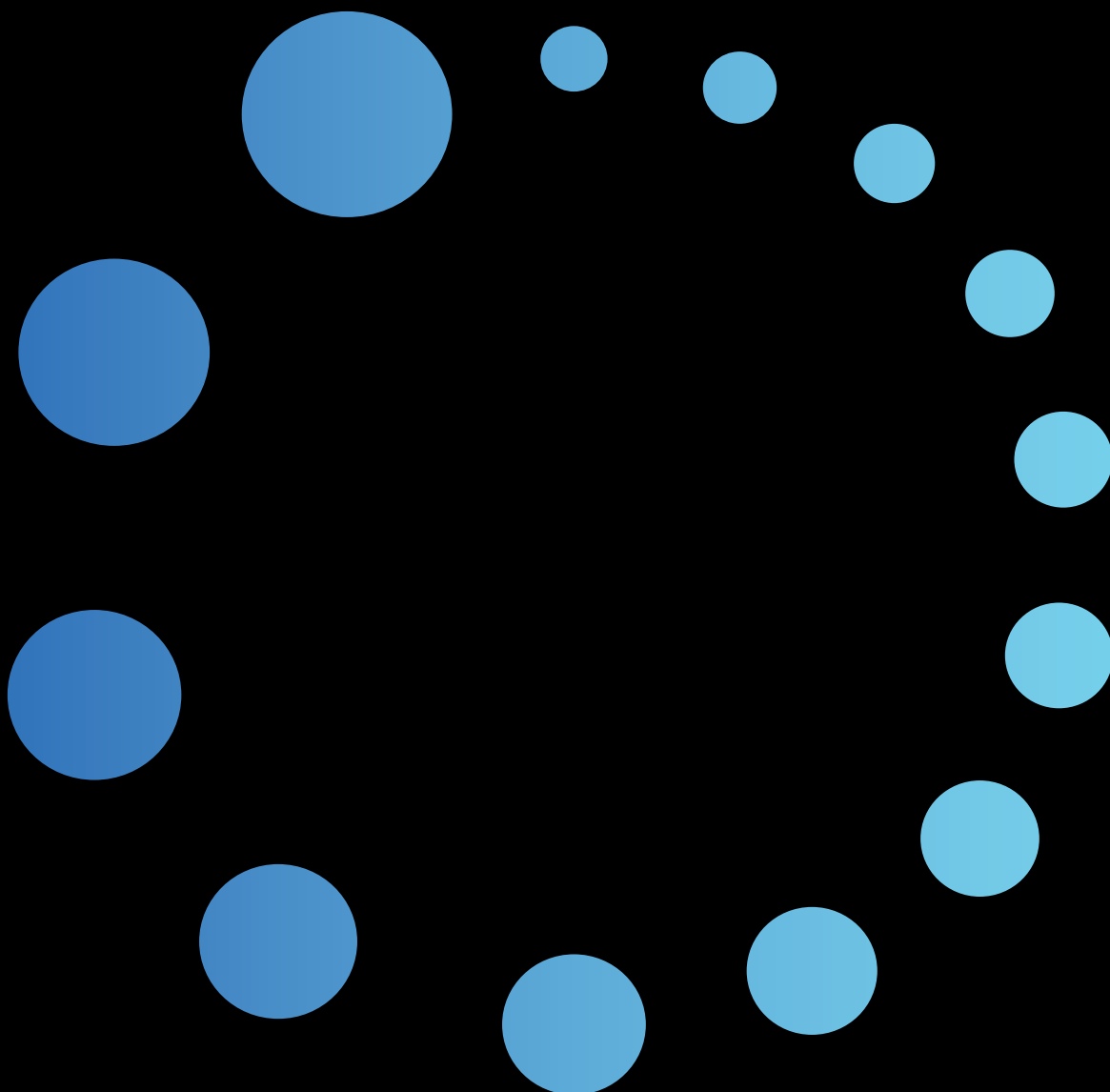


# Power BI 专业人员指南之 Azure Synapse Analytics



# 摘要

本指南将向 Power BI 从业者介绍 Azure Synapse，这是一种无限的分析服务，将企业数据仓库和大数据分析结合在一起。

从表面上看，Azure Synapse Analytics 是在 Azure SQL 数据仓库的基础上演变而来的。但是，它不仅仅是在 SQL 数据仓库更新中增加了几个新功能这么简单。Azure Synapse 还代表一种业内独一无二的现代、全面且统一的分析方法。Azure Synapse 是一种包含以前孤立功能（如数据集成、数据仓库和大数据处理）的集成云原生服务，可支持 Power BI 专业人员在各种用例中提供项目所需的规模、性能和成本管理。

本指南探讨了 Power BI 与作为数据源和开发平台的 Azure Synapse 的深度集成，并确定了将 Azure Synapse 用作新解决方案和现有解决方案的主要好处。

## 04 /

### Azure Synapse Analytics 简介

05 Azure Synapse SQL

## 06 /

### Azure Synapse 给 Power BI 带来的好处

06 单一的可信来源  
06 大规模 DirectQuery  
07 集中式安全性  
09 团队协作  
10 数据准备  
10 分页报表的灵活性

## 11 /

### 使用 Azure Synapse 构建 Power BI 解决方案

11 访问 Azure Synapse 工作区  
13 工作区与资源访问  
13 在 Azure Synapse Studio 中连接到 Power BI  
15 通过 Azure Synapse Studio 创建 Power BI 数据集  
17 在 Azure Synapse Studio 中创建报表  
20 创建分页报表  
20 Power BI 数据集与 SQL 池  
21 连接至 SQL 资源  
24 开发数据流  
27 AI 预测分析集成  
27 复合模型和聚合  
28 通过聚合实现目标性能  
31 表存储模式  
32 混合源与连接

## Azure Synapse Analytics 简介

Azure Synapse 是一款端到端的云原生分析平台，它将数据接收、数据仓库和大数据整合到一个服务中。借助它，你可以自由地使用无服务器资源或预配资源大规模地按条件查询数据。它将数据仓库和大数据分析结合在一起，呈现统一的体验，用于接收、准备、管理和提供数据，以满足即时的商业智能和机器学习需求。

Azure Synapse 平台与链接服务相集成，这些服务包括 Power BI、Azure 机器学习和 Azure 数据共享。用户可以在 Azure Synapse Studio 中开发交互式 Power BI 报表和企业级语义模型，该程序是一种用于开发和管理各种 Azure Synapse 工件的全新通用 Web 门户。

借助以下体系结构，Azure Synapse 可以接收结构化数据和非结构化数据，并提供提取 - 转换 - 加载 (ETL)、大数据和数据仓库技术，而所有这些功能都通过一项统一服务来实现：

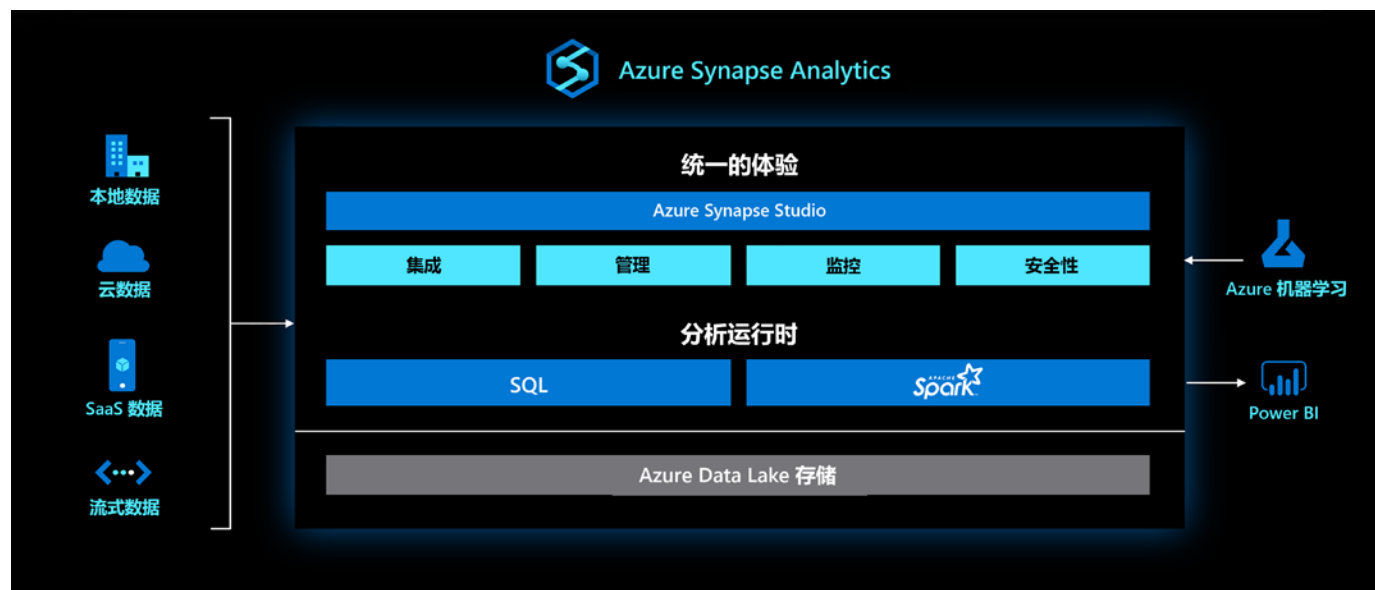


图 1：Azure Synapse Analytics

## Azure Synapse SQL

敏捷性以及对于数据湖中大型数据集的快速数据探索功能是现代数据平台中具有极高价值的功能。[Azure Synapse SQL](#) 是用户使用 SQL 技术分析数据的一站式地点。

借助 Synapse SQL，你可以自由地使用以下两种形式查询数据：

- 附带 SQL 池的预配数据仓库
- 对数据湖进行无服务器查询

为了满足按需计算能力的需求，Synapse SQL 使数据工程师能够在无需预配任何基础结构的情况下运行无服务器查询。

在下面的 Azure Synapse Studio 图片中，无服务器终结点用于对存储在 Azure Data Lake Storage 中的 Parquet 文件集合执行查询：

