EsgynDB 数据库规划文档 2.8.0

2021/03

版权

公告

本文档包含的信息如有更改，恕不另行通知。

易鲸捷产品和服务附带的正式担保声明中规定的担保是该产品和服务享有的唯一担保。本文中的任何信息均不构成额外的保修条款。

声明

Microsoft® 和 Windows® 是美国微软公司的注册商标。Java® 和 MySQL® 是 Oracle 及其子公司的注册商标。Bosun 是 Stack Exchange 的商标。Apache®、Hadoop®、HBase®、Hive®、openTSDB®、Sqoop® 和 Trafodion® 是 Apache 软件基金会的商标。Esgyn，EsgynDB 和 QianBase 是易鲸捷的商标。

目录

关于本文档

本指南为搭建 EsgynDB 系统以及系统规模规划提供建议。

目标读者

本指南适用于 EsgynDB 系统规划人员。

版本

版本	日期	描述
2.8.0	2021/03	更新 3.参考架构指南中的图片
2.7.0	2020/04	第一版

相关文档

本指南为 EsgynDB 文档库的一部分，EsgynDB 文档库**包括但不限于**以下文档:

文档名称	说明
EsgynDB 安装部署指南	本文介绍安装 EsgynDB，包括安装前准备、安装 Hadoop 发行版、故障排除、配置、启用安全功能、提高安全性和卸载 EsgynDB 等。
EsgynDB 命令行工具指南	本指南介绍了如何在客户端工作站上使用

	EsgynDB 命令界面（trafci）连接和查询 EsgynDB。trafci 使您可以交互地或从脚本文件运行 SQL 语句。
EsgynDB 管理指南	资料库（Repository）实例是一个数据仓库，用于从 EsgynDB 实例收集可管理数据。
易鲸捷加载转换指南	本指南介绍如何将数据加载转换到易鲸捷数据库。
易鲸捷 Designer 用户指南	本文介绍易鲸捷图形化数据库管理工具
易鲸捷迁移工具用户指南	本文介绍如何安装和使用易鲸捷迁移工具。
易鲸捷 DTM 技术白皮书	本文介绍 EsgynDB 技术架构，组件介绍，技术特点等。
EsgynDB 数据库规划文档	本文介绍节点数量规划、数据目录和安装部署目录规划、集群角色分配规划等。
EsgynDB 常见问题提排查与解决	本文介绍如何排查和解决 EsgynDB 的常见问题。
EsgynDB 灾难恢复手册	本文介绍 EsgynDB 灾难恢复设计原理，方案建议以及使用手册。
EsgynDB 备份恢复手册	本文介绍 EsgynDB 备份恢复设计原理，方案建议以及使用手册。
EsgynDB 数据库扩容指南	本文介绍 EsgynDB 如何更换节点，增加节点，

	删除节点等操作。
EsgynDB 客户端安装手册	本文介绍 EsgynDB JDBC， ODBC 以及 Trafci 驱动安装。
EsgynDB JDBC 程序员参考指南	本文介绍 EsgynDB JDBC 驱动连接设置，开发人员指南。
EsgynDB ODBC 程序员参考指南	本文介绍 EsgynDB ODBC 驱动连接设置，开发人员指南。
EsgynDB SPSQL 存储过程用户手册	本文介绍 EsgynDB SPSQL 存储过程的使用。
EsgynDB Manager 用户手册	本文介绍图形化数据库监控运维工具 DB Manager 的使用。
EsgynDB 数据库迁移指南	本文介绍如何将常见关系型数据库（Oracle、MySQL、SQL server 等）迁移至 EsgynDB。
EsgynDB LOB 大对象用户指南	本文介绍如何使用 EsgynDB 大对象。
EsgynDB SQL 用户手册	本文是 EsgynDB 的 SQL 使用手册。

批评与建议

我们支持您对本指南做出的任何批评与建议，并尽力提供符合您需求的文档。

若您发现任何错误、或有任何改进建议，请发邮件至 support@esgyn.cn。

1. 概览

EsgynDB 引擎使用 HBase 提供存储服务。因此，它依赖于 HBase 的配置和调整来获得最佳性能。EsgynDB 的集群配置必须考虑到 HBase 的配置问题。

1.1. 硬件架构

EsgynDB 集群有两种类型的机器：主节点和从节点。

- 主节点 -- HDFS NameNode、 ZooKeeper 和 HBase Master.
- 从节点-- HDFS DataNodes、 HBase RegionServers、Monitor、 Distributed Transaction Manager、 Resource Manager、 DCS Master、 DCS Server、Master Executor 等

DataNodes、NodeManagers 和 HBase RegionServers 位于同一位置或共同部署，以实现最佳数据本地性。

易鲸捷公司建议分离主节点和从节点，因为：

- 从节点的任务/应用程序业务场景应与主节点隔离。
- 从节点经常因维护而被停用。

有三台或以上机器的集群通常使用单个的 NameNode 和 ResourceManager，所有其他节点都作为从节点。高可用性集群（HA）会使用一个主要的 NameNode 和备份的 NameNode。

