数据仓库实验报告

实验名称：	实验二 使用 SSIS 创建项目和基本包
实验日期：	2023-10-08
实验地点：	文宣楼 B313
提交日期：	2023-10-10

学号：	22920212204392
姓名：	黄勖
专业年级：	软工 2021 级
学年学期：	2023-2024 学年第一学期

1 实验环境

SQL Server 2019；

先决条件：已安装 AdventureWorksDW**2022** 示例数据库

2 实验目的

（1）掌握使用 SSIS 设计器创建项目和基本 ETL 包的方法；

（2）掌握使用 SSIS 添加循环的方法；

（3）掌握使用 SSIS 添加日志记录的方法；

（4）掌握使用 SSIS 添加错误流重定向的方法。

3 实验内容和步骤（SSIS 教程的第 1-4 课）

3.1 第一课问题：

3.1.1 “平面文件连接管理器”在集成服务中的作用？

平面文档连接管理器（Flat File Connection Manager）用于管理与平面文档的连接和数据导入，从而实现数据的导入和导出任务，这对于ETL（提取、转换、加载）过程以及数据集成非常重要。平面文档通常是文本文档，如CSV（逗号分隔值）文档或定宽格式文档，它们通常包含结构简单的表格数据。具体来说有如下作用：

1. **连接文档**：平面文档连接管理器允许你指定要连接的平面文档的位置和属性。可以指定文档的路径、文档名、编码方式、分隔符等信息，以确保SSIS可以正确读取和写入数据。
2. **定义数据结构**：通过平面文档连接管理器，可以定义平面文档中的数据结构，包括列的数量、名称、数据类型以及列之间的分隔符或固定宽度。这有助于SSIS正确地解析和处理文档中的数据。
3. **预览数据**：连接管理器还允许在设计时预览平面文档的数据，以确保正确地解析了文档，并且可以看到数据的样本。
4. **数据导入和导出**：一旦配置了平面文档连接管理器，可以在SSIS数据流任务中使用它来将数据从平面文档导入到数据库表中，或将数据从数据库表导出到平面文档中。连接管理器会处理文档的读取和写入操作。
5. **动态文档处理**：SSIS中的平面文档连接管理器还支持动态文档处理，这意味着可以在运行时基于参数或变量来指定要连接的平面文档，从而实现灵活的数据导入和导出方案。

3.1.2 “OLE DB 连接管理器”在集成服务中的作用？

OLE DB 连接管理器（OLE DB Connection Manager）是SSIS中的一种连接管理器，它用于管理与各种OLE DB（Object Linking and Embedding Database）数据源的连接。OLE DB是一种通用的数据访问技术，它允许SSIS与各种数据库系统和数据源进行通信。

以下是OLE DB 连接管理器在集成服务中的主要作用：

1. 连接到不同的数据源：OLE DB 连接管理器使你能够连接到各种不同的数据源，包括但不限于：

   - 关系数据库管理系统（RDBMS），如SQL Server、Oracle、MySQL等。
   - 数据仓库，如Microsoft Azure SQL Data Warehouse、Snowflake等。
   - 文档系统，如Excel文档、Access数据库等。
   - ODBC（Open Database Connectivity）数据源。

2. 配置连接属性：通过OLE DB 连接管理器，你可以配置与特定数据源相关的连接属性，包括服务器名称、数据库名称、身份验证方式、用户名、密码等。这些属性是连接到数据源所必需的信息。

3. 运行时连接字符串：你可以使用表达式和变量来动态配置连接字符串，从而在运行时根据需要更改连接属性。这允许你创建更灵活的数据导入和导出方案。

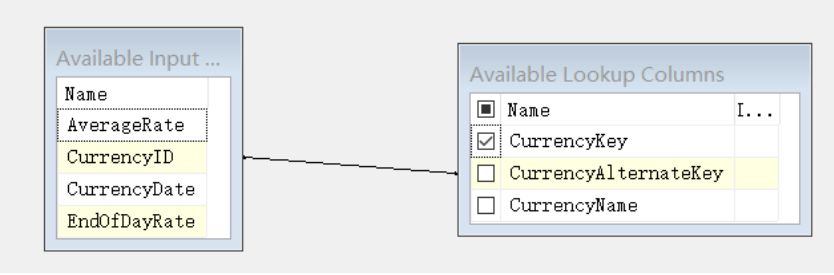
4. 数据源预览：连接管理器通常提供一个选项，允许你在设计时预览数据源的结构和数据，以确保连接配置正确。