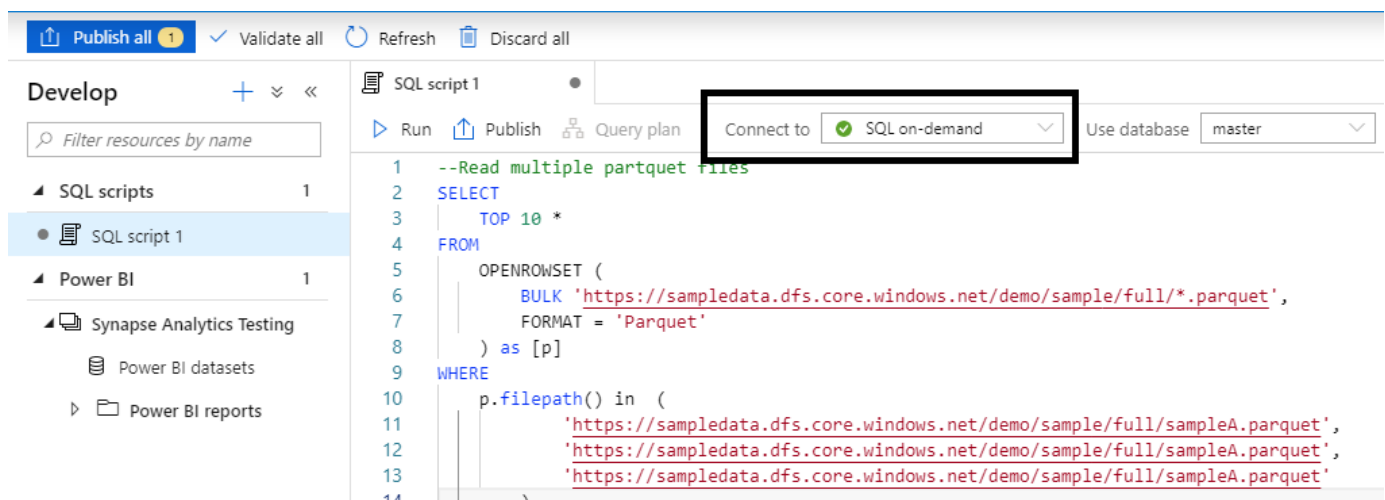


图 2：按需 SQL 分析

通过 Azure Synapse 工作区中提供的按需 SQL 终结点，数据开发人员还可以将 SQL Server Management Studio (SSMS) 和 Azure Data Studio 等工具与按需计算引擎结合使用。

Azure Synapse 非常灵活，可以预配并弹性扩展计算资源池，也可以将无服务器功能用于 Azure SQL 数据库的按需计算资源。借助 Azure Synapse，组织可以显著简化数据环境的管理，并让数据专业人员团队（包括数据工程师、数据科学家、BI 专业人员和 IT 管理员）进行合作，从而改善协作并提高工作效率。

## Azure Synapse 给 Power BI 带来的好处

负责提供解决方案以形成可行见解、打造数据探索体验的 Power BI 专业人员可以在几个不同方面从 Azure Synapse 中受益。以下各节总结了将 Azure Synapse 用于新的和现有 Power BI 解决方案时将带来的一些机会和好处。

### 单一的可信来源

在成功的原有 Azure SQL 数据仓库的基础上，组织可以将 Azure Synapse 部署为 Power BI 和其他应用程序的经过验证的单一可信来源。通过利用存储在预配 SQL 池中经过正式认可的数据仓库对象，Power BI 开发人员和 Power BI 解决方案的使用者可以确信所提供的数据在质量、一致性和准确性方面经过验证。

例如，Power BI 管理员和其他 BI 利益相关者可能坚持认为，只有专门针对 Azure Synapse 构建的 Power BI 数据集才有资格被标记为 Power BI [认证数据集](#)或发布到生产 Premium 容量中。访问其他不太受信任的源（包括文件和旧系统）的 Power BI 数据集可能仅适用于较小的临时场景。

### 大规模 DirectQuery

大多数支持适用于 Power BI 的 DirectQuery 连接的数据源一直以来都难以提供企业 Power BI 解决方案所需的高用户并发性和低查询响应时间。Power BI 报表专为打造交互式数据探索用户体验而设计，这意味着每个用户会话都需要执行大量查询，才能实时更新不同的可视化效果。随着并发用户参与的数量增长到数千个（例如广泛采用的企业 BI 解决方案），AWS Redshift 和 Google BigQuery 等常见数据仓库系统要么将传入的查询放入队列中，从而延迟执行，要么强制用户查询失败。

Azure Synapse 支持性能优化（包括实例化视图和结果集缓存），从而使 DirectQuery 模型更加适合海量源数据集并支持数千个并发用户。借助独立的弹性计算和存储资源，IT 专业人员可以应用标准 Azure 资源管理实践来扩展预配的 SQL 池，以符合工作负载的要求。例如，可以安排简单的 Azure 自动化运行手册，以便在上午 8:00 将 SQL 池扩展到 DW3000 数据仓库服务级别，为 Power BI 的峰值使用提供支持，然后在下午 3:00 缩减到 DW1000 级别，从而实现成本管理。

Azure Synapse 还为 Power BI 模型开发提供了良好的替代方案。假设采取了数据源、模型和报表层的建议做法，那么有权访问 Azure Synapse 的 Power BI 专业人员可以与其他数据团队协作来大规模部署 DirectQuery 模型。作为这种协作的一个示例，数据工程师可以分析 Power BI 解决方案访问的查询模式和源表，并通过保留（存储和检索）所需的业务逻辑和实现[有序群集列存储索引](#)来优化这些结构。

*组织自然希望避免与导入模型的计划更新和管理开销相关的数据移动或复制。但是，对大规模处理性能的需求促使许多组织追求大型内存中模型，以将其部署到具有足够 RAM 的资源，如 Azure Analysis Services。出于并发性和 BI 性能要求的原因，2017 年，对于 Azure SQL 数据仓库使用 Power BI DirectQuery，SQL 客户咨询团队将其确定为一种反模式。*

## 集中式安全性

Power BI 专业人员通常在数据模型中实现行级安全角色并控制哪些用户或组有权访问工作区、应用程序和数据集，从而为解决方案提供保护。Azure Synapse 在其他安全功能层（包括透明数据加密）为用户和组提供行级和列级安全性。虽然 Power BI 中的行级安全性非常可靠，并且通常为包含导入数据的数据模型所必需，但企业 IT 组织通常更喜欢充分利用数据仓库来实现查询处理（即 DirectQuery）和数据安全性。

由于 Power BI 身份验证是通过 Azure Active Directory (Azure AD) 处理的，并且 Azure Synapse 支持并建议使用 Azure AD 身份验证，因此组织可以选择在 Azure Synapse 的数据层为 Power BI 解决方案强制实施数据安全。Power BI 用户的身份及其在 Azure AD 中特定安全组中的成员身份可以传递给 Azure Synapse，以便实施在 Azure Synapse 中为指定组和源对象定义的安全策略。

如下所示，Power BI 开发人员可以轻松地配置发布的基于 Synapse 的 DirectQuery 模型，从而将用户的凭据传递给数据源：

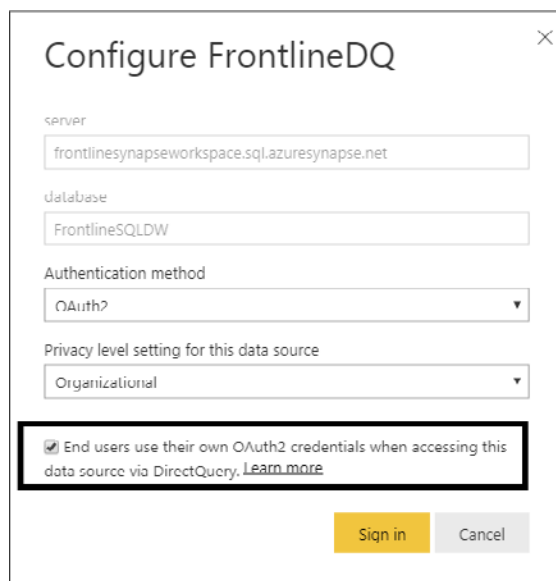


图 3：DirectQuery 连接的单一登录

凭借由 Azure Synapse 处理的数据安全策略，在完整 DirectQuery 模式下消除了 Power BI 数据模型得不到适当保护的风险。此外，由于大型 Power BI 环境通常涉及许多处于不同范围和成熟度级别的数据模型，因此这些模型的开发人员和所有者不必复制和测试行级安全角色。

每个表包含多个存储模式（如 DirectQuery 和导入）和（可以选择）多个数据源的复合模型无法通过单一登录到单个 DirectQuery 数据源的方式获得保护。例如，为了优化常见查询的性能，Power BI 团队可以选择导入聚合表，同时让详细的大型表保持在 DirectQuery 模式下。本指南末尾包含有关复合模型和聚合的更多详细信息。



## 团队协作

一直以来，不同团队运用不同技术协作来实现一个共同目标时都面临着自身固有问题，阻碍了商业智能的发展。例如，负责数据转换流程的团队通常不熟悉这些流程对 Power BI 等下游应用程序造成的影响。能够在团队之间清晰地沟通，对于及时实现预期成果至关重要。

Azure Synapse 将数据工具和团队整合在一起，提高了公司内部的透明度和工作效率。具体来说，所有使用 Azure Synapse 的团队都可以访问 Azure Synapse Studio 中的通用用户界面，因此所有用户（无论使用何种主要工具或掌握何种技能）都能查看和分析相同的数据。

在 Azure Synapse Studio 中，用户可通过 Azure 中的 Azure Synapse 工作区访问基于 Web 的门户，并可获得多种数据开发体验，包括 Power BI 报表和数据集：

