Data Loading

Prepared by WeCloudData

Data Loading Intro

Agenda.

Two types of Data Loads:

1. **Full Loads**
   ○ Refresh all tables every time.
   ○ For very simple project, no historical data requirements.

1. **Incremental Loads**
   ○ Common in projects.
   ○ At first time, data is loaded completely.
   ○ After the first time, data comes everyday or every week, we need to handle new and changed data every time.
   ○ The new and changed data include:
      i. new transactions in sales and other transaction or log tables(FACT tables).
      ii. new entry in dimension tables, like new product or vendors(DIM tables).
      iii. information changed in dimension tables, like vendor name changes.
   ○ For Incremental Loads there are two steps:
      i. **Initial Loads**: Load data in the first time.
      ii. **Delta Loads**: Load the difference after the initial loads.

WeCloud**Data**

Incremental Data Loading
Data Loading Intro

Initial Loading

The initial data load (IDL) is the very first time that data is loaded into a newly-created, empty base object. During the initial data load, all records are inserted into the base object as new records.

Delta Loading

Once the initial data load has occurred for a base object, any subsequent load processes are called delta loads (or incremental loads) because only new or updated data is loaded into the base object.

WeCloudData

Delta Loading
Data Loading Intro

Fact tables

- Mostly datetime based.
- Append the **new** data after the **max date** of the current FACT table.

2021-12-29
2021-12-30
2021-12-31

⇧ ⇧ ⇧ **new data**

2022-01-01
2022-01-02

Dimension tables

- Add new entries from the original system.
- Update the modified records.

product 1	Name=PRD-1

⇧ ⇧ ⇧ **update change**

product 1	Name=PRODUCT1

- -

product N	existing entry

⇧ ⇧ ⇧ **new data**

product N+1	new entry

WeCloud**Data**

Agenda.

Loading Method
Load Fact Table
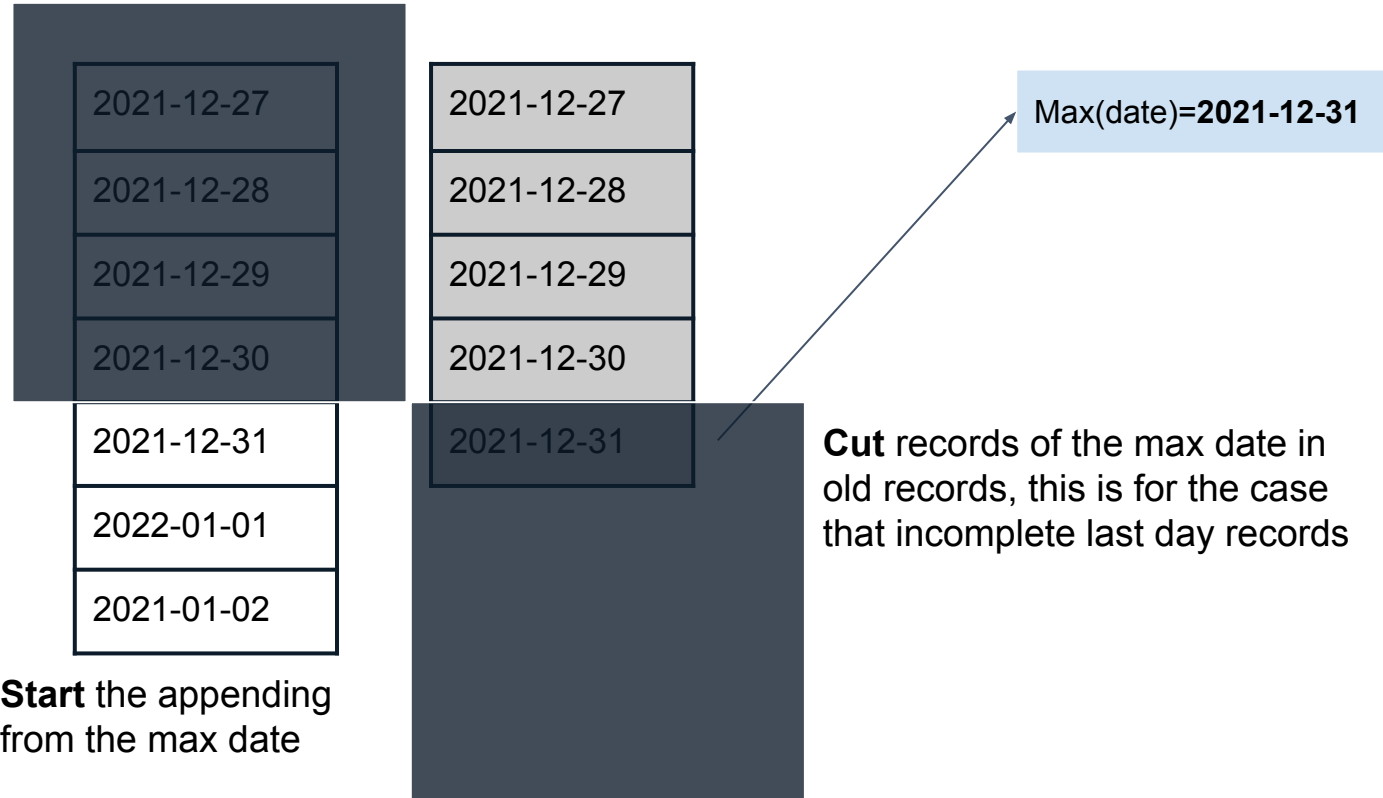
2021-12-27
2021-12-28
2021-12-29
2021-12-30
2021-12-31
2022-01-01
2021-01-02

Start the appending from the max date

2021-12-27
2021-12-28
2021-12-29
2021-12-30
2021-12-31

Max(date)=**2021-12-31**

Cut records of the max date in old records, this is for the case that incomplete last day records

2021-12-27
2021-12-28
2021-12-29
2021-12-30
2021-12-31
2022-01-01
2021-01-02

WeCloud**Data**

Agenda.