5. 数据流任务中的数据传输：一旦配置了OLE DB 连接管理器，你可以在SSIS数据流任务中使用它来执行数据传输操作，包括将数据从数据源提取到SSIS包中，以及将数据从SSIS包传送到目标数据源。这对于ETL（提取、转换、加载）和数据集成任务非常重要。

3.1.3 在步骤 6 中，“lookup”转换工具的作用是什么，为什么选择 dbo.dimCurrency。除了“lookup”转换，系统还提供了哪些转换功能？

（1）在步骤 6 中，"Lookup" 转换工具的作用是执行查找操作，通过将平面文档中的数据与引用数据集中的数据进行匹配，从而获取 CurrencyKey 和 DateKey 值，这些值将用于后续的数据转换和加载。具体来说，"Lookup" 转换在此上下文中的作用如下：

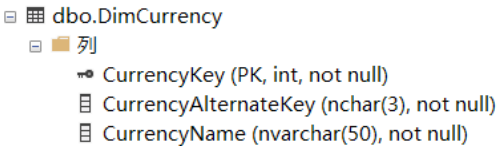
1. **匹配数据**: "Lookup" 转换将平面文档中的数据中的 "CurrencyID" 列值与 "DimCurrency" 维度表中的 "CurrencyAlternateKey" 列值进行匹配。这样可以找到与文档中的货币标识符相对应的 CurrencyKey。



2. **数据提取**: 一旦匹配成功，"Lookup" 转换将返回 "DimCurrency" 表中的相关数据，包括 CurrencyKey。这些数据将被用于构建最终的数据流。

3. "Lookup Date Key" 的功能与此类似。

（2）为什么选择 "dbo.DimCurrency"？ 在此情况下，选择 "dbo.DimCurrency" 作为引用数据集的原因是 "DimCurrency" 表包含了与货币相关的信息，包括 CurrencyKey、CurrencyAlternateKey 等。通过查找这个表，可以将平面文档中的货币标识符与 CurrencyKey 相关联，以便后续处理。



（3）此外，系统还提供了许多其他类型的转换功能，用于在SSIS包中执行不同的数据操作。一些常见的转换包括：

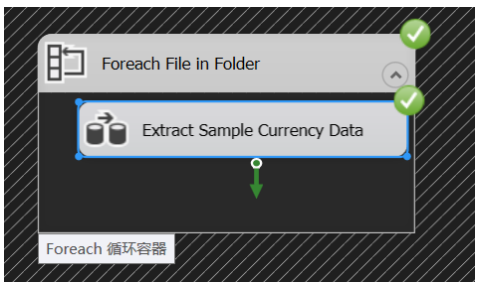
1. **数据流转换（Data Flow Transformations）**：这些转换用于数据的转换和整理，例如数据列的转换、数据列的聚合、数据列的拆分等。常见的数据流转换包括洗牌转换（Sort Transformation）、洗牌拆分转换（Splitter Transformation）、聚合转换（Aggregate Transformation）等。
2. **洗牌和合并（Merge and Union）**：这些转换用于将多个数据流合并或连接在一起，以便在后续处理中使用。常见的洗牌和合并转换包括合并联接（Merge Join Transformation）和联合转换（Union All Transformation）等。
3. **条件分支（Conditional Split）**：这个转换用于根据指定条件将数据流中的数据拆分成多个分支。通常用于根据不同的条件路由数据。
4. **聚合（Aggregate）**：这个转换用于在数据流中执行聚合操作，如对数据进行分组并计算总和、平均值等。
5. **查找（Lookup）**：除了上述步骤中提到的 Lookup 转换，还可以使用其他 Lookup 转换来执行不同的查找操作。
6. **导入和导出（Import and Export）**：这些转换用于将数据导入到目标系统或从目标系统导出数据。例如，可以使用 Excel 导入导出转换来与 Excel 文档交互。

3.2 第二课问题：

3.2.1 本课完成什么功能？

本课的主要目标是向第 1 课的示例 ETL 包中添加循环功能，具体来说，是使用 Foreach 循环容器，以便从数据文档夹中迭代访问并处理多个文档。该课程主要完成以下功能：

