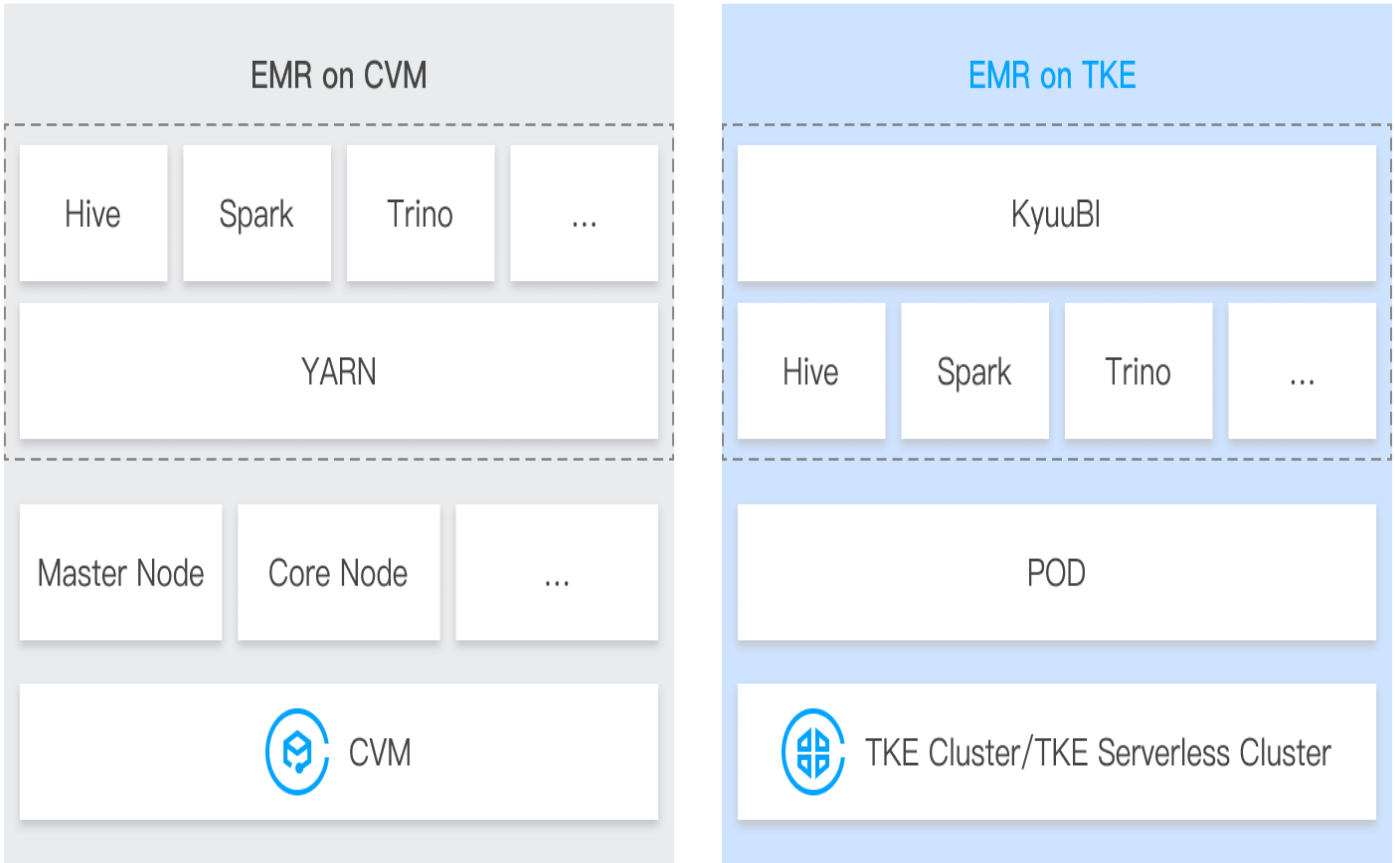
Product Introduction
Overview

Last updated: 2023-12-21 15:31:52

Elastic MapReduce (EMR) is a secure, cost-effective, and highly reliable open-source big data platform, founded on cloud-native technology and the expansive Hadoop open-source ecosystem. It offers easily deployable and manageable open-source big data components such as Hive, Spark, HBase, Flink, StarRocks, Iceberg, Alluxio, and more, assisting clients in efficiently constructing cloud-based, enterprise-level data lake architectures. EMR supports deployment based on two types of resources: CVM and TKE.

Supported formats

Tencent Cloud's EMR offers two deployment and operation methods based on Cloud Virtual Machine (CVM) and Tencent Kubernetes Engine (TKE):



Modali	Description

ties	
EMR on CVM	EMR is responsible for installing and deploying open-source big data components on CVM, and initiating the corresponding services. You can carry out operations and maintenance tasks for the cluster CVM and services through the EMR console.
EMR on TKE	If you have prepared a standard TKE cluster or a Serverless cluster, EMR will install and deploy open-source big data components based on TKE resources, realizing the containerized operation of the open-source big data platform, thereby reducing your operational focus on underlying resources.

