

Azure ETL Pipeline

Adfolks Project

Amal Tom Sunny[†]

November 30, 2020

[†]amaltomsunny@gmail.com



Contents

1. Introduction

2. Architectural Design



Problem Statement

Create an ETL pipeline using Azure Cloud Services and provide the key reporting information via API calls.

- Facts:- Sales receipts
- Dimension:- Product Information

Services Used

- Azure Blob Storage - Staging Layer
- Azure Data Factory - ETL and Orchestration
- Azure Event Grid - Event management
- Azure SQL - Storing the Reporting table
- Python Flask - To provide API end point

ADF Flows

1. Sales Data - Simple Aggregated Dump

Data load is event based trigger. Upon a new file creation the load will execute



Figure: Mapping

ADF Flows

1. Dimension Data - SCD -2

SCD-II with historical snapshots

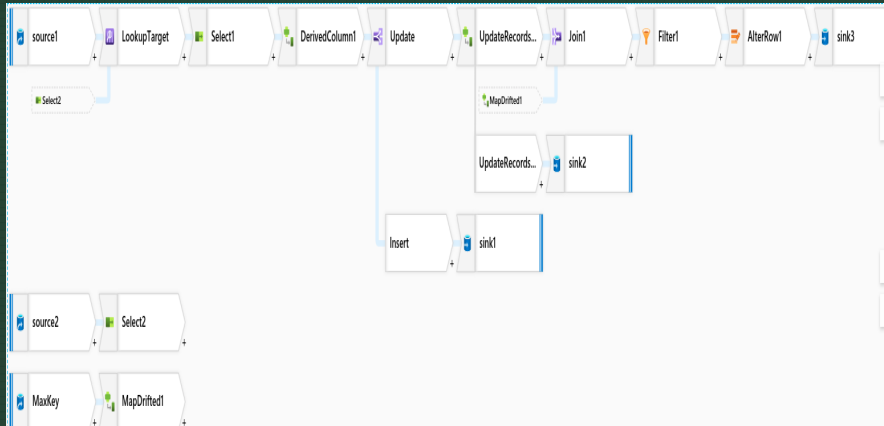


Figure: Mapping

Azure SQL

1. Reporting Layer is View in Azure SQL which has data in required form. Design helps to handle business changes in requirements.

Results

Messages

Search to filter items...

pid	product_name	pyear	qty_total	revenue_total
1	Asus Watch 3	2018	78	7800
1	Asus Watch 3	2019	987	98700
1	Asus Watch 3	2020	108	10800
2	Impex 8975 speaker	2020	19	4750
3	Asus R558UQ	2016	40	3360
3	Asus R558UQ	2020	44	3696
4	Iphone 11	2019	100	100000
4	Iphone 11	2020	5	5000

Figure: Query result

Azure SQL

1. SCD - II Dimension table with Active Flags

Results Messages				
Search to filter items...				
sl_no	product_id	product_name	active_flag	load_date
7	1	Asus Watch 3	1	2020-11-29T00:00:00.0000000
1	1	Asus Watch 2	0	2020-11-29T00:00:00.0000000
2	2	Impex 8975 speaker	0	2020-11-29T00:00:00.0000000
8	2	Impex 8975 speaker	1	2020-11-29T00:00:00.0000000
9	3	Asus R558UQ	1	2020-11-29T00:00:00.0000000
3	3	Asus R558UQ	0	2020-11-29T00:00:00.0000000
4	4	Iphone 11	0	2020-11-29T00:00:00.0000000
10	4	Iphone 11	1	2020-11-29T00:00:00.0000000

Figure: Query result

API- Flask and Python

1. Different API End points for the different questions

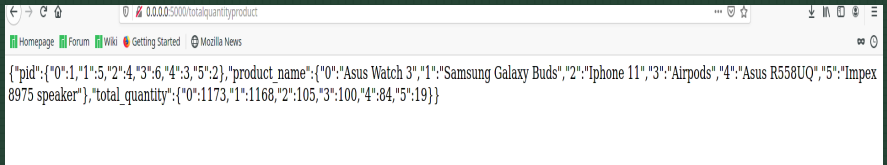


Figure: API for total quantity by product