LUMENDATA

Performing Data Quality in Snowflake using IDMC

The One-Stop Guide

5201 GREAT AMERICAN PARKWAY, SUITE 320

SANTA CLARA, CA 95054

Tel: (855) 695-8636

E-mail: info@lumendata.com

Website: www.lumendata.com

Overview:

This data sheet explains how Informatica Data Management Cloud (IDMC) helps us to achieve accurate, consistent data.

Let's consider a scenario:
We will get the data into Snowflake RAW ZONE from different sources like AWS S3, Virtual machines. Only the GOOD RECORDS in the RAW ZONE will be moved to the DERIVED ZONE.

For this, we will use **Informatica Data Management Cloud (IDMC)** Data Quality. It quickly identifies, fixes, and monitors data quality problems.
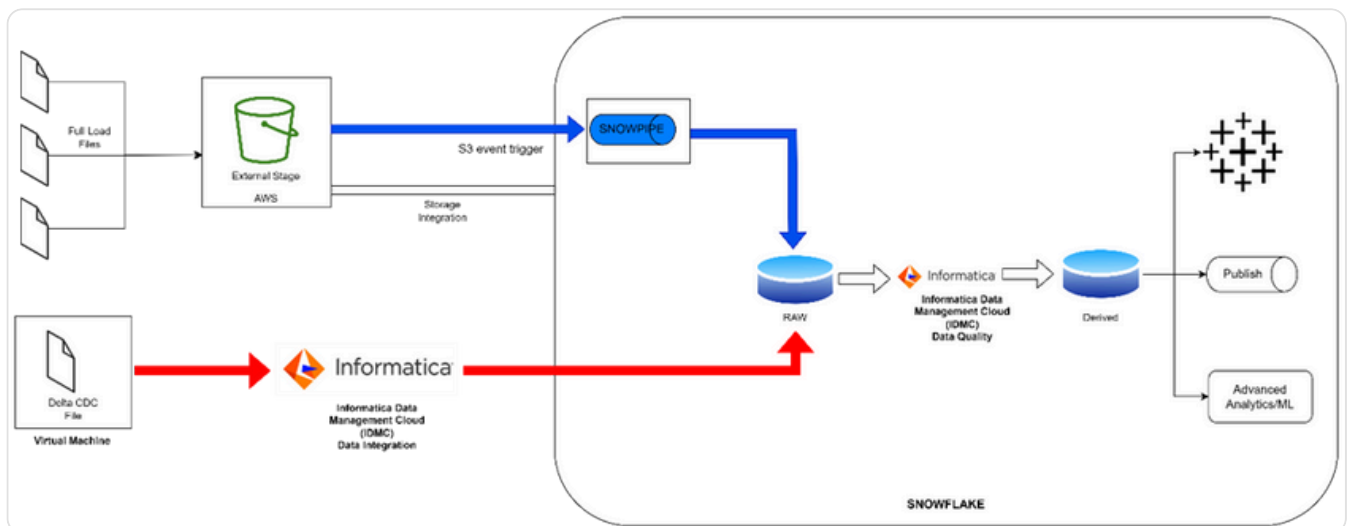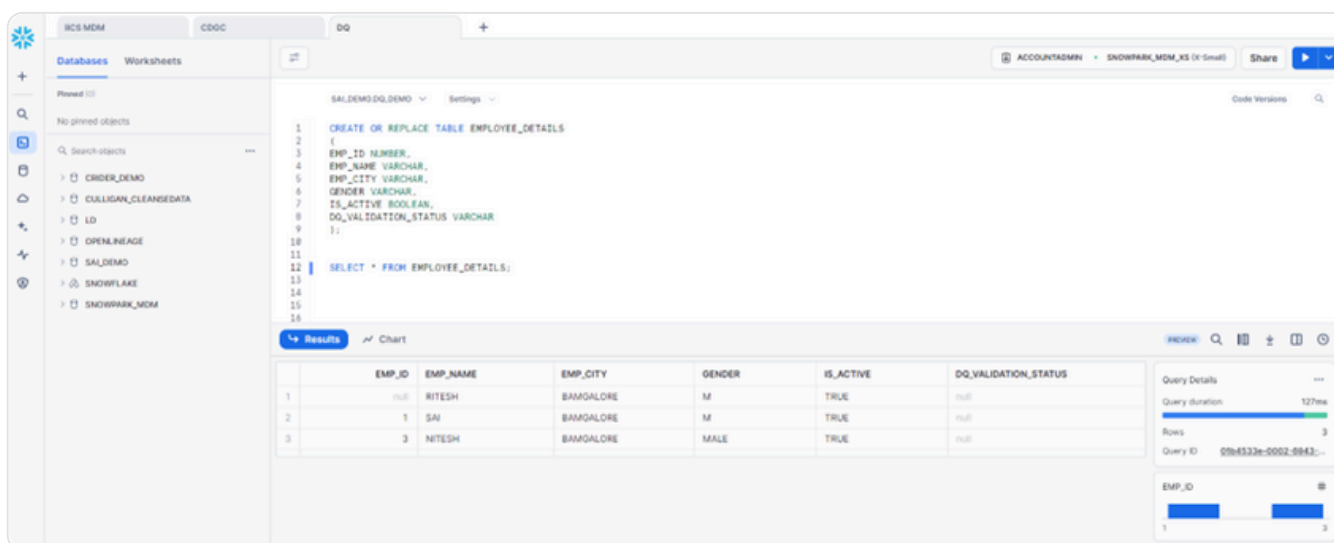

