弹性 MapReduce

产品简介

产品文档

腾讯云

【联系我们】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100。

文档目录

产品简介

产品概述

最近更新时间：2020-11-09 15:04:05

弹性 MapReduce（EMR）结合云技术和 Hadoop、Hive、Spark、Hbase、Presto、Flink、Druid、ClickHouse 等社区开源技术，提供安全、低成本、高可靠、可弹性伸缩的云端泛 Hadoop 大数据架构。您可以在数分钟内创建安全可靠的专属 Hadoop 集群，以分析位于集群内数据节点或对象存储 COS 上的 PB 级海量数据。

功能特性

弹性 MapReduce 的组件完全源于开源社区，您可以将现有的大数据集群无缝平滑迁移至腾讯云上。弹性 MapReduce 产品中集成了社区中常见的热门组件，包括但不限于 Hadoop、Hive、Hbase、Spark、Presto、Sqoop、Hue、Druid、ClickHouse 等，可以满足您对大数据在线业务、离线/近线数据仓库、实时流式计算等全方位场景的需求。

弹性 MapReduce 无缝集成了腾讯云对象存储（COS）服务，您可将原本存储于 HDFS 中的文件放置在可无限扩展、存储成本低且高可靠的 COS 中，实现计算存储分离。依托于 COS，您可以在需要的时候创建集群，并在任务完成后销毁集群。与此同时，您无需担心数据的丢失。按需创建的集群，可以大幅度降低您的大数据处理成本。

弹性 MapReduce 定义了5种节点类型：Master 节点、Core 节点、Task 节点、Router 节点和 Common 节点。各类型节点作用，请参见 节点类型说明。

弹性 MapReduce 目前支持多种资源规格，您可以采用标准型、标准网络优化型、内存型、高 IO 型、计算型、计算网络增强型及大数据机型实例作为计算资源。若您需要在黑石物理主机上部署集群，请 提交工单 联系我们。

产品优势

最近更新时间：2020-10-19 16:45:10

与自建 Hadoop 集群相比，弹性 MapReduce 能提供更方便、更安全、更可靠的云端 Hadoop 服务。

> ⚠ 注意：
> 除提供 Hadoop 集群类型外，还支持 Druid 和 ClickHouse 大数据集群，提供更丰富的大数据架构。

灵活

- 只需几分钟即可获得一个安全可靠的 Hadoop 集群，以运行 Hive、Spark、Presto 、Impala、ClickHouse、Druid、Flink 等主流开源大数据计算框架，覆盖用户**交互式 BI、数仓场景、实时计算**等场景的需求。
- 提供对现有弹性 MapReduce 集群进行快速弹性伸缩的能力，实时调配云端计算资源以应对业务数据的快速波动，节省高昂的预留 IT 硬件成本。

可靠

- Master 节点容灾设计，备节点秒级拉起，保障大数据服务可用性。
- 完善的监控体系建设，您可以通过短信渠道秒级感知集群组件及任务的运行异常状况。
- 支持将 Hive 元数据存放于 MetaDB，元数据可靠性达99.9996%。
- 支持分析存放于 COS 的高存储耐久性的 PB 级数据。
- 集群默认开启回收站功能，提供误删除设备的找回机制。

安全

- 可通过便捷的 VPC 网络安全隔离手段规划托管 Hadoop 集群网络策略，支持网络 ACL 和安全组，可从子网和节点维度筛选流量，全方位满足网络安全需求。
- 腾讯云品质的安全加固服务为 EMR 集群提供一体化的安全服务，涵盖网络防护、入侵检测、漏洞防护等。
- 提供集群级别的 Kerberos 认证，保障集群访问安全。

易用

- 可以响应业务需求创建不同版本的集群分析 COS 上的同一份数据。
- 可以借助开箱即用的 Hue、Oozie 等社区组件随心分析位于数据节点或 COS 上的 PB 级数据，无需担心产生任何知识迁移成本。
- 近千项集群级、组件级监控指标，搭配监控概览页面，提供丰富且清晰易用的监控系统。