Associated Cloud Products

Feature Name	Note
Cloud Virtual Machine (CVM)	Under the product form of EMR on CVM, CVM instances are utilized for EMR cluster nodes.
Cloud Block Storage (CBS)	CVM nodes and POD nodes can be paired with cloud block storage of varying specifications for data storage as required.
Container Service (TKE)	Under the product configuration of EMR on TKE, one can opt for PODs within the TKE cluster to serve as EMR cluster nodes.
Cloud Database MySQL (TencentDB for MySQL)	When components such as Hive, Hue, and Ranger are deployed in the EMR cluster, the cluster will concurrently purchase a Cloud Database MySQL for storing component metadata.
Object Storage (COS)	In scenarios of storage-computation separation, EMR can be utilized to read and write data in object storage.
Cloud HDFS (CHDFS)	In storage-computation separation scenarios, EMR can be employed to read and write data in Cloud HDFS.
Cloud Load Balancer (CLB)	Under the product form of EMR on TKE, certain services can be configured with a load balancer to facilitate external access.
Access Management (CAM)	Grant service role permissions to the EMR product or authorize operations for collaborators/sub-users through Access Management.
Tencent Cloud	Metric and event monitoring data from the EMR cluster are reported

Observability Platform (TCOP)	to TCOP. Through TCOP, cluster monitoring data can be retrieved and alarm notification strategies can be configured for key observability metrics.

Strengths

Last updated: 2023-12-21 15:33:09

Tencent Cloud's EMR offers an enterprise-grade open-source big data service that is easy to deploy and manage, enabling rapid construction of open-source big data services such as Hadoop, Spark, HBase, Trino, and StarRocks. Compared to self-built open-source big data platforms, it possesses the following product advantages:

A wealth of reliable open-source components.

- **Abundant Components**: Offers a rich array of high-performance, highly stable open-source big data components such as Hive, Spark, Presto, StarRocks, HBase, Flink, Iceberg, Alluxio, etc., which can be flexibly combined as per demand.
- **Continuous Iteration**: With the upgrade of open-source versions, it adapts to open-source components, avoiding version compatibility issues between them.
- **Open-Source Enhancement**: Based on deep optimization of the performance and functionality of open-source components, it offers optimization technologies such as Alluxio transparent acceleration and Iceberg Z-Order algorithm.

Ease of Deployment and Maintenance

- **Convenient Deployment**: It only takes a few minutes to build an open-source big data cluster based on CVM or TKE.
- **Convenient Operations and Maintenance**: Supports automated capacity management based on time and load, visual cluster parameter configuration, resource scheduling, federation, and other application-level policy configurations.
- **Comprehensive Monitoring**: Supports full-scale monitoring from resources to service operations, enabling rapid diagnosis of fundamental cluster operation issues through features such as trend analysis of operational metrics, key event monitoring, and log search.
- **Application Analysis**: Supports application-level analysis of key services such as HDFS, YARN, Hive, HBase, Impala, etc., enhancing the efficiency of application-level problem identification.

Cost Savings

- **Resource Elasticity**: Purchase on demand, automatically scale the cluster according to business characteristics, reducing the cost of idle resources.
- **Cluster Federation**: By integrating a unified Hive metadata database and unified object storage, a cross-cluster data set analysis architecture is achieved. Clusters are created or

destroyed on demand, saving flexible cluster costs.

- **Compute-Storage Separation Architecture**: Compute resources and storage resources are purchased separately, and different storage solutions can be chosen based on access frequency, reducing storage and computation costs; supports warm and cold data object storage COS/CHDFS, effectively reducing costs by 28% – 50%.
- **On-Premises and Hybrid Deployment**: Supports deployment based on the container service TKE, staggered reuse of computing power, reducing resource costs.

