Elastic MapReduce

Product Introduction

Product Documentation

Contents

Product Introduction
Overview

Last updated：2019-07-26 17:34:23

Tencent Cloud Elastic MapReduce (EMR) is a cloud-hosted Hadoop service that features Hadoop cluster deployment, software installation, configuration modification, monitoring and alarming, and elastic scaling, providing individual and enterprise users with a secure and stable big data processing solution.

Features

- With open-source Hadoop, you can use Tencent Cloud EMR to move vast amounts of data form your local data center to the Cloud. Tencent Cloud Elastic MapReduce (EMR) is integrated with popular components (e.g., Hive, Hbase, Spark, Presto, Sqoop, and Hue) to simplify data processing and management, such as offline big data processing and stream computing.

- EMR seamlessly incorporates the Tencent Cloud Cloud Object Storage (COS) service, allowing you to migrate files stored in HDFS to COS with unlimited scalability, low storage costs and high reliability for separation of computation and storage. With COS, a cluster can be created whenever needed and terminated after the task is completed without the concerns over data loss. In addition, the on-demand clusters can significantly reduce your big data processing costs.

- EMR has five node types: master, core, task, router, and common nodes. For the purposes of each type, see Node Type Descriptions.

- Currently, EMR supports a wide variety of resource specifications, including Standard, MEM-optimized, High IO, Compute, and Big Data models. If you need to deploy a Hadoop cluster on a CPM instance, contact us by submitting a ticket.

Strengths

Last updated：2019-07-26 17:34:29

Compared to self-built Hadoop, Tencent Cloud EMR provides simpler, more stable, and more reliable Hadoop services that simplify big data processing and analysis.

Flexibility

- With Tencent Cloud Elastic MapReduce (EMR), you can create a secure and reliable Hadoop cluster in just a few minutes to run many mainstream open-source big data frameworks such as Hive, Spark, and Presto.
- Additionally, you can elastically scale up your EMR cluster to meet increasing computing needs for data analytics and scale down to reduce IT hardware costs as the needs drop.

Reliability

- The master node is designed with disaster recovery in mind, and if it fails, a slave node will be started in seconds to ensure the availability of big data services.
- A comprehensive monitoring system is in place, which can send SMS messages for exceptions in cluster components and tasks in a matter of seconds.
- Hive metadata can be stored in TencentDB with a metadata reliability of 99.9996%.
- Petabytes of high-persistence data stored in COS can be analyzed.
- The recycle bin feature is enabled for clusters by default

Security

- The network policy for managed Hadoop clusters can be well planned through the convenient network isolation enabled by VPCs. Network ACLs and security groups can be created to filter traffic at the subnet and server levels, helping meet your network security needs in all aspects.
- Tencent Cloud security reinforcement service provides an integrated security solution for EMR clusters, ranging from network protection and intrusion detection to vulnerability protection.

Ease of use

- Different clusters versions can be created to analyze the same data in COS in response to the actual business needs.
- Petabytes of data stored in data nodes or COS can be analyzed with the aid of out-of-the-box community components such as Hue and Oozie, eliminating your concerns over any knowledge migration costs.

Reduced Costs

- EMR allows elastic scaling of your managed Hadoop cluster based on the business curve to reduce the high hardware costs.
- It comes with a rich set of OPS tools which greatly improves efficiency and enable you to focus on the business itself without having to worry about infrastructure concerns such as monitoring and security.

Features

Last updated：2019-07-26 17:34:44

Tencent Cloud Elastic MapReduce (EMR) is integrated with open-source frameworks and projects, such as Apache Hadoop, Apache Hive, Apache Spark, and Apache Storm. The cloud-based Hadoop services allow you to process vast amount of data securely, cost-efficiently, and elastically at scale. It has the following advantages:

Auto Scaling

Creating a cluster in minutes
You can create a secure, stable, cloud-managed Hadoop cluster in just a few minutes in the console.

Scaling in minutes
Your EMR cluster can be smoothly scaled up or down just a few minutes as your computing needs change.

API support
EMR clusters can be easily created, scaled, and terminated in programs through APIs.

Separation of storage and computation

Intra-cluster separation of storage and computation
At the cluster level, cloud-based Hadoop clusters can be planned in a manner where storage nodes and compute nodes are separated, so that you can scale the compute nodes as needed to lower the hardware costs.

COS-based separation of storage and computation
Massive amounts of data to be analyzed can be stored in COS. While the storage costs is reduced through COS, different versions of EMR clusters can be created to analyze the same data, which brings out extreme architectural flexibility.

