EsgynDB 企业版平台参考架构 2.0

本文件概述了基于 Apache Trafodion™（孵化）实现的 EsgynDB Enterprise 平台参考架构，该平台提供软件许可证支持和扩展。本文件概述不同供应商的服务器配置，并试图描述如何确定 EsgynDB 应用程序规格时的一些注意事项。

1. 简介

Apache Trafodion（孵化）项目将完整的事务型 SQL 数据库整合到 Apache Hadoop™生态系统，以支持运营型任务流。EsgynDB Enterprise 2.0 基于 Apache Trafodion 提供全面支持，适合企业的版本和附加特性扩展，包括 EsgynDB Enterprise Advanced 2.0 中的跨数据中心支持。

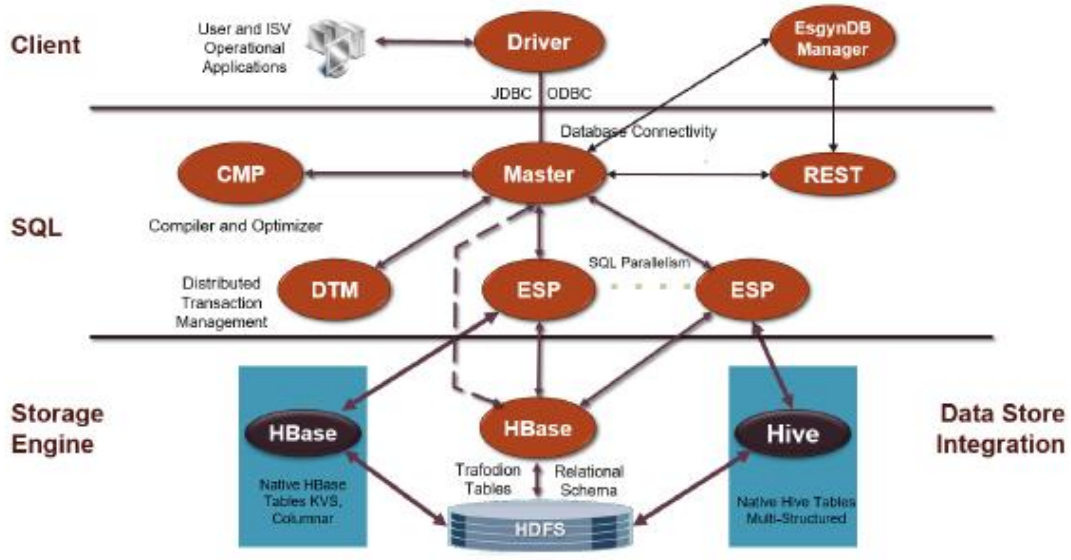
参考架构说明描述了特定用途的 EsgynDB Enterprise 安装。特别描述了架构和配置目的是运行一个或多个 EsgynDB 应用程序任务流的集群。

本参考架构说明中描述不涉及 EsgynDB Enterprise 作为范围更大的 Hadoop 集群的配置的一部分，该集群运行 MapReduce 等生态系统其他应用程序。对于运行混合任务流的集群，本文给出规模规划/配置信息。但是最终规模规划/配置还必须包含集群中其他任务流的要求。此类情况不在本文件范围内。

2. 架构

Apache Trafodion 提供 Hadoop 生态系统中的大型互联网企业级数据库引擎。另外，Trafodion 使得原生 Apache HBase™和 Apache Hive™表格可以使用 SQL 查询语言和事务语义。Trafodion 为存储在 HBase 中的数据提供事务支持。它支持跨多个语句、表格和行之间的完全分布式 ACID 事务，使得 EsgynDB Enterprise 能够支持通常超出大部分 Hadoop 生态系统组件工作范围的运营型任务流。

EsgynDB Enterprise 2.0 版通过提供额外的功能，比如支持多数据中心对 Apache Trafodion 进行了扩展，其使用的架构如下图所示。



架构包括一个或多个客户端，通过一个驱动程序（ODBC/JDBC/ADO.NET）并发使用 SQL 查询访问 EsgynDB 管理的数据。驱动程序库为应用程序（可能会或者不会在同一集群中运行）查询和 SQL 引擎层之间提供连接和会话。