- 灵活支撑云端多机型集群，实现对异构配置集群在扩容、配置下发等场景下的轻松应对，以更优硬件配置应对业务分析挑战。

节约成本

- 通过 EMR 服务，可以按业务曲线随心伸缩托管 Hadoop 集群，缩减高昂的硬件成本。
- 丰富的运维工具支持，大幅提升运维工作效率，让工程师更专注于业务本身的商业价值，摆脱重复搭建监控、安全、运维工具等基础设施。
- 支持温冷数据的对象存储 COS/CHDFS 存储，成本有效降低28% - 50%。
- 结合统一 Hive 元数据库以及统一对象存储，实现跨集群的同数据集分析架构，集群按需创建或销毁，节省集群柔性成本。

产品功能

最近更新时间：2020-09-30 16:01:27

弹性 MapReduce 结合云技术和 Hadoop、Hive、Spark、Storm 等社区开源技术，为您提供安全、低成本、高可靠、可弹性伸缩的云端 Hadoop 服务。其主要功能体现在以下方面：

弹性伸缩

分钟级集群创建

通过控制台数分钟就可创建一个安全、稳定的云端托管 Hadoop 集群。

分钟级集群扩缩容

仅需数分钟即可对现有 EMR 集群进行平滑扩缩容，以适应互联网业务需求的快速变化。

API 支持

支持通过 API 方式便捷的在程序中创建、扩缩容、销毁 EMR 集群。

存储计算分离

集群内存储计算分离

集群内支持按照存储节点、计算节点的模式来规划云端 Hadoop 集群，以支持客户对计算节点的随意伸缩来降低硬件成本。

基于 COS 的存储计算分离

支持把待分析海量数据存放于 COS，在通过 COS 规模化效应降低存储成本的同时，您还可以创建不同 EMR 版本分析同一份数据，这将为您带来极度的架构灵活性。

运维支撑

监控与多渠道告警

提供完善的监控运维体系，对包含 Spark、Hive、Presto 等在内的组件异常和任务异常的秒级感知，以保障大数据集群的稳健运行。

技术服务支持

在提供完善技术文档之外，还支持包含邮件、QQ、微信等渠道在内的技术服务体系，为客户提供完备的技术支持。

安全

EMR 创建的 CVM 子机同时会创建安全组来限制外网访问。各组件 Web UI 均通过其中一台有外网 IP 的子机进行访问，并且通过用户名和密码进行验证，有外网 IP 的子机安全组只开放 SSH 端口和代理访问端口。