OPS Support

Monitoring and multi-channel alarming
A comprehensive monitoring and OPS system is provided, which can detect exceptions in components

such as Spark, Hive, and Presto and tasks within seconds after they occur to ensure the robust operations of big data clusters.

Technical support
In addition to comprehensive technical documentation, Tencent Cloud also has a technical service system where complete technical support is provided through various channels such as email and WeChat.

Security

EMR uses security groups to control inbound and outbound traffic to your CVM instances. Components' web UIs can only be accessed through one specified instance assigned with public IP, and the access requests must be authenticated by username and password. In addition, the security group of this instance only allows SSH ports and proxy access ports.

> Changing the project will cause the CVM instance to lose its security group.

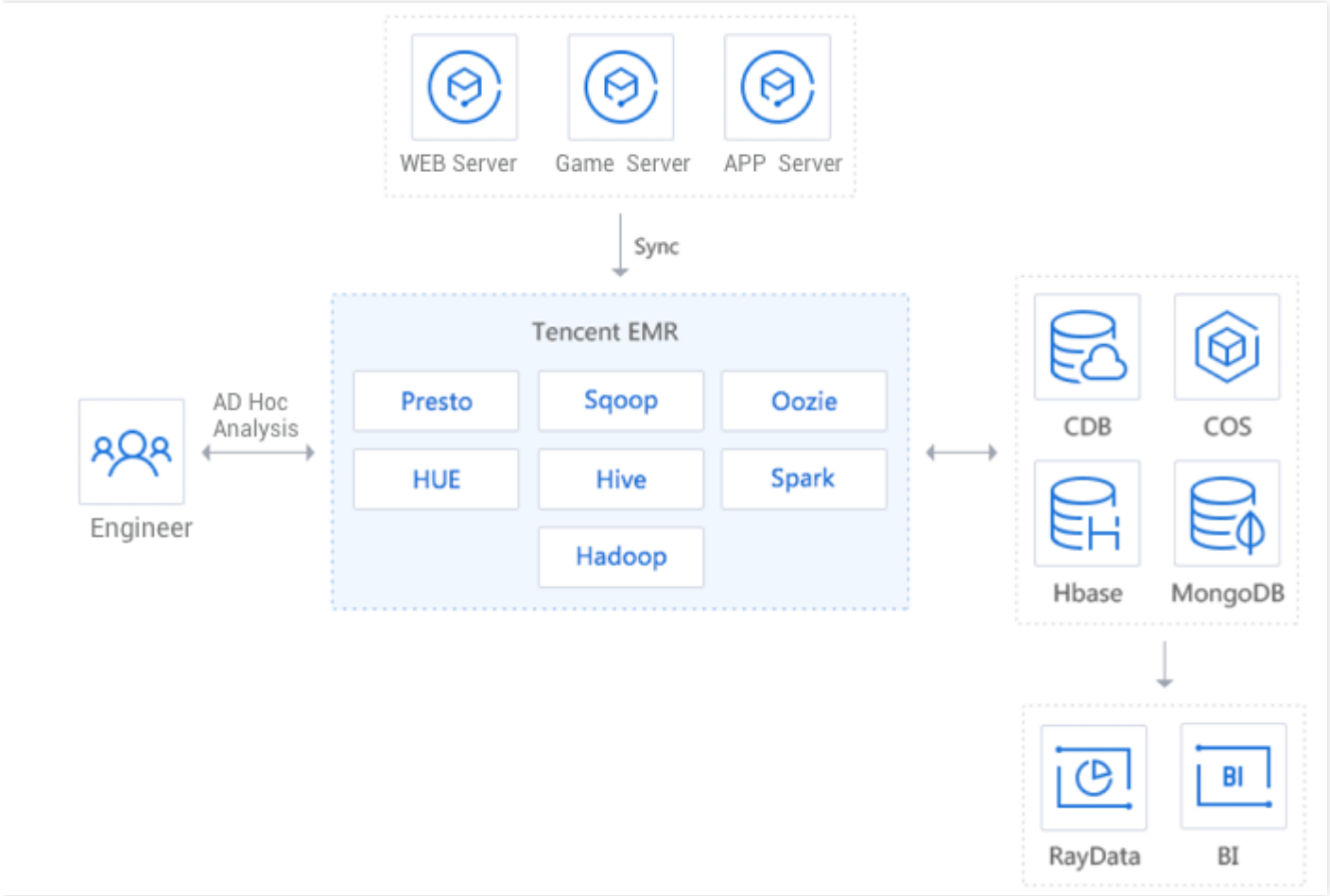
Use Cases

Last updated：2019-07-26 17:34:53

Elastic MapReduce (EMR) clusters supports many application scenarios by running big data frameworks Hadoop and Spark. Below are some typical EMR application scenarios:

