# 使用数据库迁移工具将数据导入到 DocumentDB

本文演示了如何使用开源 DocumentDB 数据迁移工具将数据从各种源导入到 [Microsoft Azure DocumentDB](https://azure.microsoft.com/services/documentdb/)，包括 JSON 文件、CSV 文件、SQL、MongoDB、Azure 表存储、Amazon DynamoDB 和 DocumentDB 集合。

阅读本文之后，你将能够回答以下问题：

* 如何将 JSON 文件数据导入到 DocumentDB？
* 如何将 CSV 文件数据导入到 DocumentDB？
* 如何将 SQL Server 数据导入到 DocumentDB？
* 如何将 MongoDB 数据导入到 DocumentDB？
* 如何将数据从 Azure 表存储导入到 DocumentDB？
* 如何将数据从 Amazon DynamoDB 导入到 DocumentDB？
* 如何将数据从 HBase 中导入到 DocumentDB？
* 如何在 DocumentDB 集合之间迁移数据？

## 先决条件

在按照本文中的说明操作之前，请确保已安装下列项：

* [Microsoft .NET Framework 4.51](https://www.microsoft.com/download/developer-tools.aspx) 或更高版本。

## DocumentDB 数据迁移工具概述

DocumentDB 数据迁移工具是一个开源解决方案，它将数据从多个源导入到 DocumentDB，包括︰

* JSON 文件
* MongoDB
* SQL Server
* CSV 文件
* Azure 表存储
* Amazon DynamoDB
* HBase
* DocumentDB 集合

导入工具将包括图形用户界面 (dtui.exe)，还可从命令行 (dt.exe) 中驱动。实际上，有一个选项可以在通过用户界面设置导入后输出关联的命令。可以转换表格源数据（例如 SQL Server 或 CSV 文件），以便可以在导入过程中创建层次结构关系（子文档）。继续阅读，以了解有关源选项、用于从每个源导入的示例命令行以及目标选项的详细信息，并查看导入结果。

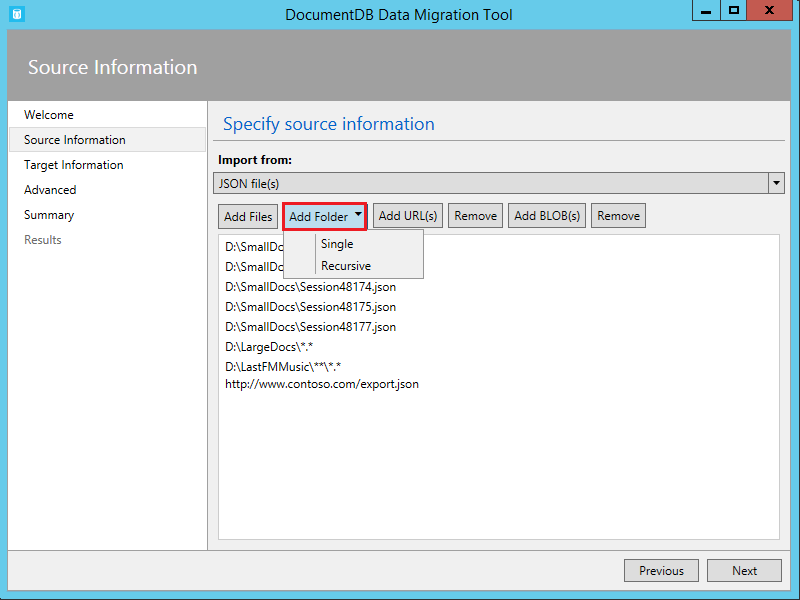
## DocumentDB 数据迁移工具的安装

迁移工具源代码可从[本存储库](https://github.com/azure/azure-documentdb-datamigrationtool)中的 GitHub 上下载，编译版本可从 [Microsoft Download Center](http://www.microsoft.com/downloads/details.aspx?FamilyID=cda7703a-2774-4c07-adcc-ad02ddc1a44d) 上下载。你可以编译解决方案，或者只下载并将编译版本解压缩到所选的目录中。然后运行以下任一文件：

* **Dtui.exe**︰该工具的图形界面版本
* **Dtui.exe**︰该工具的命令行版本

## 导入 JSON 文件

借助 JSON 文件源导入程序选项，可以导入一个或多个只包含单个文档的 JSON 文件或者每个都包含一组 JSON 文档的 JSON 文件。添加包含 JSON 文件的文件夹以供导入时，可以选择递归搜索子文件夹中的文件。



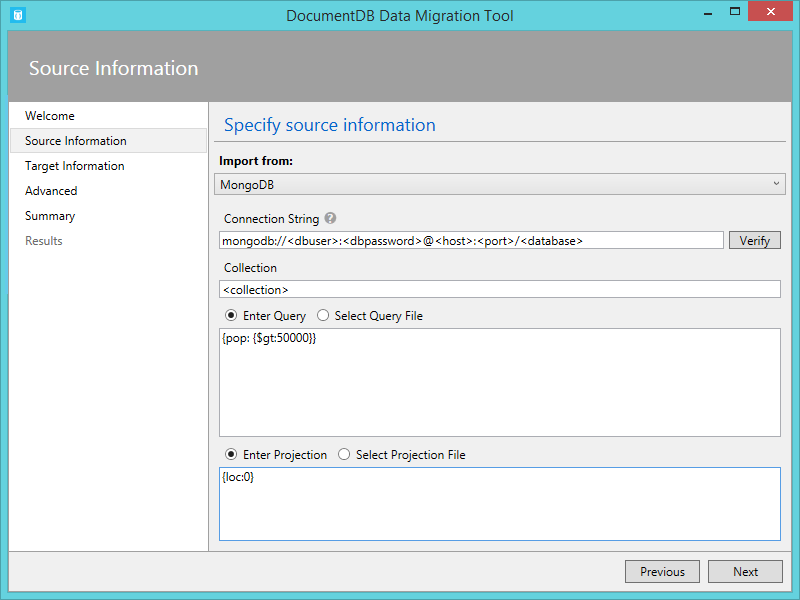
JSON 文件源选项的屏幕截图 - 数据库迁移工具

下面是一些导入 JSON 文件的命令行示例︰

#Import a single JSON file  
dt.exe /s:JsonFile /s.Files:.\Sessions.json /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:Sessions /t.CollectionTier:S3  
  
#Import a directory of JSON files  
dt.exe /s:JsonFile /s.Files:C:\TESessions\*.json /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:Sessions /t.CollectionTier:S3  
  
#Import a directory (including sub-directories) of JSON files  
dt.exe /s:JsonFile /s.Files:C:\LastFMMusic\*\*\*.json /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:Music /t.CollectionTier:S3  
  
#Import a directory (single), directory (recursive), and individual JSON files  
dt.exe /s:JsonFile /s.Files:C:\Tweets\*.\*;C:\LargeDocs\*\*\*.\*;C:\TESessions\Session48172.json;C:\TESessions\Session48173.json;C:\TESessions\Session48174.json;C:\TESessions\Session48175.json;C:\TESessions\Session48177.json /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:subs /t.CollectionTier:S3  
  
#Import a single JSON file and partition the data across 4 collections  
dt.exe /s:JsonFile /s.Files:D:\\CompanyData\\Companies.json /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:comp[1-4] /t.PartitionKey:name /t.CollectionTier:S3

## 从 MongoDB 中导入

借助 MongoDB 源导入程序选项，可从单个 MongoDB 集合中导入，并且选择使用查询筛选文档和/或使用投影来修改文档结构。



MongoDB 源选项的屏幕截图 - documentdb 与 mongodb 对比

连接字符串是标准 MongoDB 格式：

mongodb://<dbuser>:<dbpassword>@<host>:<port>/<database>

[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 MongoDB 实例。

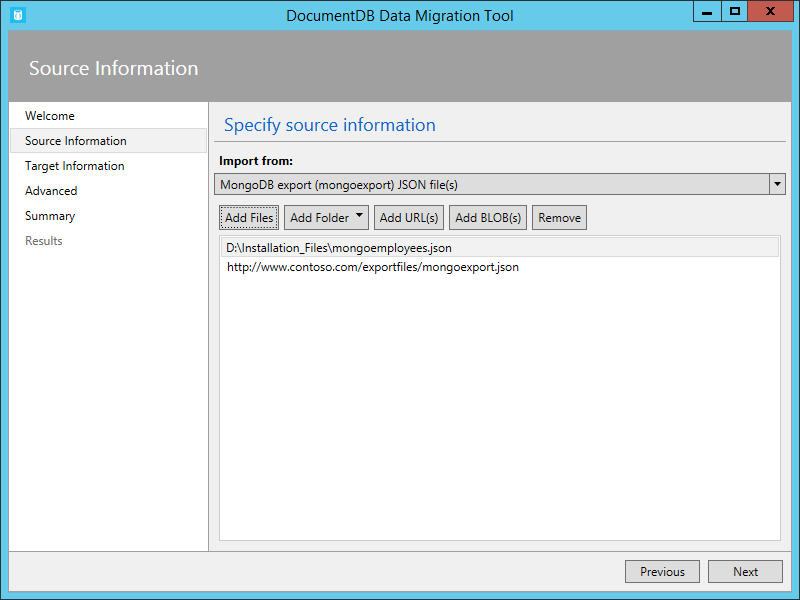
输入将从其中导入数据的集合的名称。可以选择为查询（例如 {pop: {$gt: 5000}}）和/或投影（例如 {loc:0}）指定或提供一个文件来筛选和形成要导入的数据。