在 SQL 引擎层，一个主查询执行服务器进程负责处理查询准备和执行查询处理。根据具体的任务流，可能包括分布式的事务管理器或一组或多组 Executor Server Process（ESP），它们并行执行查询计划中的一部分。这些 ESP 组（在给定查询中，可能没有 ESP 组，也可能有多个 ESP 组），体现查询的并行度。

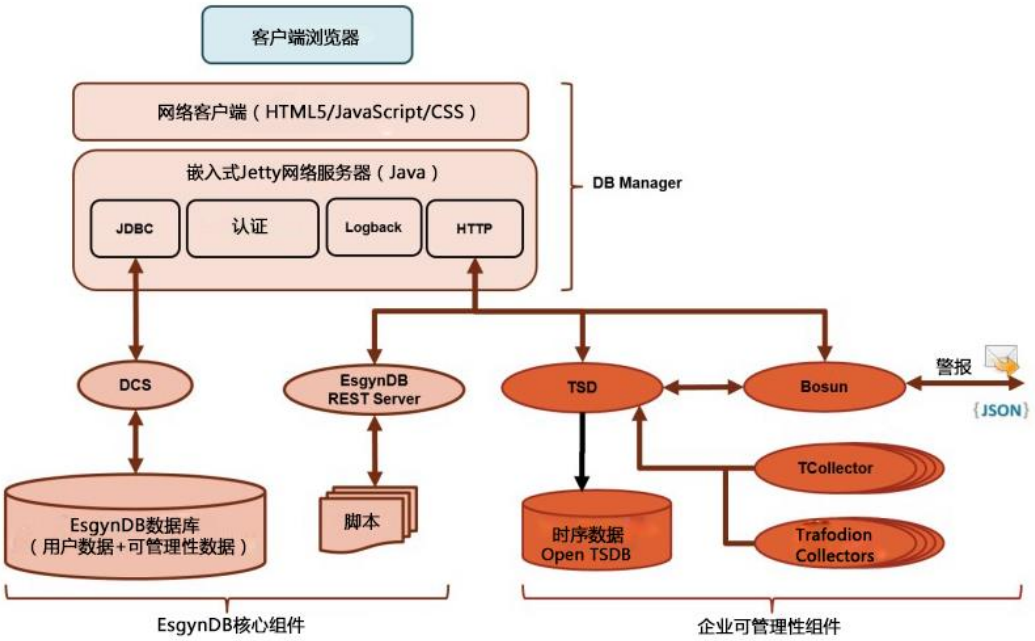
查询也可以引用原生 HBase 或 Hive 表格。最后，EsgynDB 使用 HDFS 作为存储层基础，在节点失效时，利用合适的复制因子（通常使用 3，但在一些云配置中，2 是合适的复制因子），提供可用性。

查询处理使用的主要进程包括：

进程名称	描述	分布	数量
DCS Master	用于分配保持会话的 mxosrvr 的初始连接点	位于单个节点	每个集群中只有一个有效，通常配置 floatingIP，以便获得高可用性。
DCS Server	这是管理 mxosrvr 进程状态和连接使用的进程	位于每个节点	在 mxosrvrs 运行的每个节点有一个。
Master executor (mxosrvr)	这是主执行程序进程，主持 SQL 会话、进行查询编译、并执行根运算符。	在实例中，在所有数据节点多个分布	计数器定义了并行会话的最大数量。
Executor Server Process (EXP)	执行并行的分片 SQL 计划	以各种规模的分组，多个进程在集群中的所有数据节点运行。	根据任务流确定：由并行查询、查询计划和并行度决定
DTM	维护事务的事务状态和日志结果信息。	在实例中所有数据节点运行	每个数据节点一个进程

对于 EsgynDB Enterprise 2.0 版本，通过 DTM 在对等数据中心集群中的 Transaction Manager 进程相互通信，在两个集群中复制事务，实现多数据中心支持。

