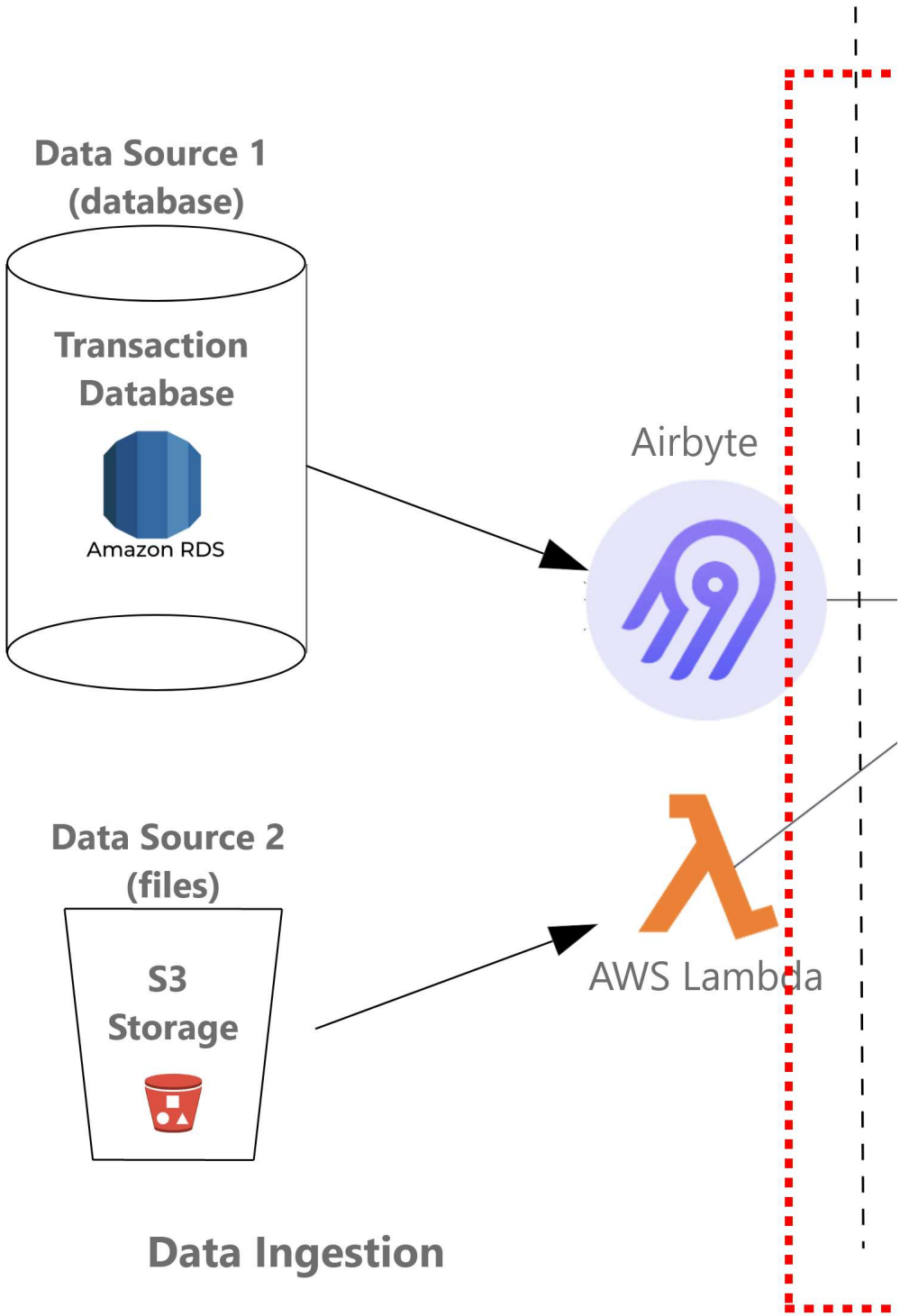
WeCloudData

Week Plan - Week 7

Data Engineering Diploma

Content developed by: WeCloudData Academy

1. Project and Skills



This week, we are going to build a data model in Snowflake and write ETL scripts with SQL. This is the traditional way to build ETL processes. In the next week, we will use the more modern tool -- DBT to build ETL.

2. Prepare For This Week

Before we start this week, it is important we know our project data.

Firstly, since we have ingested data from RDS and S3 bucket to Snowflake RAW schema, so please go to Snowflake to explore the dataset, especially the fact tables. Please explore the dataset from the following aspects:

- The earliest and latest date of the sales and inventory (you need to join date_dim to see the exact date instead of date id)
- Row numbers of each table
- Pick one item to know how frequently it is ordered by customers and how frequently it is recorded in the inventory
- How many individual items
- How many individual customers
- etc.