下面是一些从 MongoDB 中导入的命令行示例︰

#Import all documents from a MongoDB collection  
dt.exe /s:MongoDB /s.ConnectionString:mongodb://<dbuser>:<dbpassword>@<host>:<port>/<database> /s.Collection:zips /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:BulkZips /t.IdField:\_id /t.CollectionTier:S3  
  
#Import documents from a MongoDB collection which match the query and exclude the loc field  
dt.exe /s:MongoDB /s.ConnectionString:mongodb://<dbuser>:<dbpassword>@<host>:<port>/<database> /s.Collection:zips /s.Query:{pop:{$gt:50000}} /s.Projection:{loc:0} /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:BulkZipsTransform /t.IdField:\_id/t.CollectionTier:S3

## 导入 MongoDB 导出文件

借助 MongoDB 导出 JSON 文件源导入程序选项，可以导入一个或多个通过 mongoexport 实用程序生成的 JSON 文件。



MongoDB 导出源选项的屏幕截图 - documentdb 与 mongodb 对比

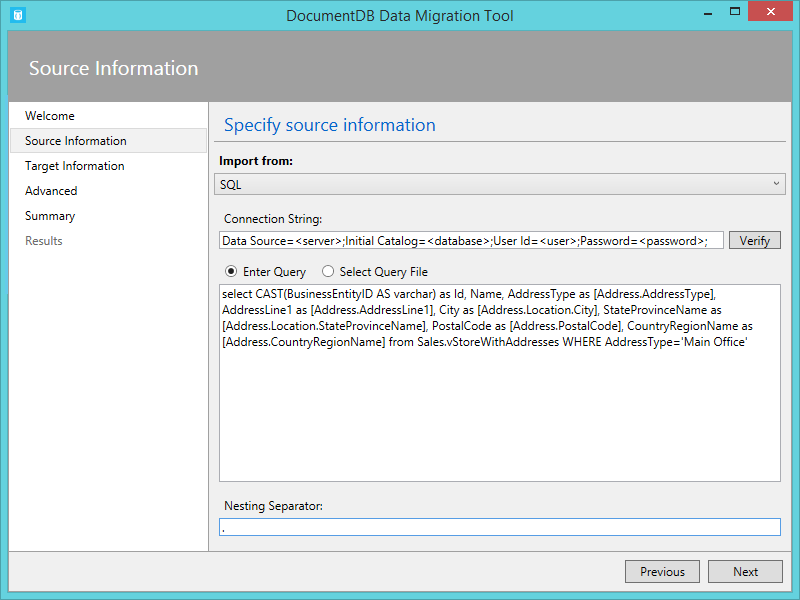
添加包含 MongoDB 导出 JSON 文件的文件夹以供导入时，可以选择递归搜索子文件夹中的文件。

下面是一个用于从 MongoDB 导出 JSON 文件中导入的命令行示例︰

dt.exe /s:MongoDBExport /s.Files:D:\mongoemployees.json /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:employees /t.IdField:\_id /t.Dates:Epoch /t.CollectionTier:S3

## 从 SQL Server 中导入

借助 SQL 源导入程序选项，可以从单个 SQL Server 数据库中导入，并且选择筛选要使用查询来导入的记录。此外，可以通过指定嵌套分隔符（稍后进行详细介绍）来修改文档结构。



SQL 源选项的屏幕截图 - 数据库迁移工具

连接字符串的格式是标准 SQL 连接字符串格式。

[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 SQL Server 实例。

嵌套分隔符属性用于在导入过程中创建层次结构关系（子文档）。请考虑下列 SQL 查询：

select CAST(BusinessEntityID AS varchar) as Id, Name, AddressType as [Address.AddressType], AddressLine1 as [Address.AddressLine1], City as [Address.Location.City], StateProvinceName as [Address.Location.StateProvinceName], PostalCode as [Address.PostalCode], CountryRegionName as [Address.CountryRegionName] from Sales.vStoreWithAddresses WHERE AddressType=‘Main Office’

它将返回以下（部分）结果：



SQL 查询结果的屏幕截图

请注意 Address.AddressType 和 Address.Location.StateProvinceName 等别名。通过指定嵌套分隔符“.”，导入工具会在导入过程中创建 Address 和 Address.Location 子文档。下面是在 DocumentDB 中生成文档的示例︰

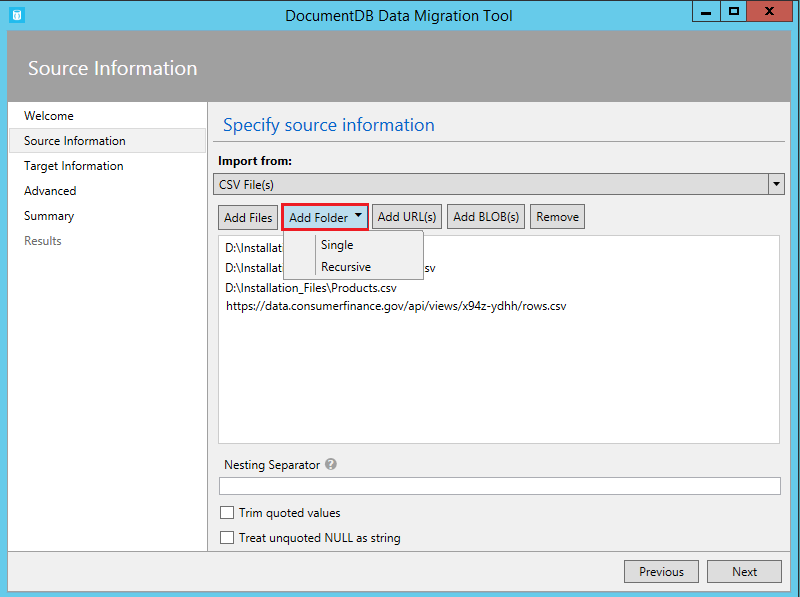
{ “id”: “956”, “Name”: “Finer Sales and Service”, “Address”: { “AddressType”: “Main Office”, “AddressLine1”: “#500-75 O’Connor Street”, “Location”: { “City”: “Ottawa”, “StateProvinceName”: “Ontario” }, “PostalCode”: “K4B 1S2”, “CountryRegionName”: “Canada” } }

下面是一些从 SQL Server 中导入的命令行示例︰

#Import records from SQL which match a query  
dt.exe /s:SQL /s.ConnectionString:"Data Source=<server>;Initial Catalog=AdventureWorks;User Id=advworks;Password=<password>;" /s.Query:"select CAST(BusinessEntityID AS varchar) as Id, \* from Sales.vStoreWithAddresses WHERE AddressType='Main Office'" /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:Stores /t.IdField:Id /t.CollectionTier:S3  
  
#Import records from sql which match a query and create hierarchical relationships  
dt.exe /s:SQL /s.ConnectionString:"Data Source=<server>;Initial Catalog=AdventureWorks;User Id=advworks;Password=<password>;" /s.Query:"select CAST(BusinessEntityID AS varchar) as Id, Name, AddressType as [Address.AddressType], AddressLine1 as [Address.AddressLine1], City as [Address.Location.City], StateProvinceName as [Address.Location.StateProvinceName], PostalCode as [Address.PostalCode], CountryRegionName as [Address.CountryRegionName] from Sales.vStoreWithAddresses WHERE AddressType='Main Office'" /s.NestingSeparator:. /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:StoresSub /t.IdField:Id /t.CollectionTier:S3

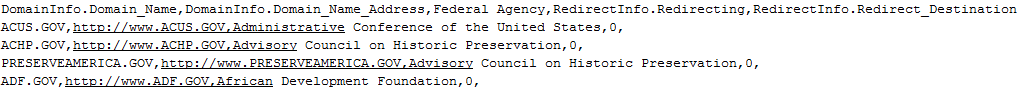
## 导入 CSV 文件 - 将 CSV 转换为 JSON

CSV 文件源导入程序选项可用于导入一个或多个 CSV 文件。添加包含 CSV 文件的文件夹以供导入时，可以选择递归搜索子文件夹中的文件。



CSV 源选项的屏幕截图 - CSV 到 JSON

与 SQL 源相似，嵌套分隔符属性可用于在导入过程中创建层次结构关系（子文档）。请考虑以下 CSV 标头行和数据行︰



CSV 示例记录的屏幕截图 - CSV 到 JSON

请注意 DomainInfo.Domain\_Name 和 RedirectInfo.Redirecting 等别名。通过指定嵌套分隔符“.”，导入工具会在导入过程中创建 DomainInfo 和 RedirectInfo 子文档。下面是在 DocumentDB 中生成文档的示例︰