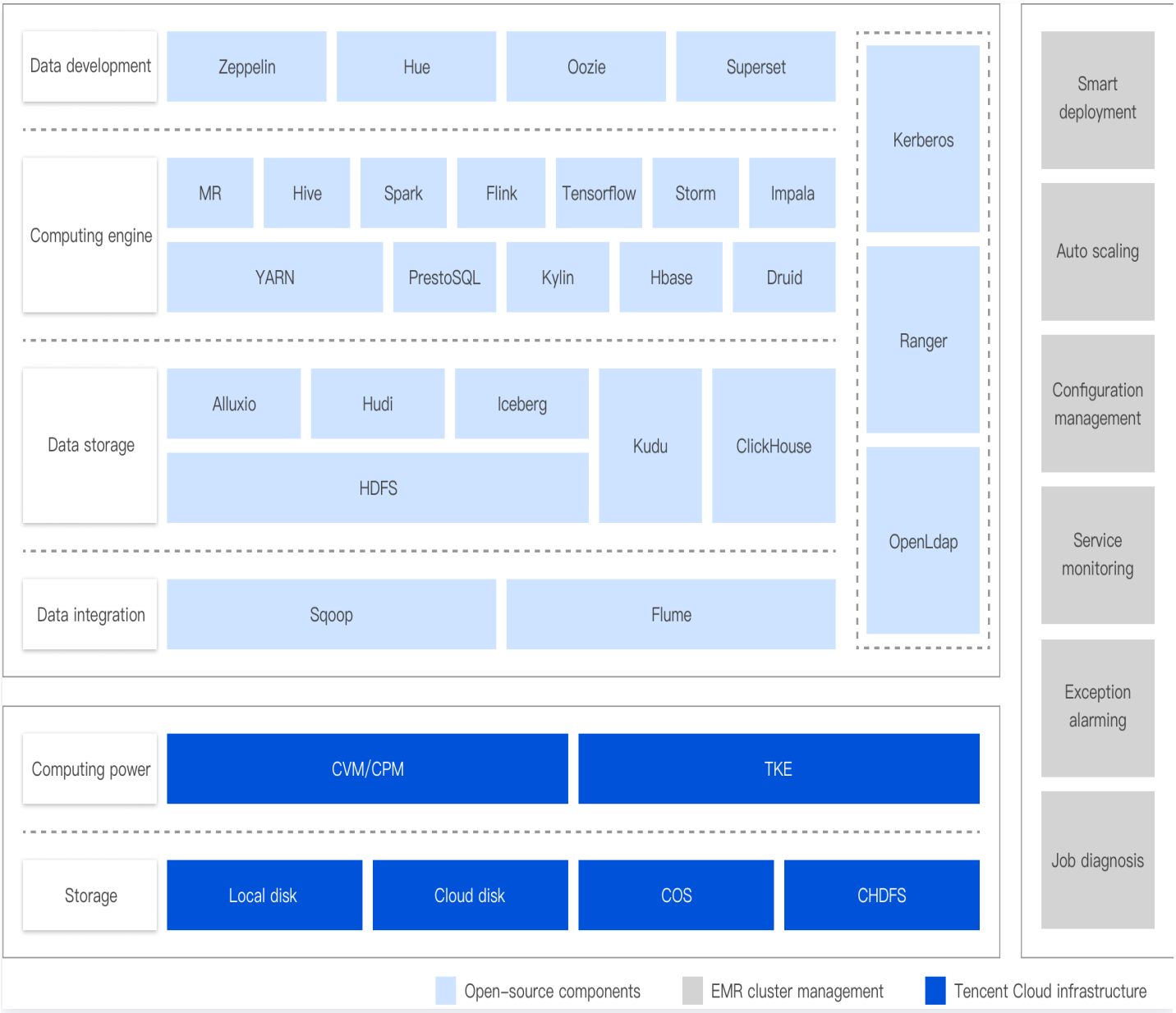
High Security and Reliability

- **Network Security**: VPC network isolation and security groups ensure trustworthy network security.
- **Access Security**: Provides cluster-level Kerberos authentication to ensure secure cluster access; supports fine-grained permission control for local and COS data based on Ranger.
- **Protection Services**: Tencent Cloud's security reinforcement service provides an integrated security service for EMR clusters, covering network protection, intrusion detection, and vulnerability prevention.
- **Disaster Recovery Architecture**: Master node disaster recovery design, backup nodes are launched in seconds, ensuring the availability of big data services.
- **Reliable Storage**: Supports storing Hive metadata in MetaDB, with metadata reliability reaching 99.9996%; supports analyzing PB-level data with high storage durability in COS.

Architecture

Last updated: 2023-12-21 15:33:16

The logical architecture of Tencent Cloud's Elastic MapReduce product is depicted as follows:



EMR primarily consists of three components: open-source elements, Tencent Cloud infrastructure, and cluster management.

- Open Source Components
  - It integrates dozens of rich and cutting-edge open-source big data components from the Apache community, including Hadoop, Hive, Spark, Hbase, Presto, Flink, Alluxio, Iceberg, among others (refer to the Product Distribution and Component Version Preview Table for details).

- Optimized based on open-source components such as Iceberg and Alluxio, it offers performance and functionality enhancements like the Iceberg Z-Order algorithm and Alluxio's transparent URI.

- Tencent Cloud Infrastructure
  - It can be deployed on a variety of underlying computing resources such as Cloud Virtual Machines (CVM), Bare Metal Cloud Servers, and supports containerized deployment.
  - Data can be stored on local disks, cloud hard drives, or on Tencent Cloud Object Storage (COS) and Cloud HDFS services.
  - Private networks (VPC), Network ACLs, and security groups provide a securely isolated network environment for EMR.
- Cluster Management
  - Rapid creation, flexible expansion and contraction, and intelligent deployment management on the cloud with automatic scaling.
  - Service configuration management, bulk node management, and service operation and maintenance visual operations are among the rich and convenient operation and maintenance tools available.
  - Comprehensive cluster monitoring and diagnostic capabilities, including multi-dimensional metric monitoring, events, inspections, alarms, and log searches.

Features

Last updated: 2023-12-21 15:33:22

Elastic MapReduce, integrating cloud technology with community open-source technologies such as Hive, Spark, Presto, StarRocks, HBase, Flink, and Iceberg, offers you a secure, cost-effective, highly reliable, and elastically scalable cloud-based Hadoop service. Its primary functionalities are manifested in the following aspects:

Cluster Management

Through Tencent Cloud's Elastic MapReduce service, you can efficiently configure and manage your open-source big data clusters, seamlessly integrating with cloud-based infrastructure services.

Module	Feature
Cluster Configuration	• Supports over 30+ open-source big data components, flexibly deployed as needed. • Supports the deployment of open-source big data components based on Cloud Virtual Machine (CVM) or Tencent Kubernetes Engine (TKE) services. • Supports the deployment of third-party application components through guided operations. • Supports the pre-setting of key software configuration parameters. • Hive metadata supports the association with existing metadata to achieve multi-cluster sharing of metadata. • Supports the configuration of object storage or cloud HDFS for business data storage.
Manage cluster	• Supports on-demand upgrades of node hardware configurations. • Expand or contract Task nodes or Router nodes as needed. • Supports automatic scaling of Task nodes based on the characteristics of business operation time and load. • Convenient recovery and reuse are possible after the local disk is replaced. • Supports automatic detection of anomalies in Task nodes or Router nodes, and automatically replaces the nodes exhibiting abnormalities. • Supports expansion of cloud hard drives used as data disks. • Supports unified operations and maintenance actions on multiple nodes through cluster scripts.

Service management

Service management provides basic operations and maintenance actions for deployed open-source big data components, while also supporting advanced operations and service strategy configuration management for certain key components.

