Microsoft

让 IT 和数据专业人员充分利用数据获得更多成果

将现有流程迁移到 Azure Synapse Analytics 指南

引言

如今，许多公司在其数据中心内长期设立分析数据仓库，以支持其不同业务部门的决策。销售、营销和财务部门尤其依赖此类系统制作标准报告和仪表板。他们还聘请业务分析人员对数据集市中的数据执行临时查询和分析，这些数据集市旨在使用自助业务智能 (BI) 工具进行多维分析。

但是，尽管数据仓库支持决策的某些关键方面，但在过去几年中，在数字技术转型的推动下，数据仓库发生了很多变化，对传统分析系统产生了重大影响。其中包括迁移到云的来源事务处理系统，例如 CRM、HR 和 ERP 系统。这样做通常是为了合并类似的系统，以扩大规模，并使客户（例如，移动银行）、合作伙伴、供应商和员工能够通过移动自助服务访问事务处理系统。因此，数据仓库现在需要从云 SaaS 应用程序采集数据。此外还大量出现了企业正在采集的新数据，例如在线点击流、社交网络数据、物联网 (IoT) 传感器数据、开放政府数据、天气数据、图像、音频和视频数据。但是，这类数据极少能够进入数据仓库。实际上，这类数据经常在云中得到大规模单独处理和分析，而数据科学家使用其构建机器学习预测性分析和规范性分析.

此外，由于现在许多新数据和分析技术都是首先出现在云上，因此，将分析工作负载迁移到云，以快速利用这些技术，变得越来越具有吸引力。在这种背景下，许多公司正在考虑将其数据仓库迁移到云中，作为数据仓库现代化工作的一部分，也就不足为奇了。

如今，许多组织正在将其旧数据仓库解决方案迁移到 Azure Synapse Analytics，以利用端到端分析平台的优势，后者可以为企业数据仓储工作负载提供较高的可用性、安全性、速度和可伸缩性，并节省成本，获得行业领先的性能。

随着技术的发展，拥有基于云的数据仓库解决方案的优势远远超过同类本地解决方案。Azure Synapse 不仅为在云中运行企业数据仓库工作负载提供了业界领先的性能，同时还是一个端到端分析平台，可以将数据接收、数据仓储和大数据分析整合到单个服务中。凭借其可扩展性以及独立的计算和存储架构，Azure Synapse 可以通过传统系统（如 Teradata、Netezza 或 Exadata）无法使用的方式实现即时扩展。

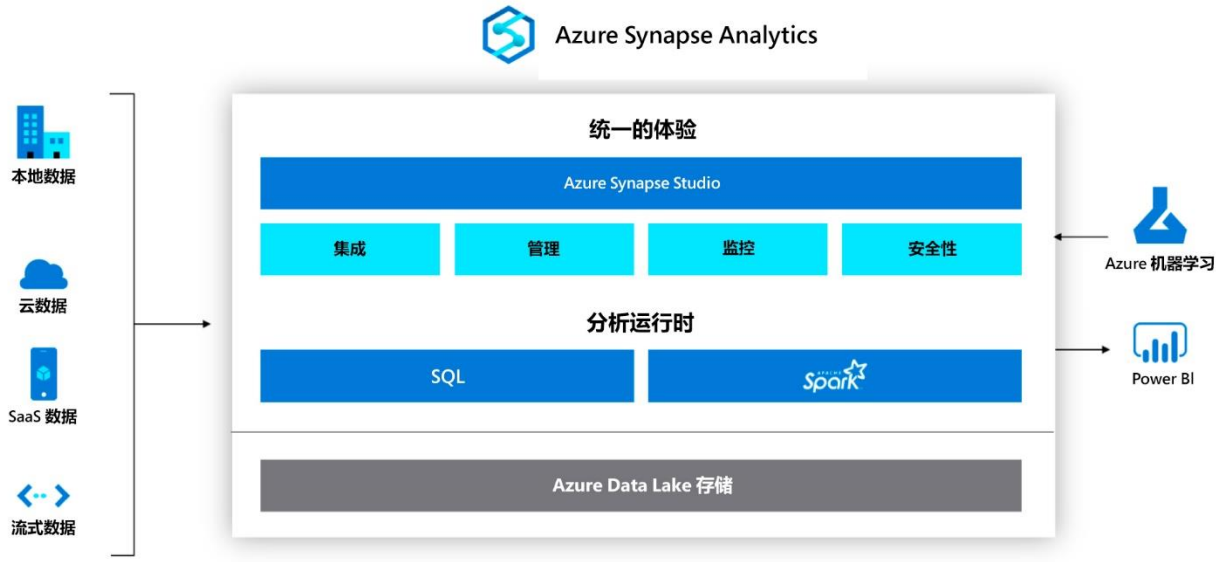
您应该考虑迁移到 Azure Synapse，原因不外乎其带来的几个业务优势，比如帮助您降低总体拥有成本，提高性价比，利用丰富的附加数据和分析技术生态系统来帮助实现数据仓库现代化，以及缩短实现价值的时间。

业务优势包括：

- 降低实施和维护成本 - 只需为使用的资源付费
- 不用管理基础架构，因此可以专注于竞争性见解
- 仅在需要时为数据和分析工具付费，不使用时则无需消费
- 缩短分析项目开发时间并提高创新能力
- 能够从计算中独立扩展存储
- 用于暂存和生产数据的低成本存储

- 避免因为数据量增加和 ELT 处理占用容量导致的昂贵升级
- 随着数据量的增长降低存储成本
- 提高安全级别和业务连续性
- 通过本机集成 Apache Spark 和 SQL 引擎的分析平台，加快获取见解的速度
- 轻松融入最新创新的面向未来的架构