1. **循环遍历多个文档**：在第 1 课中，我创建了一个包，用于从单个平面文档源中提取数据。然而，现实世界中的 ETL 任务通常需要处理来自多个源的数据。本课程的目标是了解如何处理这种情况，通过循环遍历多个文档来提取数据。



2. **使用 Foreach 循环容器**：本课程使用了 Foreach 循环容器，这是 SSIS 中的一种控制流容器，用于在循环中执行控制流任务。通过 Foreach 循环容器，我可以指定一个枚举器（enumerator），以便它可以迭代遍历一组对象，在这里指的是多个文档。

3. **设置用户定义的包变量**：在本课程中，我需要设置一个用户定义的包变量（package variable）。这个变量用于存储当前迭代中要处理的文档的路径。每次循环迭代时，这个变量的值会更新为不同的文档路径。

4. **循环访问示例文档夹中的文档**：通过 Foreach 循环容器和用户定义的包变量，我可以实现循环访问示例文档夹中的每个文档。这意味着我的控制流将会多次执行，每次处理一个不同的文档。

3.2.2 Foreach 循环容器起什么作用，该例中，什么控件被加入容器？

- **作用**：Foreach 循环容器的作用是循环遍历一个文档夹（*example*：*E:\大三上\数仓\实验\lab_2 设计创建简单ETL包\Creating a Simple ETL Package\Creating a Simple ETL Package\Sample Data*）中的多个文档，每次迭代处理一个文档。
- **加入容器的控件**：在该例子中，加入到Foreach 循环容器内的控件是一个名为 "Extract Sample Currency Data" 的数据流任务（Data Flow Task）。这个数据流任务用于处理文档文档中的数据。这个数据流任务会在Foreach 循环容器内的每次迭代中执行，每次处理一个不同的文档。

3.2.3 变量 varFileName 什么时候被定义，变量的作用域是哪里，在哪个步骤被赋予什么值？

1. **变量的定义**：变量 `User::varFileName` 在第 2-2 课中的 "Foreach 循环容器" 中被定义。在该课程的 "为 Foreach 循环容器配置枚举器" 步骤中，我创建了这个变量并将其映射到 "Foreach File in Folder" 枚举器，以存储每次迭代中要处理的文档的路径。
2. **变量的作用域**：变量 `User::varFileName` 的作用域是整个 SSIS 包。这意味着它可以在包的不同控制流任务之间传递和使用。
3. **赋值时机**：变量 `User::varFileName` 在 Foreach 循环容器的每次迭代中被赋予不同的值。在枚举器配置中，我把这个变量映射到了文档名，每次迭代时，枚举器会识别下一个文档，然后将文档的路径赋值给 `User::varFileName` 变量（如*...\Currency_ARS.txt*）。

3.3 第三课问题：

3.3.1 日志文件的记录将起到什么作用？

1. **故障排除和调试**：日志文档允许开发人员和管理员在ETL（提取、转换、加载）过程中的各个阶段跟踪和记录信息，包括成功和失败的操作。这有助于快速识别和解决任何问题，例如数据导入失败、连接问题、转换错误等。通过查看日志文档，可以确定导致问题的具体步骤。
2. **性能优化**：日志文档可以记录ETL任务的性能数据，如执行时间、资源消耗等。这些数据有助于分析和优化包的性能。如果发现某个任务耗时过长或资源消耗过高，可以采取适当的措施来改进性能。
3. **合规性和审计**：在某些情况下，特别是在处理敏感数据或需要遵守法规的情况下，日志文档记录是强制性的。它可以用于跟踪数据的来源、目的地和处理方式，以满足合规性和审计要求。日志文档提供了对数据处理活动的可追溯性，以便证明合规性。

4. **监控和报警**：通过监视日志文档，可以实时跟踪包的执行状态。如果发生错误或任务失败，可以设置警报以及自动化通知，以便及时采取纠正措施。这有助于减少数据处理中的停机时间。
5. **历史记录**：日志文档可以存储历史运行的信息，以便回顾以前的ETL任务执行情况。这对于长期性能分析、问题趋势分析以及制定数据处理策略非常有用。
6. **报告和文档**：SSIS日志文档中的信息可以用于生成报告和文档，以便与团队、管理层或相关利益相关者分享有关数据处理任务的详细信息。这些报告可以提供可视化的数据处理统计和趋势。

总之，SSIS日志文档记录是ETL过程中的关键组成部分，它们为监视、调试、性能优化、合规性和审计提供了必要的信息。通过详细记录和分析日志文档，可以确保数据处理任务在各个方面都按预期运行，并及时采取措施来处理任何问题。

3.3.2 日志文件的记录由哪些事件负责?