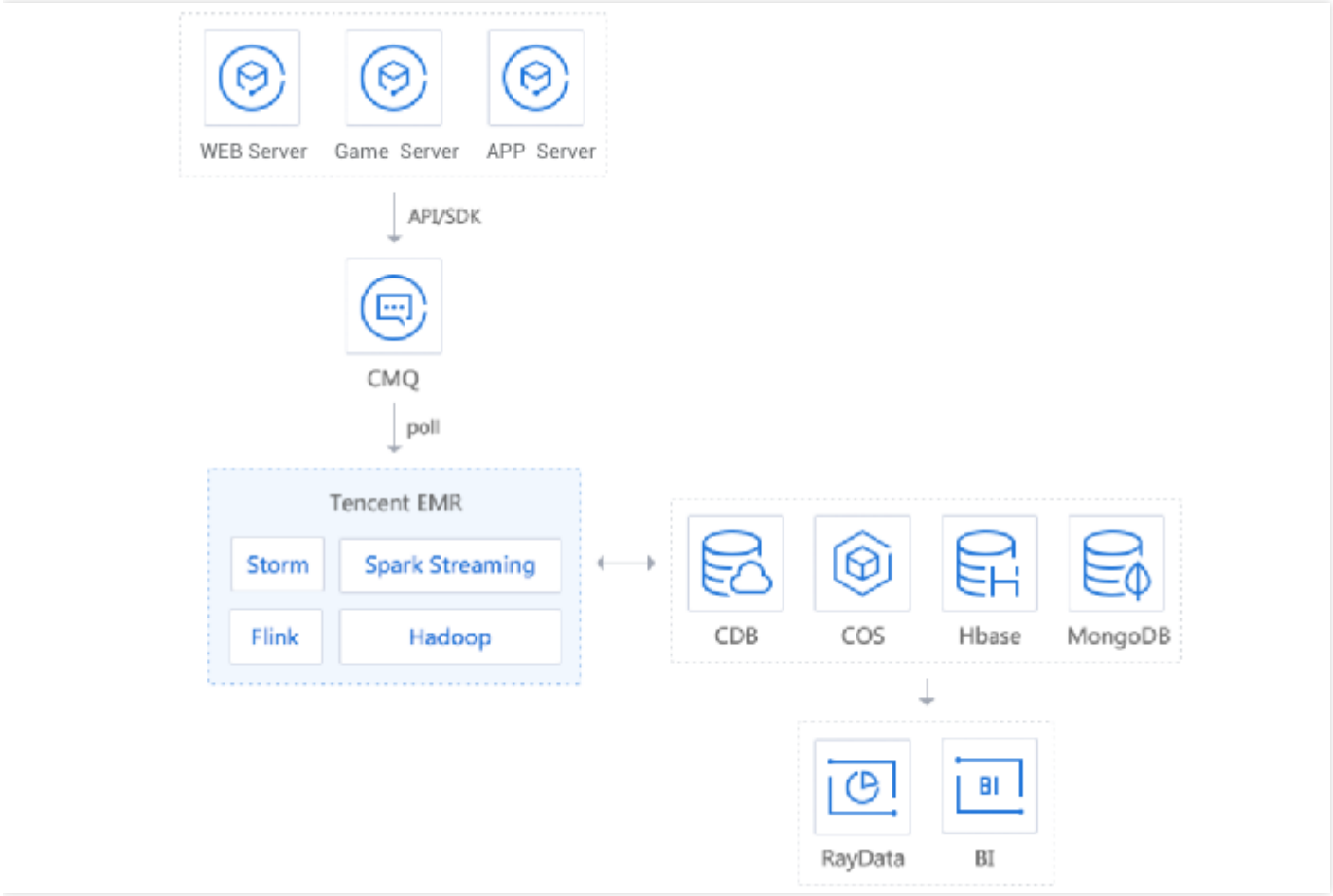
Offline Data Analytics

EMR synchronizes vast amounts of log data from business applications (e.g. games, webs, and mobile Apps) servers to nodes or COS. With Hue, an open source SQL Workbench for Data Warehouses, you can use big data frameworks such as Hive, Spark, and Presto to analyze data to derive business insights. In addition, you can use Sqoop to integrate and analyze the data scattered across TencentDB and other storage engines. Sqoop synchronizes the analyzed data back to TenentDB to support data visualization services like RayData.



