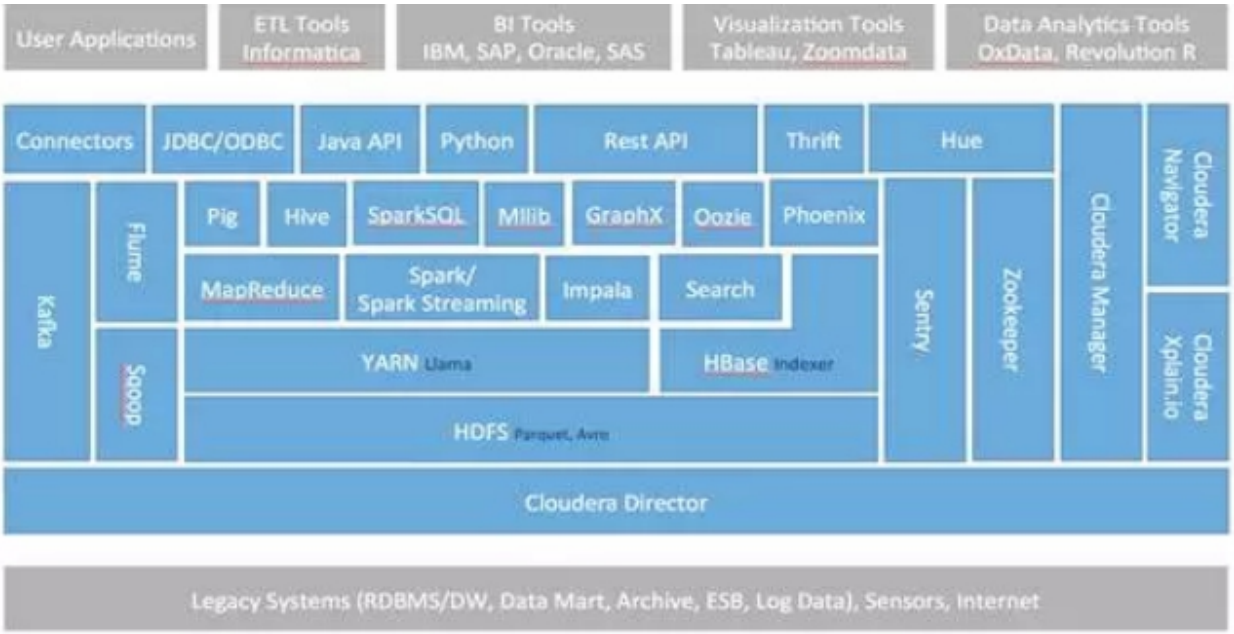
Cloudera平台软件体系结构



Cloudera的软件体系结构中包含了以下模块：系统部署和管理，数据存储，资源管理，处理引擎，安全，数据管理，工具库以及访问接口。一些关键组件的角色信息：

模块	组件	管理角色	工作角色
系统部署和管理	Cloudera Manager	Cloudera Manager Server	Cloudera Manager Agent
		Host Monitor	
		Service Monitor	
		Reports Manager	
		Alert Publisher	
		Event Server	
	Cloudera Director		
	HDFS	NameNode	DataNode
数据存储		Secondary NameNode	
		JournalNode	
		FailoberController	
	HBase	HBase Master	RegionServer

资源管理	YARN	ResourceManager	NodeManager
		JobHistory Server	
处理引擎	Spark	History Server	
	Impala	Impala Catalog Server	Impala Daemon
		Impala StateStore	
	Search		Solr Server
安全、数据管理	Sentry	Sentry Server	
	Cloudera Navigator	Navigator Metadata Server	
		Navigator Audit Server	
		Navigator KeyTrustee	
工具库	Hive	Hive Metastore	
		Hive Server2	

硬件配置

集群服务器按照节点承担的任务分为管理节点和工作节点。管理节点上一般部署各组件的管理角色，工作节点一般部署有各角色的存储、容器或计算角色。根据业务类型不同，集群具体配置也有所区别：

1. 实时流处理服务集群：Hadoop实时流处理性能对节点内存和CPU有较高要求，基于Spark Streaming的流处理消息吞吐量可随着节点数量增加而线性增长。

	管理节点	工作节点
处理器	两路 Intel®至强处理器，可选用 E5-2630 处理器	两路 Intel®至强处理器，可选用 E5-2660 处理器
内核数	6 核/CPU（或者可选用 8 核/CPU），主频 2.3GHz 或以上	6 核/CPU（或者可选用 8 核/CPU），主频 2.0GHz 或以上
内存	128GB ECC DDR3	128GB ECC DDR3
硬盘	2 个 2TB 的 SAS 硬盘（3.5 寸），7200RPM， RAID1	4-12 个 4TB 的 SAS 硬盘（3.5 寸），7200RPM，不使用 RAID
网络	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。
硬件尺寸	1U 或 2U	1U 或 2U
接入交换机	48 口千兆交换机，要求全千兆，可堆叠	
聚合交换机（可选）	4 口 SFP+万兆光纤核心交换机，一般用于 50 节点以上大规模集群	