通常，中型到大型集群由两级或三级架构组成，该架构由机架式服务器构建。每个服务器机架都使用一个 10 千兆位以太网（GbE）交换机进行互连。每个机

架式交换机都连接到一个集群交换机（通常是一个端口密度为 40GbE 的较大交换机)。这些集群交换机还可以与其他集群交换机互连，甚至可以向上传输到另一级交换架构。

1.2. 软件架构

EsgynDB 使用 HDFS 作为底层存储，当节点失效时，EsgynDB 会使用适当的复制因子（通常设置为 3）确保可用性。

查询处理的重要进程包括:

进程名称	描述	分布	数目
DCS Master	用于分配 mxosrvr 的初始连接点	在单个节点上	每个集群一个活跃的 DCS Master，通常配置一个浮动 IP 以获得高可用性。
DCS Server	管理 mxosrvr 进程的状态和连接使用的进程	在每个节点上	每个运行 mxosrvrs 的节点对应一个。
Master executor (mxosrvr)	主执行器进程，它承载 SQL 会话，执行根操作符的查询编译和执行	每个数据节点上有多个	可根据业务需要自定义每个节点的最大并发会话数。
Executor Server	SQL 计划的并行	运行在集群中的所	业务场景相关：由并

Process (ESP)	分段执行	有数据节点上，是可变大小的组	发并行查询、查询计划和并行度决定。
DTM	维护事务状态和日志结果信息	在实例中的所有数据节点上运行	每个数据节点上一个。

EsgynDB Manager 进程包括:

进程名称	**描述**	**分布**	**数目**
DB Manager	浏览器连接到的 Web 应用程序服务器	单个或多个节点上	和 DCS Master 个数相等， 和 DCS Master 运行在相同节点上。
OpenTSDB	用于收集时间序列度量的轻量级服务进程	在每个节点上	每个节点上一个
TCollectors	按时间间隔收集基于时间的度量的收集脚本	实例中所有数据节点上有多个;每个节点上进程数有所不同	在每个节点上收集系统和 HBase 指标 EsgynDB 指标从第一个数据节点上的进程收集集群范围的数据
REST Server	处理来自集群内外客户端的 REST 请求的进程	单个或者多个节点上。	和 DCS Master 个数相等， 和 DCS Master 运行在相同节点上。

除了列出的用于查询处理和管理性的进程之外，作为 EsgynDB 软件栈一部分，还有其他一些进程支持运行时执行环境。这些进程通常使用较少的资源，并且对系统规模和调配的实质性影响很小。

HBase 进程可以分为两类：控制进程和数据进程。控制进程是管理存储子系统及其元数据的一次性进程。数据进程是指服务于数据本身的进程，包括读取、更新和写入。

ZooKeeper：

ZooKeeper	HBase 服务依赖此服务，用于跨节点的信息管理和协调。

HBase 控制进程包括：

进程	**描述**
HMaster	元数据和表的创建/删除

HBase 数据进程包括：

进程	**描述**
RegionServer	控制数据服务，包括服务获取/输出以及将数据分离到各个区域 （region）。

HBase 使用 HDFS 服务在集群中实现可扩展性、可用性和恢复（复制）。因此，EsgynDB 集群配置结合了 HDFS 配置考虑因素，包括复制。控制进程是管理 HDFS 文件系统的单一进程。在 HDFS 中，它们控制单个数据块的位置。数据处理涉及到数据的读取和访问。

HDFS 控制进程包括：

进程	描述
NameNode	管理用于将块映射到单个文件并选择复制位置的元数据文件。如果是HA模式,存在主NameNode和备份的NameNode,提供服务的高可用性。
备份 NameNode	每个时间间隔（默认为小时）内，从 NameNode 获取一次所有元数据的检查点。如果 NameNode 丢失，可以使用此数据重新创建块->文件映射。但是，它不仅仅是 NameNode 的热备份。如果是 HA 模式，由备份 NameNode 进程代替此进程。

HDFS 数据进程包括:

进程	描述
DataNode	提供对单个文件的读写,并定期发送 I'm-alive 信息包括它管理的文件/块到 NameNode。

除了上面列出的 HBase 和 HDFS 控制进程外，其他控制节点进程包括:

进程	描述
Management Server Process	Cloudera Manager 等网页节点。一些管理服务器做详细的数据库和分析功能。

在较小的集群中，控制进程和数据进程可能驻留在同一个节点上。对于较大的集群，管理进程具有明显不同的配置需求，因此常常在不同的节点上进行隔离。我们的参考架构假设有独立的主节点和从节点。

2. 硬件建议

2.1. 典型的工作负载模式

磁盘空间、I/O 带宽、内存和计算能力是精确规划硬件规模最重要的参数。

2.2. 服务器节点硬件建议

必须将服务器节点硬件建议用作选择节点数、每个节点的存储选项（磁盘数、磁盘大小、平均故障间隔时间和磁盘故障的复制成本)、每个节点的计算能力（套插槽、核数、晶振频率)、每个节点的内存和网络容量（端口数量、速度）的最佳实践。