除了使用 Azure Synapse 在云中运行企业数据仓库带来的业务优势之外，所有数据专业人员现在都可以在数据准备、数据管理、数据仓储、大数据和 AI 任务方面使用统一的体验。客户可以通过两种方法分析数据，即通过预配工作负载或通过提供按查询付费功能的无服务器消费模型，因此可以为每个用例选择最具成本效益的选项。此外，在数据方面，安全性和隐私性至关重要，也是通过分析发现见解的基础。Azure Synapse 的结构中内置高级安全性和隐私性，例如不间断数据加密。对于精细化访问控制，企业可以使用列级别安全性和本机行级别安全性以及动态数据掩码来帮助确保数据的安全和隐私，以实时自动保护敏感数据。



本指南提供了用于规划、准备和执行迁移的方法，以便将旧数据仓库系统成功迁移到 Azure Synapse Analytics。本指南不用作全面的分步迁移手册，而是实践概述，可帮助您进行迁移规划和项目范围确定。

本指南末尾的附录确定了一些常见的迁移问题和可能的解决方案。

本指南的目标受众是数据仓库架构师、解决方案架构师、CTO 和项目经理，他们需要使用明确定义的方法将现有的本地数据仓库迁移到 Azure Synapse Analytics。

为什么要迁移

数据仓库迁移流程

3 步迁移流程（准备、迁移、发布）

2 种迁移类型（升级和迁移、重新设计）

在迁移前降低复杂性

迁移现有架构

迁移历史数据

迁移现有 ETL

迁移 BI/查询

安全考虑事项和工具

结语

为什么应该将旧数据仓库迁移到 Azure Synapse Analytics?

通过迁移到 Azure Synapse Analytics，拥有旧数据仓库系统的公司可以利用云技术的最新创新，并将基础架构维护和平台升级等任务委托给 Azure。

迁移到 Azure Synapse 的客户已经获得了许多好处，包括：

性能

Azure Synapse Analytics 使用大规模并行处理 (MPP) 和自动内存中缓存等技术，提供一流的关系数据库性能。例如，GigaOm 进行的独立基准测试中展示了这方面的成果（参考：https://gigaom.com/report/data-warehouse-cloud-benchmark/）。本报告将 Azure Synapse 与其他常见云数据仓库产品进行了比较。

快捷

数据仓储是一种进程密集型技术。这涉及数据接收、数据转换、数据清理、数据聚合、数据集成以及进行数据可视化和生成报告。将数据从原始源迁移到数据仓库所涉及的许多过程都较为复杂且相互依赖。单个瓶颈会减慢整个管道的速度，而数据量的意外激增会增加对速度的需求。在数据及时性意义重大的情况下，Azure Synapse Analytics 可以满足快速处理的需求。

提高安全性和合规性

Azure 是一个全球可用的高度可扩展安全云平台。位于 Azure 生态系统中的 Azure Synapse Analytics 继承了上述所有优点。

弹性和成本效益

在数据仓库中，工作负载处理的需求可能会发生波动。有时，峰值和谷值间的波动可能会非常巨大。例如，在节假日期间，销售数据量可能会突然激增。借助云弹性，Azure Synapse 能够根据需求快速增加和减少容量，而不会影响基础结构的可用性、稳定性、性能和安全性。最重要的是，你只需为实际使用量付费。

托管基础架构

通过消除数据中心管理和数据仓库运营的开销，公司可以将宝贵的资源重新分配到可以带来价值的地方，并专注于使用数据仓库来提供实用信息和见解。这还可降低总体拥有成本，并能够更好地控制运营支出。

可伸缩性

数据仓库中的数据量通常随着时间的推移和历史记录的收集而增长。Azure Synapse Analytics 可以通过随着数据和工作负载的增加逐步添加资源来进行扩展，以适应这种增长。

成本节省

运行本地旧数据中心成本高昂（服务器和硬件、网络、物理空间、电力、冷却和人员的成本）。使用 Azure Synapse Analytics 可以大幅减少这些费用。

Azure Synapse Analytics 可为你提供真正的即用即付云可扩展性，而无需随着数据或工作负载的增长而进行复杂的重新配置。

最大限度提高技能

将整个企业中的所有现有技能结合起来，利用数据完成更多工作。借助 Azure Synapse 中深度集成的 Apache Spark 和 SQL 引擎，喜欢或熟悉 SQL 的数据专业人员可以与喜欢 Spark 的人无缝协作，反之亦然。

例如，熟悉或喜欢 SQL 的人可以使用 T-SQL 语言查询 Spark 表。喜欢使用 Python、Scala、SparkSQL 或 C＃等语言的数据工程师或数据科学家，可以在包含数据管道，数据湖和数据仓库相同的服务中转换数据、训练模型并创建概念证明。

数据湖到数据仓库

管理、保护和分析所有类型的数据。Azure Synapse 可以使用数据仓储资源查询结构化或半结构化数据，还可以对数据湖提供的非结构化数据快速执行无服务器查询。使你的数据专业人员能够构建端到端分析解决方案，而无需将多种服务结合在一起。

利用 Azure Synapse Link 将消除云中数据孤岛的工作提升到新的水平 — 这是一种云原生混合事务分析处理 (HTAP) 实施，现已在公共预览版中提供。这项技术消除了 Azure 数据库服务和 Azure Synapse 之间的障碍，因此客户只需点击一下即可从存储在其运营数据库中的实时事务数据中获取见解，而无需管理数据移动或给操作系统带来负担。

数据仓库迁移流程

成功的数据迁移项目始于精心设计的计划。一个有效的计划会考虑许多需要考虑的组成部分，并特别关注架构和数据准备。下面是 3 步迁移流程计划。

3 步迁移流程

准备

- 定义要迁移的范围
- 构建要迁移的数据和流程清单
- 定义数据模型更改（如果有）
- 定义源数据提取机制
- 确定要使用的合适 Azure（和第三方）工具和服务
- 在新平台上对员工进行早期培训
- 设置 Azure 目标平台

