

Chinook OLTP → Snowflake Data Warehouse Integration using Azure Data Factory

1. Objective

The goal of this project is to replicate the *Chinook OLTP to Data Warehouse* data integration workflow using **Azure Data Factory (ADF)**, **Azure SQL Database**, **Azure Blob Storage**, and **Snowflake**.

The deliverable demonstrates an end-to-end ETL process that extracts data from the Chinook transactional database hosted in Azure SQL, stages it in Azure Blob Storage as Parquet files, and then loads it into Snowflake's **STAGE** and **DW** schemas using ADF pipelines and data flows.

2. Architecture Overview

The architecture follows a layered ELT approach:

1. Extract Layer (ADF Pipeline 1)

- Source: Azure SQL Database (Chinook OLTP)
- Destination: Azure Blob Storage (Parquet format)

2. Stage Layer (ADF Pipeline 2)

- Reads Parquet data from Blob Storage
- Loads into Snowflake **STAGE** schema tables

3. Transformation Layer (ADF Data Flow)

- Performs transformation logic (hashing, deduplication, merge updates)
- Loads transformed data into Snowflake **DW** schema tables

4. Data Warehouse Layer

- Snowflake hosts final **DATE_DIM**, **TIME_DIM**, **CUSTOMER_DIM**, **ARTIST_DIM**, and **SALES_FACT** tables.
- These form a star schema suitable for business analytics.

3. Tools and Technologies Used

Component	Purpose
Azure SQL Database	Stores source Chinook OLTP tables (Customer, Invoice, Artist, Album)
Azure Data Factory (ADF)	Orchestrates extraction, staging, and loading via pipelines and data flows
Azure Blob Storage	Intermediate storage for Parquet files
Azure Key Vault	Secures Snowflake and SQL credentials
Snowflake	Target cloud data warehouse hosting STAGE and DW schemas
DBeaver / Snowflake UI	Used for running SQL DDL, validation, and testing

4. Azure Setup

Screenshot: [Setup.png](#)

Created the following resources:

- SQL Server: [damg7370fall2025.database.windows.net](#)
- SQL Database: [DAMG7370FALL2025](#)
- Storage Account: [stgchinookdamg](#)
- Azure Data Factory: [adfchinookdamg](#)
- Azure Key Vault: [AzureKeyVault1](#)

Networking setup:

- Added client IP address to firewall
- Enabled “Allow Azure services to access server”
- Verified connection via SSMS and ADF

5. Datasets

Screenshots: [Dataset1.png](#), [Dataset2.png](#)

Dataset Name	Type	Linked Service	Purpose
sqlserverdb_chinook	Azure SQL Database	Sql_db_Chinook	Source dataset
Parquet_ds	Parquet (Blob)	Storage_Chinook	Staging Parquet output
SnowChinook_Ds	Snowflake	snow_chinook	Target dataset for loading into STAGE schema

6. Pipeline 1 — Extract SQL DB to Parquet

Screenshot: [1stpipeline.png](#)

Pipeline Name: [extract_SQLDB_PL](#)

- Uses a **ForEach** activity looping through the array:
[\["Customer", "Artist", "Album", "Invoice"\]](#)
- Inside loop: Copy Activity ([sql_2_parquet](#))
 - Source: Azure SQL Database
 - Sink: Azure Blob Storage (Parquet)
- Managed identity access configured for Blob Storage.
- All activities succeeded with matching row counts.

8. Pipeline 2 — Parquet to Snowflake Stage

Screenshot: [2nd pipeline.png](#)

Pipeline Name: [Parquet_2_SnowStage_PL](#)

- Reads Parquet files from Blob container [/stage_data/](#)
- Writes into Snowflake [STAGE](#) tables:
 - [STAGE.CUSTOMER](#)
 - [STAGE.ARTIST](#)
 - [STAGE.ALBUM](#)
 - [STAGE.INVOICE](#)
- Connected using SAS-based Snowflake Linked Service for stage loading.

