Orchestrating Big Data Solutions with Azure Data Factory

Lab 5 – Final Challenge

Overview

In this lab, you will use Azure Data Factory to implement a data orchestration solution for sales order data. You must create and run a pipeline that successfully transfers daily sales records to a database, separating the data into sales orders and line items as it is transferred.

You can choose to perform the required data transformations by using either U-SQL in an Azure Data Lake Analytics service, or by using Hive in an HDInsight cluster.

What You'll Need

To complete the labs, you will need the following:

- A web browser
- A Microsoft account
- A Microsoft Azure subscription
- A Windows, Linux, or Mac OS X computer
- The lab files for this course

Note: To set up the required environment for the lab, follow the instructions in the Setup document for this course.

Challenge 1: Prepare Azure Resources

The source data for your orchestration solution is provided as a collection of tab-delimited text files. Your solution must transfer these from Azure Storage to an Azure SQL Database.

1. Create the Azure resources you will need. These include:
    - An Azure Data Factory
    - An Azure SQL Database (hosted in an Azure SQL Database Server)
    - An Azure Storage Account
    - An Azure Data Lake Analytics service with an associated Azure Data Analytics Store (only required if you plan to use U-SQL to transform the data – if you plan to use Hive, your pipeline should provision an HDInsight cluster on-demand)
2. Configure your Azure Storage account so that you can use it to store blobs. Then upload the **sales** folder from the folder where you extracted the lab files to your Azure storage account. These files

are organized into a *year* and *month* hierarchy for the first quarter of 2016, and each day's transactions are in a separate file, which is named after the *day*.

3. Configure your Azure SQL Database Server so that it can be accessed by your Azure Data Factory and by a SQL Server client tool on your local computer. Then in your Azure SQL Database, use the following script (which you can copy and paste from **CreateTables.sql** in the **sales** folder) to create the **Orders** and **LineItems** table into which your solution must load the sales data:

```
CREATE TABLE Orders
(OrderDate datetime,
 OrderID int,
 CustomerID int,
 CustomerName nvarchar(100),
 Total money);

CREATE TABLE LineItems
 (OrderID int,
  OrderLineID int,
  StockItemID int,
  StockItemName nvarchar(100),
  Quantity int,
  UnitPrice money,
  LineTotal money);
```

4. If you plan to use U-SQL, in your Azure Data Lake Analytics service, add a data source that enables Azure Data Lake to access your Azure Storage account.

Challenge 2: Implement a Pipeline

Your pipeline must process the sales data in your Azure Storage account daily to generate two datasets each day – one containing order-level data, including a calculated total for the order; and one containing line item level data, including a calculated line total. It must then copy the order and line item data to the corresponding tables in your Azure SQL Database.

1. If you plan to use U-SQL to transform the data, examine the **sales.usql** script provided in the sales folder. If you plan to use Hive, examine the **sales.hql** script. These scripts have been provided to help you transform the data – you may adapt them as required for your specific solution.
2. Create a pipeline, together with the necessary linked services and datasets. The pipeline should run daily from January 1st 2016 to March 31st 2016, and include activities to:
    - Transform the sales data in your Azure Storage Account using either U-SQL or Hive, and generate the intermediary data files.
    - Load the orders data into the **Orders** table in Azure SQL Database
    - Load the line items data into the **LineItems** table in Azure SQL Database
3. Deploy and run the pipeline, monitoring the activity windows as they run.
4. After the pipeline has loaded data from January 1st 2016 to March 31st 2016, view the Activity Windows for the U-SQL or Hive activity that have a status of **Waiting: Dataset dependencies** to identify days on which no sales data is available.

Challenge 3: Explore the Loaded Data

After your pipeline has successfully loaded the tables in your Azure SQL database, you can query the data to find order totals and line item totals.

1. Use the following Transact-SQL queries (which you can copy and paste from **TestQueries.sql** in the **sales** folder) to check your transferred data:

```
-- Count orders by day
SELECT OrderDate, COUNT(*) AS Orders
FROM dbo.Orders
GROUP BY OrderDate
ORDER BY OrderDate;

-- Check a specific order
SELECT * FROM dbo.Orders
WHERE OrderID = 68417;

-- Check line items for a specific order
SELECT * FROM dbo.LineItems
WHERE OrderID = 68417;
```