迁移

- 开始小规模简单试点
- 在所有可能的位置实现自动化
- 利用 Azure 内置工具和功能减少迁移工作量
- 迁移表和视图的元数据
- 迁移要维护的历史数据
- 迁移或重构存储过程和业务流程
- 迁移或重构 ETL/ELT 递增负荷流程

迁移后

- 监控并记录流程的所有阶段
- 利用获得的经验来为将来的迁移构建模板
- 如果需要，重新设计数据模型（利用新的平台性能和可伸缩性）
- 测试应用程序和查询工具
- 进行基准测试并优化查询性能

两种类型的迁移策略

对现有数据仓库进行评估，确定哪种迁移策略最适合您的情况，开始进行迁移规划。有两种类型的迁移策略可供考虑：

升级和迁移策略

对于提升和迁移策略，现有数据模型将原样迁移到新的 Azure Synapse Analytics 平台。这可以将更改范围减至最小，从而最大限度降低迁移的风险和所需的时间。

提升和迁移是一个很好的策略，适合有以下情况的旧数据仓库环境（例如 Netezza）：

- 需要迁移的单个数据集市，或
- 数据已经采用精心设计的星型或雪花架构，或者
- 面临着迁移到现代云环境的紧迫时间和成本压力

重新设计策略

在旧数据仓库随时间变化的场景中，可能必须要进行重新设计，以维持最佳性能水平或支持新数据类型。这可能包括更改基础数据模型。

为了最大限度降低风险，建议先使用提升和迁移策略进行迁移，然后再使用重新设计策略逐步实现 Azure Synapse Analytics 上数据仓库数据模型的现代化。对数据模型进行完整更改将增加风险，因为这将影响从源到数据仓库的 ETL 作业和下游数据集市。

在迁移前降低现有旧数据仓库的复杂性

在上一节中，我们介绍了两种迁移策略。作为最佳实践，在初始评估步骤中，请留意简化现有数据仓库的任何可能性并进行记录。目的是在迁移之前降低现有旧数据仓库系统的复杂性，以简化迁移过程。

以下是有关如何降低现有旧数据仓库复杂性的一些建议：

在迁移前删除未使用的表格并存档

- 避免迁移不再使用的数据

将实体数据集市转化为虚拟数据集市

- 最大限度减少必须迁移的内容
- 降低总拥有成本
- 提高灵活性

在下一节中，我们将详细谈论为什么应考虑将物理数据集市转化为虚拟数据集市。

将物理数据集市转化为虚拟数据集市

在迁移旧数据仓库之前，请考虑将当前的物理数据集市转化为虚拟数据集市。通过使用虚拟数据集市，可以消除数据集市的物理数据存储和 ETL 作业，同时不会在迁移之前失去任何功能。这样做的目的是减少要迁移的数据存

储的数量、减少数据副本、降低总拥有成本并提高灵活性。要实现这个目标，您将需要从物理数据集市切换到虚拟数据集市，然后再迁移数据仓库。您可以将其视为迁移之前的数据仓库现代化步骤。

物理数据集市的缺点

- 同一数据多个副本
- 总拥有成本较高
- 难以更改，因为 ETL 作业会受到影响

虚拟数据集市的优势

- 简化数据仓库架构
- 无需存储数据副本
- 灵活性更高
- 总拥有成本较低
- 使用下推优化以利用 Azure Synapse Analytics 的功能
- 易于更改
- 易于隐藏敏感数据

将现有数据仓库架构迁移到 Azure Synapse Analytics

接下来，计划如何迁移现有旧数据仓库的架构。架构迁移涉及迁移现有暂存表、旧数据仓库和从属的数据集市架构。

为了帮助您理解架构迁移的规模和范围，我们建议您为现有的旧数据仓库和数据集市创建目录。

这是一份清单，可帮助您收集必要的信息：

- ✓ 行计数
- ✓ 暂存、数据仓库和数据集市的数据规模
  - 表和索引
- ✓ 数据压缩率
- ✓ 当前硬件配置
- ✓ 表（包括分区）
  - 识别小尺寸表
- ✓ 数据类型

- ✓ 视图
- ✓ 索引
- ✓ 对象相关性
- ✓ 对象使用
- ✓ 功能
  - 开箱即用的功能和 UDF
- ✓ 存储过程
- ✓ 可扩展性要求
- ✓ 增长预测
- ✓ 工作负载需求
  - 并发用户数