Streaming Data Processing

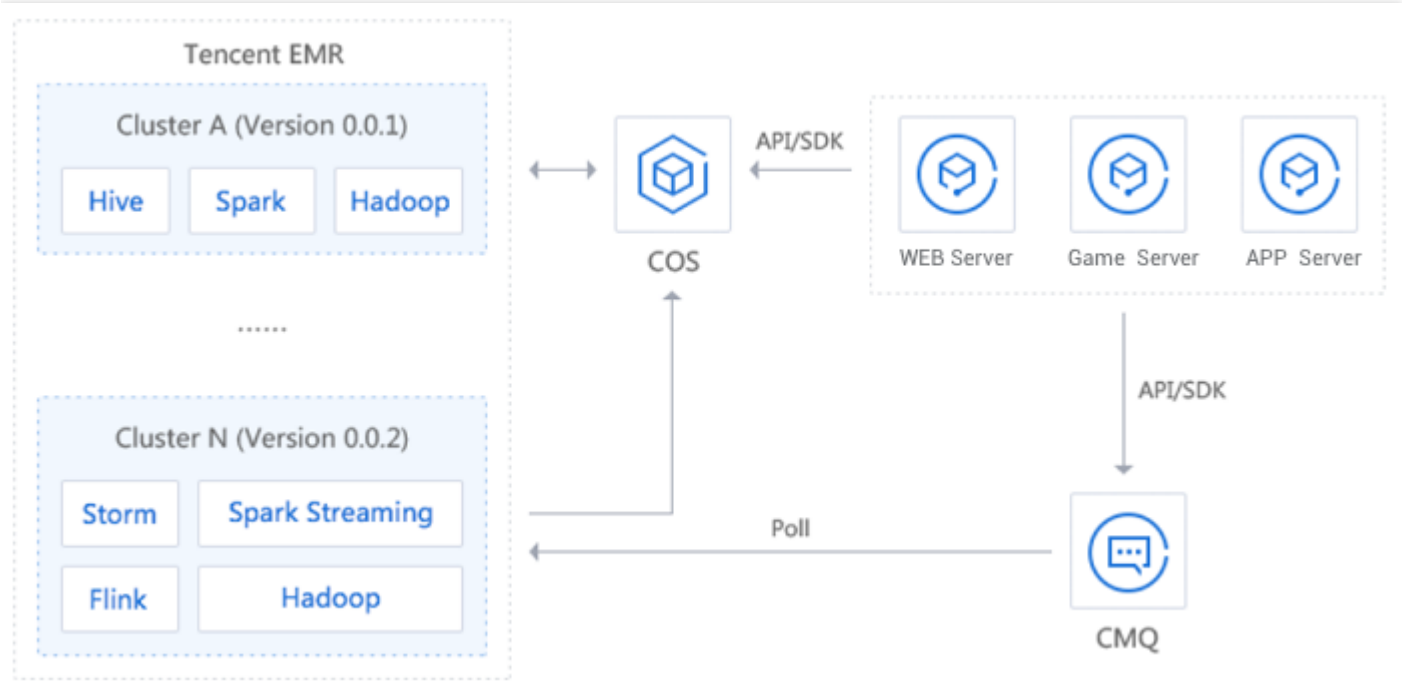
After pushing the real-time data generated from business servers to the CMQ messaging middleware through APIs and SDKs in programs/tools, you can select an appropriate streaming data processing engine in EMR to analyze the data for real-time alarming in respond to business changes. The analysis results can be synced to storage engines such as TencentDB in real time, which helps you monitor your business by using data visualization services like RayData.



COS Data Analytics

You can use EMR to process and analyze vast amounts of data stored in COS so that computation and storage can be completely separated. This architecture allows you to use various data synchronization tools in COS. At the same time, you can use multiple Hadoop cluster versions to analyze the same data to achieve data consistency and resolve legacy issues caused by the coexistence of multi-version Hadoop

clusters.