Module	Feature
Basic service operations and maintenance	• Management of service role status and start-stop operations. • Configuration management for adjusting service parameter settings. • Support for viewing the service's native WebUI. • Support for viewing the service client configuration list.
Advanced service operations and maintenance.	• Support for visual operations of HDFS primary and secondary switchovers and data balancing. • Support for visual operations of YARN primary and secondary switchovers and queue refresh. • Support for HBase RIT repair.
Service policy configuration.	• Support for setting HDFS federation management policies. • Support for setting YARN resource scheduling policies. • Support for OpenLDAP user management.

Monitoring and alarming

Elastic MapReduce provides comprehensive monitoring and alert services for nodes and services within the cluster, particularly offering application layer analysis for certain core components, thereby facilitating more efficient diagnosis of application issues.

Module	Feature
Basic monitoring	• Support for viewing operational metrics of nodes and services, with the ability to retrieve these metrics via API. • Support for monitoring events of nodes and services. • Support for monitoring system operational events and configuring event monitoring strategies. • Support for searching operational logs of service roles. • Support for proactive inspection of cluster operations.
Application Analysis	• HDFS supports storage file analysis.

	• YARN supports job queries, offering insights into the operational status of MR, Tez, and Spark type jobs. • Hive supports query management and data table analysis. • Impala supports query management. • HBase supports data table analysis. • Kudu supports data table analysis.
Alarm management	• Supports default metrics and time-based alarm strategies. • Supports setting alarms for changes in node and service operation metrics. • Supports setting alarms for node and service events.

Scenarios

Last updated: 2023-12-21 15:33:34

Elastic MapReduce (EMR) clusters have a multitude of applications. Every scenario that Hadoop and Spark can support, EMR can also accommodate, as EMR is fundamentally a cluster service for Hadoop and Spark. The following text presents classic scenarios for EMR application.

Offline Data Analysis

Massive logs from game servers, web applications, mobile apps, and other business servers can be synchronized to EMR data nodes or COS. With the aid of tools like Hue, mainstream computing frameworks such as Hive, Spark, and Presto can be used to quickly gain data insights. Tools like Sqoop can be employed to load data scattered across various TencentDB or other storage engines, and the analyzed data can be synchronized back to TencentDB, providing data support for data visualization products like RayData.

Stream Data Processing

After pushing real-time data generated on business servers to CMQ message middleware through APIs or SDKs in programs/tools, an appropriate stream data processing engine can be selected in the EMR product to analyze the data, enabling real-time alerts for business changes. Additionally, the analysis results can be synchronized in real-time to storage engines like TencentDB, facilitating real-time visual inspection of business status through data visualization products like RayData.

Analyzing COS Data

Massive data stored on COS can be swiftly analyzed through the EMR product, achieving thorough storage-compute separation. With such a design, the rich data synchronization tools provided by COS can be fully utilized. Simultaneously, it allows multiple Hadoop clusters of different versions to analyze the same data, addressing the issue of coexistence of multiple Hadoop clusters due to data consistency and historical reasons.

Constraints and Limits

Last updated: 2023-12-21 15:34:42

Prior to utilizing Tencent Cloud's EMR service, we kindly request that you thoroughly peruse and comprehend the following usage restrictions:

- In the interest of ensuring network security for newly established clusters, these clusters will be situated within the same VPC. Please refrain from arbitrarily altering the VPC of existing clusters or nodes to prevent the occurrence of network disconnection within the cluster.
- During the cluster creation process, EMR can assist in establishing a new security group, or you may manually select an existing EMR security group. Please ensure that the manually selected security group possesses the necessary inbound and outbound rules for EMR, and refrain from arbitrarily deleting or modifying the security group in use post cluster creation, to prevent any disruption in cluster communication that could adversely affect the service.
- Please plan the storage space of nodes in advance according to business needs, and timely expand storage nodes to avoid risks to data and node operation due to insufficient storage space. Currently, the three types of EMR cluster nodes – Core, Task, and Router – support mounting multiple cloud disks, with a maximum of 20 cloud disks accumulated per node. The Blackstone 2.0 model and local disk model clusters (IO series and D series) do not currently support multiple cloud disks.
- While utilizing the EMR service, we kindly request that you avoid operations on the CVM console as much as possible, such as shutdown, restart, switching private networks, adjusting security group rules, etc., to prevent cluster anomalies. Operations such as OS reinstallation, instance destruction/resizing, renewal, and changing billing types will also be restricted. You can perform necessary cluster maintenance operations on the EMR console.
- Public IP addresses can, to a certain extent, increase the likelihood of Master nodes being subjected to network attacks. We kindly request that you manage and monitor these associated risks. Elastic public IP addresses (including those on auxiliary network cards) will continue to be retained after the cluster is destroyed. Idle IPs will incur costs, so if retention is not required, please proceed to the corresponding resource manager page for release.
- During the cluster creation process, EMR provides initialization parameters for components that cater to general scenarios. Prior to utilizing component services, we recommend that you verify parameters for components such as HDFS/HBase to ensure they align with your business scenario. Should you require a guide for component initialization, please feel free to reach out to our technical support team.

- We kindly request that you securely store the host login password for your EMR cluster. After you configure password-free login between nodes, Tencent Cloud Security may detect potential vulnerabilities and provide you with corresponding alerts.
- In the event of an abnormal cluster status, billing will continue. We recommend promptly contacting our technical support team for assistance. If there is a need to log into the cluster for troubleshooting, our technical support team will request your account credentials, but only with your consent.