9. Data Flow — Load Customer Dimension

Screenshot: [Dataflow.png](#)

Data Flow Name: [DF_Load_Customer_DIM](#)

Steps:

1. **Source:** [STAGE.CUSTOMER](#)
2. **Derived Column:** Generate [CUSTOMER_HASH](#) using SHA-256 for change detection.
3. **Join:** Compare incoming vs existing customers.
4. **Sink:** Upsert into [DW.CUSTOMER_DIM](#)
 - Merge logic based on hash value.
 - Insert new and update changed records.

10. Snowflake Schemas and Tables

Screenshot: Dimtables.png

Executed SQL scripts in Snowflake:

- 1. create_stage_schema.sql
- 2. create_dw_schema.sql
- 3. load_date_dim.sql
- 4. load_time_dim.sql
- 5. merge_artist_dim.sql
- 6. load_sales_fact.sql
- 7. validation_counts.sql

Final Star Schema Tables:

- **Dimensions:**
DATE_DIM, TIME_DIM, CUSTOMER_DIM, ARTIST_DIM
- **Fact Table:**
SALES_FACT (references DATE_DIM_KEY and CUSTOMER_KEY)

11. SQL Script Summary

Script	Purpose
create_stage_schema.sql	Defines staging layer tables
create_dw_schema.sql	Creates DW layer (DIM & FACT)
load_date_dim.sql	Populates DATE_DIM
load_time_dim.sql	Populates TIME_DIM

merge_artist_dim.sql

Incremental load for ARTIST_DIM

load_sales_fact.sql

Inserts transactional data into SALES_FACT

validation_counts.sql

Validates record counts across all layers

All Screenshots

SETUP

Microsoft Azure

Search resources, services, and docs (G+)

Copilot

loya.ar@northeastern.edu

Home >

Recent

Browse all resources in this subscription

Review service health across resources

Identify performance issues in resources

Manage view Refresh Export to CSV Clear Assign tags

Group by none

Filter for any field...

Subscription equals allResource Group equals allType equals allLocation equals allAdd filter

<input type="checkbox"/>	Name	Type	Location	Resource Group	Subscription	Last accessed
<input type="checkbox"/>	adfchinookdamg	Data factory (V2)	West US 3	DAMG7370FALL2025	Azure for Students	2 minutes ago
<input type="checkbox"/>	stgchinookdamg	Storage account		DAMG7370FALL2025	Azure for Students	2 minutes ago
<input type="checkbox"/>	DAMG7370FALL2025	Resource group	East US	DAMG7370FALL2025	Azure for Students	3 minutes ago
<input type="checkbox"/>	damg7370fall2025	SQL server	West US 3	DAMG7370FALL2025	Azure for Students	14 hours ago
<input type="checkbox"/>	DAMG7370FALL2025	SQL database	West US 3	DAMG7370FALL2025	Azure for Students	16 hours ago
<input type="checkbox"/>	adf-chinook-damg	Key vault	West US 3	DAMG7370FALL2025	Azure for Students	4 days ago