以下是一些常见的事件，它们负责生成日志文档：

1. **Pre-Execute事件**: 在数据包执行之前触发，用于记录与数据包执行相关的信息。
2. **Post-Execute事件**: 在数据包执行之后触发，用于记录与数据包执行结果相关的信息。
3. **OnError事件**: 在数据包中发生错误时触发，用于记录错误信息以及错误处理步骤。
4. **OnWarning事件**: 在数据包生成警告时触发，用于记录警告信息。
5. **OnInformation事件**: 在数据包执行过程中生成信息消息时触发，用于记录有关数据包执行的附加信息。
6. **OnTaskFailed事件**: 在数据包中的任务失败时触发，用于记录任务失败的详细信息。
7. **OnProgress事件**: 在数据包执行过程中生成进度信息时触发，用于记录数据包执行的进度。

这些事件可以通过配置数据包的日志设置来启用，并选择将日志信息写入文档，数据库表，或其他目标。通常，可以在SSIS包的控制流中设置日志，以便根据需要记录不同类型的事件。然后，可以使用SSIS提供的工具和报告来分析这些日志，以监视和调试数据包的执行过程。

3.3.3 除了文本类型，系统还支持哪些其他格式的文件吗?

SSIS支持多种文档格式，不仅仅限于文本类型。具体来说，可以处理以下常见的文档格式：

1. **文本文档**：SSIS可以轻松处理各种文本文档，包括逗号分隔值 (CSV)、制表符分隔值 (TSV)、定长格式等。文本文档通常用于数据导入和导出。
2. **Excel文档**：SSIS可以读取和写入Microsoft Excel文档，包括XLS和XLSX格式。这使得可以与Excel电子表格进行数据交互。
3. **XML文档**：SSIS可以处理XML文档，包括读取XML数据、转换XML结构以及将数据写入XML文档。这对于与Web服务、数据交换以及处理XML格式的数据非常有用。
4. **数据库文档**：SSIS可以连接到各种关系型数据库管理系统，包括SQL Server、Oracle、MySQL等，以进行数据抽取、加载和转换。
5. **JSON文档**：虽然SSIS不直接支持JSON文档，但可以通过脚本任务或第三方组件来处理JSON数据。
6. **平面文档**：除了文本文档外，SSIS还可以处理其他平面文档格式，如逗号分隔的值 (CSV)、制表符分隔的值 (TSV) 等。

7. **日志文档**：SSIS可以用于处理日志文档，包括文本日志文档、事件日志、应用程序日志等，以进行日志分析和数据提取。
8. **二进制文档**：虽然处理二进制文档可能需要自定义组件或脚本任务，但SSIS也可以处理二进制文档，例如处理图像、音频或视频文档。
9. **压缩文档**：SSIS可以处理压缩文档，如ZIP、GZIP等，以从这些文档中提取数据或将数据写入这些文档。

对于其他不支持的特定格式，也可以使用自定义脚本任务或第三方组件来实现相应的处理。

3.4 第四课问题：

3.4.1 错误流重定位起到什么作用？

错误流重定向用于处理数据流任务中发生的错误数据行。它的主要作用是将错误行从正常数据流中分离出来，并将它们发送到一个独立的数据流，以便进行进一步的处理、记录或报告。错误流重定向有以下几个作用：

1. **错误数据的隔离**：错误流重定向允许将包含错误的数据行从正常数据中分离出来。这可以确保错误数据不会影响正常数据的处理，从而维护了数据的完整性。
2. **错误数据的处理**：一旦错误数据被重定向到错误流中，你可以对其进行特殊处理。例如，你可以将错误数据写入到一个错误日志文档中，或者将其发送到一个专门用于处理错误数据的目标表。
3. **错误数据的记录**：错误流重定向允许你记录错误数据的详细信息，例如哪些列包含了错误数据、错误的类型是什么，以及发生错误的原因。这有助于后续的故障排除和数据质量改进。
4. **继续执行**：在处理错误数据时，你可以选择是否继续执行数据流任务。这意味着即使在数据流中发生了错误，你仍然可以继续处理其他数据，而不会中断整个包的执行。
5. **定制错误处理逻辑**：通过错误流重定向，你可以编写自定义逻辑来处理不同类型的错误。这可以根据业务需求来定制处理方式，例如跳过特定错误、自动修复错误数据等。