备注:

EsgynDB 集群节点不需要企业数据中心服务器中的许多常见功能。

2.3. 从节点硬件

在为 EsgynDB 集群中的从属节点部署硬件时，必须考虑服务器平台、存储选项、内存大小、内存配置、处理能力、功耗和网络等因素。

易鲸捷公司建议:

双插槽服务器通常是银行核心应用程序的最佳选择。与入门级服务器相比，使用这些服务器是更好的选择，因为它们具有负载平衡和并行处理功能。在密度方面，选择适用于少量机架单元的服务器硬件。通常情况下，1U 或 2U 服务器用于 19 英寸机架或机柜中。

2.3.1. 处理能力

在为 EsgynDB 集群处理能力规划时，请考虑以下因素:

- 在典型的高性能配置中，管理节点与数据节点分开配置。对于存储（大小、配置）以及网络和内存，这两种类型的配置通常是不同的。
- 在非常小或测试配置中，管理节点和数据节点之间的区别是模糊的，而且大多数管理进程都与数据处理共存。特别是对于基本开发和测试集群来说，只要这个配置满足性能和可用性目标，它就是有效的。

评估所需节点数量时，请考虑以下因素：

- 对于典型的生产业务场景来说，只要每个节点的核数是主流的（例如 16 核或更多），具有较少核数而有更多节点的配置比分布在较少节点上的同等核数的配置具有优势。横向扩展（增加节点数以获得所需的核数）比纵向扩展（增加每个节点的核数以获得所需的核数）更为可取，因为：
  - 核数少节点多通常比核数多节点少的配置便宜
  - 当丢失具有更多节点的集群上的节点或磁盘时，故障域较小。
  - 可用的 I/O 带宽和并行性随着节点的增多而提高。
- 考虑到核心银行的可用性和可恢复性要求,不建议使用小于 3 个节点的集群。
- 同时使用的用户数（并发用户）决定公司外部网络连接的节点数量，数据到达/刷新的接收率也是如此。该数字决定了 mxosrvr 进程的总数。基于 mxosrvr 进程分布，实际连接分布在集群上。多个 mxosrvr 进程可以在同一节点上运行。
- 业务场景类型是节点数量的另一个关键考虑因素。节点和核数反映了集群上运行的应用程序并发用户可用的并行数量。如果典型的业务是高并发短查询，那么可接受低配置的节点。如果典型的业务场景涉及大规模扫描，那么就需要更强的处理能力。最好是尽可能通过原型化业务场景和查询来了解应用程

序的类型、频率、计划和典型并发数量。

2.3.2. 存储选项

为 EsgynDB 集群磁盘能力规划时，请牢记以下注意事项：

- 对于数据节点，SSD 只对高并发写入有利。一般来说，HDD 就足够了。对于控制节点，SSD 同样性价比不高——其目标是在内存中缓存大多数控制信息。
- 对于数据节点，HDD 数据磁盘在 JBOD（只是简单一组磁盘）的设置中配置直接转载存储设备。RAID 分条降低了 HDFS 的速度，实际上降低了并发性和可恢复性。对于控制节点，可以将数据磁盘配置为 JBOD、RAID1 或 RAID10。
- 与处理能力一样，磁盘是一个并行单元。对于给定的每个节点的总磁盘数，如果业务场景包括许多大规模扫描，那么在数据节点上，每个节点拥有更多小容量磁盘通常比拥有较少大容量磁盘更为有效。系统参考架构假设大多数业务场景都包含大规模扫描。
- 强烈建议使用 HBase SNAPPY 或 GZ 压缩。SNAPPY 的 CPU 开销较少，但 GZ 的压缩效果更好。压缩程度随数据和业务场景模式的不同变化很大，但公认的计算数据表明，压缩率大约降低 30%-40%。压缩会增加读取和写入的路径长度，这可能会对数据的增长和加载产生影响。压缩发生在 HBase 文件块级别，限制了读取时所需的解压缩量。
- 在计算每个节点的总磁盘空间和数据磁盘空间时，务必考虑每个节点的工作空间和预期的数据加载/流出量。同样需要记住的是，HDFS 文件块带有一个复制因子（通常设置为 3，所以数据有 3 个副本）。这意味着每个 10GB 的文件实际上在磁盘上占用 30 GB。易鲸捷公司建议留出大约 33%的空磁盘空间

作为工作空间开销。

我们建议每台服务器使用相对多数量的硬盘驱动器（通常为 8 到 12 个 SATA LFF 驱动器）。目前，生产环境中的典型容量约为每个驱动器 2 TB。高 I/O 密集型环境可能需要使用 12 x 2 TB SATA 驱动器。成本和性能之间的最佳平衡通常是采用每分钟 7200 转的 SATA 驱动器。如果您当前或预测的存储增长显著，您还应该考虑使用 3 TB 的磁盘。