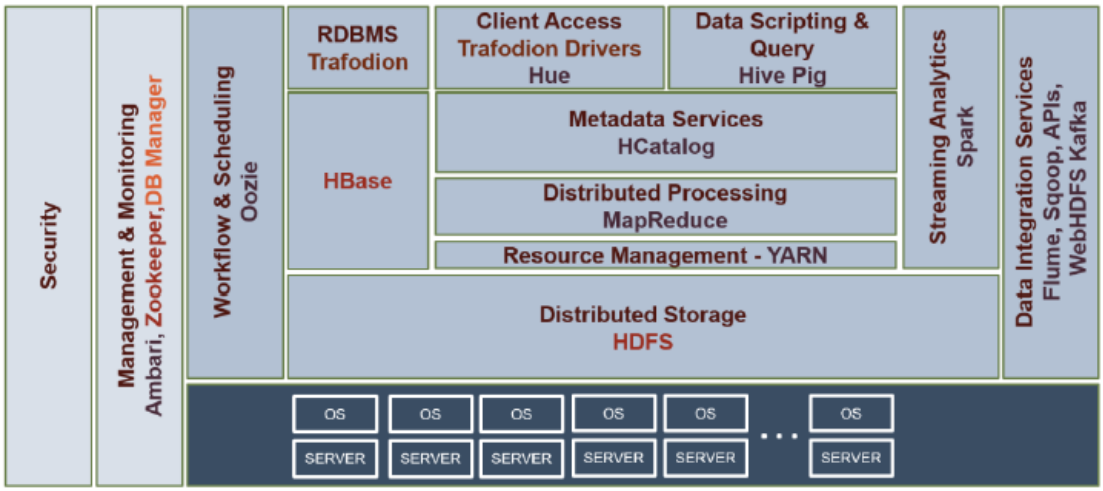
下面的架构图是简化的 EsgynDB Manager 架构，说明了与查询处理引擎的关系。EsgynDB Manager 子系统架构扩展到多个进程，如下图所示：



EsgynDB Manager 进程包括：

进程名称	描述	分布	数量
DB Manager	这是与浏览器连接的 Web 应用程序服务器	位于单个节点	在第一个数据节点，每个集群有一个进程
OpenTSDB	这是收集时间序列度量的轻量级服务进程	位于每个节点	每个节点有一个进程
TCollectors	这是按照一定时间间隔收集基于时间的度量的收集脚本	在实例中，在所有数据节点分布多个进程；每个节点的进程数不同	在每个节点收集系统和 HBase 度量 从第一个节点的进程中，收集整个集群范围中 EsgynDB 度量
REST Server	这是处理登陆和注销集群客户端的 REST 请求的进程	每个集群一个进程	在第一个节点，每个集群一个进程

除了已经列出的查询处理和管理进程外，EsgynDB 协议中还有其他进程，支持其运行时执行环境。这些进程通常使用的资源较少，对平台规模规划和配置几乎没有重大影响。EsgynDB Enterprise 和 Hadoop 生态系统的集成，如下图所示：



EsgynDB 数据库引擎使用 HBase 提供存储服务。同样依靠 HBase 配置和调整，达到最佳性能。在进行 EsgynDB 集群配置时必须考虑 HBase 配置。

HBase 进程可以分成两类：控制进程和数据进程。控制进程是一次性的，对 HBase 系统及其元数据进行管理。数据进程是为数据本身提供服务，包括数据读取、更新和写入（HBase scan，get 和 put 操作）。

HBase 控制进程包括：

进程	描述
HMaster	创建/删除元数据和表格
ZooKeeper	这不是一个 HBase 进程，但是用于信息管理和处理节点之间的协调。

HBase 数据进程包括：

进程	描述
RegionServer	控制数据服务，包括服务 get/put，将数据分布到单独的 region。

HBase 反过来利用 HDFS 服务实现集群内部的可扩展性、可用性和恢复（复制）。同样，在进行 EsgynDB 集群配置时，也必须考虑 HDFS 的配置，包括复制。控制进程是管理 HDFS 文件系统的单例进程。在 HDFS 中，控制进程控制单个数据块的位置。数据进程则负责读取和访问该数据。

HDFS 控制进程包括：

进程	描述
NameNode	该进程是对用于将数据块映射到单个文件并选择复制位置的元数据文件进行管理。
Secondary NameNode	为 NameNode 中的所有元数据获得一个检查点，每个时间间隔一次（默认每隔 1 小时）。可以使用该数据重新创建数据块->如果 NameNode 丢失，文件映射。但是，它不仅是 NameNode 热备份。