清单完成后，你可以决定要迁移的架构的范围。从本质上说，有四个选项可用于设置旧数据仓库架构迁移的范围。

1.一次迁移一个数据集市

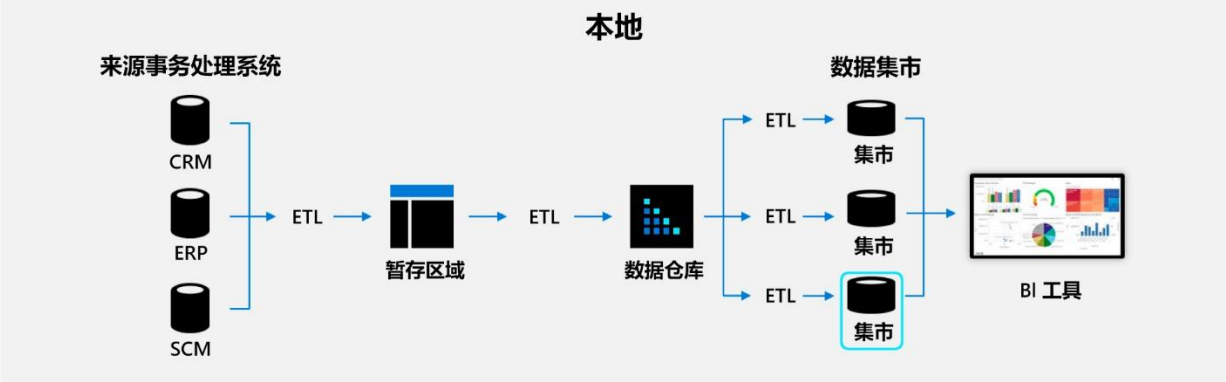


图 3 - 一次迁移一个数据集市

2.一次迁移所有数据集市，然后迁移数据仓库

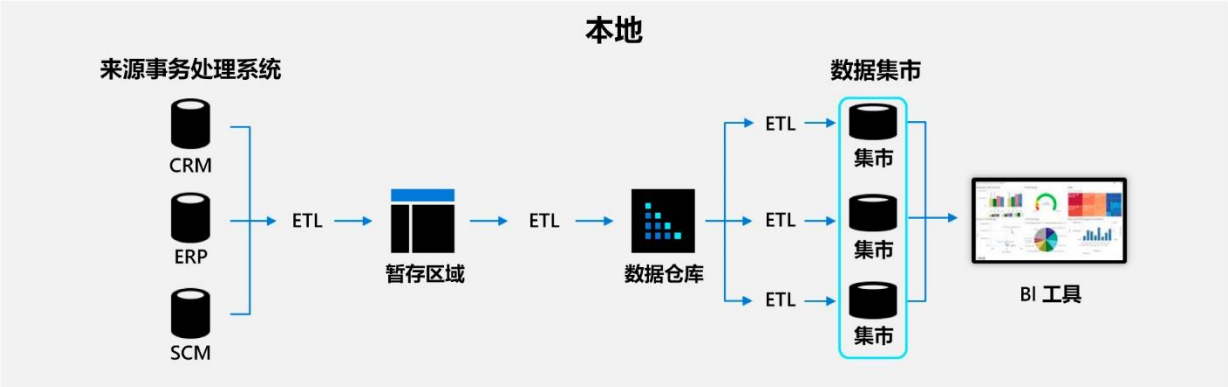


图 4 - 一次迁移所有数据集市，然后迁移数据仓库

3.同时迁移数据仓库和暂存区域

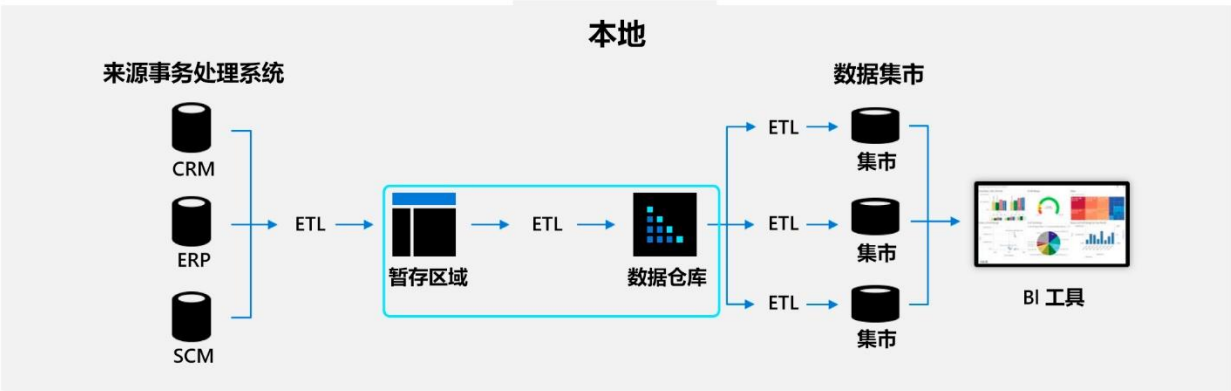


图 5 - 同时迁移数据仓库和暂存区域

4.一次性迁移所有内容

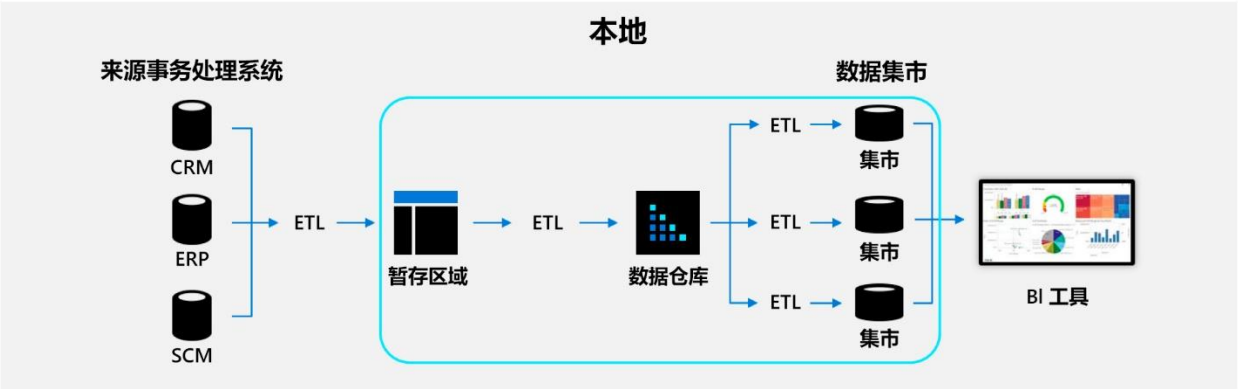


图 6 - 一次性迁移所有内容

在选择选项时，请记住，你的目标是实现在性能上与当前旧数据仓库系统相当或有所超越的物理数据库设计，并且最好以较低的成本实现。

作为总结，以下是对架构迁移的一些建议：

- 避免迁移不必要的对象或进程
- 考虑使用虚拟数据集市来减少物理数据集市的数量或将其消除
- 尽可能实现自动化
- 使用旧数据仓库系统中系统目录表内的元数据为 Azure Synapse Analytics 生成 DDL
- 在 Azure Synapse Analytics 上执行任何所需的数据模型更改或数据映射优化