为了获得更好的磁盘带宽，某些配置会采用 SFF 磁盘。我们建议您监控自己的集群是否存在任何潜在的磁盘故障，因为更多的磁盘会增加磁盘故障率。如果每台服务器确实有大量的磁盘，我们建议您使用两个磁盘控制器，以便可以跨多个核分担 I/O 负载。我们强烈建议只使用 SATA 或 SAS 互连。

Hadoop 是一款存储密集型、寻求高效的产品，但不需要快速而昂贵的硬盘。如果您的业务模式并非 I/O 密集型，每个节点配置四个或六个磁盘是安全的。请注意，电力成本与磁盘数量成正比，而不是每个磁盘的存储容量。因此，我们建议您通过添加磁盘来增加存储空间，而不仅仅是寻求高效。

您的磁盘驱动器应该具有良好的 MTBF 能力，因为数据节点会受到常规概率故障的影响。

Hadoop 是为适应数据节点磁盘故障而设计的。可以在不将服务器从机架中取出的情况下换出磁盘，尽管将其关闭（短暂地）也是一种成本低廉的操作。

2.3.3. 内存大小

提供足够的内存以使处理器在不交换内存的情况下保持忙碌是至关重要的。根据内核数量，您的数据节点通常需要 256 到 512 GB 的内存来处理核心银行应用

程序。

在为 EsgynDB 集群的内存大小进行规划时，请牢记以下注意事项:

- 许多 Hadoop 生态系统进程都是 Java 进程。由于 JVM 的内存效率优化，一个重要的限制是在 32 GB 以下。实际上,超过这个阈值会导致可用内存减少，因为指针的内部表示方式发生了变化，从而消耗更多的内存空间。
- 数据节点的大内存消耗者包括:
  - HDFS DataNode 进程
  - HBase RegionServers

  在控制进程中，大内存消耗者有:
  - HDFS NameNode 进程

推荐这些进程使用 16-32 GB 大小的内存堆，以便在大规模集群上获得最佳性能。减少这些组件的内存会显著影响性能，因此在选择较小的内存值之前，请小心分析并调整。

- EsgynDB 引擎内存的主要使用者是 mxosrvr。对于节点上的每个并发连接，计划为每个节点的每个连接配置 128 MB (0.5 GB)。

为了检测和纠正由热力学效应和宇宙射线引起的随机瞬态误差,我们强烈建议使用纠错码（ECC）内存。纠错内存使您更加信任计算的质量。一些部件（芯片猎杀/芯片备件技术）已被证明能比传统设计提供更好的保护，因为它们可以达到更少的重复位错误。

如果您希望保留将来向服务器添加更多内存的可能,请确保在初始内存模块旁边有空间进行此操作。

2.3.4. 电源

鉴于应用程序的重要性，我们建议使用冗余电源设备（PSU）。

备注:

在现代数据中心，电力和冷却设备占设备总寿命周期成本的 33.33%至 50%。

2.3.5. 网络

在为 EsgynDB 集群的网络规模规划时，请牢记以下注意事项:

- 10GigE 是集群内数据通讯网络的标准。为数据流使用较慢的网络会显著影响性能。2 个 10GigE 绑定网卡为 I/O 密集型应用提供了更高的吞吐量。
- 在某些情况下，为集群维护配置了第二个较慢的网卡，以便将该网络流量与运营数据工作流分离。
- 考虑从不同机架连接节点时的故障场景。如果复制因子大于或等于 3，HDFS 块放置算法偏向于为块位置选择至少 2 个不同机架上的节点。
- 如果使用跨数据中心功能，两个数据中心之间必须有高速连接。
- 如果使用跨数据中心功能，则必须配置两个集群，以便应用程序可以在运行和访问两个集群时通过 EsgynDB 驱动程序主动连接到任何一个对等集群。此功能确保当应用程序与这两个集群中的一个失去通信时，能访问另一个集群。

一个好的网络设计应该考虑到在实际负载下，在网络的临界点出现不可接受的拥塞的可能性。通常可接受的过载率在服务器访问层约为 4:1，在访问层和聚合层或内核之间约为 2:1。如果需要更高的性能，可以考虑降低过载率。此外，我们还建议机架之间有 1 GigE 的过载。

为集群配置专用交换机，而不是试图在现有交换机中分配 VC，这一点至关重

要——EsgynDB 集群的负载将影响交换机的其他用户。同样重要的是，需要与网络团队合作，确保交换机同时适用于 EsgynDB 及其监控工具。

网络设计方案应保留增加 QianBse 服务器更多机架的可能性。修正错误网络方案代价是昂贵的。您不太可能复制交换机的引用带宽。在交换机中，“深度缓存”优于低延迟。跨集群启用巨型帧可以通过更好的校验和提高带宽，还可能提供包完整性。

备注：

EsgynDB 集群的网络策略：分析网络与计算机成本的比例。确保网络成本始终在总成本的 20%左右。网络成本应该包括您的整个网络、核心交换机、机架交换机、所需的任何网卡等等。