HDFS 数据进程包括：

进程	描述
DataNode	提供单个文件的读写服务，定期向 NameNode 发送在运行的消息，包括其管理的文件/块。

除了上述 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	Ambari、Cloudera Manager 等网页节点。一些管理服务器 具有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位于同一个节点上。而在较大的集群中，管理进程有明显不同的配置要求，时常位于不同的节点上。参考架构假设控制节点与数据节点分离。

3. 容量规划

本节讨论在确定 EsgynDB Enterprise 数据库规模时需要考虑的问题和规模建议。

3.1 处理能力使用

当规划 EsgynDB Enterprise 集群的处理能力大小时，考虑以下几点：

- 在典型的高性能配置中，管理节点与数据节点分开配置。这两种类型一般在存储（大小、配置）、网络和内存方面有不同的配置。
- 在非常小的配置或测试配置中，数据节点与控制节点之间的差别比较模糊，Hadoop/HBase 和 EsgynDB 中的大部分管理进程与数据进程在在同一节点并发使用。只要该配置符合性能和可用性目标，特别是针对基础开发和测试集群，该配置是适当的。在评估所需节点数量时，考虑以下因素：
- 只要每个节点的核数对于典型生产任务流相对主流（例如 8 核或以上），与相同数量的核心分布在较少的节点上相比，最好选用增加节点，减少核心的方案。向外扩展（增加节点的数量，以便获得预期数量的节点）比按比例放大（增加每个节点的核心，以便获得预期数量的节点）更好，因为：
  - 增加节点，减少核心的成本一般比减少节点，增加核心的成本低。

- 当有更多节点的集群失去一个节点或磁盘时，失效域更小。
- 节点增加时，可用 I/O 宽带和并行度更高。

- 考虑到 HDFS 为了提供可用性和可恢复性的复制要求，不建议使用小于 3 个节点的集群。
- 并发用户（并发性）的数量和数据到达/刷新的提取速度推动公司外部网络连接节点的数量。该数量决定 mxosrvr 进程的总数。实际连接根据 mxosrvr 进程分布情况分布在集群中。多个 mxosrvr 进程可以在同一个节点运行。
- 在确定节点数量时，任务流类型是其他关键考虑因素。节点和核数反映集群中运行的应用程序并发用户的可用并行数量。如果典型任务流是高并发短时间查询，则可以接受较细的节点。如果典型任务流涉及大量数据扫描，则需要更大的处理能力。理想情况下，尽可能建立任务流和查询原型，以便了解应用程序的类型、频率、计划和典型并发性。

3.2 内存使用

当根据内存使用情况确定 EsgynDB Enterprise 集群大小时，要牢记以下注意事项：

- 很多 Hadoop 生态系统进程是 Java 进程。由于 JVM 的内存效率优化，内存低于 32GB 有很大限制。超过这个阈值，实际可用内存减少，因为指针的内部表示法发生变化，因此明显消耗更多的内存空间。
- 数据节点中消耗内存较大的包括：
  - HDFS DataNode 进程
  - HBase RegionServers

  在控制进程中，消耗内存较大的是：
- HDFS NameNode 进程

为了在大集群中实现最佳性能，这些进程中规划每个进程使用 16-32GB 堆内存。减少这些组件的内存会对性能造成明显影响，因此在选择较小的值之前，要谨慎地做好调整和分析。

- 主要占用 EsgynDB 数据库引擎内存的是 mxosrvrs。对于一个节点上的每个并发连接，预留 512MB（0.5 GB）内存给一个节点上的每个连接。

3.3 硬盘使用

当根据硬盘使用情况确定 EsgynDB Enterprise 集群大小时，要牢记以下注意事项：

- 对于数据节点，SSD 仅有利于高并发写入。一般情况下，HDD 足够。对于控制节点，SSD 同样性价比不高——目标是在内存中缓存最多的控制信息。
- 对于数据节点，HDD 数据磁盘配置磁盘，如同 JBoD（只是一批磁盘）配置中的直接附加存储。RAID striping 降低 HDFS 速度，实际上降低了并发性和可恢复性。对于控制节点，数据磁盘可以配置成 JBOD 或 RAID1 或 RAID10。