While using and maintaining your EMR cluster, certain unexpected actions may lead to cluster unavailability or instability. Prior to executing some operations on the console, you will be provided with corresponding risk warnings. This document also enumerates some prohibited and high-risk operations for your reference:

Prohibited Operations

Action	Operational Risks
Modifying the EMR Node's Internal IP in CVM	Node Communication Anomalies, Cluster Unavailability
Modifying the Security Group of a CVM Node During Cluster Operation	Node Communication Anomalies, Component Service Unavailability
Deleting Existing Processes/Applications/Files on the Node	Cluster/Component Service Unavailability
Deleting or Modifying the Hosts File in the /etc Directory	The Cluster's Inability to Associate with the Node's Service, Resulting in Service Anomalies
Deleting or Modifying HDFS Metadata File Edit Log	Resulting in HDFS Cluster Unavailability
Manually Modifying the Data of the Hive Metadata Database	Hive Data Parsing Error, Resulting in Service Anomalies
Deleting the Relevant Data Directory of ZooKeeper	Associated Dependency Components Unable to Operate

High-risk Operation

Action	Operational Risks	Suggestions
Shutdown and Restart of EMR	Service Unavailability due to Restart or Shutdown	Confirm the Necessity of the Operation and Thoroughly

Cluster Nodes within CVM		Review the CVM Operation Restrictions
Mounting Disks to EMR Nodes via the CVM Console	EMR's Inability to Recognize and Initialize, Resulting in Disk Unavailability	It is Recommended to Implement through Core Node Expansion or Contraction, or Under the Guidance of Technical Personnel
Unmounting Disks from EMR Nodes via the CVM Console	This May Lead to Data Loss or Cluster Unavailability	It is Recommended to Implement through Core Node Expansion or Contraction, or Under the Guidance of Technical Personnel
Directly Modifying Component Configuration File Parameters on CVM	After Issuing Configurations and Restarting Services from the EMR Console, Parameters Modified on CVM Will Be Overwritten	Modifying Parameter Configurations on the EMR Console is Recommended, Special Cases Should be Conducted Under the Guidance of Technical Support Personnel
Deleting or Modifying the resolv.conf File in the /etc Directory	The Cluster's Inability to Associate with the Node's Service, Resulting in Service Anomalies	Confirm the Necessity of the Operation and Proceed Under Technical Guidance
Modifying the Hostname of EMR Nodes	The Cluster's Inability to Associate with the Node's Service, Resulting in Service Anomalies	Confirm the Necessity of the Operation and Proceed Under Technical Guidance
Modifying the MetaDB Password	EMR Relies on the Password Configured in MetaDB, Modifications May Result in Services Such as Hive/Ranger Becoming Unavailable	Synchronize Configuration Modifications via the EMR Console and Proceed Under the Guidance of Technical Personnel
Modifying the Floating IP of MetaDB	EMR Depends on the IP Configured in MetaDB, Alterations May Lead to Services Like Hive/Ranger Becoming Unavailable	Synchronize Configuration Modifications via the EMR Console and Proceed Under the Guidance of Technical Personnel
Modifying the MetaDB Security	This May Impede Communication Between	Proceed Under the Guidance of Technical Personnel

Group	MetaDB and the Cluster, Rendering Services Like Hive/Ranger Unavailable	

Technical Support Scope

Last updated: 2023-12-21 15:34:54

Tencent Cloud's Elastic MapReduce (EMR) cluster resources are owned by the user, with EMR providing semi-managed cloud service capabilities based on these resources. Users have full administrative rights over the cluster, and are responsible for its daily operation and maintenance. To better support usage, we will now provide a detailed description of the technical support service standards for the EMR product.

Supported Services

- Cluster Purchase, Creation, and Termination Process
  We assist customers in successfully completing the entire purchase and creation process, including software configuration, regional and hardware setup, and basic configuration. We also support the termination of the cluster.
- Cluster Expansion and Reduction Process
  We facilitate customers in selecting different node types to successfully complete the entire process of cluster expansion and reduction.
- Cluster Configuration Modification Process
  We support customers in modifying their machine configurations after successful creation, including the process of individually selecting cloud data disk expansion for configuration changes.
- Cluster Service Functionality
  We support the addition of new component features within the range of optional components, as well as the start-stop service functionality and related management features.
- Cluster Alert Monitoring Functionality
  We enable customers to view the operational status of cluster nodes from the console, set monitoring event rules and inspection times, view alert history, and utilize log search functionality.
- Cluster Auto-Scaling Functionality
  We provide the ability to enable or disable auto-scaling. Once enabled, customers can choose between custom scaling or managed scaling.

Assisted Support Services

- Assistance in Troubleshooting EMR Product Open Source Component Deficiencies or Requirements
  We will address these based on product planning and iteration schedules. This includes,

but is not limited to, active communication with the open source community, providing the community and industry with verified, feasible solutions. However, due to the nature of open source components, Tencent Cloud cannot promise solutions that exceed the progress of the community. For details on open source components, please refer to the Product Release and Component Version Preview Table.

- Assistance in Troubleshooting Deficiencies or Requirements of Other Tencent Cloud Basic Products Dependent on EMR Products.
  We will liaise with the respective product teams for resolution. The other dependent products are as follows:
  - Underlying computational products such as Cloud Virtual Machine (CVM) and Bare Metal Cloud Servers.
  - Local disks, cloud hard drives, or storage products such as Tencent Cloud Object Storage (COS).
  - Private Network (VPC), Network Access Control List (ACL), Security Groups, and other network environments.