1. Data Background

This dataset is from TPCDS, a famous dataset for database testing. The business background of the dataset is Retail Sales. The data contains the sales records from the website and Catalog. And also, the inventory level of each item in each warehouse. In addition to these, there are 15 dimensional tables that contain the information of customers, warehouse, items, etc. This entire dataset is not stored in one place; instead, the dataset was split into 2 parts:

Fact tables	Dimention tables
Catalog_Sales	Data_Dim
Web_Sales	Customer
Inventory	Item
From S3	Promotion
	Customer_Gemographics
	Call_Center
	Customer_Address
	Catalog_Page
	Warehouse
	Time_Dim
	Ship_Mode
	Household_Demographics
	Icome_Band
	Web_page
	Web_Site

- **RDS:** All the tables except for the inventory tables are stored in the Postgres DB in AWS RDS. The tables will be refreshed every day and updated with the newest data so for sales data, so in order to get the newest data, you need to run ETL processes every day.
- **S3 Bucket:** The single Inventory table is stored in an S3 bucket, every day there will be a new file containing the newest data dump into the S3 bucket. BUT, be aware that the inventory table usually only records the inventory data at the end of each week, so usually each week you can only see one entry for each item each warehouse (Please go to your RAW schema in Snowflake to explore the data). But you also need to ingest the inventory file from the S3 bucket every day.

2. Tables in the Dataset

You can review the tables' schema from this link. Also, you can find the schema in the Snowflake tables.

In this sheet, you can see there are several tables correlated to the customer; these tables' schema is arranged horizontally. This means when you are doing ETL consider putting integrate all these tables into one customer dimension table.

Dimention Tables

Customer (c)

Column	Datatype	NULLs	Primary Key	Foreign Key
c_customer_sk	identifier	N	Y	
c_customer_id (B)	char(16)	N		
c_current_cdemo_sk	identifier			cd_demo_sk
c_current_hdemo_sk	identifier			hd_demo_sk
c_current_addr_sk	identifier			ca_addres_sk
c_first_shipto_date_sk	identifier			d_date_sk
c_first_sales_date_sk	identifier			d_date_sk
c_salutation	char(10)			
c_first_name	char(20)			
c_last_name	char(30)			
c_preferred_cust_flag	char(1)			
c_birth_day	integer			

Customer_address (ca)

Column	Datatype	NULLs	Primary Key	Foreign Key
ca_address_sk	identifier	N	Y	
ca_address_id (B)	char(16)	N		
ca_street_number	char(10)			
ca_street_name	varchar(60)			
ca_street_type	char(15)			
ca_suite_number	char(10)			
ca_city	varchar(60)			
ca_county	varchar(30)			
ca_state	char(2)			
ca_zip	char(10)			
ca_country	varchar(20)			
ca_gmt_offset	decimal(5,2)			

Customer_demographics (cd)

Column	Datatype
cd_demo_sk	identifier
cd_gender	char(1)
cd_marital_status	char(1)
cd_education_status	char(20)
cd_purchase_estimate	integer
cd_credit_rating	char(10)
cd_dep_count	integer
cd_dep_employed_count	integer
cd_dep_college_count	integer

Date_dim (D)

3. Study of This Week

Business Requirements

You need to build a new fact table in your data model; in your fact table, these metrics are required:

- **sum_qty_wk:** the sum sales_quantity of this week
- **sum_amt_wk:** the sum sales_amount of this week
- **sum_profit_wk**: the sum net_profit of this week
- **avg_qty_dy:** the average daily sales_quantity of this week (= sum_qty_wk/7)
- **inv_on_hand_qty_wk:** the item's inventory on hand at the end of each week in all warehouses (=The inventory on hand of this weekend)
- **wks_sply**: Weeks of supply, an estimate metric to see how many weeks the inventory can supply the sales (inv_on_hand_qty_wk/sum_qty_wk)
- **low_stock_flg_wk**: Low stock weekly flag. During the week, if there is a single day, if [(avg_qty_dy > 0 && ((avg_qty_dy) > (inventory_on_hand_qty_wk)), then this week, the flag is True

In addition to the fact table, you also need to integrate the customer dimension.

1. **Tuesday:** We will review the data and business requirements together to see how we can build the data model.
2. **Thursday:** We will build ETL scripts to populate data from the RAW tables to the Data Model tables.
3. **Saturday afternoon:** You will put all these together to finish the data modeling and ETL scripts in your Snowflake for your project.

4. Advanced Study

Next week, we will study dbt, a tool to manage the ETL process. We have a Udemy course on pre-bootcamp. Please watch the course before we learn dbt. This is the link leading you to the pre-bootcamp page.

←
‹
[Week Plan] W7
›