DATA SET 1

Microsoft Azure

Data Factory > adfchinookdamg

Search

loya.ar@northeastern.edu

Home

Author

Monitor

Manage

Learning Center

Data Factory

Validate all

Publish all

Preview experience Off

Factory Resources

Filter resources by name

Pipelines2

extract_SQLDB_PL

pipeline1

Change Data Capture (preview)0

Datasets3

Parquet_ds

SnowChinook_Ds

sqlserverdb_chinook

Data flows0

Power Query0

Parquet_ds

SnowChinook_Ds

sqlserverdb_chinook

SQL

Azure SQL Database

sqlserverdb_chinook

ConnectionSchemaParameters

Linked service *

Sql_db_Chinook

Test connection

Edit

New

Learn more

Table

@dataset().schema_name

@dataset().table_name

Preview data

Enter manually

DATASET 2

The screenshot shows the Microsoft Azure Data Factory interface. The left sidebar contains navigation options: Home, Author, Monitor, Manage, and Learning Center. The main area displays the 'Factory Resources' pane on the left, listing Pipelines (2), Change Data Capture (0), Datasets (3), Data flows (0), and Power Query (0). The 'Datasets' section is expanded, showing 'Parquet_ds', 'SnowChinook_Ds', and 'sqlserverdb_chinook'. The 'Parquet_ds' dataset is selected, and its configuration is shown in the main pane. The 'Connection' tab is active, showing the 'Storage_Chinook' linked service. The 'File path' is configured with three fields: '@dataset().container_name', '@dataset().folder_name', and '@dataset().file_name'. The 'Compression type' is set to 'snappy'.

PIPELINE- 1

The screenshot shows the Microsoft Azure Data Factory interface with the 'Activities' pane expanded. The 'Activities' section lists various activities: Move and transform, Synapse, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The 'ForEach' activity is selected, and its configuration is shown in the main pane. The 'ForEachSourceTable' activity is expanded, showing a list of activities: 'sql_2_parquet' and 'sql_2_parquet'. The 'Output' tab is active, showing a table of activity results.

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
sql_2_parquet	✓ Succeeded	Copy data	10/14/2025, 3:28:40 AM	17s	AutoResolveIntegrationRuntime (West U
sql_2_parquet	✓ Succeeded	Copy data	10/14/2025, 3:28:21 AM	16s	AutoResolveIntegrationRuntime (West U
sql_2_parquet	✓ Succeeded	Copy data	10/14/2025, 3:28:21 AM	18s	AutoResolveIntegrationRuntime (West U
sql_2_parquet	✓ Succeeded	Copy data	10/14/2025, 3:28:21 AM	15s	AutoResolveIntegrationRuntime (West U
ForEachSourceTable	✓ Succeeded	ForEach	10/14/2025, 3:28:21 AM	37s	

PIPELINE-2

Microsoft Azure | Data Factory | adfchinookdamg

Search

Factory Resources

- Pipelines
 - extract_SQLDB_PL
 - Parquest_2_snowstage_pl
- Change Data Capture (preview)
- Datasets
- Data flows
 - DF_Load_Customer_DIM
- Power Query

Activities

- Move and transform
- Synapse
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Activities

- ForEach
 - ForEachTable
 - Activities
 - Parquet_2_Snow_Sta...

Parameters Variables Settings Output

Pipeline run ID: 5b549a9f-0432-478a-ae27-089c6d1bedaf

Pipeline status: Succeeded

Data flow activity for this debug run will start as soon as the data flow debug session is ready.

All status List

Monitor in Azure Metrics Export to CSV

Showing 1 - 5 of 5 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
Parquet_2_Snow_Stage	Succeeded	Copy data	10/14/2025, 3:33:39 AM	46s	AutoResolveIntegrationRuntime (West U
Parquet_2_Snow_Stage	Succeeded	Copy data	10/14/2025, 3:33:35 AM	39s	AutoResolveIntegrationRuntime (West U
Parquet_2_Snow_Stage	Succeeded	Copy data	10/14/2025, 3:32:56 AM	39s	AutoResolveIntegrationRuntime (West U
Parquet_2_Snow_Stage	Succeeded	Copy data	10/14/2025, 3:32:56 AM	42s	AutoResolveIntegrationRuntime (West U
ForEachTable	Succeeded	ForEach	10/14/2025, 3:32:56 AM	1m 31s	

DATAFLOW

Microsoft Azure | Data Factory | adfchinookdamg

Search

Factory Resources

- Pipelines
 - extract_SQLDB_PL
 - pipeline1
- Change Data Capture (preview)
- Datasets
- Data flows
 - dataflow1
- Power Query

Activities

- CustomerStage
 - Import data from SnowChinook_Ds
- joinSourceTgt
 - Left outer join on 'CustomerStage' and 'sourceCustomerDIM'
- derivedColumn1
 - Creating/updating the columns 'CUSTOMERID, FIRSTNAME, LASTNAME, COMPANY, ADDRESS'
- RowChanges
 - Inserting, updating, deleting, and/or upserting using
- sinkCustomerDL...