> **注意：**
> CVM 子机如果更换项目会导致 CVM 安全组丢失。

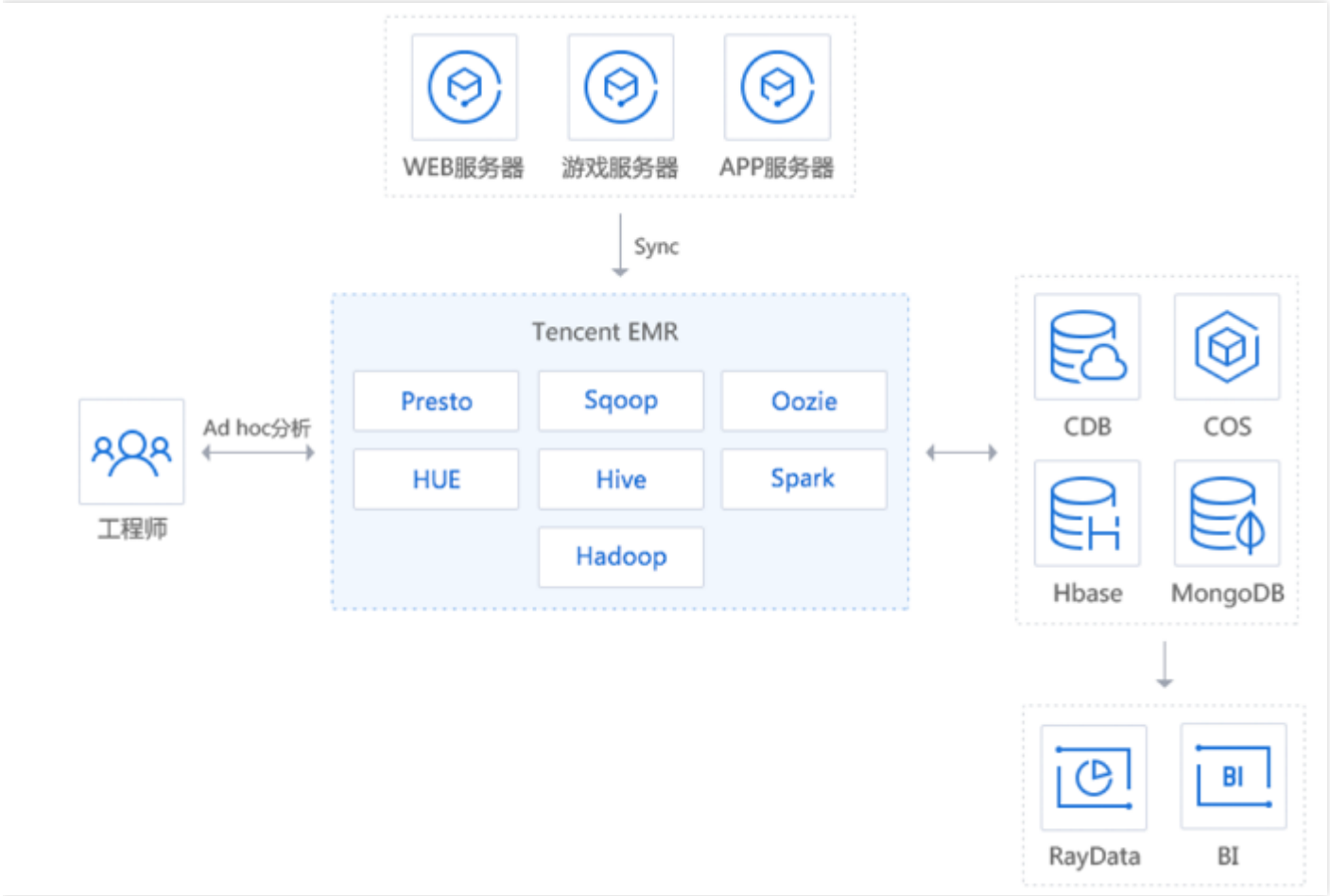
应用场景

最近更新时间：2020-06-02 17:13:25

弹性 MapReduce（EMR）集群应用场景很多，Hadoop 和 Spark 能够支持的场景 EMR 都可以支持，因为 EMR 本质就是 Hadoop 和 Spark 的集群服务。下文为 EMR 应用的经典场景。

离线数据分析

把游戏、Web 应用、手机 App 等业务服务器上的海量日志同步到 EMR 的数据节点或 COS 后，可借助于 Hue 等工具使用 Hive、Spark、Presto 等主流计算框架快速获取数据洞察力。可使用 Sqoop 等工具加载分散于各 TencentDB 或其他存储引擎的数据，并把分析后的数据同步到 TencentDB，为 RayData 这样的数据可视化产品提供数据支撑。