将历史数据和 ETL 进程从旧数据仓库迁移到 Azure Synapse Analytics

确定架构迁移范围后，我们就可以做出关于如何迁移历史数据的决策。

迁移历史数据的步骤如下所示：

1. 在 Azure Synapse Analytics 上创建目标表
2. 迁移现有历史数据
3. 迁移所需的任何函数和存储过程
4. 迁移传入数据的增量负载 (ETL/ELT) 暂存和进程
5. 应用所需的任何性能调整选项

下表概述了四个数据迁移选项及其优缺点。

数据库迁移选项	优点	缺点
首先迁移数据集市数据，然后迁移数据仓库数据	• 一次从一个数据集市迁移数据是一种低风险增量方法 • 后续 ETL 迁移仅限于所迁移依赖数据集市中的数据	• 在迁移完成之前，你将有一些数据同时存在于本地和 Azure 上 • 从数据仓库到数据集市的 ETL 处理需要绕开防火墙，并更改为以 Azure Synapse 为目标
首先迁移数据仓库数据，然后迁移数据集市数据	• 所有数据仓库历史数据已迁移	• 将依赖数据集市留在本地并不理想，因为 ETL 必须将数据回流到数据中心 • 没有真正的机会进行增量数据迁移
同时迁移数据仓库和数据集市	• 一次性迁移所有数据	• 潜在风险更高 • 意味着所有 ETL 很可能都必须一起迁移
将物理市场转化为虚拟市场，并仅迁移数据仓库	• 没有要迁移的数据集市数据存储 • 没有要从数据仓库迁移到集市的 ETL • 仅数据仓库数据要迁移 • 减少数据副本 • 功能无损失 • 总拥有成本较低 • 灵活性更高 • 整体数据基础架构更简单 • 可能使用 Azure Synapse 中的视图	• 如果嵌套视图不支持虚拟数据集市，则可能需要在 Azure 上使用第三方数据虚拟化软件。 • 在迁移数据仓库数据之前，所有集市都需要进行转换 • 虚拟集市和数据仓库到虚拟集市的映射将需要移植到 Azure 上的数据虚拟化服务器并重定向到 Azure Synapse

将现有 ETL 进程迁移到 Azure Synapse Analytics

我们可以提供许多选项，将您现有的 ETL 进程迁移到 Azure Synapse Analytics。下表根据现有 ETL 作业的构建方式概述了部分 ETL 迁移选项。

现有 ETL 作业是如何构建的?	迁移选项	迁移原因和注意事项
定制 3GL 代码和脚本	• 计划使用 Azure 数据工厂对其进行重新开发	• 代码不提供元数据沿袭 • 如果作者离开就很难维护 • 如果暂存表位于旧数据仓库中，并且使用 SQL 来转换数据，则使用 T-SQL 解决差异
在旧数据仓库 DBMS 中运行的存储过程	• 计划使用 Azure 数据工厂对其进行重新开发	• 旧数据仓库和 Azure Synapse 之间可能存在重大差异 • 无元数据沿袭
图形化 ETL 工具（例如 Informatica、Talend 等）	• 继续使用现有 ETL 工具，并将目标切换到 Azure Synapse • 可能移至现有 ETL 工具的 Azure 版本，并移植元数据以便在 Azure 上运行 ELT 作业，以确保支持对本地数据源的访问 • 使用 Azure 数据工厂控制 ETL 服务的执行	• 避免重新开发 • 最大限度降低风险并加快迁移速度
数据仓库自动化软件	• 继续使用现有 ETL 工具切换目标并将其暂存到 Azure Synapse	• 避免重新开发 • 最大限度降低风险并加快迁移速度

使用 Azure 数据工厂重新开发可扩展的 ETL 进程

处理现有旧 ETL 进程的另一种方法是使用 Azure 数据工厂 (ADF) 对其进行重新开发。ADF 是 Azure 数据集成服务，用于创建数据驱动的工作流（称为管道），以协调数据移动和数据转换并实现自动化。您可以使用 ADF 创建和调度管道，以便从不同的数据存储中提取数据。ADF 可以使用 Spark、Azure Machine Learning、Azure HDInsight、Hadoop 和 Azure Data Lake Analytics 等计算服务来处理和转换数据。