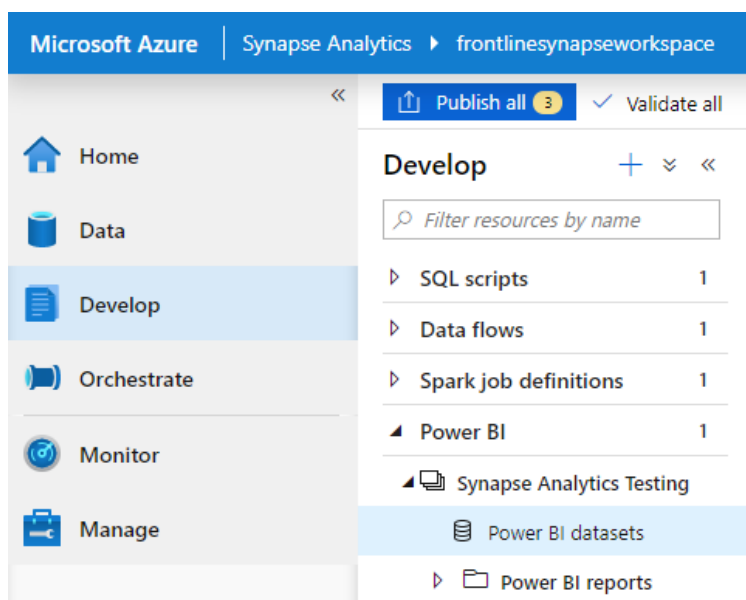


图 4：Azure Synapse Studio

例如，负责用于加载 SQL 池的数据管道的团队通常会使用“Orchestrate”（编排）页面，而数据科学家、大数据工程师和 Power BI 开发人员则可以使用“Data”（数据）和“Develop”（开发）页面来访问与其角色相关的工具和工件。借助 Azure Synapse Studio，团队和工具统一呈现在一个通用门户中，因而可实现比以往更高效的协作。

## 数据准备

Power BI 解决方案通常包含嵌入式数据转换过程和集成过程，如包含 Power Query、数据流或计算的 DAX 列和表。虽然这些转型过程适用于短期和小规模场景，但可能会给解决方案的可扩展性和可持续性带来重大风险。Azure Synapse 强大的数据处理工具以及 Azure Synapse 数据工程师的专业知识可以满足 Power BI 解决方案的数据准备需求。

Azure Synapse 包括 Azure 数据工厂的企业级数据转换和编排功能。数据工程团队可以构建可靠的数据管道、Synapse Spark 作业或 SQL 存储过程以满足各种数据准备需求，这样，Power BI 开发人员就无需在解决方案中处理这些要求。借助 Azure Synapse 丰富的数据处理功能，Power BI 开发人员能够将精力重新投入到解决方案的其他方面，如分析、用户体验和分发。

## 分页报表的灵活性

使用 Power BI Report Builder 开发的分页报表是 Power BI 环境中的一项重要服务，尤其是考虑到它在导出或打印大量数据方面所具有的优势。针对详细数据级别（如单个销售订单）的分页报表可以在更加聚合的级别为 Power BI 报表和仪表板提供良好的补充。此外，由于能够访问相同的 SQL 查询，Power BI Report Builder 中提供的精细控制让用户可以大规模复制由其他企业报表工具开发的几乎任何报表。

由于完全支持 Azure Synapse（包括基本身份验证和单一登录身份验证方法），Power BI 分页报表开发人员可以选择直接针对预配的 SQL 池创建包含常见 T-SQL 查询的报表。此选项对于加快将包含 SQL 查询的旧 SQL Server Reporting Services (SSRS) 以及其他基于 SQL 的报表工具迁移到 Power BI 非常重要。



在 Azure Synapse 工作区的 “Manage”（管理）边栏选项卡中，工作区的管理员可以向工作区中的资源和工件添加具有不同权限级别的用户或 Azure AD 安全组。

管理员应注意，用户要访问 Azure Synapse Studio，就需要将用户或组映射到工作区本身的某个角色，而不是工作区 Azure 资源。在图 6 中，通过工作区的 “Access control”（访问控制）页面向用户和用户安全组（Power BI 开发人员）授予 Azure Synapse 工作区的管理员角色：

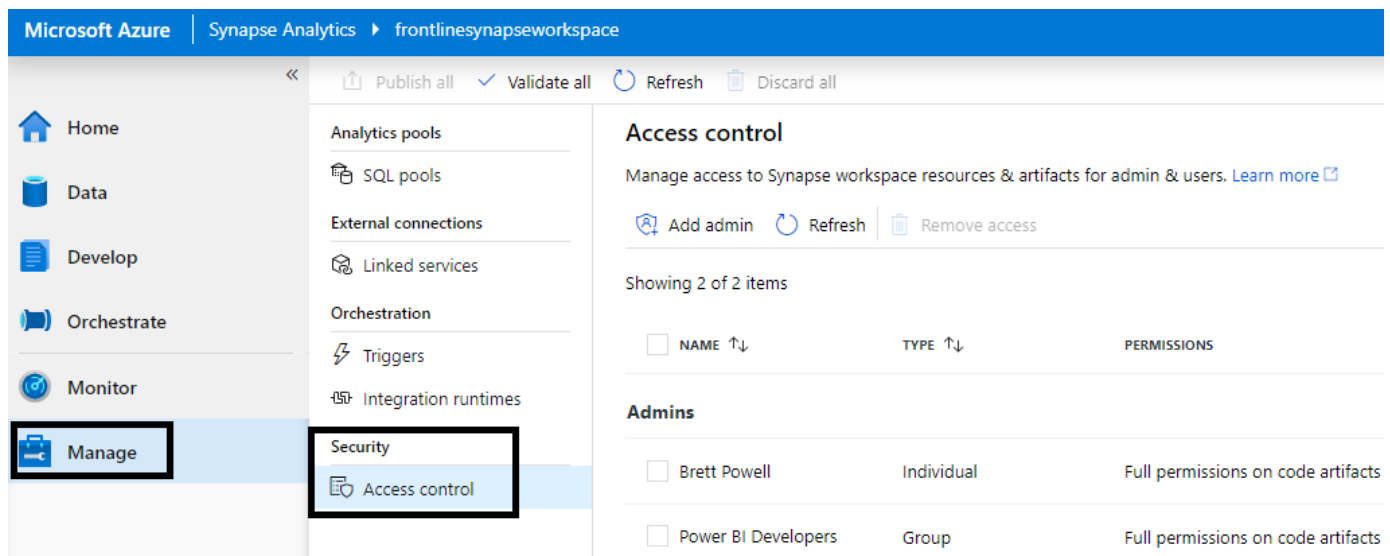


图 6：工作区访问控制

提供用户访问权限的一种常见的简便方法是将用户安全组映射到一个内置的 RBAC 角色，例如专门指定资源的参与者。授予权限的另一种更精细的常见方法是创建和管理仅包含必需的 [Azure 资源操作](#) 的自定义角色定义。具体来说，管理 Azure 资源访问权限的管理团队可以通过 Azure PowerShell (Get-AzProviderOperation) 来确定 Azure Synapse 的可用操作，并仅向 Power BI 开发所需的操作授予自定义角色。

## 工作区与资源访问

将对 Azure Synapse 工作区的访问权限与对在工作区中配置的资源（如 SQL 池）的访问权限区分开来非常重要。仅当 Power BI 用户将在 Azure Synapse Studio 中开发 Power BI 内容或利用 Azure Synapse Studio 中的其他功能（如使用 SQL、Python 或其他受支持的语言开发脚本或笔记本）时，才需要访问 Azure Synapse 工作区（如上一节所述）。

通常，负责针对数据仓库创建数据模型、报表和仪表板的 Power BI 开发人员仅被授予读取源数据库的访问权限。大多数企业 IT 组织遵循严格的最低权限策略来管理对 Azure 资源的访问，因此，至少在初次启动时，可能会继续限制 Power BI 开发人员只能访问所需的数据源，如 SQL 池中的数据库。BI 和云体系结构团队可以确定要获得本指南中介绍的 Azure Synapse Studio 为 Power BI 用户带来的好处，是否需要提供这种额外的访问权限。例如，如果 Power BI 开发人员还定期编写 SQL 查询并 / 或与数据工程师协作，那么访问 Azure Synapse Studio 可能尤其有利。

## 在 Azure Synapse Studio 中连接到 Power BI

授予对 Azure Synapse 工作区的访问权限之后，就需要建立从 Azure Synapse 工作区到相关 Power BI 应用工作区的连接。在 Azure Synapse 中，与这些工作区的连接被定义为链接服务，用户可以借助这些服务直接在 Azure Synapse Studio 中创建和修改 Power BI 工作区内容。

有两种方法可用于建立指向 Power BI 的链接服务。最直观的方法是单击工作区的主页窗格中的“Visualize”（可视化）图标，如图 7 所示：



图 7：Synapse 工作区主页窗格

“Visualize”（可视化）图标可启动一个窗体，使用户能够输入要连接到的 Power BI 应用工作区以及链接服务的名称和描述。例如，在图 8 中，创建了一个新的链接服务，并连接到 Power BI 中的“Synapse Analytics Testing app workspace”（Synapse Analytics 测试应用工作区）：

Connect to Power BI

**i** Choose a name for your linked service. This name cannot be updated later.

Connect a Power BI workspace to create reports and datasets from data in your workspace.  
[Learn more](#)

Name \*

PBI Synapse Analytics Testing Workspace

Description

Power BI direct query testing over Synapse SQL pool

Workspace name \*

Synapse Analytics Testing (b96f5546-1c68-4949-974e-40e1e8e8509e)

Annotations

+ New

图 8：为 Power BI 创建链接服务

创建链接服务后，Azure Synapse 工作区将具有针对 Power BI 应用工作区的读写功能。可以通过“Linked services”（链接服务）页面查看工作区的所有链接服务，该页面通过“Manage”（管理）窗格（工具箱图标）访问，如图 9 所示：