Unsupported Services

- EMR offers a wealth of convenient operational tools, including service configuration management, bulk node management, and visual service operation. While ensuring the reliability and availability of these tools, it does not provide operational actions for specific cluster and component management.
- In the absence of any apparent anomalies in the cluster components or clear product defects, we do not take responsibility for troubleshooting individual job issues.
- Services outside the standard product capabilities, such as Core node reduction or disk capacity cleanup, are not supported.
- We do not support handling issues related to customer business application development.
- We do not support handling issues related to third-party components installed by the user.
- We do not support handling issues resulting from user actions that destabilize or render the cluster unavailable, which are outside the product's expected operations. For more details, please refer to Constraints and Limitations.

Support

Should you require technical support while using Elastic MapReduce (EMR), please submit a ticket to contact customer service.

Product release Version Overview

Last updated: 2023-12-21 15:35:22

Product Release Version Number Format

1. EMR On CVM employs the version number format EMR-Va.b.c, as detailed below:
   1.1 The version significance represented in different clusters is as follows:
      - In Hadoop cluster types, 'a' equals 2 signifies support for Hadoop 2.x version, 'a' equals 3 indicates support for Hadoop 3.x version, and 'a' equals 4 denotes a special version of Hadoop 3.x running under the Jdk11 environment.
      - In Kafka clusters, 'a' represents the Kafka version supported by the current version, where 'a' equals 1 signifies support for Kafka 1.x, and 'a' equals 2 indicates support for Kafka 2.x.
      - In StarRocks clusters, 'a' represents the StarRocks version supported by the current version, where 'a' equals 1 signifies support for StarRocks 2.x, and 'a' equals 2 indicates support for StarRocks 3.x.
   1.2 'b' denotes the addition of new components or the upgrade of supported component versions in the release.
   1.3 'c' signifies functional optimization.
2. EMR on TKE adopts the version number format EMR-TKE-Va.b.c, with detailed explanations as follows:
   2.1 'a' represents substantial changes in the overall version.
   2.2 'b' denotes the addition of new components or the upgrade of supported component versions in the release.
   2.3 'c' signifies functional optimization.

Note

- Each version comes bundled with specific components and their respective versions, which are fixed. It does not support the selection of multiple different versions of a component, nor does it allow users to change the version of a component independently. For instance, the EMR-V2.7.0 version comes pre-installed with Hadoop 2.8.5, Spark 3.2.1, and so on.
- Once a cluster is created with a specific EMR version, the EMR version and component versions used by that cluster will not be automatically upgraded. For example, if you choose the EMR-V2.7.0 version, Hadoop will remain at version

2.8.5 and Spark will stay at version 3.2.1. Subsequent upgrades to EMR-V2.8.0 version, or higher versions of Hadoop, or Spark to version 3.3.0, will not affect the already created clusters. Only new clusters will utilize the new images.

- When you upgrade the cluster version through data migration (for example, from EMR-V2.6.0 to EMR-V2.7.0), to prevent issues such as upgrade incompatibility and environmental changes, it is imperative to test the tasks that need to be migrated to ensure they can operate normally in the new software environment.
- The JDK in the EMR release version is based on Tencent Kona (which is based on OpenJDK8). In support of cloud scenarios and features, we have developed and optimized Kona. For more details on Kona, please refer to Knoa.

Overview of Component Versions

Last updated: 2023-12-21 15:36:34

Directions on Discontinued Purchases for EMR Versions

Certain historical EMR releases have been discontinued due to their lower open-source component versions which do not support experiencing new community features. These discontinued EMR releases no longer support creating new clusters but purchased clusters can still use the scaling services normally.
Discontinued EVersions for EMR on the CVM are:
- Hadoop Cluster Types: EMR-V1.3.1, EMR-V2.0.1, EMR-V2.1.0, EMR-V2.2.0, EMR-V2.4.0, EMR-V2.5.1, EMR-V3.0.0, EMR-V3.2.0, EMR-TIANQIONG-V1.0.0.
We advise the utilization of the most recent stable release for each cluster type for the creation of clusters, in order to acquire an abundance of features and enhanced stability.

Update Log for EMR on CVM Releases

EMR on CVM supports several cluster types, including Hadoop, Kafka, StarRocks, etc. Currently, for Hadoop cluster type, there are two options available: the standard version and the Jdk11-beta version.

Component Version Support for Standard Hadoop 2.x Cluster Version

Component Name	EMR-V 2.7.0	EMR-V 2.6.0	EMR-V 2.5.0	EMR-V 2.3.0
Release Date	2022.07	2021.07	2020.09	2020.05
hdfs (Required Component)	2.8.5	2.8.5	2.8.5	2.8.5
yarn (Required Component)	2.8.5	2.8.5	2.8.5	2.8.5
zookeeper (Required Component)	3.6.3	3.6.1	3.6.1	3.5.5
openldap (Required Component)	2.4.44	2.4.44	-	-
knox (Required	1.6.1	1.2.0	1.2.0	1.2.0