流式数据处理

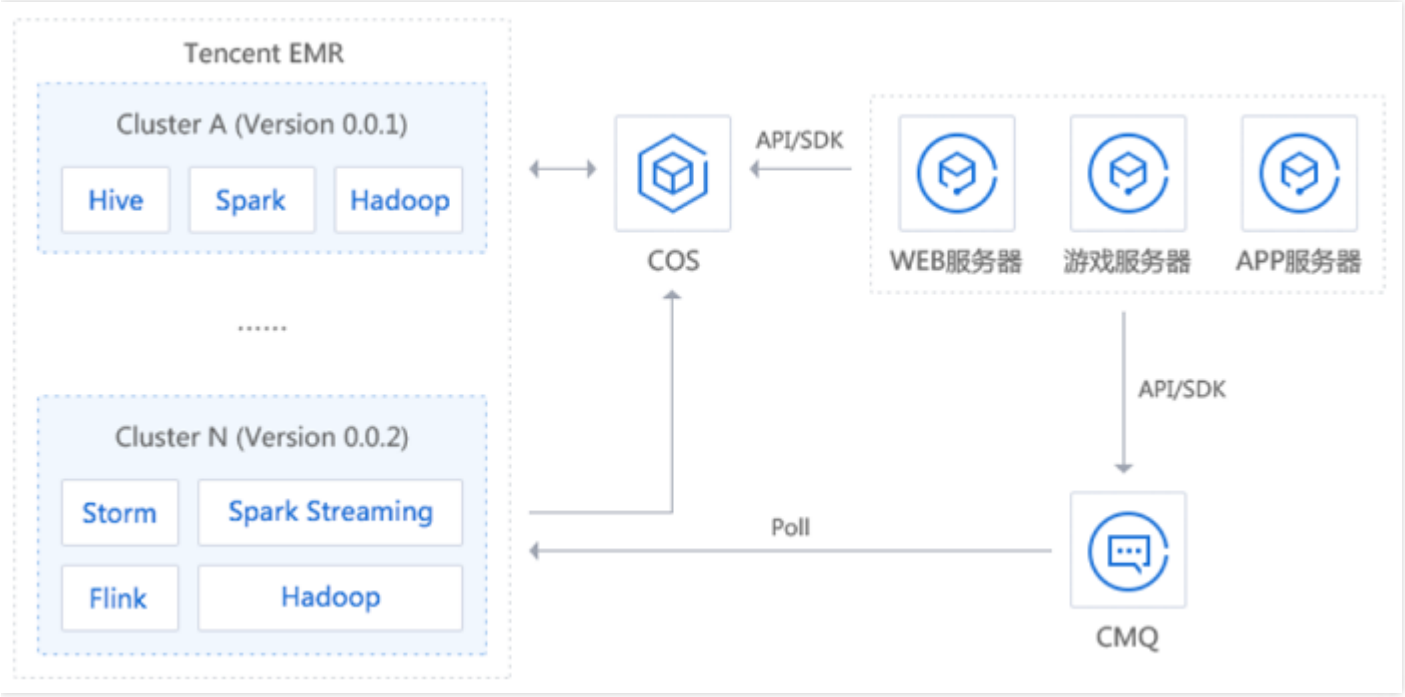
在程序/工具中通过 API、SDK 把位于业务服务器上实时产生的数据 Push 到 CMQ 消息中间件之后，可在 EMR 产品中选择合适的流式数据处理引擎来分析数据，以实现对业务变动的实时告警；还可以把分析结果实时同步到

TencentDB 等存储引擎，以便于通过 RayData 等数据可视化产品对业务状态进行实时可视化检测。

分析 COS 数据

可通过 EMR 产品快速分析存储于 COS 上的海量数据，以实现彻底的存储计算分离。通过这样的设计，可充分利用 COS 提供的丰富数据同步工具，同时还可以让多个不同版本 Hadoop 集群分析同一份数据，以满足数据一致

性及历史原因导致的多版本 Hadoop 集群共存的问题。



节点类型说明

最近更新时间：2020-11-20 15:56:25

EMR 定义了5种节点类型，您可以根据集群类型进行选择：

Hadoop 集群

节点类型	说明	HA（高可用）数量	非 HA 数量
主节点（Master）	部署 NameNode、ResourceManager、HMaster 等进程。	2	1
核心节点（Core）	部署 DataNode、NodeManager、RegionServer 等进程。	≥ 3	≥ 2
计算节点（Task）	部署 NodeManager、PrestoWork 等进程。	可随时更改 Task 节点数，实现集群弹性伸缩，最小值为0。	
通用节点（Common）	部署分布式协调器组件，如 ZooKeeper、JournalNode 等节点。	≥ 3	0
路由节点（Router）	部署 Hadoop 软件包，可选择部署 Hive、Hue、Spark 等软件和进程。	可随时更改 Router 节点数，最小值为0。	

- Master 节点为管理节点，保证集群的调度正常进行。
- Core 节点为计算及存储节点，您在 HDFS 中的数据全部存储于 Core 节点中，因此为了保证数据安全，扩容 Core 节点后不允许缩容。
- Task 节点为纯计算节点，不存储数据，被计算的数据来自 Core 节点及 COS 中，因此 Task 节点往往被作为弹性节点，可随时扩容和缩容。
- Common 节点为 HA 集群 Master 节点提供数据共享同步以及高可用容错服务。
- Router 节点用以分担 Master 节点的负载或者作为集群的任务提交机，可以随时扩容和缩容。