2. 在线分析业务集群：在线分析业务一般基于Impala等MPP SQL引擎，复杂的SQL计算对内存容量有较高要求，因此需要配置128G甚至更多的内存。

	管理节点	工作节点
处理器	两路 Intel®至强处理器，可选用 E5-2630 处理器	两路 Intel®至强处理器，可选用 E5-2650 处理器
内核数	6 核/CPU（或者可选用 8 核/CPU），主频 2.3GHz 或以上	6 核/CPU（或者可选用 8 核/CPU），主频 2.0GHz 或以上
内存	128GB ECC DDR3	128GB -256GB ECC DDR3
硬盘	2 个 2TB 的 SAS 硬盘（3.5 寸），7200RPM， RAID1	12 个 4TB 的 SAS 硬盘（3.5 寸），7200RPM，不使用 RAID
网络	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。
硬件尺寸	1U 或 2U	2U
接入交换机	48 口千兆交换机，要求全千兆，可堆叠	
聚合交换机（可选）	4 口 SFP+万兆光纤核心交换机，一般用于 50 节点以上大规模集群	

3. 云存储业务集群：云存储业务主要面向海量数据和文件的存储和计算，强调单节点存储容量和成本，因此配置相对廉价的SATA硬盘，满足成本和容量需求

	管理节点	工作节点
处理器	两路 Intel®至强处理器，可选用 E5-2630 处理器	两路 Intel®至强处理器，可选用 E5-2660 处理器
内核数	6 核/CPU（或者可选用 8 核/CPU），主频 2.3GHz 或以上	6 核/CPU（或者可选用 8 核/CPU），主频 2.0GHz 或以上
内存	128GB ECC DDR3	48GB ECC DDR3
硬盘	2 个 2TB 的 SAS 硬盘（3.5 寸），7200RPM， RAID1	12-16 个 6TB 的 SATA 硬盘（3.5 寸），7200RPM，不使用 RAID
网络	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。	至少两个 1GbE 以太网电口，推荐使用光口提高性能。 可以两个网口链路聚合提供更高带宽。
硬件尺寸	1U 或 2U	2U 或 3U
接入交换机	48 口千兆交换机，要求全千兆，可堆叠	
聚合交换机（可选）	4 口 SFP+万兆光纤核心交换机，一般用于 50 节点以上大规模集群	

角色分配

小规模集群

搭建小规模集群一般是为了支撑专有业务，受限于集群的存储和处理能力，不太适合用于多业务的环境。这可以部署成一个HBase的集群；也可以是一个分析集群，包含YARN，Impala。在小规模集群中，为了最大化利用集群的存储和处理能力，节点的复用程度往往也比较高。下图是一个典型的小规模集群部署方式：

HDFS NameNode
HDFS FailoverController
HDFS JournalNode
Zookeeper
YARN ResourceManager
HBase Master
Impala StateStore

对于那些需要两个以上节点来支持HA功能的，集群中分配有一个工具节点可以承载这些角色，并同时可以部署一些其他工具角色，这些工具角色本身消耗不了多少资源：

HDFS JournalNode
Zookeeper
HBase Master
Cloudera Manager Server
Cloudera Management Service
History Server
Job History Server
Hive Metastore
HiveServer2
Impala Catalog Server
Hue Server
Oozie Server
Gateway

其余节点可以部署为纯工作节点，包含：

HDFS DataNode
YARN NodeManager
Impala Daemon
HBase RegionServer

中等规模集群

一个中等规模的集群，集群的节点数一般在20到200左右，通常的数据存储可以规划到几百TB，适用于一个中型企业的数据平台，或者大型企业的业务部门数据平台。节点的复用程度可以降低，可以按照管理节点、主节点、工具节点和工作节点来划分。

*在YARN模式下不需要部署Spark的Master或者Worker节点

管理节点上就安装Cloudera Manager、Cloudera Management Service。

主节点上安装有个CDH服务的管理节点以及HA的组件，可以如下方式部署：

服务	主节点 1	主节点 2	主节点 3
HDFS	NameNode	NameNode	
	FailoverController	FailoverController	
	JournalNode	JournalNode	JournalNode
YARN	Resource Manager	Resource Manager	
			Job History Server
Zookeeper	Zookeeper Server	Zookeeper Server	Zookeeper Server
HBase	HBase Master	HBase Master	HBase Master
Impala	Impala StateStore	Impala Category Server	
Hive			Hive Metastore
Spark			History Server