{ “DomainInfo”: { “Domain\_Name”: “ACUS.GOV”, “Domain\_Name\_Address”: “http://www.ACUS.GOV” }, “Federal Agency”: “Administrative Conference of the United States”, “RedirectInfo”: { “Redirecting”: “0”, “Redirect\_Destination”: “” }, “id”: “9cc565c5-ebcd-1c03-ebd3-cc3e2ecd814d” }

导入工具将尝试针对 CSV 文件中不带引号的值推断类型信息（带引号的值始终作为字符串处理）。按以下顺序标识类型︰数值、日期时间、布尔值。

关于 CSV 导入，还需要注意两个事项︰

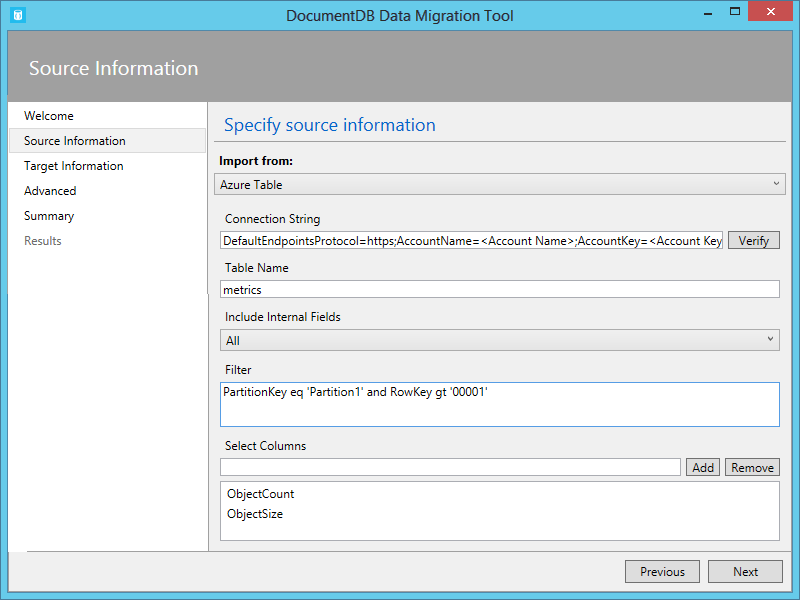
1. 默认情况下，不带引号的值总是会去掉制表符和空格，而带引号的值将保持原样。通过使用“剪裁带引号的值”复选框或 /s.TrimQuoted 命令行选项，可以重写此行为。
2. 默认情况下，不带引号的 null 视为 null 值。通过使用“将不带引号的 NULL 作为字符串处理”复选框或 /s.NoUnquotedNulls 命令行选项，可以重写此行为（即将不带引号的 null 作为“null”字符串处理）。

下面是 CSV 导入的命令行示例︰

dt.exe /s:CsvFile /s.Files:.\Employees.csv /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:Employees /t.IdField:EntityID /t.CollectionTier:S3

## 从 Azure 表存储中导入

借助 Azure 表存储源导入程序选项，可以从单个 Azure 表存储表中导入，并且选择筛选要导入的表实体。



Azure 表存储源选项的屏幕截图

Azure 表存储连接字符串的格式为：

DefaultEndpointsProtocol=<protocol>;AccountName=<Account Name>;AccountKey=<Account Key>;

[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 Azure 表存储实例。

输入将从其中导入数据的 Azure 表的名称。可以根据需要指定[筛选器](https://msdn.microsoft.com/library/azure/ff683669.aspx)。

Azure 表存储源导入程序选项具有下列附加选项︰

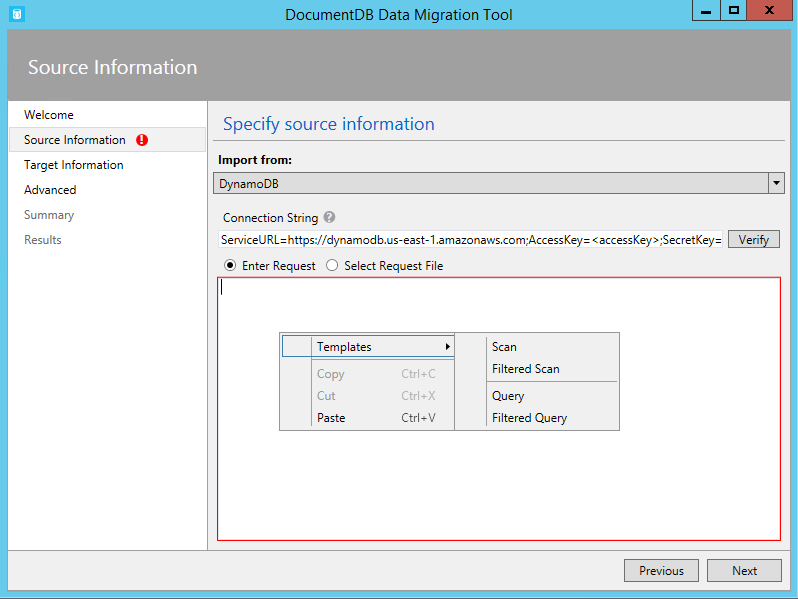
1. 包括内部字段
   1. 所有 - 包括所有内部字段（PartitionKey、RowKey 和 Timestamp）
   2. 无 - 排除所有内部字段
   3. RowKey - 仅包括 RowKey 字段
2. 选择列
   1. Azure 表存储筛选器不支持投影。如果想要仅导入特定的 Azure 表实体属性，请将它们添加到“选择列”列表中。这样将忽略所有其他实体属性。

下面是一个用于从 Azure 表存储中导入的命令行示例︰

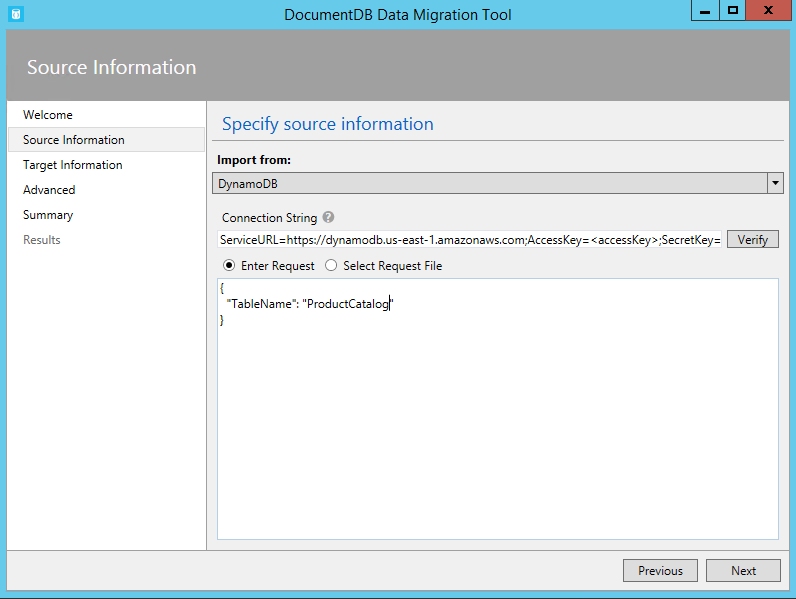
dt.exe /s:AzureTable /s.ConnectionString:"DefaultEndpointsProtocol=https;AccountName=<Account Name>;AccountKey=<Account Key>" /s.Table:metrics /s.InternalFields:All /s.Filter:"PartitionKey eq 'Partition1' and RowKey gt '00001'" /s.Projection:ObjectCount;ObjectSize /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:metrics /t.CollectionTier:S3

## 从 Amazon DynamoDB 中导入

借助 Amazon DynamoDB 源导入程序选项，可以从单个 Amazon DynamoDB 表中导入，并且可以选择筛选要导入的实体。提供多个模板，以便尽可能简化导入设置。



Amazon DynamoDB 源选项的屏幕截图 - 数据库迁移工具



Amazon DynamoDB 源选项的屏幕截图 - 数据库迁移工具

Amazon DynamoDB 连接字符串的格式为：

ServiceURL=<Service Address>;AccessKey=<Access Key>;SecretKey=<Secret Key>;

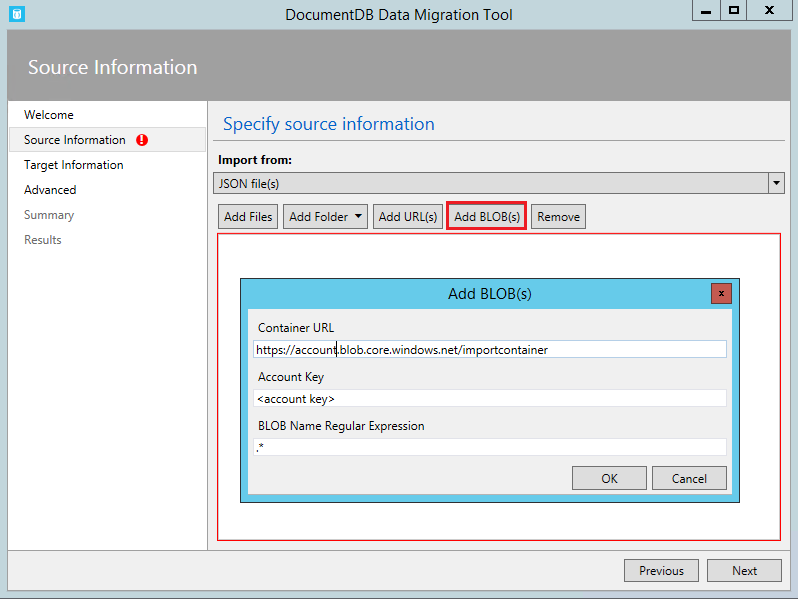
[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 Amazon DynamoDB 实例。