2.4. 管理节点硬件配置

管理节点是唯一的，与数据节点相比，具有显著不同的存储和内存需求。

我们建议使用双 NameNode 服务器，一个主服务器和一个辅助服务器。两个 NameNode 服务器的名称空间存储和编辑日志记录都应该具有高度可靠的存储。硬件 RAID 和/或可靠的网络存储通常是合理的选择。

主服务器应该至少有四个冗余存储卷，一些是本地卷，一些是网络卷，但每个卷都相对较小（通常为 1 TB）。

备注：

管理节点上的 RAID 磁盘应在出现故障时迅速更换。确保数据中心提供备用磁盘。

ResourceManager 服务器的存储选项

实际上，ResourceManager 服务器不需要 RAID 存储，因为它们将持久状态保存到 HDFS。事实上，ResourceManager 服务器可以在有一点额外剩余内存的从节点上运行。但是，对 ResourceManager 服务器使用与 NameNode 服务器相同的硬件规范，可以在 NameNode 故障时将 NameNode 迁移到与 ResourceManager 相同的服务器，并且可以将 NameNode 状态的副本保存到网络存储。

在分布式设置中，HBase 将其数据存储在 Hadoop DataNode 中。为了获得最大的读/写本地化，HBase RegionServer 和 DataNode 应该共同部署在同一台机器上。因此，所有关于 DataNode 和 NodeManager 硬件设置的建议也适用于 RegionServer。根据您的 HBase 应用程序是读/写还是面向数据处理，您必须平衡磁盘数量与可用 CPU 核心数量。通常情况下，每个磁盘至少应该对应有一个内核。

2.4.1. 内存

管理节点所需的内存量取决于 NameNode 要创建和跟踪的文件系统对象（文件和块副本）的数量。64 GB 的内存支持大约 1 亿个文件。

Hbase 使用不同类型的缓存来占用内存，通常情况下，HBase 拥有的内存越多，缓存读取请求的能力就越强。EsgynDB 集群（中的每个数据节点（RegionServer）维护多个 region（region 是内存中的数据块）。对于大型集群，确保 HBaseMaster 和 NameNode 在不同的服务器上运行是很重要的。

注意，在大规模部署中，ZooKeeper 节点不与 Hadoop/HBase 从节点共同部署在同一服务器上。

HBase Master 节点不像典型的 RegionServer 或 NameNode 服务器那样需要大量计算。因此，可以为 HBase Master 选择更适度的内存设置。RegionServer 内存

需求很大程度上取决于 HBase 集群的业务特性。虽然内存的过度配置有助于所有的业务模式，但由于内存堆非常庞大，Java 的 GC 暂停可能会导致出现问题。

2.4.2. 处理器

NameNodes 与客户端通讯繁忙。因此，我们建议提供 16 甚至 24 个 CPU 内核来处理主节点的消息传输。

2.4.3. 网络

交换机提供多个网络端口和 10 GB 带宽配置是可以接受的（如果交换机有能力处理）。

2.4.4. 其他问题

除了硬件方面的考虑外，还必须为您的 EsgynDB 集群考虑服务器机架的重量、集群的可扩展性等因素。

2.4.5. 重量

最新一代服务器的存储密度意味着需要考虑机架的重量。您应该确认机架的重量不超过数据中心基底的容量。

2.4.6. 可扩展性

通过在集群中添加新的服务器或整个服务器机架，或增加主节点中的内存可以很容易地扩展一个 EsgynDB 集群来满足增加的负载。这将在一开始产生大量“重新平衡流量”，但将提供更多的存储容量和计算能力。因为主节点至关重要，所

以我们建议您为这些机器支付更多费用。

使用以下准则扩展现有的 EsgynDB 集群:

- 确保集群所在数据中心有潜在的可用空间。该空间应该能够容纳更多机架的电源容量。
- 规划网络设备以应对更多服务器
- 如果服务器有备用插槽，则可向现有服务器添加更多的磁盘和内存以及更多的 CPU。这可以在不添加更多机架或更改网络的情况下扩展现有集群。
- 在运行集群中执行硬件升级需要花费大量时间和精力。我们建议您一次计划扩展一台服务器。
- 主服务器可能需要更多内存。

3. 参考架构指南

本节内容包含对裸金属 EsgynDB 集群的硬件配置和软件配置的建议。这些建议不依赖于硬件。与您的硬件供应商一起检查特定的配件编号和交货时间。

描述的配置适用于中型或大型 EsgynDB 的安装，具有单独的管理和数据节点。对于相同节点上包含所有进程的较小配置将在单独的章节中介绍。

对于数据节点，每个节点的基本硬件建议是:

资源	**建议**
CPU	Intel Xeon 或 AMD 64 位处理器 16≤每个节点的核数≤64
内存	整个生态系统和查询处理需要 256 GB 或者更高，加上通常的开销，

	再加上节点上每个 MxOSRVR 所需的 128 MB。 mxosrvr 进程数量计算公式: $\frac{\text{最大并发连接数}}{\text{节点数量}}$ 256 GB≤内存大小≤512 GB。最常见的值是 512 GB。
网络	10 GigE、1 GigE 或 2x10 GigE 绑定
存储器	SATA 或 SAS 或 SSD，通常在 JBOD 配置中配置 6-12 个 1 TB 磁盘

