WeCloudData

Mini Project--ETL and Data Loads

Data Engineering Diploma

Content developed by: WeCloudData Academy

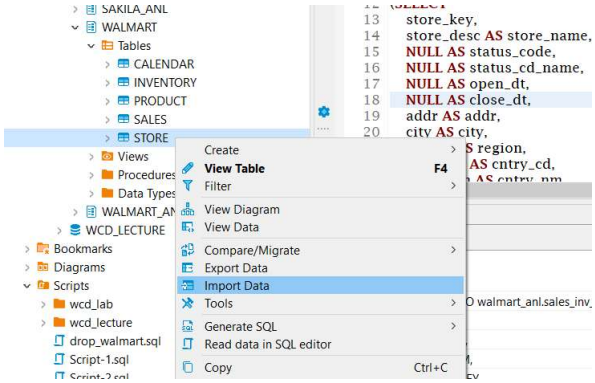
1. Project Requirements

Step 1 - Load The Original Dataset

You have a set of data. The dataset includes tables **store**, **product**, **inventory**, **sales** and **calendar**. These are the tables from operational databases.

Store		product		Inventory		sales		calendar	
store_key	*integer*	prod_key	*integer*	cal_dt	*date*	trans_id	*date*	cal_dt	*date*
store_num	*varchar(30)*	prod_name	*varchar(30)*	store_key	*int*	prod_key	*int*	cal_type_desc	*varchar(30)*
addr	*varchar(500)*	vol	*number(38,2)*	prod_key	*int*	store_key	*int*	day_of_wk_num	*integer*
city	*varchar(50)*	wgt	*number(38,2)*	inventory_on_hand_qty	*number(38,2)*	trans_dt	*date*	day_of_wk_desc	*varchar(30)*
region	*varchar(100)*	brand_name	*varchar(100)*	inventory_on_order_qty	*number(38,2)*	trans_time	*int*	yr_num	*integer*
cntry_cd	*varchar(30)*	status_code	*varchar(30)*	out_of_stock_flg	*boolean*	sales_qty	*number(38,2)*	wk_num	*integer*
cntry_nm	*varchar(30)*	status_code_name	*varchar(30)*	waste_qty	*number(38,2)*	sales_price	*number(38,2)*	yr_wk_num	*integer*
postal_zip_cd	*varchar(10)*	category_key	*integer*	promotion_flg	*boolean*	sales_amt	*number(38,2)*	mnth_num	*integer*
prov_state_desc	*varchar(30)*	category_name	*varchar(30)*	next_delivery_dt	*date*	discount	*number(38,2)*	yr_mnth_num	*integer*
prov_state_cd	*varchar(30)*	subcategory_key	*integer*			sales_cost	*number(38,2)*	qtr_num	*integer*
store_type_cd	*varchar(30)*	subcategory_name	*varchar(30)*			sales_mgrn	*number(38,2)*	yr_qtr_num	*integer*
store_type_desc	*varchar(100)*					ship_cost	*number(38,2)*		
frnchs_flg	*boolean*								
store_size	*numeric*								
market_key	*integer*								
market_name	*varchar(150)*								
submarket_key	*integer*								
submarket_nam	*varchar(150)*								
latitude	*numeric(19, 6)*								
longitude	*numeric(19, 6)*								

You need to download the dataset from this link and load the data into a schema of the Snowflake data warehouse. (It is easier to load the csv files from DBeaver)



You can use this script to build the schema for the original system dataset tables.

Step 2 - Read and Understand the Business Requirements and data model

We have create the data model. You need to populate the data from the original schema to the data model schema.

First of all, The data model has 3 dimensions : **store**, **date**, **product**.

For the fact table, we need 2 fact tables:

- **Daily fact table**: On the daily level, the row grain is date + store + product. In the daily fact table, in addition to the existing columns from the sales and inventory tables, in the fact table we need one more column called "low_stock_flg". This flag is True when sales_qty in the sales table lower than the stock_on_hand_qty in the inventory at that date.
- **Weekly fact table**: We need the second fact table on weekly base. This week table contains all the aggregate values from the daily fact table and also includes some new columns:

- eop_stock_on_hand_qty: This is the on hand stock qty at the end of week (Saturday), which means the on hand stock qty of the last day in the week. So you can't simply aggregate it.
  - eop_stock_on_order_qty: This is the on order stock qty at the end of week (Saturday).
  - out_of_stock_times: During one week, how many times when the out_of_stock_flg is True.
  - in_of_stock_times: During one week, how many times when the in_of_stock_flg is True.
  - low_stock_times: During one week, how many times when the low_stock_flg is True.