下面是一个用于从 Amazon DynamoDB 中导入的命令行示例︰

dt.exe /s:DynamoDB /s.ConnectionString:ServiceURL=https://dynamodb.us-east-1.amazonaws.com;AccessKey=<accessKey>;SecretKey=<secretKey> /s.Request:"{ """TableName""": """ProductCatalog""" }" /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:catalogCollection /t.CollectionTier:S3

## 从 Azure Blob 存储中导入文件

借助 JSON 文件、MongoDB 导出文件和 CSV 文件源导入程序选项，可以从 Azure Blob 存储中导入一个或多个文件。在指定 Blob 容器 URL 和帐户密钥后，只需提供正则表达式来选择要导入的文件。



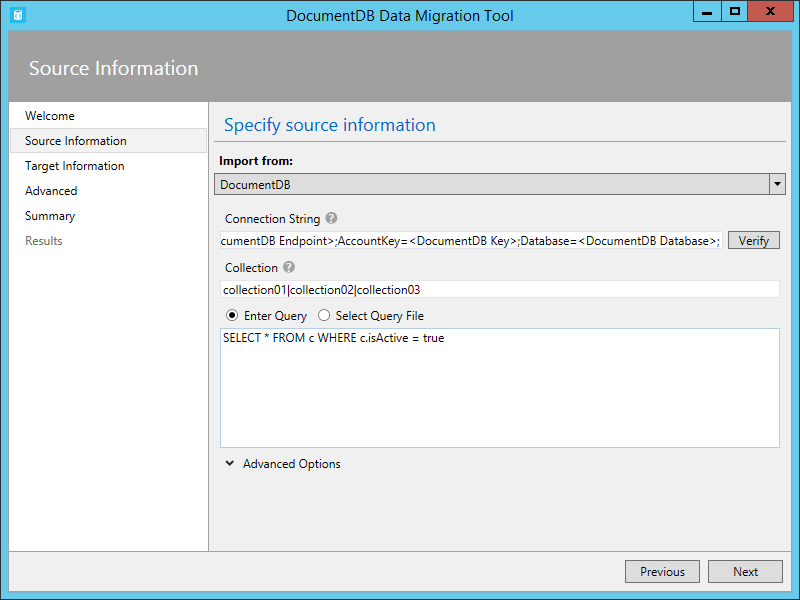
Blob 文件源选项的屏幕截图

下面是一个用于从 Azure Blob 存储中导入 JSON 文件的命令行示例︰

dt.exe /s:JsonFile /s.Files:"blobs://<account key>@account.blob.core.windows.net:443/importcontainer/.\*" /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:doctest

## 从 DocumentDB 中导入

借助 DocumentDB 源导入程序选项，可以从一个或多个 DocumentDB 集合中导入数据，并且选择使用查询来筛选文档。



DocumentDB 源选项的屏幕截图

DocumentDB 连接字符串的格式为︰

AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;

DocumentDB 帐户连接字符串可从 Azure 门户的“密钥”边栏选项卡中检索到（如[管理 DocumentDB 帐户的方法](/documentation/articles/documentdb-manage-account)中所述），但是还需要将数据库的名称追加到连接字符串中（格式如下）︰

Database=<DocumentDB Database>;

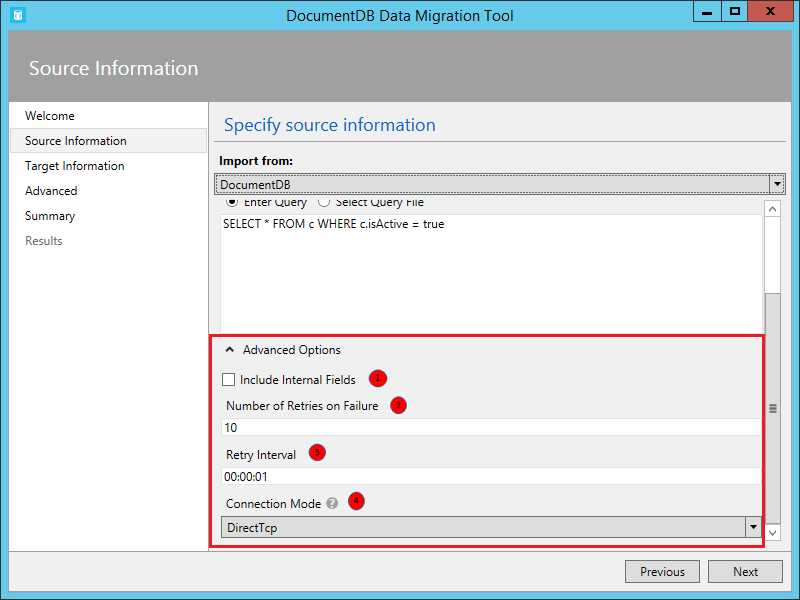
[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 DocumentDB 实例。

要从单个 DocumentDB 集合中导入，请输入将从其中导入数据的集合的名称。若要从多个 DocumentDB 集合中导入，请提供与一个或多个集合名称相匹配的正则表达式（例如 collection01 | collection02 | collection03）。可以选择为查询指定或提供一个文件来筛选和形成要导入的数据。

[AZURE.NOTE] 由于集合字段接受正则表达式，因此如果要从名称包含正则表达式字符的单个集合中导入，则必须相应地转义这些字符。

DocumentDB 源导入程序选项具有下列高级选项︰

1. 包括内部字段︰指定是否在导出中包括 DocumentDB 文档系统属性（例如 \_rid、\_ts）。
2. 失败重试次数︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接的次数。
3. 重试间隔︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接等待的时间间隔。
4. 连接模式︰指定要配合 DocumentDB 使用的连接模式。可用选项包括 DirectTcp、DirectHttps 和网关。直接连接模式速度更快，而网关模式对于防火墙更加友好，因为它仅使用端口 443。



DocumentDB 源高级选项的屏幕截图

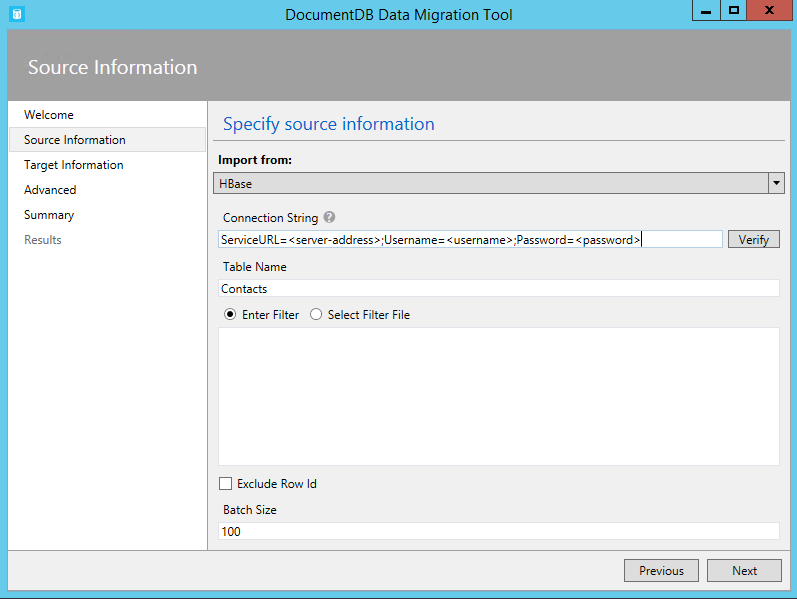
[AZURE.TIP] 导入工具默认设置为 DirectTcp 连接模式。如果遇到防火墙问题，请切换到网关连接模式，因为它只需要端口 443。

下面是一些从 DocumentDB 中导入的命令行示例︰

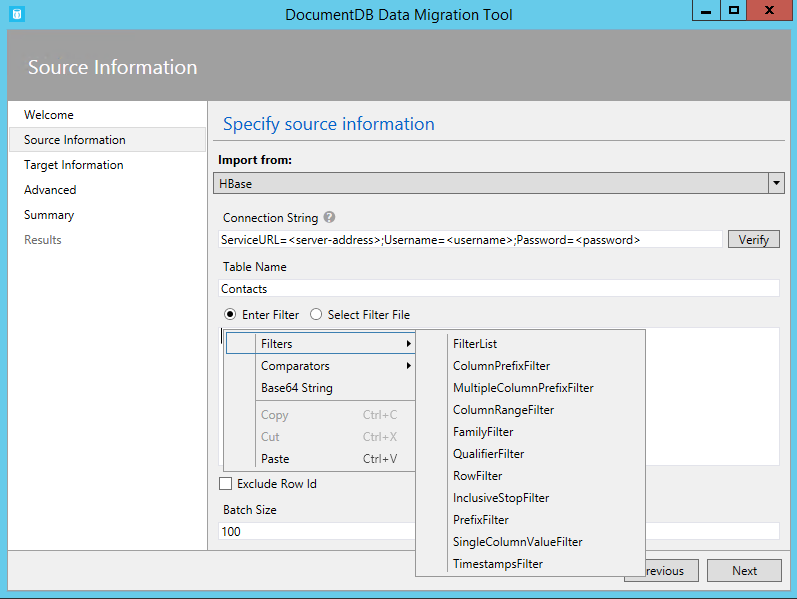
#Migrate data from one DocumentDB collection to another DocumentDB collections  
dt.exe /s:DocumentDB /s.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /s.Collection:TEColl /t:DocumentDBBulk /t.ConnectionString:" AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:TESessions /t.CollectionTier:S3  
  