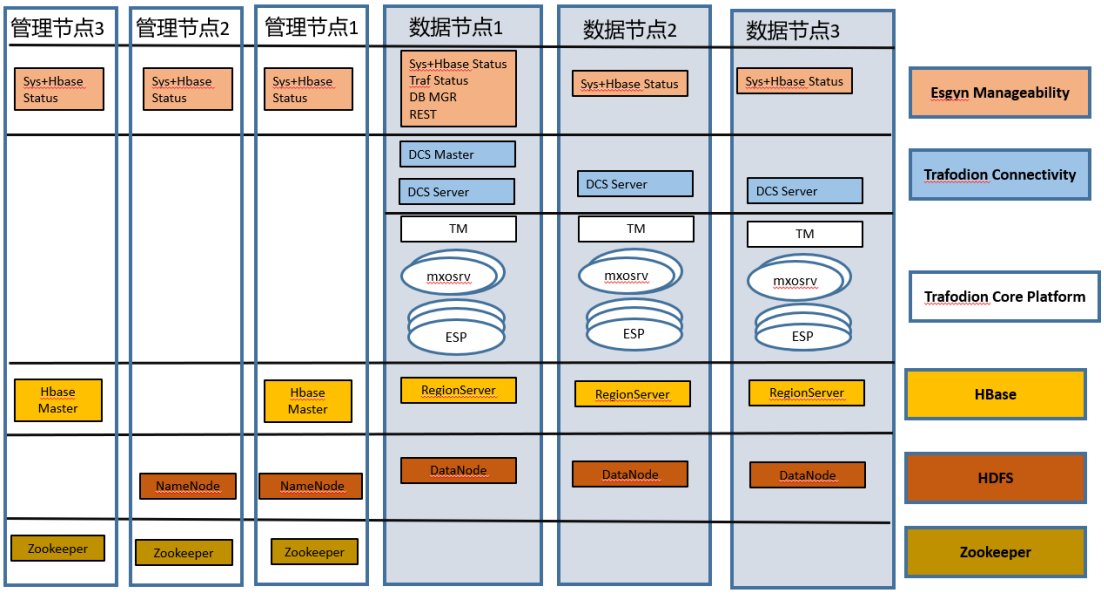
对于管理节点，每个节点的基本硬件建议是:

资源	**建议**
CPU	Intel XEON 或 AMD 64 位处理器 16≤每个节点的核数≤64 个
内存	整个生态系统需要 256 GB，再加上有可能或需要产生的用于交换和进程维护的日常费用。 128 GB≤内存大小≤256 GB。最常见的值是 256 GB。
网络	10 GigE、 1 GigE 或 2x10 GigE 绑定，加上用于连接到数据库或独立运行的合适的交换机
存储器	SATA 或 SAS 或 SSD，通常在 RAID1 或 RAID10 配置中配置 6-12 个 1 TB 磁盘

3.1. 中型/大型部署

中型或大型部署使用上述规范，包括管理节点和数据节点。节点中的进程如下

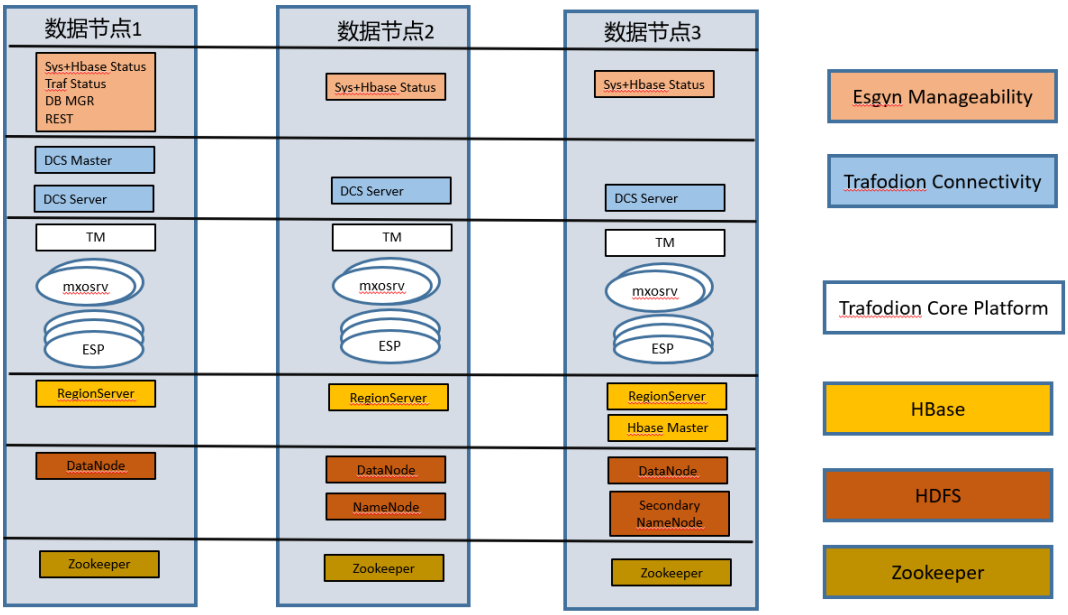
图：



在上图中，管理节点位于数据节点的侧面，仅用于 DCS 主进程。对于节点命名没有特定的约束，包括不假设节点是连续编号的。图中竖条代表单个节点，椭圆代表节点内的进程。