ClickHouse 集群

节点类型	说明	HA（高可用）数量	非 HA 数量
核心节点（Core）	部署 ClickHouseServer 进程。	≥ 2	≥ 1

通用节点（Common）	部署分布式协调器组件 ZooKeeper 节点。	⩾ 3	0

- Core 节点为计算及存储节点。
- Common 节点为 HA 集群 Master 节点提供数据共享同步以及高可用容错服务。

Druid 集群

节点类型	说明	HA（高可用）数量	非 HA 数量
主节点（Master）	部署 NameNode、ResourceManager、Ooverlord、coordinator、ZKFailoverController、JobHistoryServer 等进程。	2	1
核心节点（Core）	部署 DataNode、NodeManager、middlemanager、historical 等进程。	⩾ 3	⩾ 2
计算节点（Task）	部署 NodeManager、middlemanager 等进程。	可随时更改 Task 节点数，实现集群弹性伸缩，最小值为0。	
通用节点（Common）	部署分布式协调器组件，如 ZooKeeper、JournalNode 等节点。	⩾ 3	0
路由节点（Router）	部署 Hadoop 软件包，可选择部署 broker 等软件和进程。	可随时更改 Router 节点数，最小值为0。	

- Master 节点为管理节点，保证集群的调度正常进行。
- Core 节点为计算及存储节点，您在 HDFS 中的数据全部存储于 Core 节点中，因此为了保证数据安全，扩容 Core 节点后不允许缩容。
- Task 节点为纯计算节点，不存储数据，被计算的数据来自 Core 节点及 COS 中，因此 Task 节点往往被作为弹性节点，可随时扩容和缩容。
- Common 节点为 HA 集群 Master 节点提供数据共享同步以及高可用容错服务。
- Router 节点用以分担 Master 节点的负载或者作为集群的任务提交机，可以随时扩容和缩容。

组件版本

最近更新时间：2020-11-26 09:35:27

腾讯云弹性 MapReduce 由一系列大数据生态的开源应用程序组成。每个弹性 MapReduce 的版本，包含了一组特定版本的开源程序。当您在创建集群时，可以选择对应的 EMR 版本，以满足您对其中包含的开源组件的版本需求。

弹性 MapReduce 采用 EMR-Va.b.c 格式的版本号，详细说明如下：

- a 代表当前版本支持的 Hadoop 版本，a 等于1或2为支持 Hadoop 为2.X版本。
- b 代表版本中新增组件或支持组件版本升级。
- c 代表功能优化。

> ⚠ 注意：
>
> - 每一个版本上捆绑的组件和组件的版本都是固定的。目前还不支持组件的多个不同版本的选择，也不支持用户自行更改组件的版本。例如在 EMR-V2.0.1 版本中内置的是 Hadoop 2.7.3、Spark 2.2.1 等。
> - 一旦选择了 EMR 某个版本创建集群，该集群使用的 EMR 版本和组件版本不会自动升级，例如选 EMR-V2.0.1 版本，那么 Hadoop 就一直保持在2.7.3版本，Spark 就一直保持在2.2.1版本。后续如果版本升级到了 EMR-V2.1.0 版本，Hadoop 到了2.8.4版本，Spark 到了2.3.2版本，也不会影响已创建的集群。只有新的集群才会使用新的镜像。
> - 当您通过数据迁移的方式升级集群版本时（例如，从 EMR-V2.0.1 版本升级到 EMR-V2.1.0 版本），为防止一些升级不兼容、环境变化等问题的出现，请务必测试需要迁移的任务，以确保在新的软件环境中可以正常运行。
> - EMR-V2.4.0 版本安装 kona（基于 OpenJDK8），基于云场景的支撑及特性，我们在 kona 进行了开发及优化，kona 详情可参考 腾讯 Kona。

已支持 Hadoop 2.X 版本的 EMR 版本如下：

组件名称	EMR-V 1.3.1	EMR-V 2.0.1	EMR-V 2.1.0	EMR-V 2.2.0	EMR-V 2.3.0	EMR-V 2.4.0	EMR-V 2.5.0
发布时间	-	-	2019.05	2020.03	2020.05	2020.08	2020.09
Hadoop	2.7.3	2.7.3	2.8.4	2.8.5	2.8.5	2.8.5	2.8.5
Spark	2.0.2	2.2.1	2.3.2	2.4.3	2.4.3	3.0.0	3.0.0
Hive	2.1.1	2.3.2	2.3.3	2.3.5	2.3.5	2.3.7	2.3.7