#Migrate data from multiple DocumentDB collections to a single DocumentDB collection  
dt.exe /s:DocumentDB /s.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /s.Collection:comp1|comp2|comp3|comp4 /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:singleCollection /t.CollectionTier:S3  
  
#Export a DocumentDB collection to a JSON file  
dt.exe /s:DocumentDB /s.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /s.Collection:StoresSub /t:JsonFile /t.File:StoresExport.json /t.Overwrite /t.CollectionTier:S3

## 从 HBase 中导入

借助 HBase 源导入程序选项，可以从 HBase 表中导入数据并且选择筛选数据。提供多个模板，以便尽可能简化导入设置。



HBase 源选项的屏幕截图



HBase 源选项的屏幕截图

HBase Stargate 连接字符串的格式为︰

ServiceURL=<server-address>;Username=<username>;Password=<password>

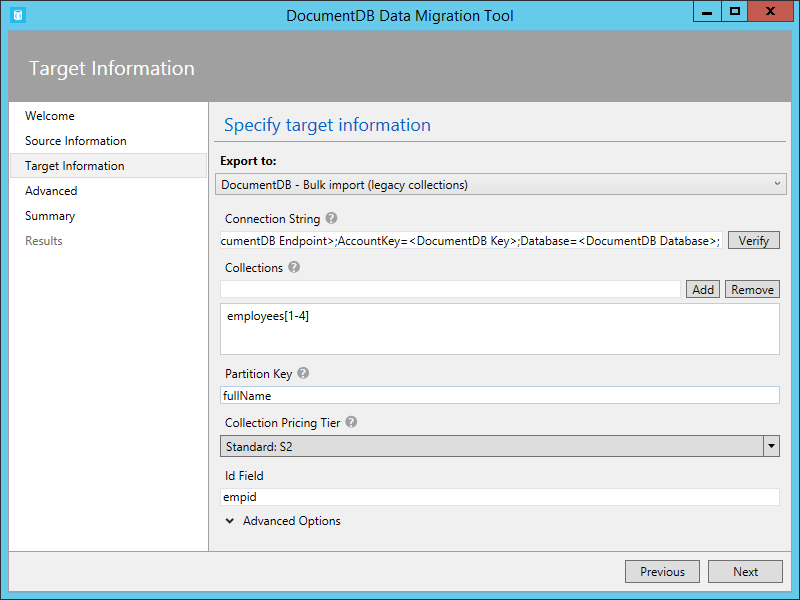
[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 HBase 实例。

下面是一个用于从HBase 中导入的命令行示例︰

dt.exe /s:HBase /s.ConnectionString:ServiceURL=<server-address>;Username=<username>;Password=<password> /s.Table:Contacts /t:DocumentDBBulk /t.ConnectionString:"AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;" /t.Collection:hbaseimport

## 导入到 DocumentDB（批量导入）

借助 DocumentDB 批量导入程序，可以使用 DocumentDB 存储的过程从所有可用的源选项中导入，以提高效率。该工具支持导入到单个 DocumentDB 集合，并支持分片导入，通过这种方法可跨多个 DocumentDB 集合对数据进行分区。有关数据分区的详细信息，请参阅 [Azure DocumentDB 中的分区和扩展](/documentation/articles/documentdb-partition-data)。该工具将创建、执行，然后删除目标集合中存储的过程。



DocumentDB 批量选项的屏幕截图

DocumentDB 连接字符串的格式为︰

AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;

DocumentDB 帐户连接字符串可从 Azure 门户的“密钥”边栏选项卡中检索到（如[管理 DocumentDB 帐户的方法](/documentation/articles/documentdb-manage-account)中所述），但是还需要将数据库的名称追加到连接字符串中（格式如下）︰

Database=<DocumentDB Database>;

[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 DocumentDB 实例。

要导入到单个 DocumentDB 集合，请输入将向其中导入数据的集合的名称，然后单击“添加”按钮。若要导入到多个集合，请分别输入每个集合的名称，或使用以下语法来指定多个集合︰*collection\_prefix*[起始索引-结束索引]。在通过前述语法指定多个集合时，请注意以下事项︰

1. 仅支持整数范围名称模式。例如，指定 collection[0-3] 将产生下列集合︰collection0、collection1、collection2 和 collection3。
2. 可以使用缩写的语法︰collection[3] 将生成与第 1 步相同的一组集合。
3. 可以提供多个替代。例如，collection[0-1] [0-9] 将生成 20 个带前导零的集合名称 (collection01，…02…03)。

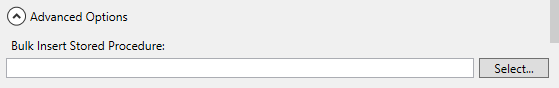
指定集合名称后，请选择集合所需的定价层（S1、S2 或 S3）。为了获得最佳导入性能，请选择 S3。有关性能级别的详细信息，请参阅 [DocumentDB 中的性能级别](/documentation/articles/documentdb-performance-levels)。

[AZURE.NOTE] 性能层设置仅适用于创建集合。如果指定的集合已经存在，将不会修改其定价层。

在导入到多个集合时，导入工具支持基于哈希的分片。在此方案中，指定要用作分区键的文档属性（如果分区键留空，文档将跨多个目标集合随机分片）。

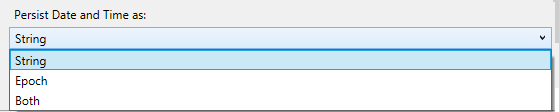
可以选择指定在导入过程中将导入源中的哪个字段用作 DocumentDB 文档 ID 属性（请注意，如果文档不包含此属性，则导入工具将生成 GUID 作为 ID 属性值）。

导入过程中可以使用多个高级选项。首先，虽然工具包含默认的批量导入存储过程 (BulkInsert.js)，但你可以选择指定自己的导入存储过程︰



DocumentDB 批量插入 sproc 选项的屏幕截图

此外，在导入日期类型时（例如从 SQL Server 或 MongoDB 中），可以选择三个导入选项中的一个︰

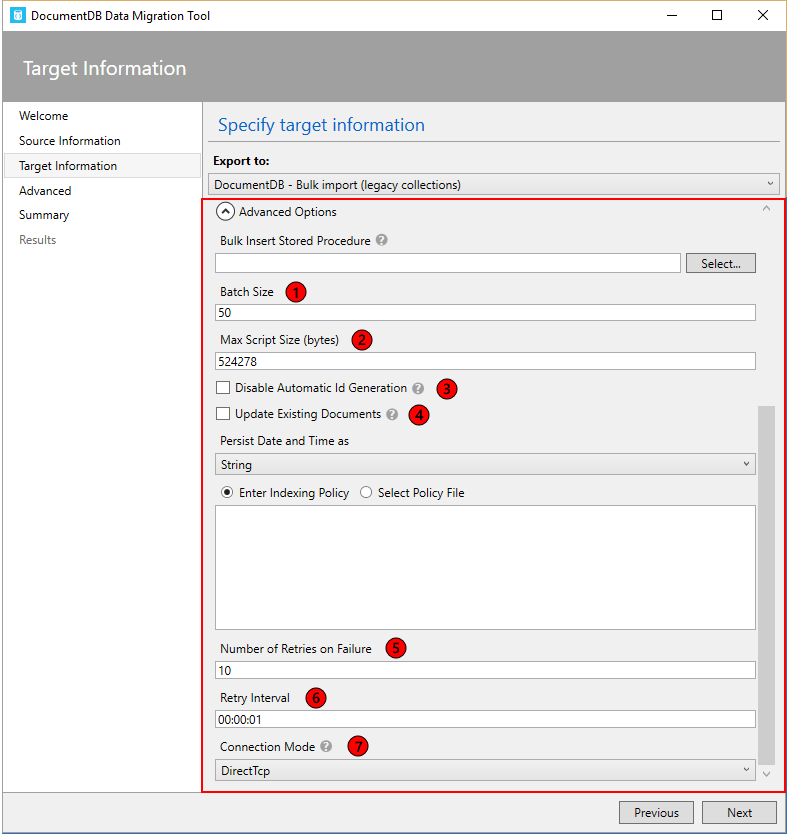


DocumentDB 日期时间导入选项的屏幕截图

* 字符串：保持字符串值
* Epoch：保持 Epoch 数字值
* 两者：保持字符串和 Epoch 数字值。此选项将创建一个子文档，例如：“date\_joined”: { “Value”: “2013-10-21T21:17:25.2410000Z”, “Epoch”: 1382390245 }