总的来说，错误流重定向可以帮助开发者更有效地管理和处理ETL过程中的错误，提高数据质量并确保包的可靠性。通过将错误数据与正常数据分离开来，可以更好地监视和处理ETL任务中的问题。

3.4.2 步骤 2 中，Currency_BAD.txt 文件为什么出现导入失败，失败的级别是个别文件还是整个包？

在步骤 2 中，Currency_BAD.txt 文档导致导入失败的原因是该文档中的数据与期望的数据不匹配，导致了查找转换的失败。具体来说，文本文档 Currency_BAD.txt 中的 "CurrencyID" 列的值被编辑为 "BAD"，而查找转换旨在查找与 "CurrencyID" 匹配的值，但在源数据中找不到匹配的数据。

Currency_BAD.txt 文档导致**整个包**失败，因为 Lookup Currency Key 转换的错误输出会将错误行导向脚本转换操作，进而影响了整个包的执行。

3.4.3 ErrorOutput.txt 文件中有哪些内容?

经过实际实验，ErrorOutput.txt 文档的内容如下：

```
0.00158,BAD,2005-07-01,0.00158,-1071607778,0,Row yielded no match during
lookup.
……
```

其中，每一行代表一个失败的行，包括以下列：

- 失败的行的内容
- `ErrorCode`：- 1071607778，用于标识错误的代码或编号。
- `ErrorColumn`：0，非特定列错误。
- `ErrorDescription`：提供了关于错误的详细描述。

具体的内容会根据实际的错误数据和错误情况而变化，但这个文档的目的是记录有关导致错误的失败行的详细信息，以便后续的处理和调试。

4 实验总结(完成的工作、对实验的认识、遇到的问题及解决方法)

4.1 完成的工作：

1. 学习创建简单的 ETL 包，该包从单个平面文档源提取数据，使用两个查找转换组件转换数据，并将转换后的数据写入 `AdventureWorksDW2022` 示例数据库中的“FactCurrencyRate”事实数据表的副本。
2. 学习使用 SSIS 中的 Foreach 循环容器和用户定义的包变量来处理多个文档，而不需要修改数据流。这对于需要从多个源提取数据并将其集成到单一目标中的 ETL 任务非常有用。通过循环遍历文档，你可以有效地处理多个数据源，实现更灵活和通用的数据导入和处理方案。
3. 学习添加和配置日志记录，以在包执行过程中监控特定事件。
4. 创建了损坏的示例平面文档（Currency_BAD.txt），并通过编辑其中的数据引入错误；配置了 Lookup Currency Key 转换以处理损坏的文档，导致查找失败；添加了平面文档目标并配置平面文档连接管理器，以便将失败行的详细信息记录到 ErrorOutput.txt 文档中。

