Springboard

Azure Data Factory Hands-On Mini-Project

End-to-End Azure ELT solution

Estimated Time: 7-10 hours

Objective

In this mini-project, you will deploy an End-to-End Azure ELT solution. You will use Azure Data Factory and Mapping Dataflows to perform Extract Load Transform (ELT) using Azure Blob storage and Azure SQL DB. You will also use Azure DevOps repositories to perform source control over ADF pipelines. Optionally, you may use Azure DevOps pipelines to deploy across multiple environments including Dev, Test, and Production.

This hands-on mini-project is designed to provide exposure to many of Microsoft's transformative line of business applications. The goal is to show an end-to-end solution, leveraging many of these technologies, but not necessarily doing work in every component possible. The lab architecture is below and includes:

- Azure Data Factory (ADF)
- Azure Storage
- Azure Data Factory Mapping Dataflows
- Azure SQL Database
- Azure Key vault
- (optional) Azure DevOps

Check the deliverables at the end of this document before beginning your work.

Overview

Wide World Importers (WWI) imports products then resells them both to retailers and directly to the public. In an increasingly crowded market, they are always looking for ways to differentiate themselves and provide additional value to their customers.

They are looking to pilot a data warehouse to provide additional information useful to their internal sales and marketing agents. They want to enable their agents to perform as-is and as-was analysis in order to price the items accurately and predict the product demand throughout the year.

To extend their physical presence WWI recently acquired Smart Food, a supermarket business who provides comprehensive nutritional information to customers so they can make healthy decisions. SmartFoods runs a loyalty program where customers accumulate points on their purchases. WWI CIO is hoping to use the loyalty program information and the food nutrients database of SmartFoods to provide customers with a HealthSmart portal. The portal will be showing aggregated information on customers' important food nutrients (carbs, saturated fats, etc.) to promote healthy shopping.

In this hands-on mini-project, you'll build an end-to-end solution for data warehousing using data lake methodology.

Solution architecture

Below is a diagram of the solution architecture you will build in this mini-project. Please study this carefully so you understand the solution as a whole, before building various components.

Data sources

1. SmartFoods Rest API:

Type	Rest API
Authentication	Oauth2
Data Endpoints	1. Order line Transactions (CSV) 2. Customers (JSON) 3. Auth Token (JSON)
Frequency	Daily
Documentation	https://github.com/Mmodarre/retailDataGeneratorAzureFunction

2. SmartFoods Items

Type	On-premises local file system
Authentication	NA
Data Endpoints	1. Food (CSV) 2. Food-Nutrition (CSV) 3. Nutrition (CSV)
Frequency	NA – One Off

3. WWI OLTP

Type	SFTP
Authentication	Username/Password
Data Endpoints	1. Orderline Transactions (Parquet) 2. Orders Transactions (Parquet) 3. Customers (Parquet)
Frequency	Daily

Follow the instructions provided in the guided documents below to complete the lab:

1. Before the hands-on lab (Prepare the environment)
2. Linked Services Datasets and Integration Runtimes
3. Copy Activity Parameters Debug and Publishing
4. Lookup activity ForEach loop and Execute Pipeline activity
5. Get Metadata activity filter activity and complex expressions
6. Self-hosted Integration Runtime - decompress files and Delete activity

Azure Data Factory Mapping Data Flows:

7. SmartFoodsCustomerELT
8. ELT with Mapping Data Flows – Practice exercises

Springboard

By the end of this mini-project, you will learn to:

- Deploy Azure Data Factory including an Integration Runtime.
- Build Mapping Data Flows in ADF.
- Create Blob Storage and Azure SQLDB Linked Services.
- Create Azure Key Vault and Linked Services in ADF.
- Create an ADF parameterized pipeline.
- Install Azure Data Factory self-hosted integration runtime to ingest from on-premises data systems.

It is **strongly** advised that you take screenshots of steps and settings that you make in order to complete each part of the lab. This will serve as a revision guide for yourself and will facilitate a more meaningful discussion with your mentor if you require help.

Deliverables

- A written document that includes your reflections on the following questions. Submit it to your mentor and discuss during your weekly call.
- (Optional) A document with screenshots of each step as performed according to this lab.

  Answer these questions to demonstrate your understanding at the end of this lab:

1. Why should one use Azure Key Vault when working in the Azure environment? What are the pros and cons? What are the alternatives?
2. How do you achieve loop functionality within a Azure Data Factory pipeline? Why would you need to use this functionality in a data pipeline?
3. What are expressions in Azure Data Factory? How are they helpful when designing a data pipeline? Please explain with an example.
4. What are the pros and cons of parametrizing a dataset's activity in Azure Data Factory?
5. What are the different supported file formats and compression codecs in Azure Data Factory? When will you use a Parquet file over an ORC file? Why would you choose an AVRO file format over a Parquet file format?