So to wrap up, the data model of the project has 3 dimension tables and 2 fact tables (one for daily, one for weekly). The data model can be found from here.

Dimensions

calendar_dim	
cal_dt	date
cal_type_name	varchar(30)
day_of_wk_num	integer
year_num	integer
week_num	integer
year_wk_num	integer
month_num	integer
year_month_num	integer
qtr_num	integer
yr_qtr_num	integer
update_time	timestamp

store_dim	
store_key	integer
store_name	varchar(150)
status_code	varchar(10)
status_cd_name	varchar(50)
addr	varchar(50)
city	varchar(100)
region	varchar(30)
cntry_cd	integer
cntry_nm	varchar(30)
postal_zip_cd	integer
prov_name	varchar(50)
prov_code	integer
market_key	integer
market_name	varchar(150)
submarket_key	integer
submarket_name	varchar(150)
latitude	number(38,2)
longitude	number(38,2)
tlog_active_flg	boolean
start_dt	date
end_dt	date
update_time	timpstamp

product_dim	
prod_key	integer
prod_name	varchar(30)
vol	number(38,2)
wgt	number(38,2)
brand_name	varchar(30)
status_code	integer
status_code_name	varchar(30)
category_key	integer
category_name	varchar(30)
subcategory_key	integer
subcategory_name	varchar(30)
tlog_active_flg	boolean
start_date	date
end_date	date
update_time	timpstamp

Fact tables

Daily Fact

sales_inv_store_dy		Explanation
cal_dt	date	
store_key	integer	
prod_key	integer	
sales_qty	number(38,2)	sum(sales_qty) per day
sales_price	number(38,2)	avg(sales_price) per day
sales_amt	number(38,2)	sum(sales_amt) per day
discount	number(38,2)	avg(discount) per day
sales_cost	number(38,2)	sum(sales_cost) per day
sales_mgrn	number(38,2)	sum(sales_mgrn) per day
stock_on_hand_qty	number(38,2)	stock_on_hand_qty of the date
ordered_stock_qty	number(38,2)	stock_ordered_qty of the date
out_of_stock_flg	boolean	out_of_stock_flg of the date
in_stock_flg	boolean	in_stock_flg of the date
low_stock_flg	boolean	true when sales_qty > stock_on_hand_qty of the date
update_time	timestamp	

Weekly Fact

sales_inv_store_wk		
yr_num	integer	
wk_num	integer	
store_key	integer	
prod_key	integer	
wk_sales_qty	number(38,2)	sum(sales_qty) per week
avg_sales_price	number(38,2)	avg(sales_price) per week
wk_sales_amt	number(38,2)	sum(sales_amt) per week
wk_discount	number(38,2)	avg(discount) per week
wk_sales_cost	number(38,2)	sum(sales_cost) per week
wk_sales_mgrn	number(38,2)	sum(sales_mgrn) per week
eop_stock_on_hand_qty	number(38,2)	the on hand stock of the last day in the week
eop_ordered_stock_qty	number(38,2)	the on order stock of the last day in the week
out_of_stock_times	integer	count(out_of_stock_flg) per week
in_stock_times	integer	count(in_stock_flg) per week
low_stock_times	integer	count(low_stock_flg) per week
update_time	timestamp	

Step 3 - Create data model in Snowflake schema and Populate data in the data model for the first time.

- Create a schema on Snowflake.
- Create physical data model tables in the schema.
- Create ETL script to populate data from the original dataset to the data model.
- Be aware of your ETL scripts, make sure they are also available for the later Delta data loading.

Step 4 - Update the original dataset with the following scripts.

Update FACT raw tables

```
INSERT INTO walmart.SALES
VALUES
(302836,540260,3220,'2012-12-31',18,37.80,3.58,129.42,0.01,300.72,-150.37,5.47),
(312076,399912,3220,'2013-01-01',7,29.00,145.45,3773.59,0.02,3486.71,731.32,17.85),
(337584,135665,1104,'2013-01-02',18,11.00,41.32,447.09,0.09,365.40,89.03,8.66);


INSERT INTO walmart.INVENTORY
VALUES
('2012-12-31',1103,540260,26.46,75.60,1,0.00,'FALSE','2012-12-31'),
('2012-12-31',1103,904715,27.09,21.07,0,1.00,'FALSE','2012-12-31'),
('2013-01-01',1104,135665,11.00,14.30,1,0.00,'FALSE','2012-12-31'),
('2013-01-01',1104,200147,6.72,5.88,0,0.00,'TRUE','2012-12-31'),
('2013-01-02',1104,399912,7.83,46.98,1,1.00,'TRUE','2012-12-31');
```