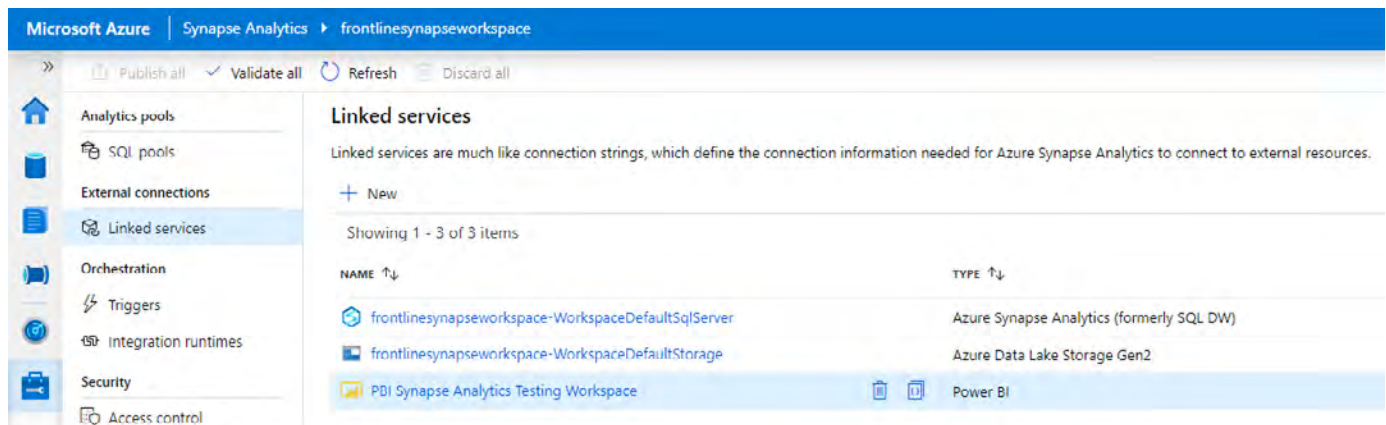


图 9：管理链接服务

创建指向 Power BI 的链接服务的另一种方法是通过“Linked services”（链接服务）页面中的“New”（新建）图标来完成，如图 9 所示。截至撰写本文时，只能从 Azure Synapse 工作区创建指向 Power BI 的单个链接服务。因此，如果需要访问另一应用工作区，则当前需要删除现有的链接服务，并为另一应用工作区创建新的链接服务。

## 通过 Azure Synapse Studio 创建 Power BI 数据集

在 Power BI 中将分析数据模型定义为数据集是 BI 解决方案和整体 BI 体系结构的核心，因为对于许多报表、仪表板和临时分析场景，它们可以作为经过认证的高性能来源。对于 Azure Synapse，Power BI 开发人员可以更轻松地与其他数据专业人员就影响其模型的数据源和流程进行协作。

指向 Power BI 应用工作区的链接服务部署就绪后，即可借助 Azure Synapse Studio 轻松创建 Power BI 数据集文件 (.pbids)，其中包含 Azure Synapse 中配置的所需数据源的元数据。在 Power BI Desktop 中打开数据集文件将以熟悉的 Power Query 编辑器体验显示数据源对象。

如图 10 所示，与链接服务关联的工作区将显示在“Develop”（开发）窗格中，并提供在此工作区中创建新数据集的选项：

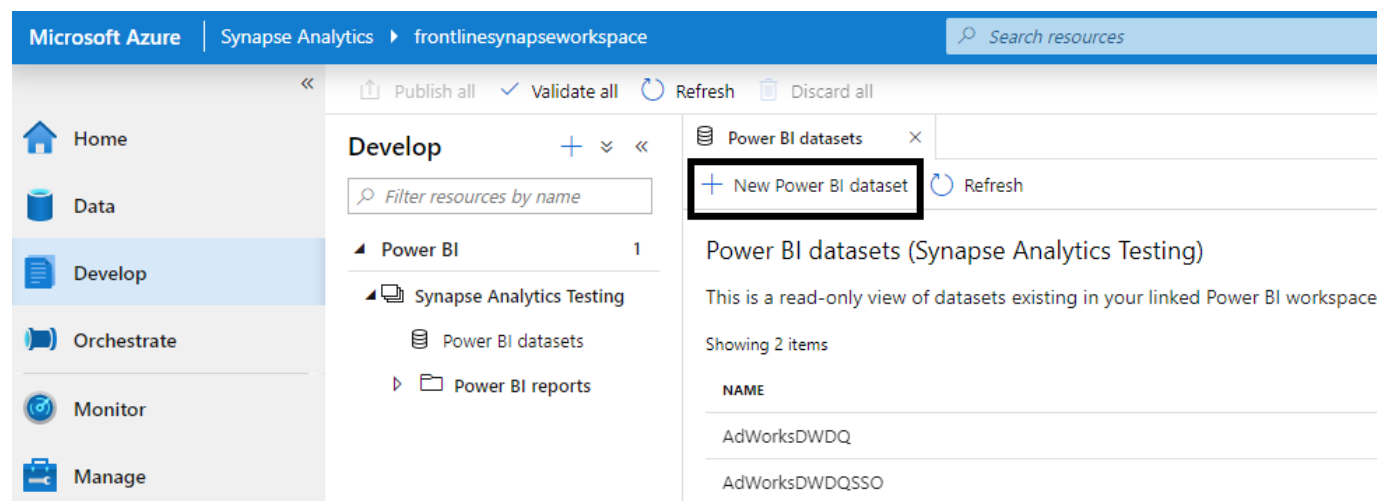


图 10：创建 Power BI 数据集



“New Power BI dataset”（新建 Power BI 数据集）表单需要从工作区中选择数据源，并在选择源后提供下载数据集文件的链接。在图 11 中，托管在预配 SQL 池资源中的 FrontlineSQLDW 数据库被标识为新 Power BI 数据集的来源：

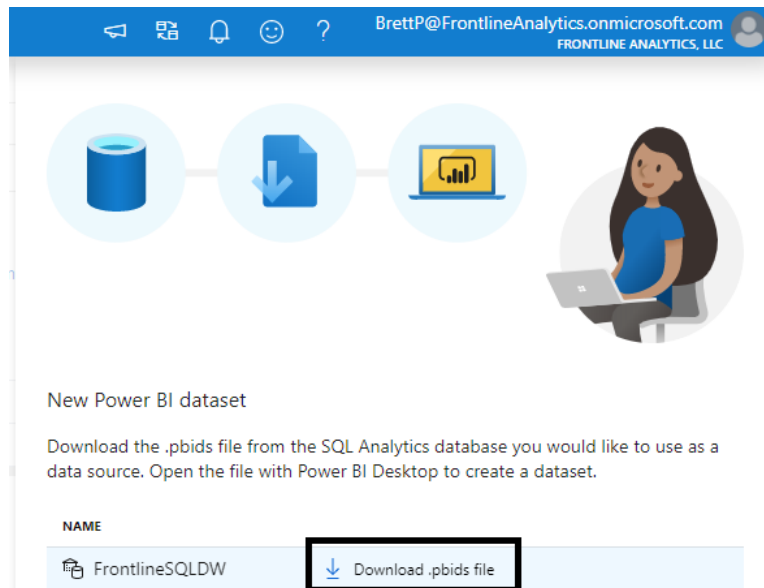


图 11：下载数据集文件

使用 Power BI Desktop 在本地打开 .pbids 文件会自动启动指定数据源的导航器，如图 12 所示：

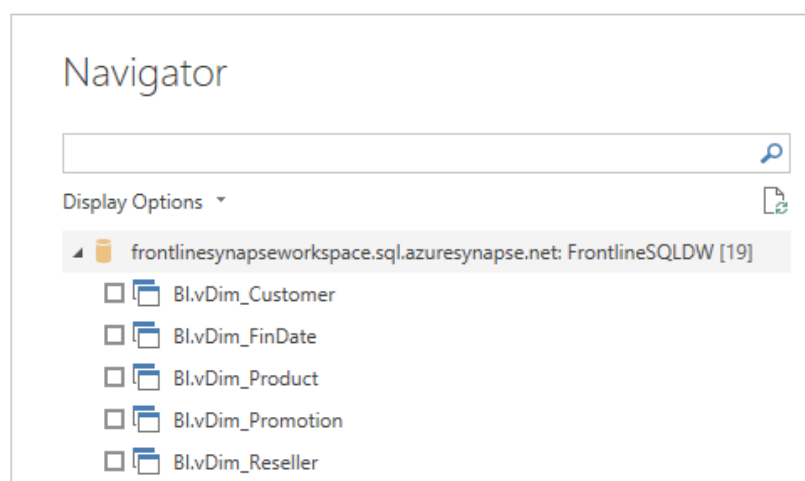


图 12：在 Power BI Desktop 中打开 .pbids 文件



然后，Power BI 模型开发人员可以使用常见的 Power BI Desktop 控件来修改表的存储模式，并进一步开发模型的关系、指标和其他元数据。可以将新模型发布回在 Azure Synapse 中配置为链接服务的同一应用工作区，也可以发布到用户拥有权限的 Power BI 中的任何其他应用工作区。

作为从 Azure Synapse 工作区下载数据集文件 (.pbids) 的替代方法，此示例中的数据建模师还可以使用 Power BI Desktop 中的“Get Data”（获取数据）体验来定义自己的源连接。具体来说，将选择在 Azure 数据源组中发现的 Azure SQL 数据仓库连接器，并要求用户手动输入服务器和数据库名称。

### 在 Azure Synapse Studio 中创建报表

可以直接在 Azure Synapse Studio 中创建和编辑 Power BI 交互式报表。在此示例中，已创建了名为 FrontlinedQ 的数据模型并且发布到 Synapse Analytics 测试 Power BI 应用工作区 - 也就是在 Azure Synapse 中配置为链接服务的工作区。目的是将此模型用作新的 Power BI 交互式报表的来源。

如图 13 所示，单击 Azure Synapse Studio 的“Develop”（开发）页面顶部的加号 (+) 图标可将 Power BI 报表显示为可开发的工件：