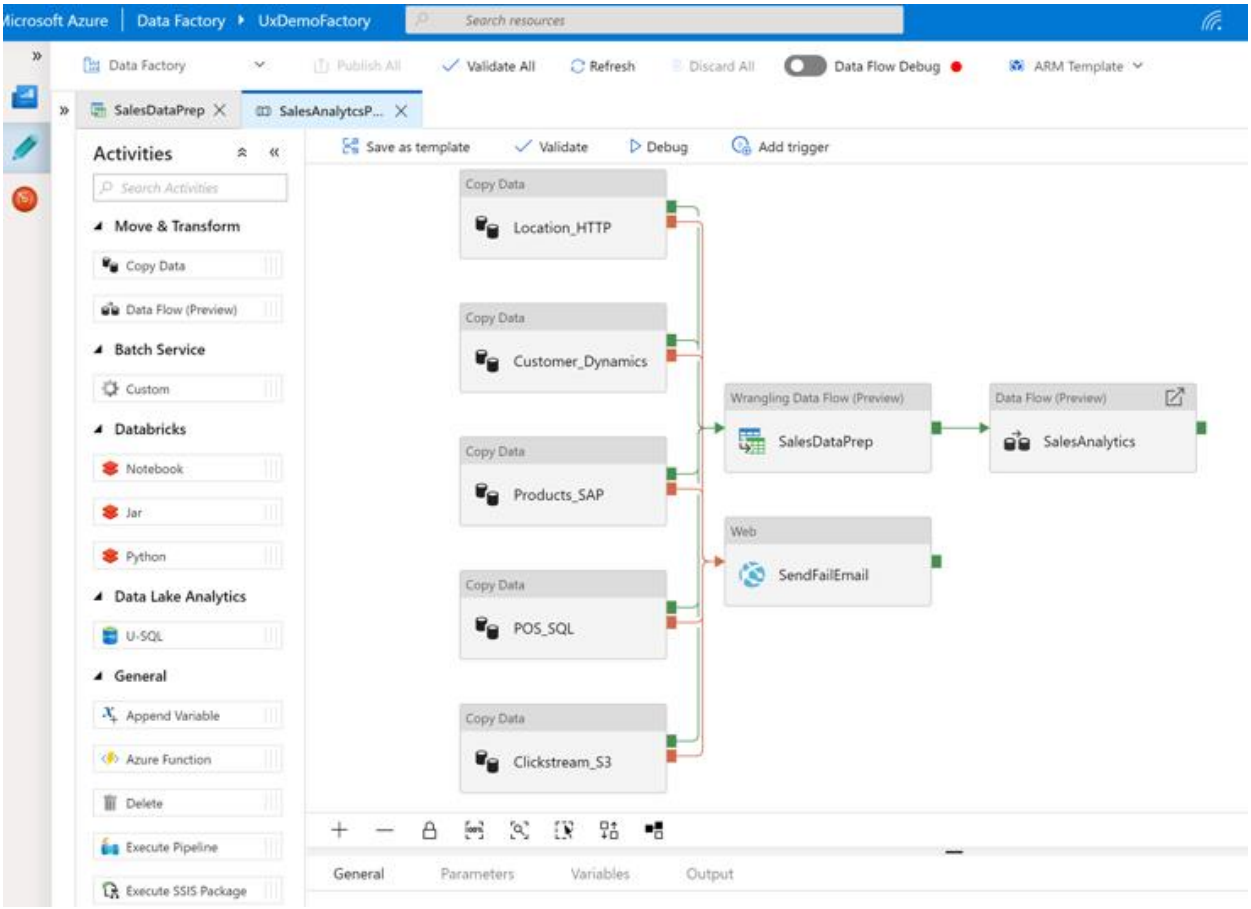


图 10 - 使用 Azure 数据工厂重新开发可扩展的 ETL 进程

有关迁移查询、BI 报告、仪表板和其他可视化内容的建议

如果旧系统使用标准 SQL，则可以轻松地将查询、BI 报告、仪表板和其他可视化内容从旧数据仓库迁移到 Azure Synapse Analytics。

然而情况通常不是这样。在这种情况下，必须使用另一个策略：

- 确定首先迁移的高优先级报告。
- 利用使用情况统计信息确定哪些报告从未使用过。
- 避免迁移任何不再使用的内容。
- 生成要迁移的报告列表、其优先级以及要跳过的未使用报告后，请与利益相关方确认此列表。
- 对于正在进行迁移的报告，请尽早发现不兼容问题，以估计迁移工作量
- 某些不兼容问题可能是由于不支持的数据类型导致的。参见**附录 A - 常见迁移问题和解决方法。**

- 考虑数据虚拟化，以保护 BI 工具和应用程序，避免迁移期间可能发生的数据仓库和/或数据集市数据模型的结构性更改对其造成影响

安全考虑事项和工具

在任何数据仓库系统中，保护数据资产的安全都是至关重要的。在计划数据仓库迁移项目时，还必须考虑安全性、用户访问管理、备份和还原。例如，HIPAA、PCI 和 FedRAMP 等行业和政府法规以及在非监管行业中，可能强制要求进行数据加密。

Azure 包括许多标准特性和功能，这些特性和功能传统上必须进行定制，才能在旧数据仓库产品中使用。作为标配，Azure Synapse 支持静态和动态数据加密：

静态数据

- 可以启用透明数据加密 (TDE)，以动态加密和解密 Azure Synapse 数据、日志以及关联的备份。
- Azure Data Storage 也可以自动加密非数据库数据。

动态数据

- 默认情况下，使用行业标准协议（例如 TLS 和 SSH）对与 Azure Synapse Analytics 的所有连接进行加密

另外，可以使用动态数据屏蔽 (DDM)，根据数据屏蔽规则为给定类别的用户进行数据混淆处理。

最佳作法是，如果您的旧数据仓库包含权限、用户和角色的复杂层次结构，请考虑在迁移过程中使用自动化技术。您可以使用旧系统中的现有元数据来生成必要的 SQL，以迁移 Azure Synapse Analytics 上的用户、组和特权。

帮助迁移到 Azure Synapse Analytics 的工具

前面我们已经介绍了迁移过程的计划、准备和概述，接下来我们将介绍可用于将旧数据仓库迁移到 Azure Synapse Analytics 的工具。我们将讨论的工具有：

- Azure 数据工厂 (ADF)
- Azure 数据仓库迁移实用程序
- Microsoft 物理数据传输服务
- Microsoft 数据接收服务

Azure 数据工厂 (ADF)