Properties

General Related

Name: dataflow1

Description:

Sink Settings Errors Mapping Optimize Inspect Data preview

Schema

Input Output

Number of columns Updated: 0 Dropped: 3 Unchanged: 25 Total: 25

Order	Column	Type	Updated	Input column
1	CUSTOMERID	decimal		CUSTOMERID
2	FIRSTNAME	string		FIRSTNAME
3	LASTNAME	string		LASTNAME
4	COMPANY	string		COMPANY
5	ADDRESS	string		ADDRESS

Cancel

LINKED SERVICES

Microsoft Azure

Data Factory > adfchinookdamg

Search

Preview experience

Off

Home

Connector upgrade advis...

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

ADF in Microsoft Fabric

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoi...

Workflow orchestration manager

Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

Filter by name

Annotations : **Any**

Showing 1 - 5 of 5 items

Name ↑↓	Type ↑↓	Related ↑↓	Annotations ↑↓
AzureKeyVault1	Azure Key Vault	2	
snow_chinook	Snowflake V2	1	
Sql_db_Chinook	Azure SQL Database	1	
storage_sas_auth	Azure Blob Storage	1	
Storage_Chinook	Azure Blob Storage	1	

VALIDATION

The screenshot displays the Snowflake web interface. On the left is a navigation sidebar with sections like 'Work with data', 'Horizon Catalog', and 'Manage'. The main area is divided into three panes. The left pane shows the 'Database Explorer' with a tree view of the 'CHINOOK_DB' database, including schemas like 'DW', 'INFORMATION_SCHEMA', 'PUBLIC', 'STAGE', and 'MEDIA_DB'. The middle pane shows a SQL query in a text editor:

```

324
325 SELECT 'CUSTOMER' AS TBL, COUNT(*) CNT FROM STAGE.CUSTOMER
326 UNION ALL SELECT 'ARTIST', COUNT(*) FROM STAGE.ARTIST
327 UNION ALL SELECT 'ALBUM', COUNT(*) FROM STAGE.ALBUM
328 UNION ALL SELECT 'INVOICE', COUNT(*) FROM STAGE.INVOICE;
329

```

The right pane shows the 'Results' table, which has 4 rows and 2 columns: 'TBL' and 'CNT'. The data is as follows:

TBL	CNT
CUSTOMER	236
ARTIST	550
ALBUM	694
INVOICE	412

At the bottom, a 'Query History' pane shows the current query and its execution time of 1.3s.

DIMTABLES

+

+

+

Work with data

Projects

Ingestion

Transformation

AI & ML

Monitoring

Marketplace

Horizon Catalog

Catalog

Data sharing

Governance & security

Manage

Compute

Admin

\$393 credits left

Trial ends in 111 days

Upgrade

Arjun Loya

CHINOOK_ROLE

My Workspace

Worksheets

Database Explorer

Search for files

Owned

Shared with you

Scratchpad

Objects

Data Products

Search

Filter

CHINOOK_DB

DW

INFORMATION_SCHEMA

PUBLIC

STAGE

Tables 6

MEDIA_DB

SNOWFLAKE

SNOWFLAKE_LEARNING_DB

SNOWFLAKE_SAMPLE_DATA

USERS\$ARJUNLOYA

Untitled 4.sql

Untitled 7.sql

Untitled 8.sql

CHINOOK_ROLE

COMPUTE_WH (X-Small)