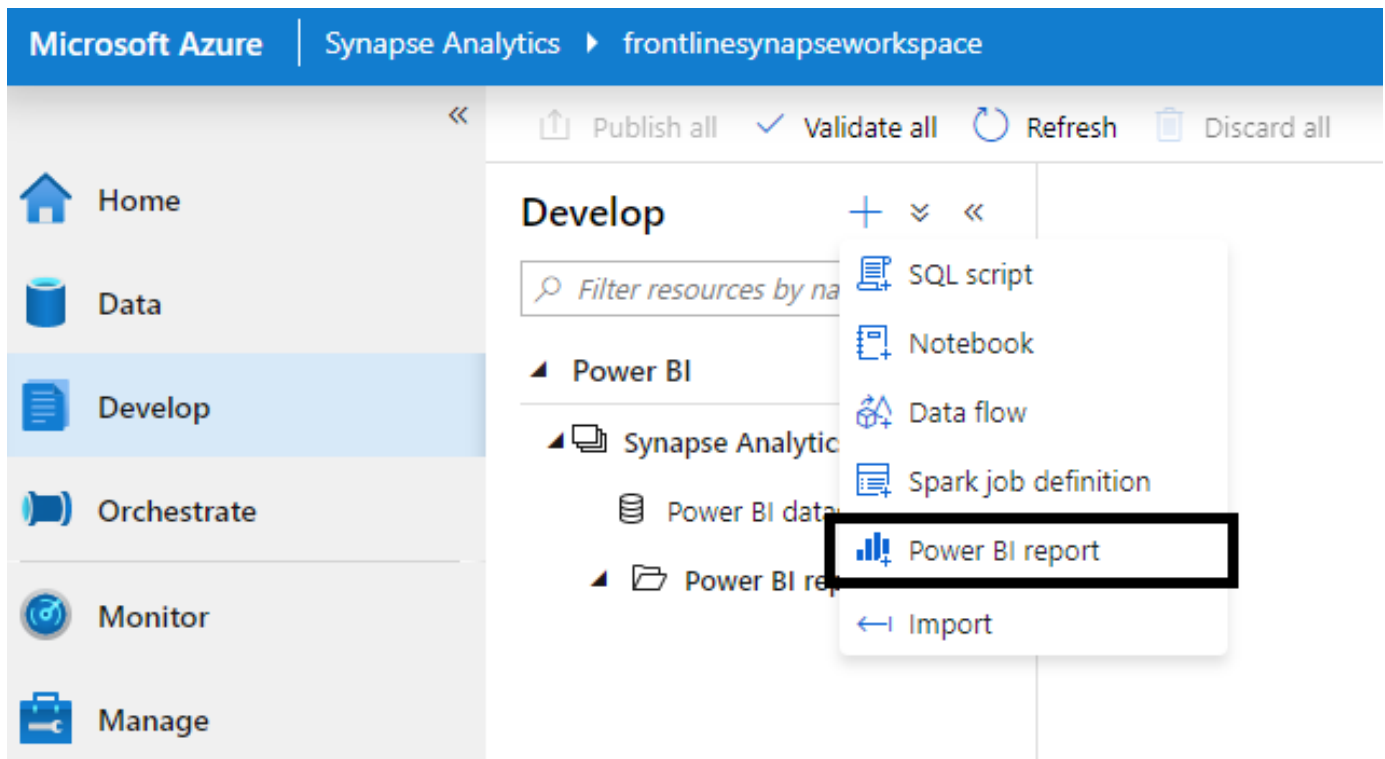


图 13：新建 Power BI 报表

选择 Power BI 报表后，开发人员必须确定将用作新报表源的 Power BI 数据集，如图 14 所示：

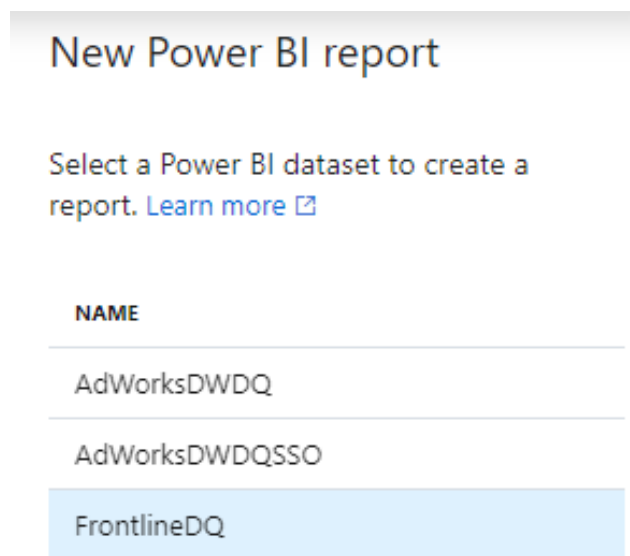


图 14：报表的可用数据集

最后，单击数据集选择表单中的“Create”（创建），就会在 Azure Synapse Studio 中调出基于 Web 的 Power BI 报表开发工具。在图 15 中，在 Azure Synapse Studio 中创建了一个名为“Internet Sales”（互联网销售）的 Power BI 报表，并将 FrontlineDQ 数据集作为来源：

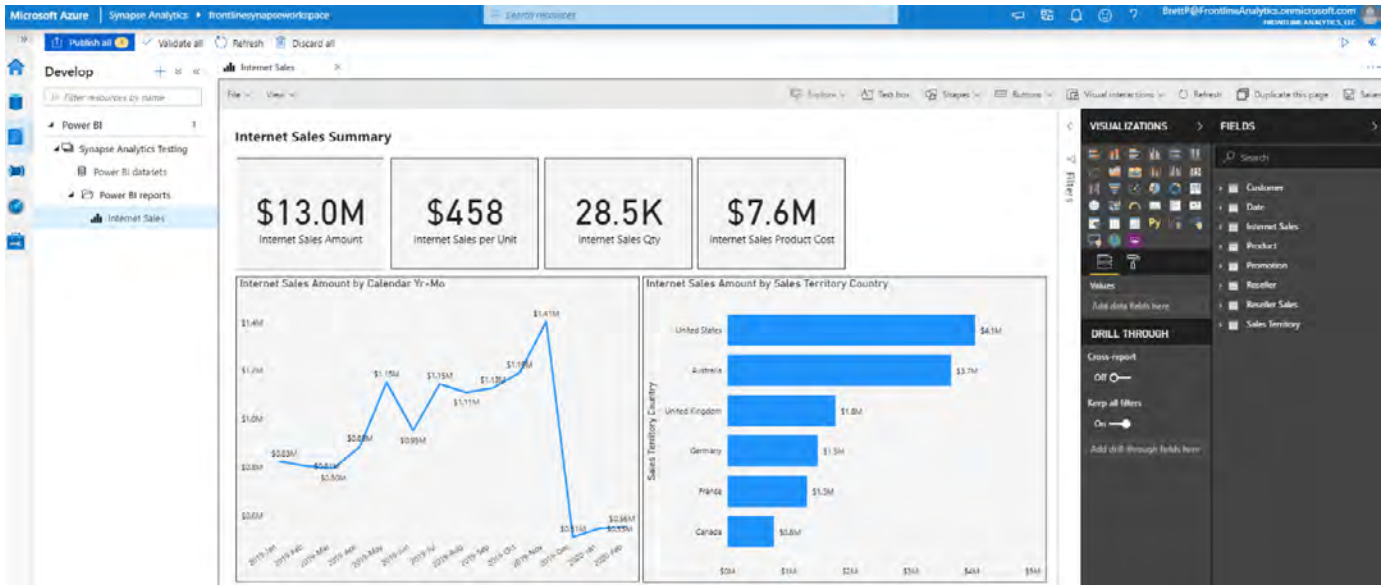


图 15：在 Azure Synapse Studio 中创建 Power BI 报表

如图 15 所示，报表作者可以看到 Power BI Desktop 中熟悉的“FIELDS”（字段）、“VISUALIZATIONS”（可视化效果）和“Filters”（筛选器）窗格。同样，Azure Synapse Studio 中的 Power BI 报表提供了与 Power BI 和 Power BI Desktop 相同的筛选和交叉高亮显示交互体验。

可以在 Azure Synapse Studio 中保存完成或修改的 Power BI 报表，以便 Power BI 应用工作区的用户可以访问这些报表，进行进一步开发或与报表的使用者共享。（截至撰写本文时，将报表副本下载为 .pbix 文件的功能是 Power BI 所独有的。）

虽然在 Azure Synapse Studio 中制作 Power BI 报表很方便，尤其是对于定期使用该工作区的 BI 开发人员，但应注意的是，这种基于 Web 的开发体验仅提供 Power BI Desktop 的一部分报表制作功能。例如，Web 体验目前不支持视觉对齐和发布格式选项以及编写报表范围内的指标的功能。

截至撰写本文时，由于 Azure Synapse Studio 的单个应用工作区链接服务存在限制以及尚未正式提供 Power BI 中的共享数据集这一事实，在 Azure Synapse Studio 创建的报表将基于与链接服务相同的工作区中的数据集，并将保存到同一工作区。

## 创建分页报表

与交互式 Power BI 报表不同的是，（截至撰写本文时）[分页报表](#)无法在 Azure Synapse Studio 中创建或编辑。但是，正如在 [Azure Synapse 给 Power BI 带来的好处](#) 一节中所述，分页报表可以利用 SQL 池资源（以前的 Azure SQL 数据仓库）作为安全、可扩展和高性能的数据源。重要的是，借助此选项，Power BI 开发人员能够使用熟悉的 T-SQL 语法，并可能使用现有的报表 SQL 查询和存储的过程。

## Power BI 数据集与 SQL 池

在开始制作分页报表之前，报表开发人员和商业智能团队应评估将现有 Power BI 数据集或 Analysis Services 模型作为报表源的可行性。如果数据模型可包含或可以对其进行更新以包含分页报表所需的数据和逻辑，那么在架构上，最好将模型中定义的关系和计算重用于交互式报表和分页报表。

但是，在某些情况下（如非常复杂或不同的报表要求），与尝试编写针对模型的 DAX 查询相比，直接使用 SQL 池资源（Azure SQL 数据仓库）可能是更好或者必需的选择。可将包含可能比较复杂的多步逻辑流程或公开用户筛选参数的存储过程创建为 SQL 池资源中的可重用数据库对象。在选择 SQL 池作为报表源时，使用存储的过程和丰富的 T-SQL 语法的功能可能是一个重要因素。

## 连接至 SQL 资源

要对 SQL 池资源创建分页报表，报表开发人员必须首先在 Power BI Report Builder 中创建数据源。配置此数据源时，报表开发人员可以连接到 Azure SQL 数据仓库数据源。他们需要使用 SQL Server 身份验证凭据，因为 Power BI Report Builder 目前不支持 Azure AD 身份验证。

在图 16 中，定义的数据源的连接类型为 “Azure SQL Data Warehouse”（Azure SQL 数据仓库），名称与 SQL 池中的数据库名称相同：