Update DIM table product

```
UPDATE walmart.PRODUCT SET PROD_NAME='CHANGE-1' WHERE PROD_KEY=657768;
UPDATE walmart.PRODUCT SET PROD_NAME='CHANGE-2' WHERE PROD_KEY=293693;
INSERT INTO walmart.PRODUCT VALUES (999999,'ADD-1',2.22, 88.88, 'brand-999', 1, 'active', 4, 'category-4', 1, 'subcategory-1');
```

Step 5 - Check if the raw data has been updated.

Step 6 - Run script again to load data from the original dataset to the data model tables. Check the result sample with the following queries:

```
SELECT min(cal_Dt), MAX(cal_dt) FROM walmart_anl.sales_inv_store_dy;
SELECT * FROM walmart_anl.product_dim WHERE PROD_KEY=657768;
```

5. Help

If you are facing big challenges please find help from here.

Week Plan

[Week Plan] W1

Lectures and Lab

[Lecture] 1 : Linux (2023-07-25)

[Lecture] 2 : AWS Basics (2023-07-27)

[Lab] 1 : AWS and Linux Workshop (2023-07-29)

Practice

[Exercise] Linux Basics

[Exercises] 42 Linux Exercises

[Exercise] Hackerrank Linux Practice

[Workshop] AWS EC2

[Workshop] AWS S3

[Workshop] DBT Installation

[Lab] EC2 and Linux

Self Study

[Slides] Advanced Linux

[Mini Project] 1: Linux -- Toronto Climate Data

Mini Project Solution

Previous Videos

1.1 Linux System

1.2 Linux Commands and Shell Scripts

[Lecture] 2 AWS S3 and EC2 (2023-03-11)

4

[Week 2] <Data Ingestion> Docker

Chapter overview

Week Plan

[Week Plan] W2

Lectures and Lab

[Lecture] Docker Basic (2023-08-01)

[Lecture] Docker Compose and Demo(2023-08-03)

Practice

[Exercise] 1: Run Docker

[Workshop] Docker Compose --Flask

[Workshop] Docker Compose -- Spark Cluster

[Exercise] 2: Install Zepplin with Docker

[Lab] Install Airbyte and Metabase with Docker

Self Study

[Learning Material] Dockerfile

[Mini Project] Build Docker container to Process data

Docker Cheat Sheet

Previous Videos

Lecture 1 Docker Basics (Wed 2023-03-01)

[Lecture] 2 Dockerfile and Docker Compose （Sat 2023-03-04)）

5

[Week 3] <Data Ingestion> Python in Data Engineering and Cloud

Chapter overview

Week Plan

[Week Plan] W3

Lectures and Lab

[Lecture] Python in DE (2023-08-08)

[Lecture] AWS Lambda (2023-08-10)
[Lab] Lambda
Practice
[Mini Project]: Python in Cloud
[Workshop] Lambda
[Mini Project] Lambda Project
Self Study
[Tutorial] AWS SAM (Create Lambda from CLI)
Previous Videos
3.1 Python Libraries
3.33 AWS Lambda
6
[Week 4] <Data Ingestion> Airbyte, Data Ingestion and Snowflake
Chapter overview
Week Plan
[Week Plan] W4
Lectures and Lab
[Lecture] Snowflake (2023-08-15)
[Lecture] Airbyte, Lambda and Project Data Ingestion(2023-08-17)
[Lab] Project part one
Practice
[Exercise] Snowflake Exercise
[Workshop] Airbyte
[Exercise] Lambda Function Ingest Data
[Lab] Project Capstone 1: Data Ingestion
Self Study
Previous Videos
[Lecture] 2 Snowflake (WED 2023-05-10)
7
[Week 5] <Data Transformation> Data Warehouse
Chapter overview
Week Plan
[Week Plan] W5
Lectures and Lab
[Lecture] Data Warehouse (2023-08-22)
[Lecture] Data Model and ETL (2023-08-24)
[Lab] Data Model and ETL
Practice
[Mini Project] 1 Sakila data - Data Modelling and ETL
Self Study
Previous Videos
[Lecture] 1 Data Warehouse Intro (SAT 2023-05-06)
[Lecture] 3 Data Modeling and ETL Design (SAT 2023-05-13)
8
[Week 6] <Data Transformation> SQL in ETL and Data Loading
Chapter overview
Week Plan
[Week Plan] W6
Lectures and Lab
[Lecture] SQL in ETL (2023-08-29)