CHINOOK_DB

PUBLIC

```

330  -- DW dimension counts
331  SELECT 'DATE_DIM', COUNT(*) FROM DW.DATE_DIM
332  UNION ALL SELECT 'TIME_DIM', COUNT(*) FROM DW.TIME_DIM
333  UNION ALL SELECT 'CUSTOMER_DIM', COUNT(*) FROM DW.CUSTOMER_DIM
334  UNION ALL SELECT 'ARTIST_DIM', COUNT(*) FROM DW.ARTIST_DIM;
335
336  SELECT COUNT(*) AS SALES_FACT_ROWS FROM DW.SALES_FACT;
337

```

Results (just now)

Table Chart

1 row 122ms


#	SALES_FACT_ROWS
1	329

Query History

Current file All files

just now	122ms	SELECT COUNT(*) AS SALES_FACT_ROWS FROM DW.SALES_FACT;	01bf74a-3203-5ff6-0006-b9ea0005a0d6
just now	693ms	SELECT 'DATE_DIM', COUNT(*) FROM DW.DATE_DIM UNION ALL SELECT '...	01bf749-3203-6012-0006-b9ea000590a2
8 minutes ago	1.3s	SELECT 'CUSTOMER' AS TBL, COUNT(*) CNT FROM STAGE.CUSTOMER UNIO...	01bf741-3203-5f01-0006-b9ea0005f0be
9 minutes ago	102ms	SELECT 'CUSTOMER' AS TBL, COUNT(*) CNT FROM STAGE.CUSTOMER UNIO...	01bf741-3203-5fd2-0006-b9ea000567e6

FACT_TABLE_VALIDATION



My Workspace

+

+

+

Work with data

Projects

Ingestion

Transformation

AI & ML

Monitoring

Marketplace

Horizon Catalog

Catalog

Data sharing

Governance & security

Manage

Compute

Admin

\$393 credits left

Trial ends in 111 days

Upgrade

Arjun Loya

CHINOOK_ROLE

Worksheets

Search for files

Owned

Shared with you

Scratchpad

Database Explorer

Objects

Data Products

Search

Filter

CHINOOK_DB

DW

INFORMATION_SCHEMA

PUBLIC

STAGE

Tables 6

MEDIA_DB

SNOWFLAKE

SNOWFLAKE_LEARNING_DB

SNOWFLAKE_SAMPLE_DATA

USERSARJUNLOYA

My Workspace

Untitled 4.sql

Untitled 7.sql

Untitled 8.sql

CHINOOK_ROLE

COMPUTE_WH (X-Small)

CHINOOK_DB

PUBLIC

My Workspace

Untitled 7.sql

```
-- Dw dimension counts
330 SELECT 'DATE_DIM', COUNT(*) FROM DW.DATE_DIM
331 UNION ALL SELECT 'TIME_DIM', COUNT(*) FROM DW.TIME_DIM
332 UNION ALL SELECT 'CUSTOMER_DIM', COUNT(*) FROM DW.CUSTOMER_DIM
333 UNION ALL SELECT 'ARTIST_DIM', COUNT(*) FROM DW.ARTIST_DIM;
334
335 | SELECT COUNT(*) AS SALES_FACT_ROWS FROM DW.SALES_FACT;
336
337
```

Results (just now)

Table

Chart

1 row 122ms

#	SALES_FACT_ROWS
1	329

Query History

Current file

All files

just now 122ms SELECT COUNT(*) AS SALES_FACT_ROWS FROM DW.SALES_FACT; 01bf74a-3203-5ff6-0006-b9ea0005a0d6

just now 693ms SELECT 'DATE_DIM', COUNT(*) FROM DW.DATE_DIM UNION ALL SELECT '... 01bf749-3203-6012-0006-b9ea000590a2

8 minutes ago 1.3s SELECT 'CUSTOMER' AS TBL, COUNT(*) CNT FROM STAGE.CUSTOMER UNIO... 01bf741-3203-5f91-0006-b9ea0005f0ba

9 minutes ago 102ms SELECT 'CUSTOMER' AS TBL, COUNT(*) CNT FROM STAGE.CUSTOMER UNIO... 01bf741-3203-5fd2-0006-b9ea000567e6

THANKYOU