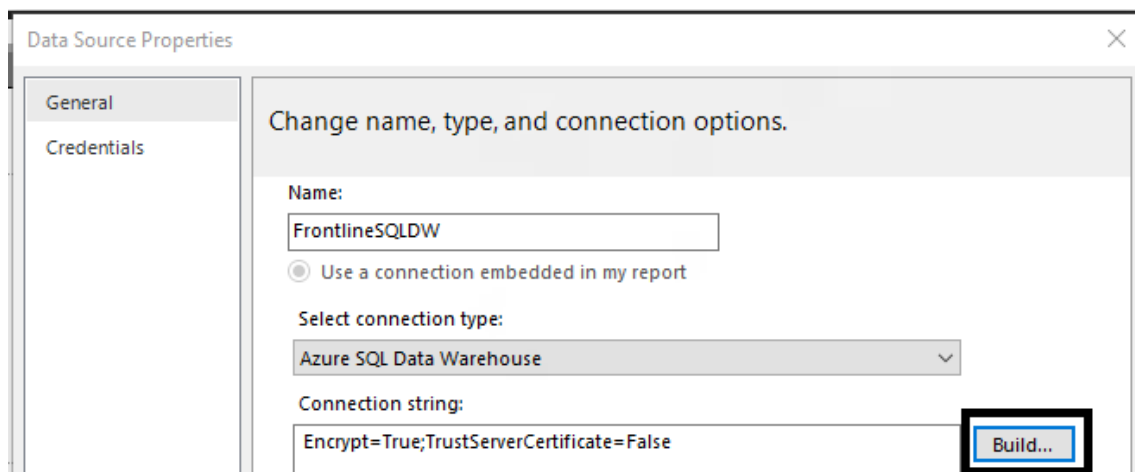


图 16：分页报表数据源

单击 “Data Source Properties”（数据源属性）页面中的 “Build”（构建），将显示用于提供要使用的服务器名称、数据库名称和凭据的选项。此身份验证信息不包含在分页报表文件 (.rdl) 中，应由管理 SQL 资源的团队以安全的方式提供。

根据图 17 所示，需要 SQL Server 身份验证凭据才能从 Power BI Report Builder 连接到 Azure SQL 数据仓库资源：

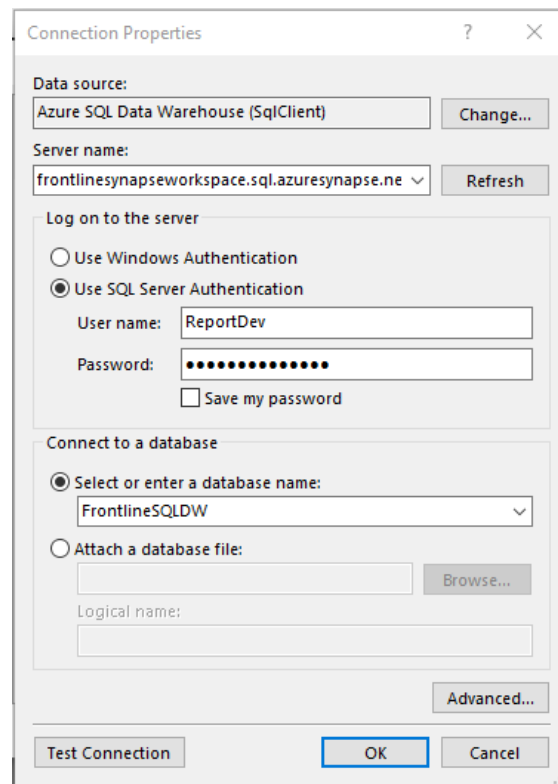


图 17：数据仓库连接属性窗口

创建和配置数据源后，分页报表开发人员可以通过 SQL 语句或通过引用源数据库中的存储过程来定义报表的数据集。在图 18 中，FrontlineSQLDW 资源中的 BI.spCustomersSalesOrders 存储过程对象被用作报表中的数据集：

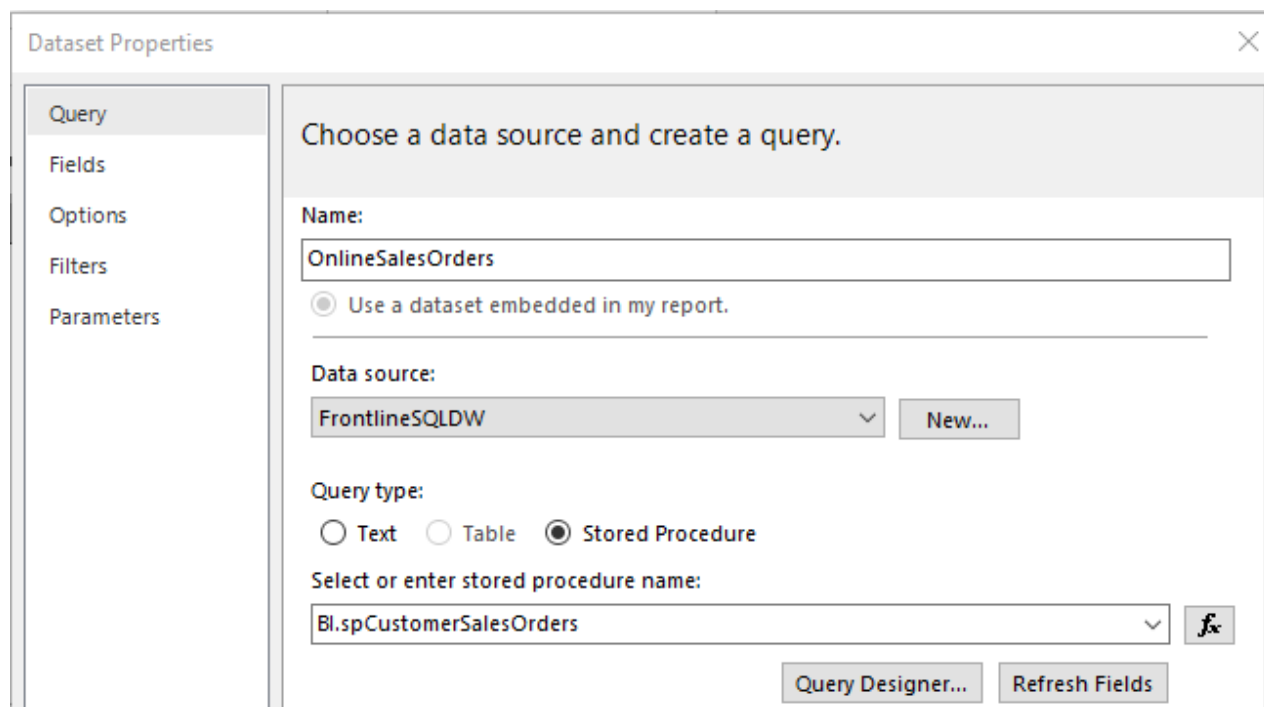


图 18：对分页报表使用存储过程

报表作者可以选择“Text”（文本），然后只需输入或粘贴到现有的 SQL 语句中，也可以打开“Query Designer”（查询设计器）以使用图形界面来帮助定义查询。通常建议尽可能在分页报表中使用存储过程，以提高报表解决方案的可管理性。

将分页报表发布到 Power BI 服务后，报表作者可以选择配置报表的身份验证，以传递查看报表的用户的身分。

在 Power BI Report Builder 中针对 Analysis Services 模型和 Power BI 数据集创建的分页报表会针对这些源发出 MDX 查询。虽然可以定义自定义 DAX 查询并 / 或使用查询设计器图形界面，但 Power BI Report Builder 针对基于 DAX 的报表创作所提供的支持相对有限。例如，在报表中针对表格模型或 Power BI 数据集简单配置多选参数需要使用自定义 DAX 代码的重要解决方法。

## 开发数据流

Power BI 数据流是一种针对商业用户的自助 ETL 功能，专门在 Power BI 中创建和管理。但是，与分页报表相似，Azure Synapse 可以作为可靠的常见数据源用于数据流中，以进一步增强和集成数据。

在针对 Azure Synapse 开发任何数据流之前，请记住，Azure Synapse 和数据仓库通常会尽力避免使用自助数据准备（以及这些过程产生的版本控制风险）。例如，比起由业务分析师创建数据流以合并、清理和增强数据源，由训练数据工程师开发的企业级管道可能是更好的长期解决方案。综上所述，资源往往非常稀缺，无法满足新的和不断变化的数据转换场景的要求，也无法构建、测试和部署必要的管道或处理作业。Power BI 数据流可以帮助弥合这一差距，以提供技术含量较低但可扩展的自助 ETL 选项。

要针对 Azure Synapse 资源创建数据流，请导航到 Power BI 服务中的应用工作区。从“Create”（创建）下拉菜单中选择“Dataflow”（数据流），然后选择“Add new entities”（添加新实体）选项，如左侧图 19 所示：



图 19：在 Power BI 中创建数据流



这将启动 Power BI 数据流支持的可用数据源。导航到 Azure 类别，然后选择 “Azure SQL Data Warehouse”（Azure SQL 数据仓库），如图 20 所示：



图 20：数据流的 Azure 源

与 Power BI Report Builder 中的数据源配置一样，截止撰写本文时，Azure SQL 数据仓库的数据流仅支持基本 SQL Server 身份验证。在图 21 中，使用 SQL 身份验证登录凭据连接到 Azure Synapse 的 SQL 池资源：