[Lecture] Data Loading(2023-08-31)

Practice

[Mini Project] Walmart data - ETL and Data Loads

[Mini Project] - Solution - ETL Data Loading

Self Study

CTE (Common Table Expression)

Previous Videos

[Lecture] 4 SQL IN ETL (WED 2023-05-17)

[Lecture] 5 Data Loads (WED 2023-05-24)

9

[Week 7] <Data Transformation> Data Modeling and ETL in the Project

Chapter overview

Week Plan

[Week Plan] W7

Lectures and Lab

[Lecture] Data Modelling in the project (2023-09-05)

[Lecture] ETL in the project(2023-09-07)

[Lab] Data Transformation with SQL in the project

Practice

[Lab] Build Data Model and ETL for Your Project

Self Study

Previous Videos

10

[Week 8] <Data Transformation> DBT for ETL

Chapter overview

Week Plan

[Week Plan] W8

Lectures and Lab

[Lecture] DBT (2023-09-12)

[Lecture] DBT in Project (2023-09-14)

[Lab] dbt 2023-09-16

Practice

[Lab] Data Model and ETL for Project with DBT

Self Study

11

[Week 9] <Data Analyzation> Data Analyzation with Metabase and Project Summary

Chapter overview

Week Plan

[Week Plan] W9

Lectures and Lab

[Lecture] Metabase (2023-09-19)

[Lecture] Project Wrap Up (Data Analyzation) (2023-09-21)

[Lab] Project TA

Practice

[Lab] Metabase in the project

Self Study

12

[Week 10] <Final Project> Project Week

Chapter overview

Project

Week Plan

Project Description
Project Submission
Project Presentation
13
PHASE TWO : BIG DATA
Chapter overview
Introduction
Project Description
Syllabus
Schedule
14
[Week 1]<Data Ingestion> Azure Basics, Virtual Machine and Storage
Chapter overview
Week Plan
[Week Plan] W1
Lectures and Lab
[Lecture] Azure Intro 2023-10-17
[Lecture] Azure Virtural Machine and Storage 2023-10-19
[Lab] Azure VM and Storage 2023-10-21
Practice
[Exercise] Azure basics
[Workshop] Virtual Machine
[Workshop] Azure Storage
Mini Project: Azure VM and Storage
Self Study
15
[Week 2] <Data Ingestion> Azure Data Factory
Chapter overview
Week Plan
Week 2 Pla2
Lectures and Lab
[Lecture] Azure Data Factory 2023-10-24
[Lecture] ADF in the Project
[Lab] Azure Data Factory
Practice
[Exercise] Azure Data Factory Exercise
[Lab] ADF in Project
[Mini Project] Incrementally load data from Azure SQL Database to Azure Blob storage using Data Factory
Self Study
16
[Week 3] <Data Transformation> BigData and Spark Intro
Chapter overview
Week Plan
[Week Plan] W3
Lectures and Lab
[Lecture] BigData Intro 2023-10-31
[Lecture] Databricks 2023-11-02
[Lab] Databricks 2023-11-04
Practice
[Lab] Databricks Workshop
Previous Videos
[Lecture] 1 Big Data Intro - 2023-05-31

[Lecture] 5 Databricks (2023-06-14)