DocumentDB 批量导入程序具有下列高级附加选项︰

1. 批大小︰工具默认将批大小设置为 50。如果要导入的文档很大，请考虑减小批大小。如果要导入的文档很小，请考虑增大批大小。
2. 最大脚本大小（字节）︰工具默认设置为 512KB 的最大脚本大小
3. 禁用自动生成 ID︰如果要导入的每个文档都包含一个 ID 字段，则选择此选项可以提高性能。将不会导入缺少唯一 ID 字段的文档。
4. 更新现有文档︰工具将默认设置为不替换存在 ID 冲突的现有文档。选择此选项将允许覆盖具有匹配的 ID 的现有文档。此功能可用于更新现有文档的计划内数据迁移。
5. 失败重试次数︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接的次数。
6. 重试间隔︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接等待的时间间隔。
7. 连接模式︰指定要配合 DocumentDB 使用的连接模式。可用选项包括 DirectTcp、DirectHttps 和网关。直接连接模式速度更快，而网关模式对于防火墙更加友好，因为它仅使用端口 443。

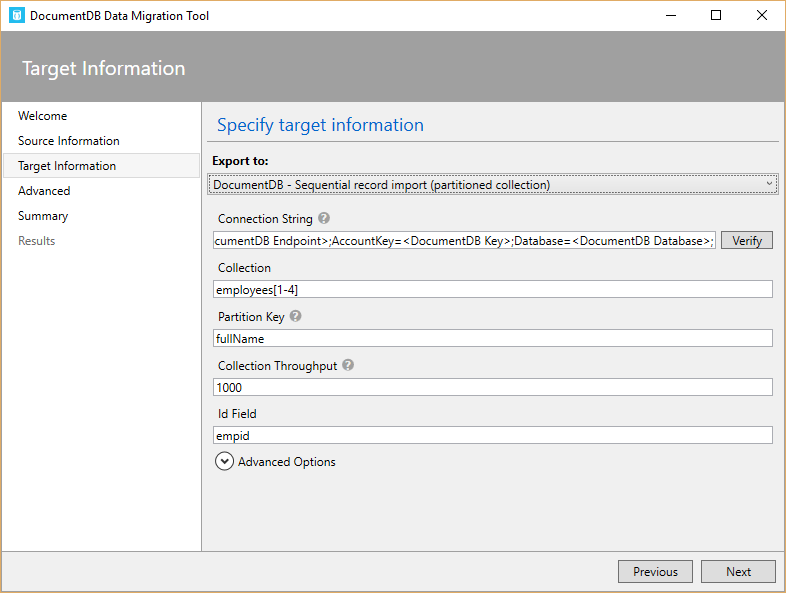


DocumentDB 批量导入高级选项的屏幕截图

[AZURE.TIP] 导入工具默认设置为 DirectTcp 连接模式。如果遇到防火墙问题，请切换到网关连接模式，因为它只需要端口 443。

## 导入到 DocumentDB（顺序记录导入）

借助 DocumentDB 顺序记录导入程序，可以从任何可用的源选项中逐条导入记录。如果要导入到已达到存储过程配额的现有集合中，可以选择此选项。该工具支持导入到单个 DocumentDB 集合，并支持分片导入，通过这种方法可跨多个 DocumentDB 集合对数据进行分区。有关数据分区的详细信息，请参阅 [Azure DocumentDB 中的分区和扩展](/documentation/articles/documentdb-partition-data)。



DocumentDB 顺序记录导入选项的屏幕截图

DocumentDB 连接字符串的格式为︰

AccountEndpoint=<DocumentDB Endpoint>;AccountKey=<DocumentDB Key>;Database=<DocumentDB Database>;

DocumentDB 帐户连接字符串可从 Azure 门户的“密钥”边栏选项卡中检索到（如[管理 DocumentDB 帐户的方法](/documentation/articles/documentdb-manage-account)中所述），但是还需要将数据库的名称追加到连接字符串中（格式如下）︰

Database=<DocumentDB Database>;

[AZURE.NOTE] 使用验证命令来确保可以访问在连接字符串字段中指定的 DocumentDB 实例。

要导入到单个 DocumentDB 集合，请输入将向其中导入数据的集合的名称，然后单击“添加”按钮。若要导入到多个集合，请分别输入每个集合的名称，或使用以下语法来指定多个集合︰*collection\_prefix*[起始索引-结束索引]。在通过前述语法指定多个集合时，请注意以下事项︰

1. 仅支持整数范围名称模式。例如，指定 collection[0-3] 将产生下列集合︰collection0、collection1、collection2 和 collection3。
2. 可以使用缩写的语法︰collection[3] 将生成与第 1 步相同的一组集合。
3. 可以提供多个替代。例如，collection[0-1] [0-9] 将生成 20 个带前导零的集合名称 (collection01，…02…03)。

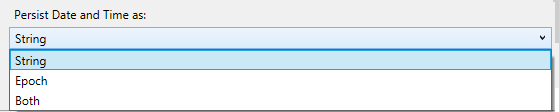
指定集合名称后，请选择集合所需的定价层（S1、S2 或 S3）。为了获得最佳导入性能，请选择 S3。有关性能级别的详细信息，请参阅 [DocumentDB 中的性能级别](/documentation/articles/documentdb-performance-levels)。

[AZURE.NOTE] 性能层设置仅适用于创建集合。如果指定的集合已经存在，将不会修改其定价层。

在导入到多个集合时，导入工具支持基于哈希的分片。在此方案中，指定要用作分区键的文档属性（如果分区键留空，文档将跨多个目标集合随机分片）。

可以选择指定在导入过程中将导入源中的哪个字段用作 DocumentDB 文档 ID 属性（请注意，如果文档不包含此属性，则导入工具将生成 GUID 作为 ID 属性值）。

导入过程中可以使用多个高级选项。首先，在导入日期类型时（例如从 SQL Server 或 MongoDB 中），可以选择三个导入选项中的一个︰

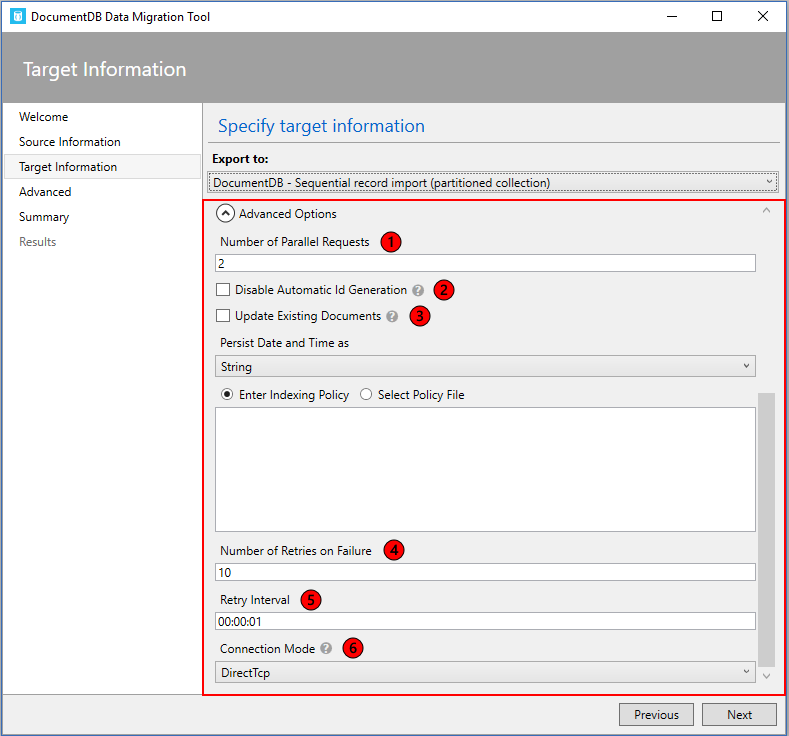


DocumentDB 日期时间导入选项的屏幕截图

* 字符串：保持字符串值
* Epoch：保持 Epoch 数字值
* 两者：保持字符串和 Epoch 数字值。此选项将创建一个子文档，例如：“date\_joined”: { “Value”: “2013-10-21T21:17:25.2410000Z”, “Epoch”: 1382390245 }

DocumentDB - 顺序记录导入程序具有下列高级附加选项︰

1. 并行请求数︰工具默认设置为 2 个并行请求。如果要导入的文档很小，请考虑增加并行请求的数量。请注意，如果此数字过大，则导入可能会遇到限制。
2. 禁用自动生成 ID︰如果要导入的每个文档都包含一个 ID 字段，则选择此选项可以提高性能。将不会导入缺少唯一 ID 字段的文档。
3. 更新现有文档︰工具将默认设置为不替换存在 ID 冲突的现有文档。选择此选项将允许覆盖具有匹配的 ID 的现有文档。此功能可用于更新现有文档的计划内数据迁移。
4. 失败重试次数︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接的次数。
5. 重试间隔︰指定在发生暂时性故障（例如网络连接中断）时重试 DocumentDB 连接等待的时间间隔。
6. 连接模式︰指定要配合 DocumentDB 使用的连接模式。可用选项包括 DirectTcp、DirectHttps 和网关。直接连接模式速度更快，而网关模式对于防火墙更加友好，因为它仅使用端口 443。