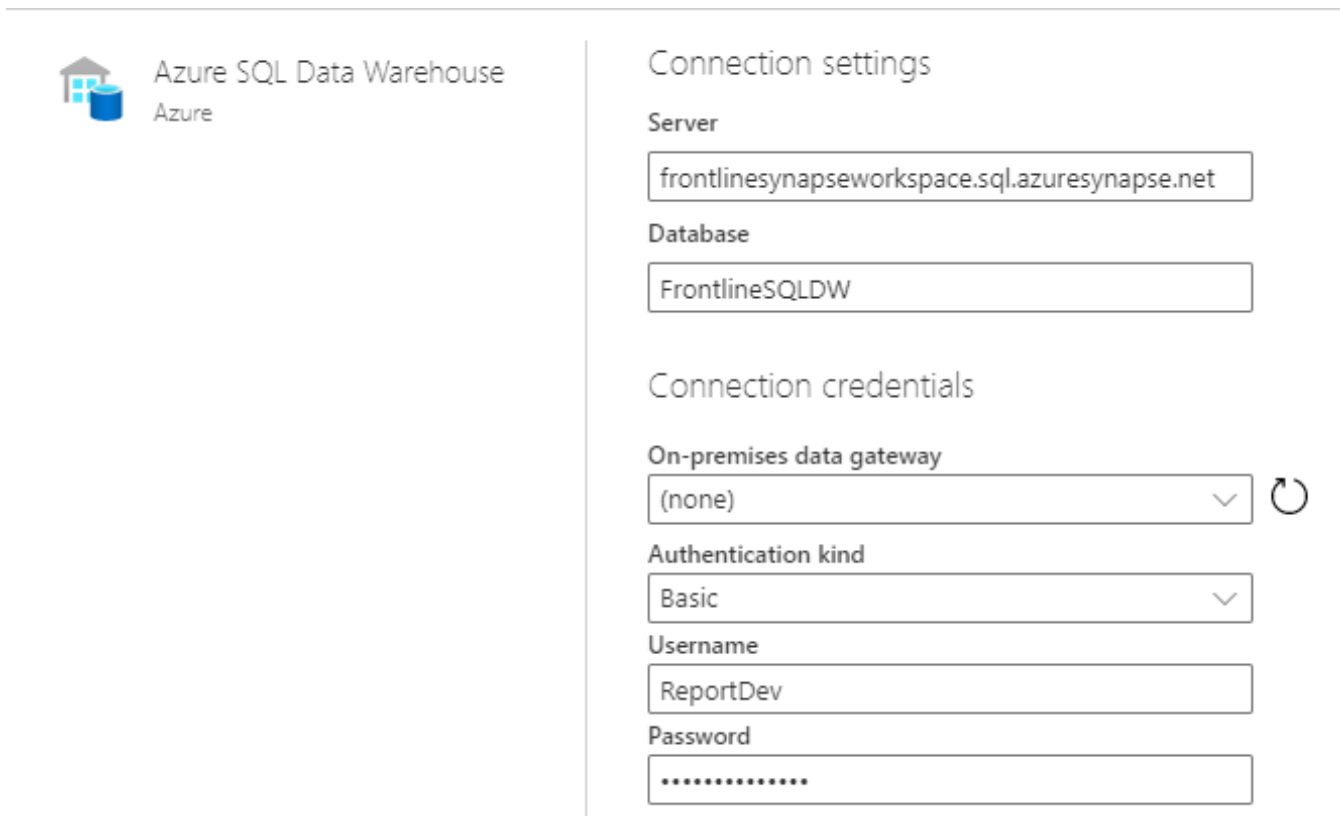


图 21：Azure SQL 数据仓库数据流身份验证

单击“Connection settings”（连接设置）页面中的“Next”（下一步）将启动 SQL 数据源的 Power Query 在线导航和转换功能，如图 22 所示：

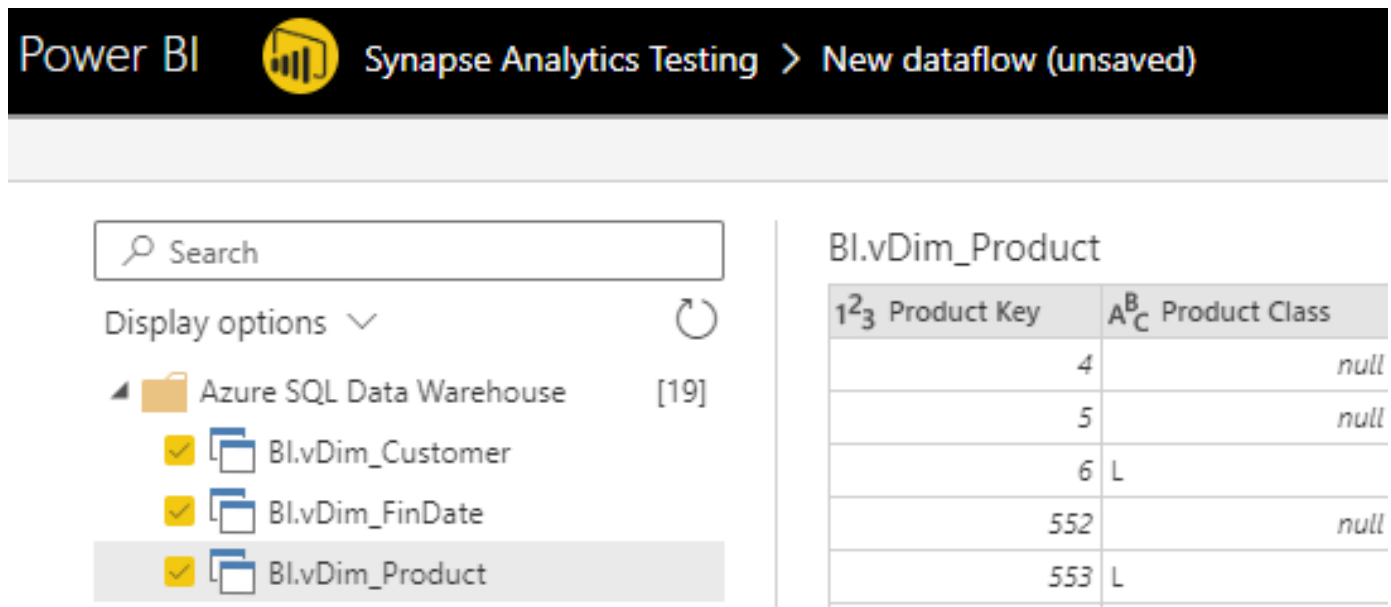


图 22：Power Query 在线导航

与 Power BI Desktop 中的“Get Data”（获取数据）体验一样，数据流作者可以选择所需的实体，然后有选择地实现任意数量的数据转换，以提升源数据的价值。如图 23 所示，这里针对 Azure Synapse 资源提供了熟悉的 Power Query 编辑器功能区及丰富的转换选项集：

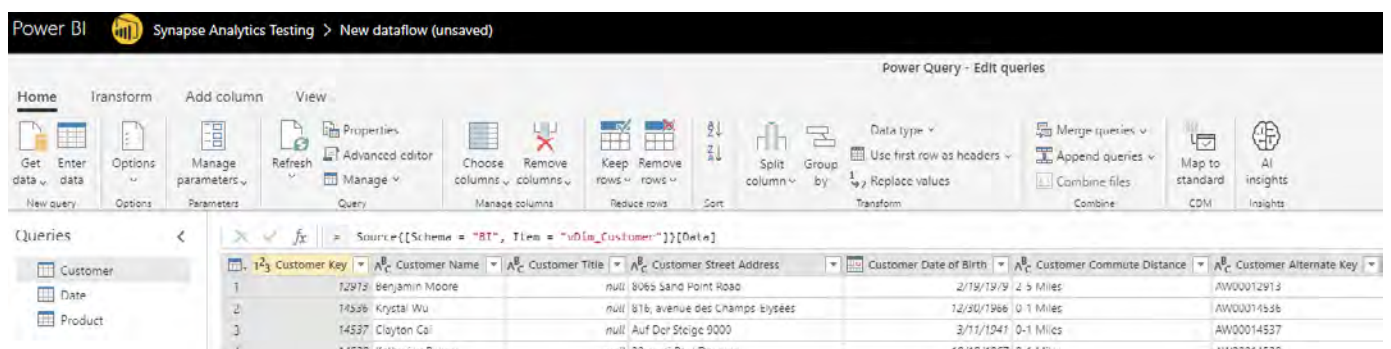


图 23：数据流实体转换

## AI 预测分析集成

除了数据集、报表、分页报表和数据流之外，Power BI 还支持 AI 支持的预测模型并集成了 Azure 机器学习。Azure Synapse 还包括 Azure 机器学习，这是 Azure Synapse Studio 中可访问的集成服务之一。

这种与 Azure 机器学习（通常供数据科学家使用）的深度集成以及 [Power BI 中的自动化机器学习](#) 使 Power BI 开发人员和分析师能够选择利用数据科学家创建的预测模型，或者利用 Power BI 中提供的自助模型创建功能。具体来说，作为训练预测模型的一种方法，Power BI 分析师可以构建或重用数据流（如上一节所述）。经过训练和验证后，可将 Power BI 中的预测模型应用于其他数据流，以向传入数据添加预测值。

## 复合模型和聚合

Power BI 数据集提供的两个最强大的数据建模功能是[复合模型](#)和[聚合](#)。无论是单独使用还是配合使用，这两个功能使 BI 团队能够灵活地在同一语义模型中的多个数据源之间平衡导入存储模式和 DirectQuery 存储模式的优势。当与数据仓库团队协作精心设计出 Power BI 模型之后，这些模型可以同时提供压缩内存中数据缓存和小型聚合表的查询性能以及源（如 Synapse SQL 池）的无限可扩展性和数据新鲜度。

Azure Synapse 适用于复合模型和聚合解决的常见数据建模场景。以下各节提供了这些场景的示例。

## 通过聚合实现目标性能

商业智能团队通常非常了解其语义模型要解决的主要业务问题以及用户的使用模式和优先级。例如，虽然表示销售数据的事实表可能与七维表相关，但通常情况下，报表或临时分析中很少使用二维或三维。此外，虽然销售事实表的粒度（两到三个）可能支持单个客户和产品级别的查询，但用户可能很少深入了解这种级别的详细信息。数据模型和它通常收到的查询类型之间的这种常见不匹配导致性能不理想，以及在内存中处理和存储数据的资源成本过高。

聚合使数据建模师能够在模型中嵌入隐藏的聚合表，从而反映模型中最常用的维度和事实的分组。通过在模型中的聚合表和维度之间定义关系，Power BI 可动态确定聚合表是否可以解析传入的查询，或者是否需要查询更精细的详细信息表。

由于聚合表比详细信息表小得多，并且由于可以选择将聚合表存储在压缩的内存缓存中，因此它解析的查询可以实现出色的性能。此外，除了聚合表对用户不可见之外，在模型中定义的用于限制用户访问的相同行级安全角色也适用于聚合。