Note Types

Last updated：2019-07-26 17:34:59

EMR offers five types of nodes:

Node Type	Description	HA Quantity	Non-HA Quantity
Master	Processes such as NameNode, ResourceManager, and HMaster are deployed here.	2	1
Core	Processes such as DataNode, NodeManager, and RegionServer are deployed here.	≥ 3	≥ 2
Task	Processes such as NodeManger and PrestoWork are deployed here.	The number of task nodes can be changed at any time to achieve elastic scalability of the cluster. The minimum value is 0.	
Common	Distributed coordinator components such as ZooKeeper and JournalNode are deployed here.	≥ 3	0
Router	Hadoop packages, including software programs and processes such as Hive, Hue, and Spark, are deployed here.	The number of router nodes can be changed at any time. The minimum value is 0.	

- A master node is a management node that ensures that the scheduling of the cluster works properly.
- A core node is a compute and storage node. All your data in HDFS is stored in core nodes. Therefore, in order to ensure data security, once core nodes are scale out, they cannot be scaled in.
- A task node is a pure compute node and does not store any data. The computed data comes from a core node or COS. Therefore, it is often used as an elastic node and can be scaled in or out at any time.
- A router is used to share the load of a master node or as the task submitter of the cluster. It can be scaled in or out at any time.

Component Versions

Last updated：2019-07-26 17:35:05

EMR consists of a series of open-source applications in the big data ecosystem. Each version of EMR contains a specific set of open-source programs. When you create a cluster, you can choose the most appropriate EMR version based on your actual needs.

- EMR is upgraded on a regular basis with versions such as EMR v1.3.1, EMR v2.0.1, and EMR v2.1.0.
- The components and their versions bundled with each EMR version are fixed. Currently, neither selecting multiple versions of a component nor changing a component version in one EMR version is supported. For example, Hadoop v2.7.3 and Spark v2.2.1 are built into EMR v2.0.1.
- Once a version of EMR is selected for cluster creation, the EMR and components used by the cluster will not be automatically upgraded. For example, if EMR v2.0.1 is selected, then Hadoop will always be v2.7.3, and Spark will always be v2.2.1. Even if you subsequently upgrade EMR to v2.1.0 where Hadoop v2.8.4 and Spark v2.3.2 are included, the previously created cluster will not be affected, and only new clusters will use the new versions.
- When you upgrade the cluster through data migration, for example, from EMR v2.0.1 to EMR v2.1.0, in order to avoid issues such as version incompatibility or environment changes, be sure to test the tasks to be migrated and ensure that they can work properly in the new software environment.

The components and their versions included in each EMR version are as follows:

Component Name	EMR v1.3.1	EMR v2.0.1	EMR v2.1.0
Release date	-	-	May 2019
Flink	1.2.0	1.2.0	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2
Hadoop	2.7.3	2.7.3	2.8.4
Hbase	1.2.4	1.3.1	1.3.1
Hive	2.1.1	2.3.2	2.3.3
Hue	3.12.0	3.12.0	4.4.0
Ooize	4.3.1	4.3.1	4.3.1

Component Name	EMR v1.3.1	EMR v2.0.1	EMR v2.1.0
Presto	0.161	0.188	0.215
Ranger	-	0.7.1	0.7.1
Spark	2.0.2	2.2.1	2.3.2
Sqoop	1.4.6	1.4.6	1.4.7
Storm	1.1.0	1.1.0	1.1.0
Tez	0.8.5	0.8.5	0.8.5
Zookeeper	3.4.9	3.4.9	3.4.9
Flume	-	-	1.8.0
Alluxio	-	-	1.8.1
Knox	1.2.0	1.2.0	-

Release History

Last updated：2019-07-26 17:35:16

May 17, 2019

New versions/specifications

Added EMR v2.1.0 and upgraded main components.

New features

Added the support for Kerberos secure clusters.

Optimizations

1. Optimized the monitoring metrics for the server, HDFS, Yarn, and HBase.
2. Optimized the console style for an improved interactive experience.
3. Optimized the CVM and TencentDB naming with the EMR cluster serial number for easier locating of the cluster information.

May 7, 2019

New features

Changed the public IP address for a master node to optional.
Made the number of common nodes adjustable as needed.

March 29, 2019

New versions/specifications

Added the support for the I3 model in Beijing, Shanghai, and Guangzhou regions. This model is a CVM whitelist model, and you can purchase it only if you are in the I3 model whitelist.

New regions/AZs

Made purchase available in Silicon Valley.

March 4, 2019

New features

1. Added the support for router nodes which are mainly used to relieve the load of master nodes and as task submitters.
2. Added the support for node configuration adjustment (to a higher configuration).

January 15, 2019

New features

Added the support for adding new components to existing clusters.