17

[Week 4] <Data Transformation> Spark Dataframes and SQL

Chapter overview

Week Plan

[Week Plan] W4

Lectures and Lab

[Lecture] Spark architecture and core components 2023-11-07

[Lecture] Spark SQL and Data Processing 2023-11-09

[Lab] Databricks 2023-11-11

Practice

[Exercise] PySpark Syntax Practice

[Lab] Mini Project-- CDR data with Databricks

Previous Videos

[Lecture] 2 PySpark Syntax （2023-06-03）

Self Study

[Lecture] 3 Run Spark on Local and EMR

[Mini Project] 1 NYC Taxi Data with EMR

18

[Week 5] <Data Transformation> Spark Machine Learning Concepts

Chapter overview

Week Plan

[Week Plan] W5

Lectures and Lab

[Lecture] Machine Learning 101 2023-11-14

[Lecture] Natural Language Processing (NLP) 2023-11-16

[Lab] Machine Learning Model Training

Practice

[Workshop], Mount Azure Storage to Azure Databricks

[Lab] Machine Learning Training

Self Study

[Crash course] Machine learning

19

[Week 6] <Data Transformation> Spark Machine Learning in Project

Chapter overview

Week Plan

Week 6 Plan

Lectures and Lab

[Lecture] Project -- ML Prediction + Summary 2023-11-21

[Lecture] Project -- ML Q&A 2023-11-23

[Lab] ML

Practice

[Lab] NLP prediction in Project

Self Study

NLP Crash Course

20

[Week 7] <Data Visualization> Spark Optimization and Azure Synapse

Chapter overview

Week Plan

Week 7 Plan

Lectures and Lab

[Lecture] Spark Performance Tuning and Optimization 2023-11-28
[Lecture] Synapse in Project 2023-11-30
[Lab] Synapse 2023-12-02
Practice
[Lab] Synapse in the Project
[Workshop] Synapse
Previous Video
21
[Week 8] Project Review
Chapter overview
Week Plan
[Week Plan] W8
Lectures and Lab
[Lecture] Project Review 2023-12-05
22
[Week 9] Project Week
Chapter overview
Project
Project Description
Week Plan
Project Submission
Presentation
23
PHASE THREE : LAKEHOUSE AND STREAMING
Chapter overview
Syllabus
Syllabus
Phase 3 Intro
Intro
24
[Week 1] DataOps
Chapter overview
Lecture and Labs
[Lecture] Git 2024-01-09
[Lecture] Terraform 2024-01-11
[Lab] a simple python project with CICD setup
Practice
[Lab] a simple python project with CICD setup
25
[Week 2] NoSQL
Chapter overview
Lecture and Lab
[Lecture] NOSQL and DynamoDB 2014-01-16
[Lecture] Elasticsearch 2024-01-18
[Lab] Dynamo DB
Practice
[Mini Project] DynamoDB and Climate data
Previous Videos
[Lecture] 1 NoSQL and DynamoDB
[Lecture] 2 Elasticsearch
Self-Study
[DynamoDB] DynamoDB Basics

[Workshop] 1 Set an Elasticsearch database on AWS
[Workshop] 2 Elasticsearch installation and practice
[Tutorial] Elasticsearch Complete Tutorial with Python
26
[Week 3] Webscraping
Chapter overview
Lectures and Lab
[Lecture] Web scraping 2024-01-23
[Lecture] backend API, Python Web development 2024-01-25
[Lab] Mini project: web scraping + elasticsearch + context search 2024-01-27
Practice
[Mini project] web scraping + elasticsearch + context search
27
[Week 4] Airflow
Chapter overview
Lectures and Labs
[Lecture] Airflow 1 2024-01-30
[Lecture] Airflow 2 2024-02-01
[Lab Video Recording] Airflow
Practice
[Workshop] Apache Airflow Installation
[Mini Project] Airflow
Previous Videos
[Lecture] 1 Apache Airflow 1
[Lecture] 2 Apache Airflow 2
Self-Study
Introduction of Airflow
28
[Week 5] Data Streaming
Chapter overview
Lectures and Labs
[Lecture] Kafka 2024-02-06
[Lecture] Spark Streaming 2024-02-08
[Lab video] Kafka
Practice
[Mini Project] 1 Kafka for Wiki Media
Previous Videos
Kafka
Data Ingestion and Spark Streaming
Self-study
<Kafka> Kafka in 30 Mins
<Kafka> Kafka on AWS Demo
29
[Week 6] Lakehouse
Chapter overview
Lecture and videos
[Lecture] Lake house concept 2024-02-13
[Lecture] Lake house implementation with Databricks 2024-02-15
Practice
[Mini Project]: build a lake house pipeline w/ databricks
Previous Videos

Lakehouse

30
[Week 7] AI

Chapter overview
Lectures and Lab

[Lecture] AI LLM for data engineering 2024-02-20

[Lecture] Build data application with Streamlit 2024-02-22

[Lab] AI

Practice

[Mini project]: build AI chatbot or dashboard application

Demo

demo script

31
[Week 8] Data Governance

Chapter overview
Lectures

[Lecture] Data governance 2024-02-27

[Lecture] System Design, Data Structure and algorithms 2024-02-29

[Note] Data governance

[Mini Project] Walmart data - ETL and Data Loads