例如，假设 SQL 池中的 Internet 销售事实表所包含的行超过 50 亿。在 Power BI Premium 容量中存储此表所需的内存量使得 DirectQuery 成为唯一可行的选项。但是，尽管对此源表应用了性能优化，但通过 DirectQuery 连接查询此表的方法可能无法在 Power BI 中提供所期望的用户体验。

为了逐渐优化有关客户和销售地区的常见和 / 或高度重视的查询的性能，可以在 SQL 池数据库中创建一个新表，表示按订单日期键、客户键和销售地区键划分的互联网销售表分组。新聚合表的大小仅占互联网销售表的一小部分，可以作为标准夜间数据仓库加载过程的一部分获得支持，并通过[群集列存储索引](#)获得性能优化，方法与其他事实表相同。

在图 24 中，利用 Power BI Desktop 中的“Manage aggregations”（管理聚合）窗体，模型作者可以为聚合表中的每个列定义摘要以及与相应详细信息表列的映射：

✕

## Manage aggregations

Aggregations accelerate query performance to unlock big-data sets. [Learn more](#)

Aggregation table      Precedence ⓘ

InternetSalesAgg ▼      0

AGGREGATION COLUMN	SUMMARIZATION	DETAIL TABLE	DETAIL COLUMN	
CustomerKey	GroupBy ▼	Internet Sales ▼	CustomerKey ▼	🗑️
OrderDateKey	GroupBy ▼	Internet Sales ▼	OrderDateKey ▼	🗑️
Sales_Summed	Sum ▼	Internet Sales ▼	SalesAmount ▼	🗑️
SalesTerritoryKey	GroupBy ▼	Internet Sales ▼	SalesTerritoryKey ▼	🗑️

图 24：管理聚合

默认情况下，将聚合应用于表之后，其元数据在用户界面中是隐藏的。

然后，模型作者只需将定义聚合表粒度的三个列与其对应的维度列相关联，即可形成如图 25 中所示的架构：

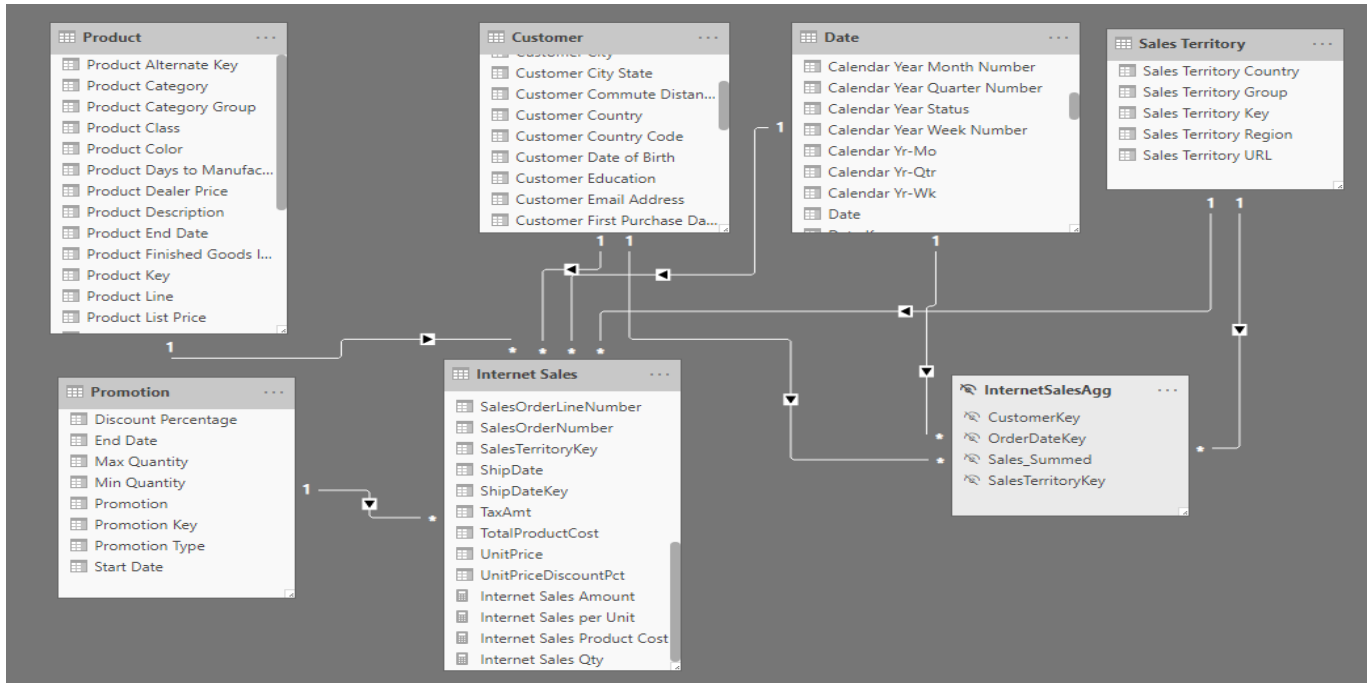


图 25：具有聚合的模型

在更新的 Power BI 数据集中，Power BI 以及需要按 “Customer”（客户）、“Date”（日期）或 “Sales Territory”（销售地区）维度或这些三维表的某种组合来统计销售金额列合计的其他工具发出的所有 BI 查询都由较小的 InternetSalesAgg 表处理，从而提高了性能。然而，请求从 Promotion 或 Product dimension 表中获取列的查询仍然使用 Internet Sales 表。

## 表存储模式

组合模型和聚合的一个主要功能是能够为 Power BI 模型中的每个表定义存储模式。通过 Power BI Desktop 的 “Modeling”（建模）视图中 “Advanced”（高级）卡中的 “Storage mode”（存储模式）属性（如图 26 所示），模型设计人员可以在 “Import”（导入），“DirectQuery” 和 “Dual”（双重）之间进行选择：

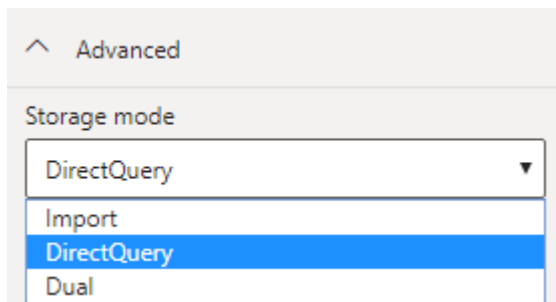


图 26：存储模式设置

对于前一个示例中的 InternetSalesAgg 聚合表，商业智能和数据仓库团队可以决定此表是否也应该像模型中的所有其他表一样采用 DirectQuery 模式，还是应该将其导入到 Power BI 中的压缩内存缓存中。此决定对性能和安全性都有重大影响。

从性能的角度来看，由于 Power BI 的内存中列式引擎具有内置优化，因此与 DirectQuery 模式下的同一个表相比，导入的表在不同程度上提高了查询性能。但是，相对于针对数十亿行事实表的查询，DirectQuery 模式下的聚合表仍然可以显著提高性能。

在安全性方面，通过让所有表保持在 DirectQuery 模式下，团队能够利用源 SQL 池和单一登录身份验证（能够将用户的身份传递给源）中内置的安全性。利用在模型中定义的导入和双重存储模式表，需要在 Power BI 模型中定义和管理行级安全角色。与往常一样，商业智能团队必须权衡与环境和要求相关的这些因素。

## 混合源与连接

与聚合动态确定要查询哪些表的方式类似，Power BI 模型还可以查询多个数据源，并且可能具有不同的存储模式。Power BI 负责查询这两个源并将结果结合在一起，以便为所需的可视化效果提供支持。例如，在项目的某个阶段，模型所需的特定表可能仅提供 .csv 格式的文件。通过复合模型和表存储模式，可以将此 .csv 文件添加到已包含多个具有与 SQL 池的 DirectQuery 连接的表的模型中。此外，可以在基于 .csv 的表与基于 DirectQuery 的 SQL 池表之间创建模型关系，以控制报表可视化效果中的交叉筛选行为。借助复合模型，团队能够灵活地利用多个数据源和替代存储模式。



## 结语

将 Azure Synapse Analytics 和 Power BI 结合使用时功能更加强大，可提供独特的现代化数据分析方法。Azure Synapse 可帮助 Power BI 专业人员在各种用例中提供项目所需的规模、性能和成本管理。用户可以在 Azure Synapse Studio 中开发交互式 Power BI 报表和企业级语义模型，该程序是一种用于开发和管理各种 Azure Synapse 工件的全新通用 Web 门户。

Azure Synapse 给 Power BI 专业人员带来的一些关键好处包括：

- 作为 Power BI 经过验证的单一可信来源
- 支持性能优化，实现大规模 DirectQuery
- 支持行级和列级安全性以及其他集成的安全功能
- 通过通用用户界面促进团队协作并提高透明度
- 包括企业级数据转换和编排功能，从而实现可靠的数据准备
- 为使用 Power BI Report Builder 创建分页报表提供灵活的支持

立即注册 Azure 免费帐户，了解将 Azure Synapse Analytics 与 Power BI 结合使用将给公司带来哪些好处。



# 关于作者

**Jack Lee** 是 Azure 高级认证顾问，也是 Azure 实践主管，对软件开发、云和 DevOps 创新充满热情。他是 Microsoft 技术社区的积极贡献者，并现身各种用户组和会议，包括 Microsoft 加拿大公司举办的全球 Azure 集训营。Jack 是经验丰富的黑客马拉松导师和评委，也是专注于 Azure、DevOps 和软件开发的用户组的主席。他因其对技术社区的贡献而被评为 Microsoft MVP。你可以在推特上关注 Jack，帐号是 [@jlee\\_consulting](#)。

**Brett Powell** 是 Frontline Analytics 的所有者。Frontline Analytics 是一家数据和分析咨询公司，也是 Microsoft Power BI 合作伙伴。自从 Power BI 技术随 Excel 2010 的 Power Pivot 外接程序首次推出以来，他就一直在使用这些技术，并为零售、制造、金融和专业服务领域的 Microsoft BI 解决方案的设计和交付做出了贡献。他还是《*掌握 Microsoft Power BI*》以及《*Microsoft Power BI 指南*》的作者，并且经常在 Microsoft 技术活动（如 Power Platform World Tour 和 Data & BI Summit）中发表演讲。他经常在博客 [Insight Quest](#) 上分享技术提示和示例。