组件名称	EMR-V 1.3.1	EMR-V 2.0.1	EMR-V 2.1.0	EMR-V 2.2.0	EMR-V 2.3.0	EMR-V 2.4.0	EMR-V 2.5.0
Tez	0.8.5	0.8.5	0.8.5	0.9.2	0.9.2	0.9.2	0.9.2
Presto	0.161	0.188	0.215	0.228	0.228	332	332
Storm	1.1.0	1.1.0	1.1.0	1.2.3	1.2.3	1.2.3	1.2.3
Flink	1.2.0	1.2.0	1.4.2	1.9.2	1.9.2	1.10.0	1.10.0
Hbase	1.2.4	1.3.1	1.3.1	1.4.9	1.4.9	1.4.9	1.4.9
Phoenix	4.8.1	4.11.0	4.13.0	4.13.0	4.13.0	4.13.0	4.13.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2
Hue	3.12.0	3.12.0	4.4.0	4.6.0	4.6.0	4.6.0	4.6.0
Sqoop	1.4.6	1.4.6	1.4.7	1.4.7	1.4.7	1.4.7	1.4.7
Ooize	4.3.1	4.3.1	4.3.1	5.1.0	5.1.0	5.1.0	5.1.0
Ranger	-	0.7.1	0.7.1	1.2.0	1.2.0	1.2.0	1.2.0
Zookeeper	3.4.9	3.4.9	3.4.9	3.5.5	3.5.5	3.6.1	3.6.1
Flume	-	-	1.8.0	1.9.0	1.9.0	1.9.0	1.9.0
Impala	-	-	-	2.10.0	2.10.0	2.10.0	2.10.0
Kylin	-	-	-	2.5.2	2.5.2	2.5.2	2.5.2
Zeppelin	-	-	-	0.8.2	0.8.2	0.8.2	0.8.2
Alluxio	-	-	1.8.1	1.8.1	1.8.1	1.8.1	2.3.0
Knox	1.2.0	1.2.0	1.2.0	1.2.0	1.2.0	1.2.0	1.2.0
Kerberos	-	-	1.15.0	1.15.0	1.15.0	1.15.0	1.15.0
Hudi	-	-	-	0.5.1	0.5.1	0.5.1	0.5.1
Superset	-	-	-	0.35.2	0.35.2	0.35.2	0.35.2
Livy	-	-	-	0.7.0	0.7.0	0.7.0	0.7.0
TensorFlow	-	-	-	-	1.4.4	1.4.4	1.4.4

组件名称	EMR-V 1.3.1	EMR-V 2.0.1	EMR-V 2.1.0	EMR-V 2.2.0	EMR-V 2.3.0	EMR-V 2.4.0	EMR-V 2.5.0
Jupyter	-	-	-	-	4.6.3	4.6.3	4.6.3

已支持 Hadoop 3.X 的 EMR 版本如下：

组件名称	EMR-V 3.0.0
发布时间	2019.11
Hadoop	3.1.2
Spark	2.4.3
Hive	3.1.1
Tez	0.9.2
Presto	0.222
Flink	1.8.1
Hbase	2.2.0
Hue	4.4
Sqoop	1.4.7
Ooize	5.1.0
Ranger	2.0.0
Zookeeper	3.4.9
Flume	1.9.0
Alluxio	1.8.1
Knox	1.2.0

DRUID 集群已支持组件产品版本如下：

组件名称	DRUID-V 1.0.0
发布时间	2020.04

组件名称	DRUID-V 1.0.0
Hadoop	2.8.5
Druid	0.17.0
Zookeeper	3.5.5
Knox	1.2.0
Superset	0.35.2
Ganglia	3.7.2

CLICKHOUSE 集群已支持组件产品版本如下：

组件名称	CLICKHOUSE-V 1.0.0	CLICKHOUSE-V 1.1.0
发布时间	2020.04	2020.05
Clickhouse	19.16.12.49	20.3.10.75
Zookeeper	3.4.9	3.4.9
Superset	-	0.35.2