Component)				
tez	0.10.1	0.9.2	0.9.2	0.9.2
hive	2.3.9	2.3.7	2.3.7	2.3.5
spark	3.2.1	3.0.2	3.0.0	2.4.3
Livy	0.8.0	0.8.0	0.7.0	0.7.0
kyuubi	1.4.1	1.4.1	–	–
kylin	4.0.1	2.5.2	2.5.2	2.5.2
presto	–	–	–	0.228
trino(prestosql)	385	332	332	–
kudu	1.15.0	1.12.0	1.12.0	–
impala	3.4.0	3.4.0	2.10.0	2.10.0
storm	1.2.3	1.2.3	1.2.3	1.2.3
flink	1.14.3	1.12.1	1.10.0	1.9.2
hbase	2.4.5	1.4.9	1.4.9	1.4.9
phoenix (Integrated into HBase)	5.1.2	4.14.3	4.14.3	4.14.3
alluxio	2.8.0	2.5.0	2.3.0	1.8.1
iceberg	0.13.0	0.11.0	–	–
hudi	0.11.0	0.7.0	–	0.5.1
Hue	4.10.0	4.6.0	4.6.0	4.6.0
oozie	5.2.1	5.1.0	5.1.0	5.1.0
zeppelin	0.10.1	0.9.1	0.8.2	0.8.2
superset	1.4.1	0.35.2	0.35.2	0.35.2
tensorFlowSpark	1.4.4	1.4.4	1.4.4	1.4.4
jupyter (Accompanies the	4.6.3	4.6.3	4.6.3	4.6.3

installation of TensorFlow)				
sqoop	1.4.7	1.4.7	1.4.7	1.4.7
flume	1.9.0	1.9.0	1.9.0	1.9.0
ranger	2.1.0	1.2.0	1.2.0	1.2.0
Kerberos (Selectable only during creation)	1.15.0	1.15.0	1.15.0	1.15.0
ganglia	3.7.2	3.7.2	3.7.2	3.7.2
goosefs	1.2.0	-	-	-

Standard Hadoop 3.x Cluster Component Versions

Component Name	EMR-V3.6.0	EMR-V3.5.0	EMR-V3.4.0	EMR-V3.3.0	EMR-V 3.2.1	EMR-V 3.1.0
Release Date	2023.08	2022.10	2022.04	2021.09	2021.07	2020.12
HDFS (Mandatory)	3.2.2	3.2.2	3.2.2	3.2.2	3.2.2	3.1.2
YARN (Mandatory)	3.2.2	3.2.2	3.2.2	3.2.2	3.2.2	3.1.2
zookeeper (Required Component)	3.6.3	3.6.3	3.6.3	3.6.1	3.6.1	3.6.1
openldap (Required Component)	2.4.44	2.4.44	2.4.44	2.4.44	2.4.44	-
knox (Required Component)	1.6.1	1.6.1	1.6.1	1.2.0	1.2.0	1.2.0
tez	0.10.2	0.10.2	0.10.1	0.10.1	0.10.0	0.9.2

hive	3.1.3	3.1.3	3.1.2	3.1.2	3.1.2	3.1.1
spark	3.3.2	3.2.2	3.2.1	3.0.2	3.0.2	2.4.3
livy	0.8.0	0.8.0	0.8.0	0.8.0	–	–
kyuubi	1.7.0	1.6.0	1.4.1	1.1.0	–	–
kylin	4.0.3	4.0.1	4.0.1	4.0.1	–	–
presto	–	–	–	–	–	–
Trino (formerly PrestoSQL)	414	389	372 (Renamed to Trino)	350	350	332
impala	4.1.1	4.1.0	4.0.0	3.4.0	3.4.0	3.4.0
kudu	1.16.0	1.16.0	1.15.0	1.15.0	1.13.0	1.13.0
hbase	2.4.5	2.4.5	2.4.5	2.3.5	2.3.3	2.3.3
phoenix (Integrated into HBase)	5.1.2	5.1.2	5.1.2	5.1.2	5.0.0	5.0.0
flink	1.16.1	1.14.5	1.14.3	1.12.1	1.12.1	1.10.0
hue	4.10.0	4.10.0	4.10.0	4.10.0	4.4.0	4.4.0
oozie	5.2.1	5.2.1	5.1.0	5.1.0	5.1.0	5.1.0
zeppelin	0.10.1	0.10.1	0.10.1	0.9.1	0.9.1	0.8.2
superset	2.0.1	1.5.1	1.4.1	1.4.1	–	–
alluxio	2.8.0	2.8.0	2.8.0	2.5.0	2.5.0	2.3.0
iceberg	1.1.0	0.13.1	0.13.1	0.11.0	0.11.0	–
hudi	0.13.0	0.12.0	0.11.0	0.8.0	–	–
flume	1.11.0	1.10.0	1.9.0	1.9.0	1.9.0	1.9.0
sqoop	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
ranger	2.3.0	2.3.0	2.1.0	2.1.0	2.1.0	2.0.0

Kerberos (Selectable only during creation)	1.15.1	1.15.1	1.15.1	1.15.1	1.51.1	1.15.1
ganglia	–	3.7.2	3.7.2	3.7.2	–	–
deltalake	2.2.0	2.0.0	–	–	–	–
goosefs	1.4.2	1.4.0	1.2.0	–	–	–

Hadoop Cluster JDK11-Beta Version Supported Component Versions

EMR v4.x is the beta version based on JDK11 environment. All components operate under the JDK11 environment. Currently, it supports only the Hadoop cluster type.