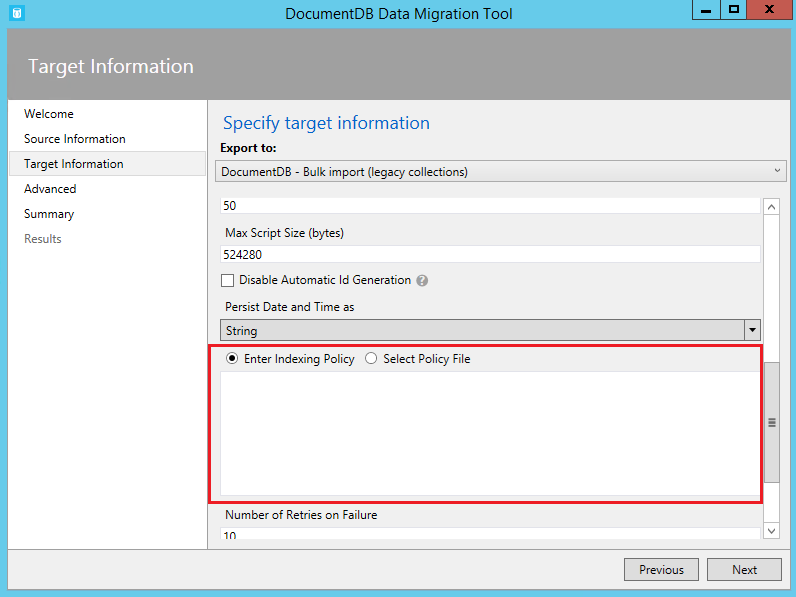


DocumentDB 顺序记录导入高级选项的屏幕截图

[AZURE.TIP] 导入工具默认设置为 DirectTcp 连接模式。如果遇到防火墙问题，请切换到网关连接模式，因为它只需要端口 443。

## 在创建 DocumentDB 集合时指定索引策略

当允许迁移工具在导入过程中创建集合时，可以指定集合的索引策略。在 DocumentDB 批量导入和 DocumentDB 顺序记录选项的高级选项部分，导航到“索引策略”部分。

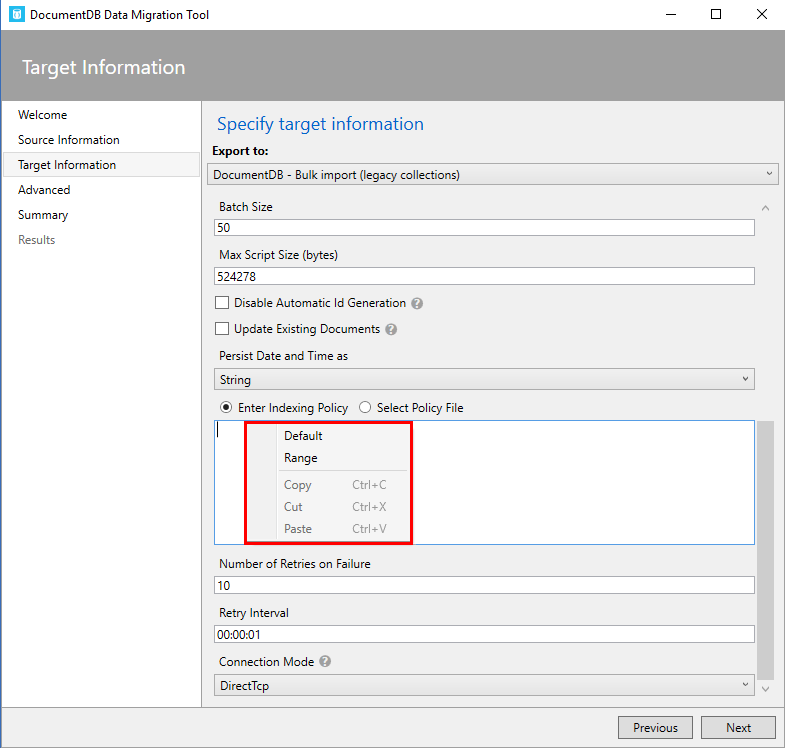


DocumentDB 索引策略高级选项的屏幕截图

使用索引策略高级选项，可以选择一个索引策略文件，手动输入索引策略，或从一组默认模板中选择（右键单击索引策略文本框）。

工具提供的策略模板包括︰

* 默认。针对字符串执行等式查询并针对数值使用 ORDER BY、范围和等式查询时，此策略最佳。与范围模板相比，此策略的索引存储开销较低。
* 范围。针对数值和字符串都使用 ORDER BY、范围和等式查询时，此策略最佳。与默认或哈希模板相比，此策略的索引存储开销较高。

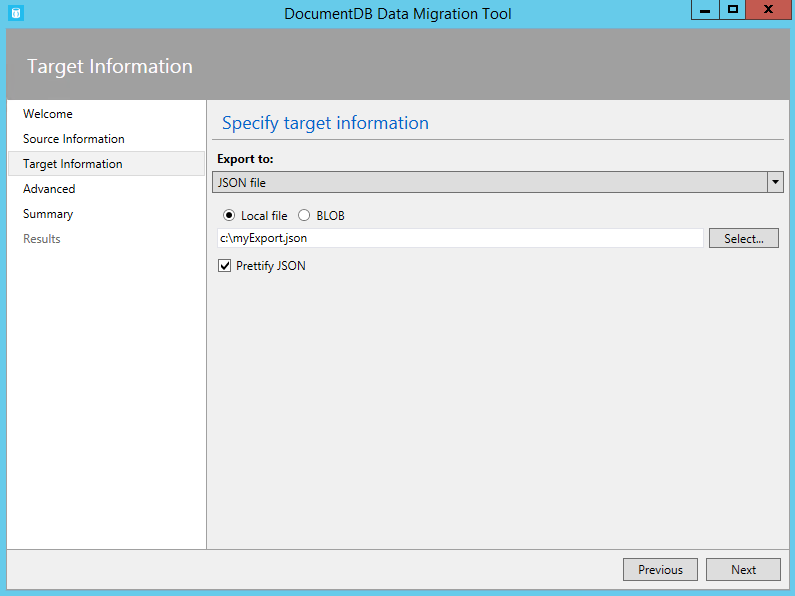


DocumentDB 索引策略高级选项的屏幕截图

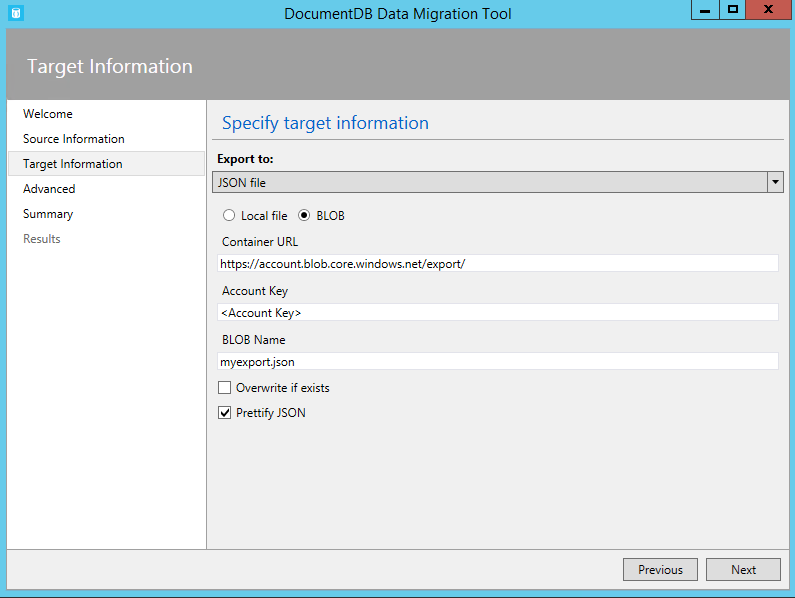
[AZURE.NOTE] 如果未指定索引策略，则将应用默认策略。有关索引策略的详细信息，请参阅 [DocumentDB 索引策略](/documentation/articles/documentdb-indexing-policies)。

## 导出到 JSON 文件

使用 DocumentDB JSON 导出程序，可以将所有可用的源选项导出到包含一组 JSON 文档的 JSON 文件。该工具可为你处理导出，你也可以选择查看生成的迁移命令并自己运行该命令。生成的 JSON 文件可能存储在本地或 Azure Blob 存储中。



DocumentDB JSON 本地文件导出选项的屏幕截图



DocumentDB JSON Azure Blob 存储导出选项的屏幕截图

可以根据需要选择整理生成的 JSON，这将增加生成文档的大小，同时提高内容的易读性。

Standard JSON export  
[{"id":"Sample","Title":"About Paris","Language":{"Name":"English"},"Author":{"Name":"Don","Location":{"City":"Paris","Country":"France"}},"Content":"Don's document in DocumentDB is a valid JSON document as defined by the JSON spec.","PageViews":10000,"Topics":[{"Title":"History of Paris"},{"Title":"Places to see in Paris"}]}]  
  