3.2. 小型部署

对于小型（4-6 个节点，通常少于一个机架）部署，管理节点被分解到常规节点架构中，如下图：



在上图中，管理节点已被删除，管理进程运行在与功能进程相同的节点上。

4. 软件配置

以下是在 EsgynDB 集群上安装软件的推荐配置。

4.1. 文件系统

以下是集群中所有节点的基本配置：

- 根分区：操作系统和核心程序文件
- 交换：2 倍系统内存

4.2. 数据节点

驱动器应按优先顺序使用 XFS 或 ext4。不要使用 LVM；它会增加延迟并导致瓶颈。

数据节点配置示例：

- /根 - 20GB（为现有文件、未来日志文件增长和操作系统升级提供充足的空间），如果/var 和/opt 不是单独的文件系统，/根建议 100G 以上。
- /磁盘 1 -本地存储驱动器
- /磁盘 2 -第二个存储驱动器
- /磁盘 3 - ...

4.3. 冗余 (RAID)

- 管理节点：配置可靠性（RAID 10、双以太网卡、双电源等）
- 数据节点：因为这些节点上的故障由集群自动管理，所以这里不需要 RAID。所有数据都存储在至少三个不同的主机上，因此冗余是内置的。数据节点应该根据速度构建。

4.4. NameNode 节点

为了确保运行中的 NameNode 主机发生故障时集群中的另一个 NameNode 始终可用，启用并配置 NameNode 高可用性。

启用 NameNode 高可用性后，至少配置 3 个有效的 JournalNode。

4.5. HBase

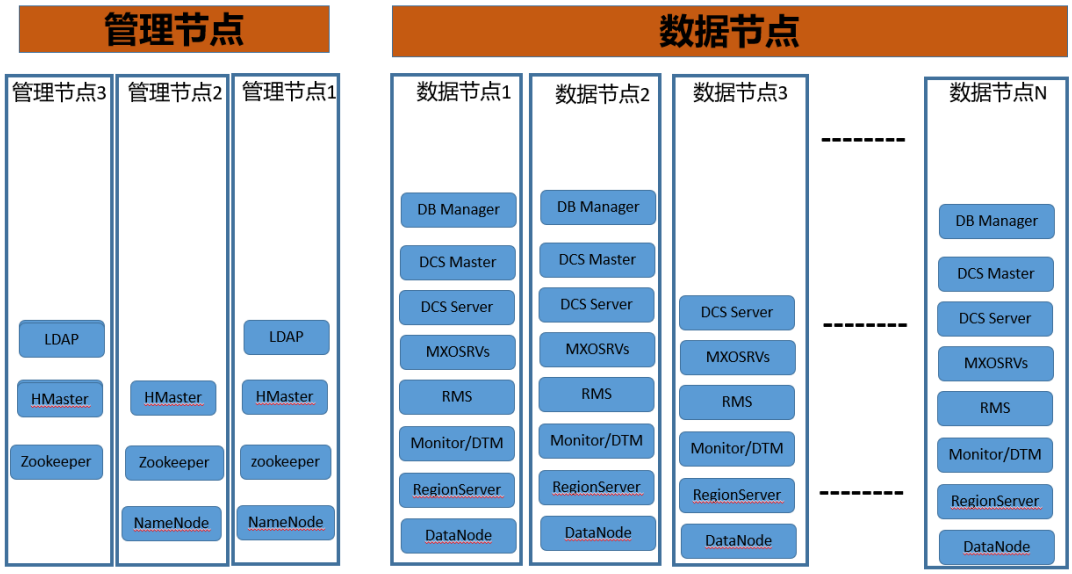
为了实现冗余，在集群中配置两个或多个 HBase Master。在运行两个以上的 HBase Master 集群中，HBase 使用 ZooKeeper 来协调运行中的 Master 。当运行中的 HBase Master 发生故障时，客户端将自动切换到备用 Master。

4.6. 连接子系统

在数据节点上配置三个或以上 DCS 主服务器，以在服务器进程或主机节点发生故障时提供冗余。通过 keepalived 和浮动 IP 结合，对于连接到 EsgynDB 的应用来说，DCS 主服务器进程的故障是透明的。

4.7. 数据库管理器

通过配置多个数据库管理器进程，与 DCS 主进程共存，以获得更高的可用性。数据库管理器也可以独立于数据库进行配置。



5. 目录结构

HDFS 有一个主/从架构。HDFS 集群由一个 NameNode 组成，NameNode 是一个主服务器，它管理文件系统名称空间并控制客户端对文件的访问。此外，集群中有许多 DataNodes，每个服务器对应一个 DataNode，它们管理这台服务器上所有的存储设备。HDFS 通过文件系统的名称空间来保存数据。在内部，一个文件被分成一个或多个块，这些块存储在不同的 DataNodes 上。