- 至于处理能力，硬盘是一个并行单元。对于给定的每节点硬盘总数值，如果任务流包括很多大量扫描，在数据节点上，每个节点使用更多的小硬盘经常比使用较少的大硬盘效率更高。参考架构假设大部分的任务流包括大面积扫描。
- 强烈建议作 HBase SNAPPY 或 GZ 压缩。SNAPPY 的 CPU 开销较少，但是 GZ 压缩效果更佳。根据数据和任务流模式不同，压缩程度变化范围很大，但是普遍接受的计算表明，根据数据，大约减少 30%-40%。通过压缩，读写路径长度增加，影响数据增长和摄取。在 HBase 文件块层次进行压缩，限制读取时需要的未压缩数量。
- 当计算总磁盘空间和每个节点的数据磁盘空间时，一定要考虑工作空间和每个节点的预期摄取/流出。另外要记住，HDFS 文件块伴有复制因子（一般设定为 3 个，因此由三份数据）。这意味着，每个 10GB 文件实际上占用了 30GB 的磁盘空间。Esgyn 建议，要留出大约 33%的自由磁盘空间，作为工作空间开销。

3.4 网络使用

当根据网络使用情况确定 EsgynDB Enterprise 集群大小时，要牢记以下注意事项：

- 一般而言，EsgynDB 集群内数据通信网络连接的标准是 10GigE。使用的数据流网络速度较慢时，明显影响性能。两个 10GigE 绑定网络为 I/O 密集应用程序提供更多的吞吐量。
- 在某些情况下，提供第二个速度较慢的网络用于集群维护（不是 Hadoop/HBase），以便使该通信与运行数据工作流分开。
- 当把不同机架的节点连接在一起时，需考虑故障场景。如果复制因子是 3 或以上，一个块位置至少需要在两个不同的机架中选择节点时，HDFS 块放置算法有偏差。
- 如果使用 EsgynDB Enterprise Advanced 2.0 的跨数据中心特性，两个数据中心之间必须有高速连接。
- 如果使用 EsgynDB Enterprise Advanced 2.0 的跨数据中心特性，必须配置两个集群，以便当两个集群都在运行和可以访问时，应用程序能够主动通过 EsgynDB 驱动程序，与任何一个对等集群连接。这种能力确保应用程序能够在与两个集群中的任何一个集群失去通信时，单独与另一个集群连接。

4. 裸机生产集群的参考架构指南

本节包含对裸机 EsgynDB 集群的硬件配置和软件配置建议。这些建议不依赖于硬件。请向您的硬件供应商查询具体的产品型号和可用性/时效性。

本节所述配置适合中型或大型 EsgynDB 安装，控制节点与数据节点分开。所有进程均在同一节点上的较小配置在单独的一节中介绍。

对于数据节点，建议每个节点使用的基本硬件是：

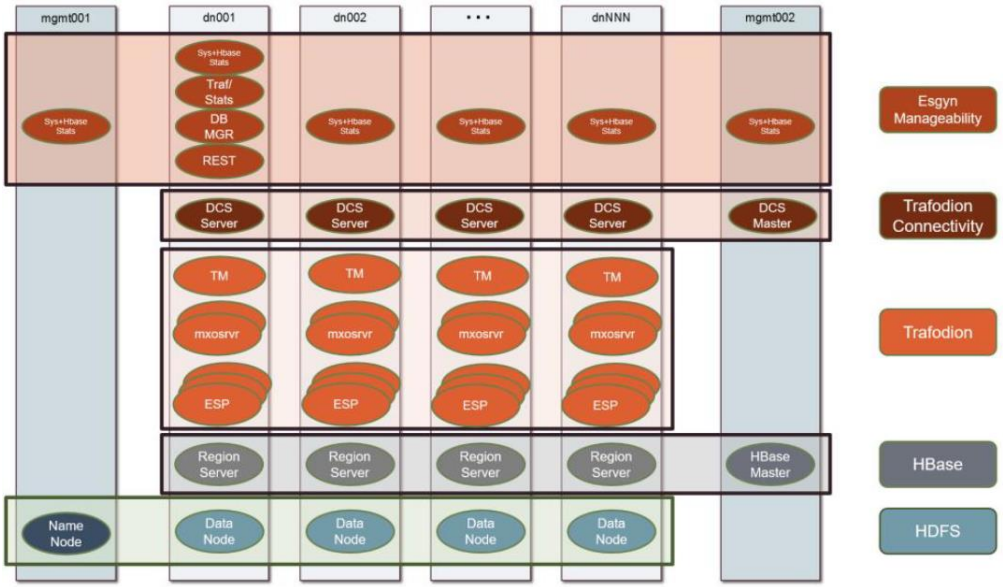
资源	建议
CPU	Intel XEON 或 AMD 64-位处理器 8≤每个节点的核数≤16
内存	整个 Hadoop 生态系统和查询处理内存+平常的开销+节点上每个 mxosrvr 进程消耗 0.5GB=64GB 计算 mxosrvr 进程的数量： $\frac{\text{并发连接的最大值}}{\text{节点数量}}$ 64GB≤内存大小≤128GB。最常用的值是 96GB。
网络	10GigE、1GigE 或 2x10GigE 绑定网络
存储	SATA 或 SAS 或 SSD，一般 JBOD 配置中使用 12-24 个 1TB 硬盘