工具节点可以部署以下一些角色：

HiveServer2
Hue Server
Oozie Server
Flume Agent
Sqoop Client
Gateway

工作节点的部署和小规模类似：

HDFS DataNode
YARN NodeManager
Impala Daemon
HBase RegionServer

大规模集群

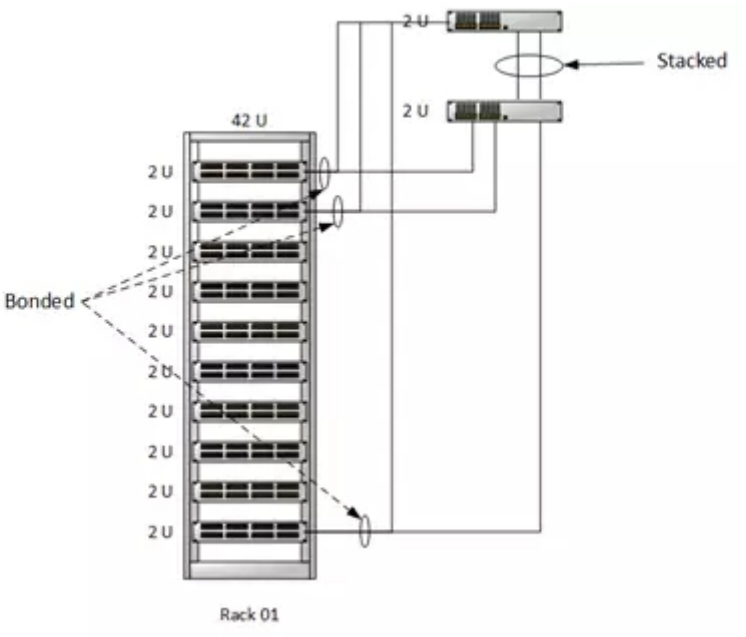
大规模集群的数量一般会在200以上，存储容量可以是大几百的TB甚至是PB级别，适用于大型企业搭建全公司的数据平台。和中等规模的集群相比，部署的方案相差不大，主要是一些主节点可用性的增强。

HDFS JournalNode由3个增加到5个，Zookeeper Server和HBase Master也由3个增加到5个，Hive Metastore的数量有1个增加到3个。

网络拓扑

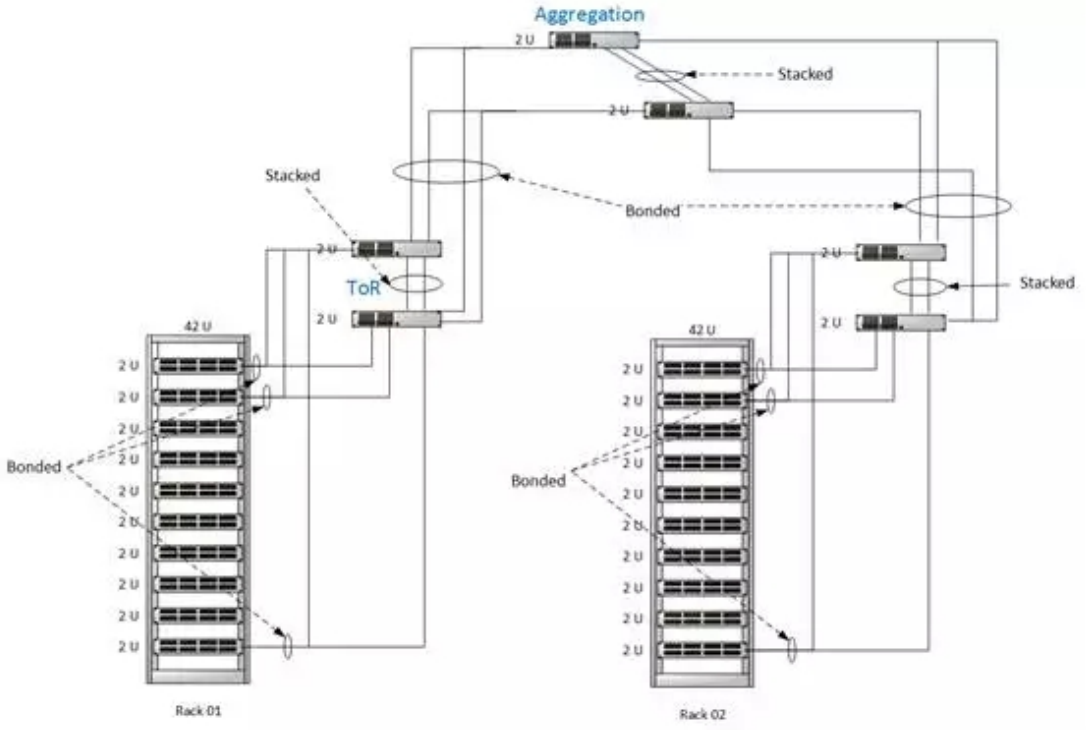
单机架部署

对于一个小规模的集群，或者一个单个rack的集群，所有的节点都连接到相同的接入层交换机。接入层交换机配置为堆叠的方式，互为冗余并增加了交换机吞吐。所有的节点两个网卡配置为主备或者负载均衡模式，分别连入两个交换机。在这种部署模式下，接入层交换机也充当了聚合层的角色。



多机架部署

在多机架的部署模式下，除了接入层交换机，还需要聚合层交换机，用于连接各接入层交换机，负责跨rack的数据存取。



实际部署样例

在机架上分配角色时，为了避免接入层交换机的故障导致集群的不可用，需要将一些高可用的角色部署到不同的接入层交换机之下（注是不同的接入层之下，而不是不同的物理rack下，很多时候，

客户会将不同物理rack下的机器接入到相同的接入层交换机下。）以下是一个80个节点的物理部署样例。