Component Name	EMR-V4.0.0
Release Date	2023.3
HDFS (Mandatory)	3.2.2
YARN (Mandatory)	3.2.2
zookeeper (Required Component)	3.6.3
openldap (Required Component)	2.4.44
knox (Required Component)	1.6.1
tez	0.10.2
hive	3.1.3
spark	3.2.2
livy	0.8.0
kyuubi	1.6.0
kylin	4.0.1
presto	–
Trino (formerly PrestoSQL)	389

impala	4.1.0
kudu	1.16.0
hbase	2.4.5
phoenix (Integrated into HBase)	5.1.2
flink	1.14.5
hue	4.10.0
oozie	5.2.1
zeppelin	0.10.1
superset	1.5.1
alluxio	2.8.0
iceberg	0.13.1
hudi	0.12.0
flume	1.10.0
sqoop	1.4.7
ranger	2.3.0
Kerberos (Selectable only during creation)	1.15.1
ganglia	3.7.2
deltalake	2.0.0
goosefs	1.3.0

Supported Component Product Versions for Kafka Cluster

Component Name	KAFKA- V2.0.0	KAFKA-V 1.0.0
Release Date	2023.03	2021.05
Kafka (Mandatory Component)	2.4.1	1.1.1

KafkaManager (Mandatory Component)	2.0.0.2	2.0.0.2
knox (Required Component)	1.2.0	1.2.0
zookeeper (Required Component)	3.6.3	3.6.1

Supported Component Product Versions for StarRocks Cluster

Component Name	STARROCKS-V2.0.0	STARROCKS-V1.4.0	STARROCKS-V1.3.0	STARROCKS-V1.2.0	STARROCKS-V1.1.0
Release Date	2023.09	2023.04	2023.03	2022.11	2022.08
starrocks (Required Component)	3.1.2	2.5.3	2.4.3	2.3.2	2.2.2
knox (Required Component)	1.2.0	1.2.0	1.2.0	1.2.0	1.2.0

EMR on TKE Release Notes

Component Name	EMR-TKE V1.0.1	EMR-TKE-DLC V1.0.0	EMR-TKE V1.0.0
Release Date	2023.11	2023.11	2023.5
spark	3.2.2	-	3.2.2
virtualspack	-	3.2.2	-

kyuubi	1.7.0	1.7.1	1.6.0
zookeeper	3.6.3	3.6.3	3.6.3
knox	1.6.1	-	1.6.1
hiveserver2	3.1.3	-	3.1.3
metastore	3.1.3	-	3.1.3
trino	389	-	389
ranger	2.3.0	-	2.3.0
hue	4.10.0	-	4.10.0
rss	0.7.1	-	0.6.0
openldap	2.4.44	-	-
presto	--	0.242	-

Mapping relationship between components and API during deployment

Last updated: 2023-12-21 15:37:25

The correlation between component names and API corresponding numerical mappings, or process names and API corresponding numerical mappings, primarily serves for the long type mapping completion of ServiceNodeInfo.N and SoftDeployInfo.N field parameters in the **Cluster Node Expansion** (ScaleOutCluster) API interface.

Component Name	Component Name Mapping ID	Process Name	Process Name Mapping ID
KYUUBI	22	KyuubiServer	87
YARN	2	ResourceManager	6
		NodeManager	7
		JobHistoryServer	14
		TimeLineServer	102
ZOOKEEPER	0	QuorumPeerMain	0
ZEPPELIN	15	Zeppelin	27
TEZ	13	Tez	25
		Tomcat	103
TENSORFLOW ONSPARK	31	-	-
SUPERSET	28	Superset	56
STORM	11	Nimbus	22
		Supervisor	23
SQOOP	7	Sqoop	13
SPARK	6	Spark	12

		SparkJobHistoryServer	17
RANGER	16	Ranger	28
		RangerUsersync	68
		Solr	95
PRESTOSQL	33	PrestoSqlWorker	66
		PrestoSqlCoordinator	67
PRESTO	5	PrestoWorker	11
		PrestoCoordinator	10
OOZIE	9	Oozie	16
LIVY	27	LivyServer	55
KYLIN	14	Kylin	26
KUDU	32	KuduMaster	64
		KuduServer	65
KRB5	21	Krb5Kdc	43
KNOX	24	Gateway	50
OPENLDAP	39	Slapd	51
IMPALA	23	ImpalaServer	47
		ImpalaServerCoordinator	45
		ImpalaServerExecutor	46
		ImpalaStateStore	48
		ImpalaCataLog	49
HUE	8	Hue	15
HUDI	26		54

HIVE	4	HiveServer2	31
		HiveMetaStore	30
		HiveWebHcat	32
HIVESERVER2	47	HiveServer2	31
		HiveWebHcat	32
METASTORE	48	HiveMetaStore	30
HDFS	1	NameNode	1
		Router	88
		DataNode	2
		zkfc	8
		JournalNode	3
HBASE	3	HMaster	4
		HRegionServer	5
		HbaseThrift	37
GANGLIA	10	Httpd	36
		Gmond	21
		Gmetad	19
FLUME	18	Flume	38
FLINK	12	Flink	24
FILEBEAT	25	Filebeat	53
DRUID	29	router	62
		coordinator	57
		overlord	58
		historical	59
		middleManager	60

		broker	61
COSRANGER	34	CosRangerServer	71
CLICKHOUSE	30	ClickHouseServer	63
ALLUXIO	20	AlluxioMaster	41
		AlluxioWorker	42
		AlluxioJobMaster	69
		AlluxioJobWorker	70
GOOSEFS	38	GooseFSMaster	82
		GooseFSWorker	83
		GooseFSJobMaster	84
		GooseFSJobWorker	85
		GooseFSProxy	101
DORIS	35	DorisFeFollower	72
		DorisBe	73
		DorisBroker	74
		DorisFeObserver	75
KAFKA	36	Kafka	76
KAFKAMANAGER	37	kafkamanager	77
ICEBERG	40	Iceberg	44
DELTA	46	Delta	44
STARROCKS	41	StarRocksFeFollower	78
TRINO	42	StarRocksBe	79
		StarRocksBroker	80

		StarRocksFeObserver	81
		StarRocksCn	104
TRINO	42	TrinoWorker	89
		TrinoCoordinator	90