Loading Method
Load Dimension Tables

Slow Change Dimension(SCD):

Type 1

Overwrite

Refresh the DIM tables, and only keep the newest version. This was used when **no history data required**.

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

change ⬇

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

Type 2

Add new row

This method tracks historical data by creating multiple records and version them. **This is the most popular way in data warehousing.**

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date	Current_Flag
123	ABC	Acme Supply Co	CA	2000-01-01T00:00:00	2004-12-22T00:00:00	N
124	ABC	Acme Supply Co	IL	2004-12-22T00:00:00	NULL	Y

WeCloud**Data**

Loading Method
Load Dimension Tables

Three types of Dimension loading:

Type 3

Add new attributes

This method tracks changes using **separate columns** and preserves limited history.

Supplier_Key	Supplier_Code	Supplier_Name	Original_Supplier_State	Effective_Date	Current_Supplier_State
123	ABC	Acme Supply Co	CA	2004-12-22T00:00:00	IL

Other types:
- **Type 0:** never change dim tables
- **Type 4:** add history table

WeCloudData

DEMO

Demo

1. Use this script to Initialize the demo environment.
    a. Load Sales and Product data into the LAND schema tables.
    b. Load Calendar data into ENTP schema Calendar_dim table.
2. Use this script to check the current status.
3. Use this script to load dimension data from LAND TO ENTP for the first time.
4. Use this script to load fact data from LAND TO ENTP for the first time.
5. Use this script to run delta loading for both dim and fact tables in LAND.
6. Use this script to load dimension data from LAND TO ENTP for the next time.
7. Use this script to load fact data from LAND TO ENTP for the next time.
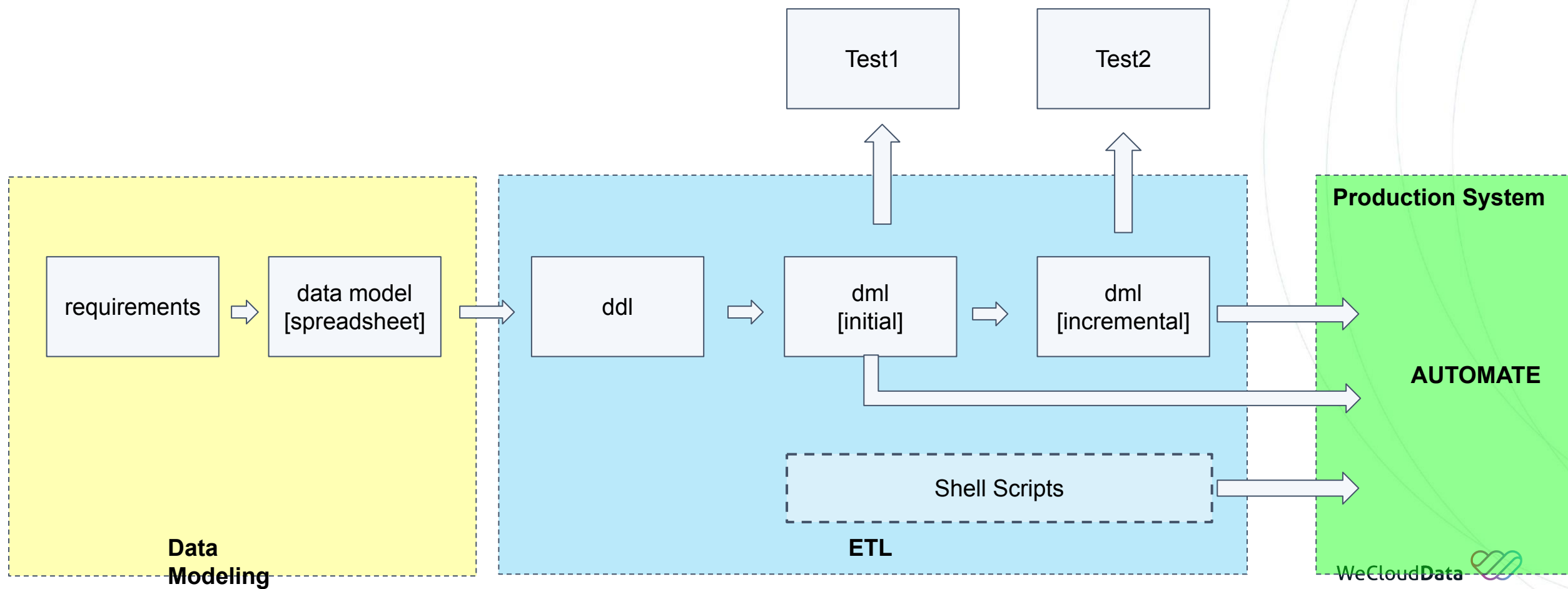8. Use this script to check the current status.

Agenda.

Steps of Data Warehouse Development

Thank you

WeCloudData

📍 500-80 Bloor Street West, Toronto
🌐 ON
www.weclouddata.com