NameNode 维护文件系统名称空间，并执行诸如打开、关闭和重命名文件和目录等操作。它还确定块到 DataNodes 的映射，并记录对文件系统名称空间或其属性的任何更改。

DataNodes 服务来自文件系统客户端的读写请求。DataNodes 还根据来自 NameNode 的指令执行块创建、删除和复制。

Hadoop 配置文件默认位于/etc/hadoop/conf/hdfs-site.xml 中。

hdfs-site.xml 中的主要项

dfs.datanode.data.dir	确定 DFS 数据节点应在本地文件系统上存储块的位置 此属性的默认值是 file://${hadoop.tmp.dir}/dfs/data
dfs.replication	用于保护 HDFS 中数据的复制数据块的数量。 默认设置为 3
dfs.namenode.https-address	NameNode URL 的位置
dfs.https.port	用于访问 HDFS 的端口

日志文件位于/var/log/hadoop-hdfs

HBase 配置文件默认位于 /etc/hbase/conf/hbase-site.xml

hbase-site.xml 中的主要项目

hbase.rootdir	区域服务器共享的目录，HBase 将一直保存到该目录中。 此属性的默认值为 file://${hbase.tmp.dir}

日志文件在 in /var/log/hbase 中

5.1. 元数据和用户表

EsgynDB 使用名称空间的概念对表进行逻辑分组。一组前缀为 TRAF_RSRVD 的保留命名空间用于存放数据库元数据。

命名空间和表位于${hbase.rootdir}/data

```
[root@nap conf]# hdfs dfs -ls /hbase/data
Found 9 items
drwxr-xr-x   - hbase hbase          0 2019-07-01 23:41 /hbase/data/TRAF_RSRVD_1
drwxr-xr-x   - hbase hbase          0 2019-06-30 04:39 /hbase/data/TRAF_RSRVD_2
drwxr-xr-x   - hbase hbase          0 2019-07-01 23:56 /hbase/data/TRAF_RSRVD_3
drwxr-xr-x   - hbase hbase          0 2019-06-29 14:53 /hbase/data/TRAF_RSRVD_4
drwxr-xr-x   - hbase hbase          0 2019-06-17 23:56 /hbase/data/TRAF_RSRVD_5
drwxr-xr-x   - hbase hbase          0 2019-06-24 22:16 /hbase/data/TRAF_RSRVD_6
drwxr-xr-x   - hbase hbase          0 2019-06-17 23:42 /hbase/data/TRAF_RSRVD_7
drwxr-xr-x   - hbase hbase          0 2019-07-01 06:34 /hbase/data/default
drwxr-xr-x   - hbase hbase          0 2019-06-16 06:44 /hbase/data/hbase
```

用户表通常位于默认的命名空间中。但是，如果出于安全原因需要，可以在其他命名空间中创建它们。

```
[trafodion@nap esgyndb]$ hdfs dfs -ls /hbase/data/default
Found 8 items
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:00 /hbase/data/default/CUSTOMER
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:06 /hbase/data/default/LINEITEM
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:16 /hbase/data/default/NATION
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:17 /hbase/data/default/ORDERS
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:23 /hbase/data/default/PART
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:30 /hbase/data/default/PARTSUPP
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:34 /hbase/data/default/REGION
drwxr-xr-x   - hbase hbase       0 2019-07-01 06:34 /hbase/data/default/SUPPLIER
```

5.2. 二进制对象

TRAF_HOME 目录包含已安装的 EsgynDB 软件。

```
[trafodion@nap conf]$ echo $TRAF_HOME
/opt/trafodion/esgyndb

[trafodion@nap esgyndb]$ tree -L 1
```

```
.
├── conf
├── dbmgr-2.7.0
├── dcs-2.7.0
├── export
├── gdb
├── hbase_utilities
├── LICENSE
├── LICENSE_Esgyn
├── mgblty
├── NOTICE
├── opt
├── rest-2.7.0
├── samples
├── sqenvcom.sh
├── sqenv.sh
├── sql
├── sysinstall
├── tmp
├── tools
├── trafci
└── wms-2.7.0
```

5.3. 日志

TRAF_LOG 目录包含 EsgynDB 产生的日志。

```
[trafodion@nap esgyndb]$ echo $TRAF_LOG
/var/log/trafodion
```

5.4. 配置文件

TRAF_CONF 目录包含 EsgynDB 配置文件。

```
[trafodion@nap esgyndb]$ echo $TRAF_CONF
/etc/trafodion/conf
```

5.5. 其它

TRAF_VAR 目录包含 EsgynDB 所需的基本元数据和临时文件。

```
[trafodion@nap esgyndb]$ echo $TRAF_VAR
```

```
/var/lib/trafodion
```