4.2 对实验的认识：

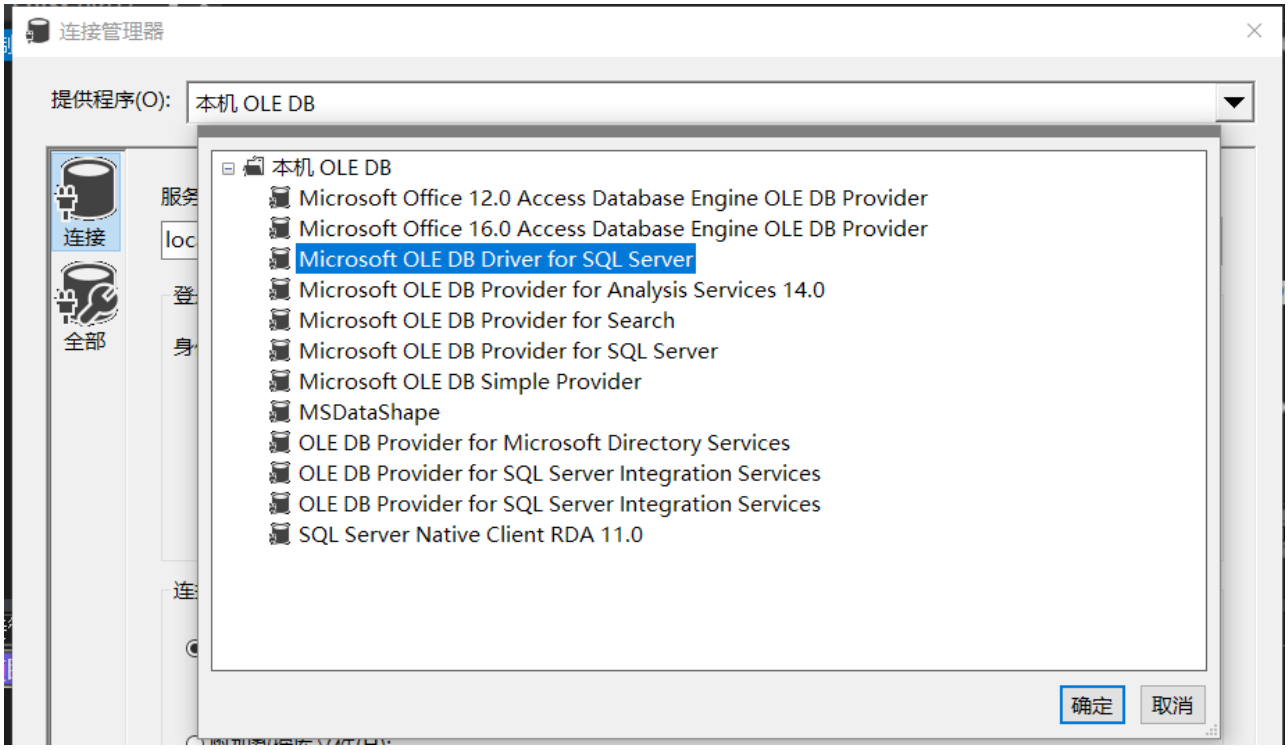
通过本次实验，我深入了解了在SSIS中处理错误数据的重要性以及如何有效地处理错误数据。以下是我的一些认识：

1. 错误处理是ETL过程中不可或缺的一部分，因为源数据可能包含各种不一致性和问题。在处理大规模数据时，错误数据的处理变得尤为重要。
2. Lookup转换是一个强大的组件，可以用于在ETL过程中与引用数据集进行匹配，但它对数据的质量和一致性要求较高。
3. 通过使用错误流和脚本转换，可以有效地捕获和记录处理错误数据的详细信息，帮助进行故障排除和数据质量改进。

4.3 遇到的问题及解决方法：

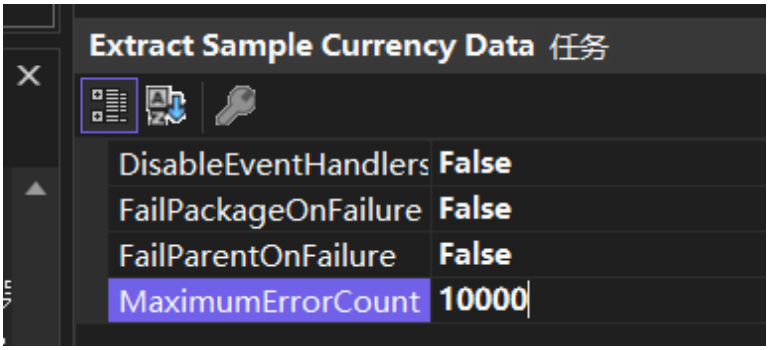
4.3.1 连接管理器由于在初始化提供程序时出错，导致连接测试失败。未在本地计算机上注册“SQLNCLI11”提供程序。

应切换提供程序



4.3.2 警告: 0x80019002，位于 Lesson 4: SSIS 警告代码 DTS_W_MAXIMUMERRORCOUNTREACHED。Execution 方法成功，但出现的错误数(3)达到了允许的最大值(1)，因此导致失败。当错误数达到 MaximumErrorCount 中指定的数目时将发生这种情况。请更改 MaximumErrorCount 或纠正这些错误。

当多次执行后，会出现这个错误。把所有容器的MaximumErrorCount修改为10000即可解决



5 附录

教程网址：https://docs.microsoft.com/zh-cn/sql/integration-services/lesson-1-create-a-project-and-basic-package-with-ssis?view=sql-server-ver15