对于控制节点，建议每个节点使用的基本硬件是：

资源	建议
CPU	Intel XEON 或 AMD 64-位处理器 8≤每个节点的核数量≤16
内存	整个 Hadoop 生态系统和查询内存+可能的/根据要求进行交换和进程维护的开销 =64GB 64GB≤内存大小≤128GB。最常用的值是 96GB。
网络	10GigE、1GigE 或 2x10GigE 绑定网络，加上注销平台到登陆平台的合适交换机。
存储	SATA 或 SAS 或 SSD，一般 RAID1 或 RAID10 配置中使用 6-12 个 1TB 硬盘

4.1 中型/大型部署

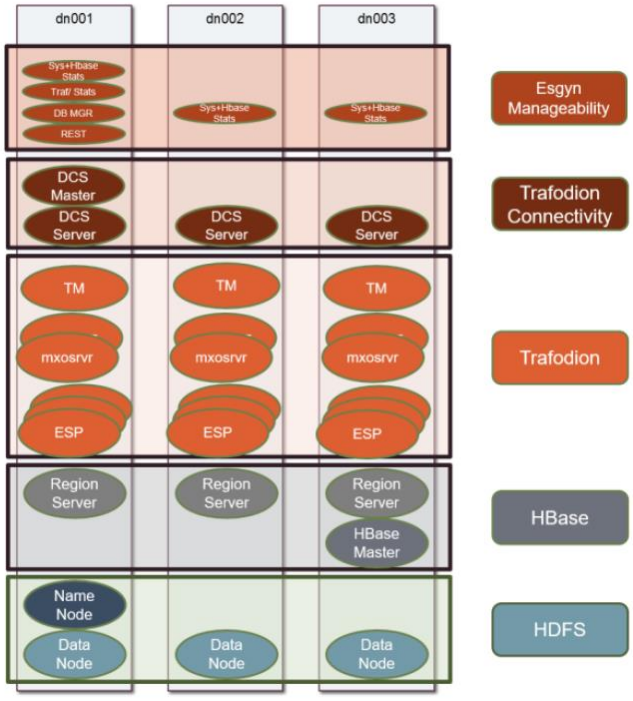
中型/大型部署使用上述规范，包括控制节点和管理节点。这些节点中的进程放置如下图所示：



在上图中，控制节点在数据节点的两侧仅用于 DCS 主进程。节点命名约定没有特定限制，包括没有假设节点是连续编号的。垂直方框表示单独的节点，椭圆形表示进程在节点内。

4.2 小型部署

对于小型部署（2-3 个节点，一般少于一个机架），控制节点重叠到普通节点基础设施结构中，如下图所示：



在上图中，控制节点被移除，控制进程运行在作为功能进程的相同节点上。

5. 云部署

当在云环境中，比如 Amazon 的 AWS，部署 EsgynDB 时，使用上述指南配置资源。如果您选择本地存储文件系统，配置时使用 HDFS 复制因子 3，否则如果您选择 EBS 容量，则使用 HDFS 复制因子 2。

6. 结论

EsgynDB 平台参考架构文件充当定义 EsgynDB 集群建立平台的起点，其中 EsgynDB 是集群的主要用途。另外，它还将有助于应用程序开发人员和用户规划 EsgynDB 应用程序的部署策略。Esgyn 建议咨询 Esgyn 技术人员，获取其他信息、培训和指导。