- ADF 是一种完全托管式混合数据集成服务，按使用量付费，用于云级别 ETL 处理

- 并行处理和分析内存中的数据，以扩展并最大限度提高吞吐量
- 创建数据仓库迁移管道，以协调数据移动、数据转换和将数据加载到 Azure Synapse Analytics 中的工作，并实现自动化
- 还可以通过将数据引入 Azure Data Lake、大规模处理和分析数据以及加载到数据仓库中来实现数据仓库的现代化
- 支持基于角色的用户界面，可供 IT 专业人员映射数据流以及业务用户处理自助数据
- 可以连接到分布在多个数据中心、云和 SaaS 应用程序的多个数据存储
- 可提供超过 90 个本机构建且免维护的连接器（参考：https://azure.microsoft.com/services/data-factory/）
- 在同一管道中混合和匹配整理和映射数据流，以大规模准备数据
- ADF 编排可以控制到 Azure Synapse Analytics 的数据仓库迁移
- 可以执行来自 Azure 数据工厂的 SSIS ETL 包

Azure 数据仓库迁移实用程序

- 将数据从基于 SQL Server 的本地数据仓库迁移到 Azure Synapse。
- 使用类似向导的方法，从基于 SQL Server 的本地数据仓库中执行架构和数据的直接迁移。
- 你可以选择包含要导出到 Azure Synapse 的表的本地数据库。然后选择你要迁移的表并迁移架构。
- 自动生成在 Azure Synapse 上创建等效空数据库和表所需的 T-SQL 代码。在为 Azure Synapse 提供连接详细信息后，可以运行生成的 T-SQL 来迁移架构。
- 创建架构后，您可以使用实用程序迁移数据。这将从基于 SQL Server 的本地数据仓库中导出数据，并生成 BCP（批量复制程序）命令以将数据加载到 Azure Synapse。

Microsoft 物理数据传输服务

Azure ExpressRoute

- Azure 与客户数据中心之间的专用连接
- 数据不通过互联网传输

AzCopy

- 通过 Internet 将数据复制到 Azure

Azure Databox

- 大容量（数十到数百 TB）

Microsoft 数据接收服务

PolyBase（推荐方法）

- 提供速度最快、可扩展性最好方法，将批量数据加载到 Azure Synapse Analytics
- 使用并行加载提供最快的吞吐
- 可以从 Azure Blob 存储中的平面文件读取或通过连接器从外部数据源读取
- 与 Azure 数据工厂紧密集成
- CREATE TABLE AS 或 INSERT…SELECT
- 将暂存表定义为 HEAP 类型以实现快速加载
- 支持最大长度 1 MB 的行

BCP (Bulk Copy Program)

- 支持长度大于 1 MB 的行
- 最初为 Microsoft SQL Server 的早期版本开发
- 可用于从任何 SQL Server 环境（包括 Azure Synapse Analytics）导入和导出数据
- （参考：https://docs.microsoft.com/sql/tools/bcp-utility）

SqlBulkCopy API

- 这是一个等效于 BCP 功能的 API
- 允许以编程方式实现加载过程
- （参考：https://docs.microsoft.com/dotnet/api/system.data.sqlclient.sqlbulkcopy）

INSERT 和 INSERT … SELECT

- Azure Synapse Analytics 支持标准 SQL
- 将单个行或 SELECT 语句的结果加载到数据仓库表中
- 可以在 PolyBase 中使用 INSERT … SELECT，通过外部数据源将提取的数据批量插入数据仓库表

结语

成功的数据迁移项目始于精心设计的计划。一个有效的计划会考虑许多需要考虑的组成部分，并特别关注架构和数据准备。

Azure Synapse Analytics 是一种无限分析服务，让你快人一步获取洞察，以加快向企业交付 BI、AI 和智能应用程序。通过将旧数据仓库迁移到 Azure Synapse Analytics，你将获得很多优势，包括性能、速度、提升的安全性和合规性、弹性、托管基础结构、可扩展性和成本节省。

本指南提供了为将现有 Netezza 系统迁移到 Azure Synapse Analytics 进行准备并将其付诸执行所需的高级方法。

我们已经介绍了 3 步迁移过程、迁移策略，了解了如何在迁移之前降低现有旧数据仓库的复杂性，以及如何将现有架构、历史数据、ETL 进程和可视化内容迁移到 Azure Synapse Analytics。我们还讨论了可帮助你成功迁移到 Azure Synapse Analytics 的安全注意事项和工具。

迁移到 Azure Synapse 后，你可以在丰富的 Azure 分析生态系统中探索其他 Microsoft 技术，以实现数据仓库的现代化。

以下是一些需要考虑的方法：

- 将暂存区和 ELT 处理卸载到 Azure Data Lake 和 Azure 数据工厂
- 以通用数据模型格式构建可靠的数据产品，并随时随地使用，而不仅仅是在数据仓库中使用
- 使用 ADF 映射和处理数据流，支持业务和 IT 数据准备管道的协作开发
- 在 ADF 中构建分析管道，以实时批量分析数据
- 构建和部署机器学习模型，为已了解内容添加更多见解
- 将数据仓库与实时流式数据集成
- 通过使用 PolyBase 创建逻辑数据仓库，简化对多个 Azure 分析数据存储中的数据和见解的访问

祝你的迁移之旅一切顺利！

要了解更多信息：

- 注册 Azure 免费帐户
- 与 Azure 销售专家联系，了解定价、分析最佳实践、安排概念证明等信息。
- 了解客户选择使用 Azure 进行分析的原因。

立即开始

立即开始使用 Azure 免费试用帐户
https://azure.microsoft.com/free/synapse-analytics/

下载免费入门工具包：
https://azure.microsoft.com/resources/azure-synapse-analytics-toolkit/

通过 Azure Synapse 文档了解更多信息：
https://docs.microsoft.com/azure/synapse-analytics/sql-data-warehouse/