Prettified JSON export  
[  
{  
"id": "Sample",  
"Title": "About Paris",  
"Language": {  
 "Name": "English"  
},  
"Author": {  
 "Name": "Don",  
 "Location": {  
 "City": "Paris",  
 "Country": "France"  
 }  
},  
"Content": "Don's document in DocumentDB is a valid JSON document as defined by the JSON spec.",  
"PageViews": 10000,  
"Topics": [  
 {  
 "Title": "History of Paris"  
 },  
 {  
 "Title": "Places to see in Paris"  
 }  
]  
}]

## 高级配置

在高级配置屏幕中，指定要向其中写入错误的日志文件的位置。本页适用的规则如下：

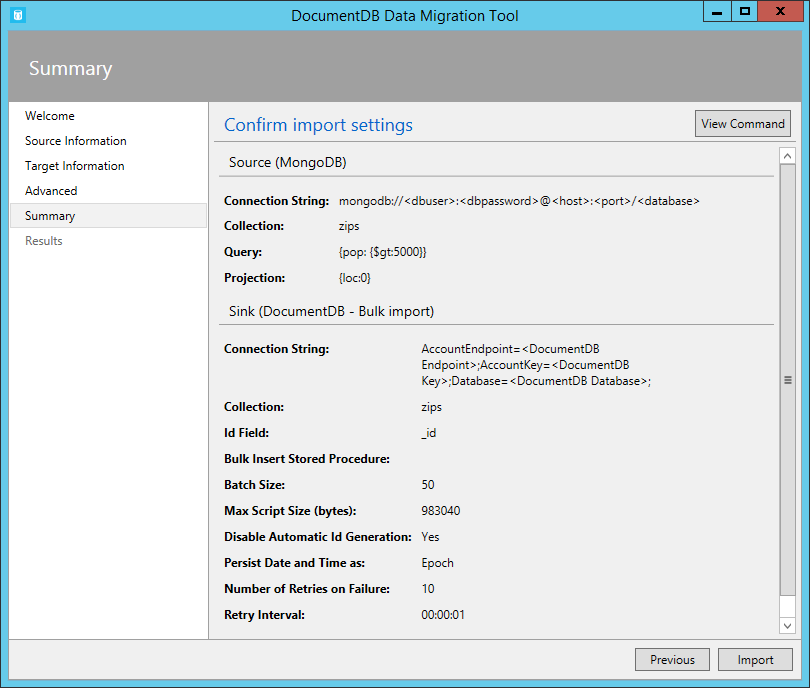
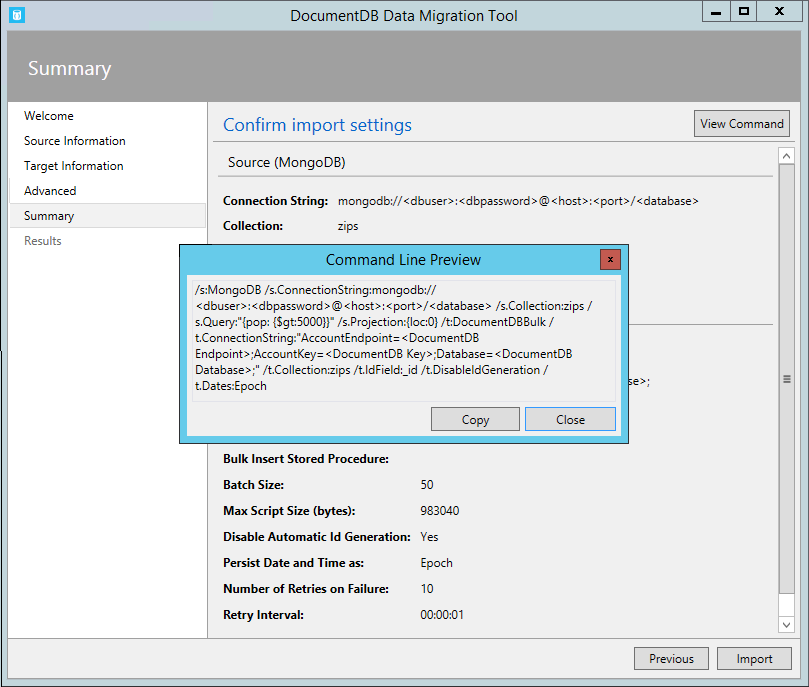
1. 如果未提供文件名，则将在结果页上返回所有错误。
2. 如果提供的文件名中不包含目录，则将在当前环境目录中创建（或覆盖）文件。
3. 如果选择一个现有文件，则该文件将被覆盖，不提供追加选项。

然后，选择是记录所有、关键还是无错误消息。最后，根据进度决定更新屏幕传输消息的频率。

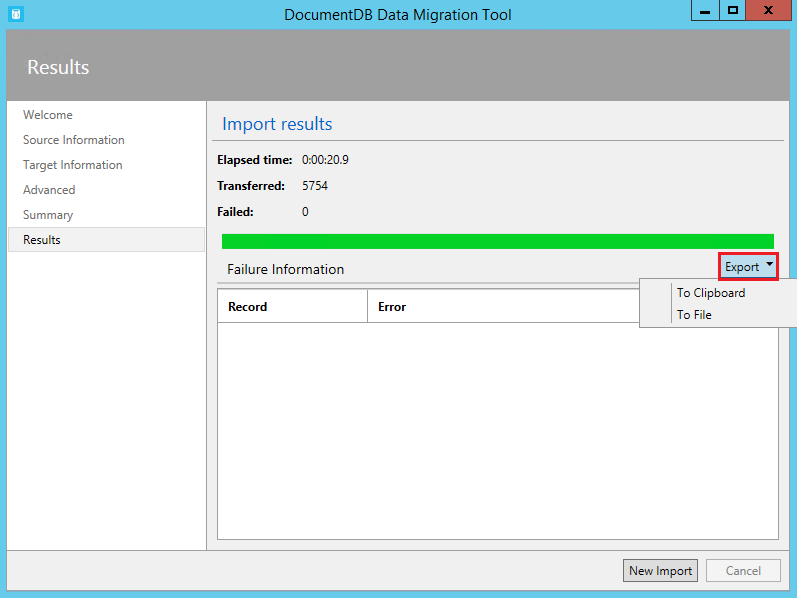
![Screenshot of Advanced configuration screen](./media/documentdb-import-data/AdvancedConfiguration.png)

## 确认导入设置并查看命令行

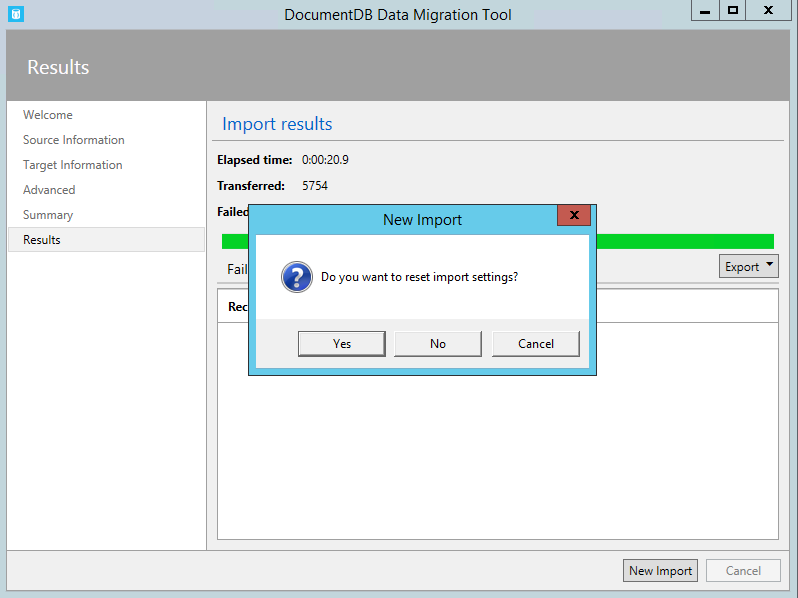
1. 在指定源信息、目标信息以及高级配置后，查看迁移摘要，并可选择查看/复制生成的迁移命令（复制命令对于自动执行导入操作非常有用）︰

* 
* 摘要屏幕的屏幕截图
* 
* 摘要屏幕的屏幕截图

1. 对源和目标选项满意后，单击**导入**。在执行导入时，将更新已用时间、传输数和失败信息（如果高级配置中未提供文件名）。完成后，可以导出结果（比如用于处理任何导入失败）。

* 
* DocumentDB JSON 导出选项的屏幕截图

1. 你也可以启动新的导入，可保留现有设置（例如连接字符串信息、源和目标选项等），也可重置所有值。

* 
* DocumentDB JSON 导出选项的屏幕截图

## 后续步骤

* 若要详细了解 DocumentDB，请访问[学习路径](/documentation/learning-paths/documentdb/)。