附录 A - 常见迁移问题和解决方案。

在迁移过程中，你可能会遇到某些需要解决的问题。在本节中，我们将重点介绍一些常见问题，并为你提供可以实施的解决方案。

问题 #1：不支持的数据类型和变通办法

下表显示了不受支持的旧数据仓库系统中的数据类型，以及适用于 Azure Synapse Analytics 的适当变通办法。

不支持的数据类型	适合 Azure Synapse Analytics 的变通办法
geometry	varbinary
geography	varbinary
hierarchyid	nvarchar(4000)
image	varbinary
text	varchar
ntext	nvarchar
sql_variant	将列拆分为几个强类型列。
table	转换为临时表。
timestamp	重做代码以使用 datetime2 和 CURRENT_TIMESTAMP 函数。
xml	varchar
用户定义的类型	可能的情况下重新转换为本机数据类型

问题 #2：完整性约束差异

密切注意旧数据仓库或数据集市与 Azure Synapse Analytics 之间的完整性约束差异。在下图中，左侧表示旧的旧数据仓库系统，右侧表示新的 Azure Synapse Analytics 环境。

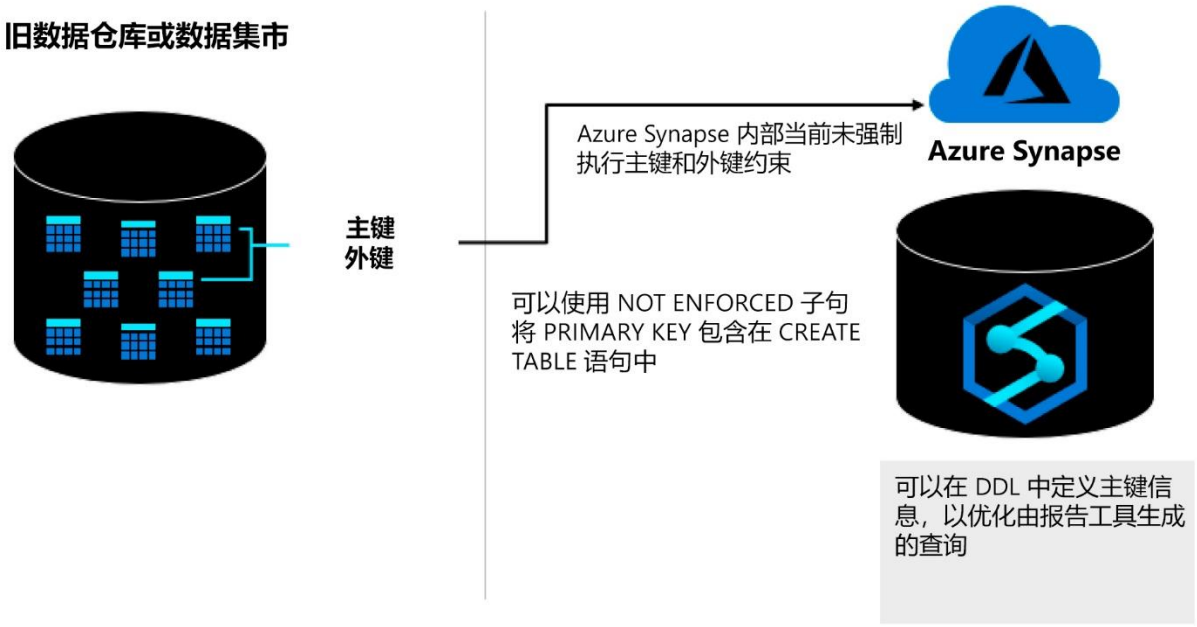


图 7 - 完整性约束差异

在后面各节中，我们将全面介绍在从旧数据仓库迁移到 Azure Synapse Analytics 的过程中，如何解决其他常见的 SQL 不兼容问题。

常见 SQL 不兼容问题和解决方案

SQL 数据定义语言 (DDL) 的差异和解决方案

专有表类型

- 在旧系统上，确定任何专有表类型的使用情况
- 解决方案：迁移到 Azure Synapse Analytics 中的标准表
- 对于时序，在日期/时间列上进行索引或分区
- 将需要向相关的临时查询中添加额外的筛选

视图

- 在旧系统上，从目录表和 DDL 脚本中识别视图
- 带有专有 SQL 扩展或函数的视图必须重新编写
- Azure Synapse Analytics 还支持实例化视图，并将自动维护和刷新这些视图

空值

- 在旧 SQL 数据库中，对 NULL 值的处理方式可能不同
  - 例如，在 Oracle 中，空字符串等效于 NULL 值

- 某些 DBMS 具有用于处理 NULL 的专有 SQL 函数
  - 例如，Oracle 中的 NVL
- 生成 SQL 查询以测试 NULL 值
- 测试包含可空列的报告

扩展的 SQL 差异和变通办法

SQL 扩展	说明	迁移方法
用户定义函数	可以包含任意代码 可以用各种语言（如 Lua、Java）进行编码 可以在 SQL SELECT 语句中调用，方式与使用内置函数（如 SUM() and AVG()）一样	使用 CREATE FUNCTION 并在 T-SQL 中重新编码
存储过程	可以包含一个或多个 SQL 语句以及围绕这些 SQL 语句的过程逻辑 以标准语言（如 Lua）或专有语言（如 Oracle PL/SQL）实现	在 T-SQL 中重新编码 一些工具可以帮助迁移 例如 Datometry、WhereScape
触发器	Azure Synapse 不支持	等效功能可以通过使用 Azure 生态系统的其他部分来实现。例如，对于流式输入数据 Azure 流分析
数据库内分析	Azure Synapse 不支持	大规模运行机器学习模型等高级分析以使用 Azure Databricks 或者，迁移到 Azure SQL 数据库并使用 PREDICT 函数
地理数据类型	Azure Synapse 不支持	将地理空间数据（如纬度/经度）和常用格式（如 WKT（熟知文本）和 WKB（熟知二进制））存储在 VARCHAR 或 VARBINARY 列中，并由地理空间客户端工具直接访问