In Snowflake, the data arrives at RAW ZONE as shown above.

For Full Load, use Snowpipe and load the data into Snowflake. Click **here** for the details.

For Incremental or Change Data Capture (CDC) files, use IDMC Data Integration. Find the details **here**.

- **Let's consider we have data in Snowflake table named "EMPLOYEE_DETAILS."**
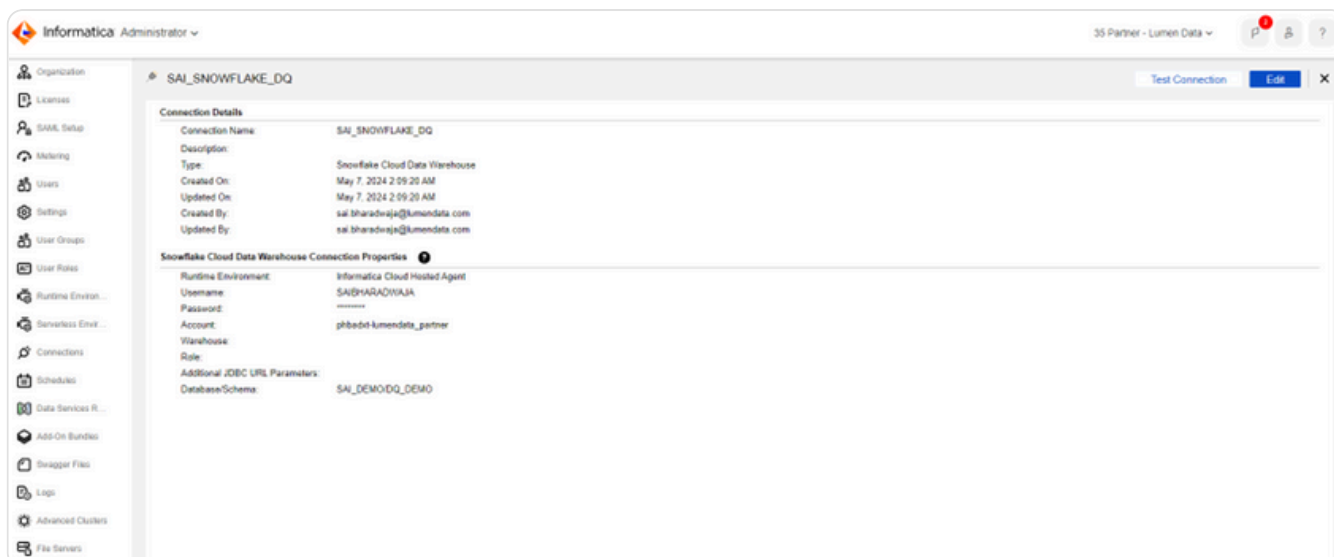
LUMENDATA

```
CREATE OR REPLACE TABLE EMPLOYEE_DETAILS
(
EMP_ID NUMBER,
EMP_NAME VARCHAR,
EMP_CITY VARCHAR,
GENDER VARCHAR,
IS_ACTIVE BOOLEAN,
DQ_VALIDATION_STATUS VARCHAR
);

SELECT * FROM EMPLOYEE_DETAILS;
```
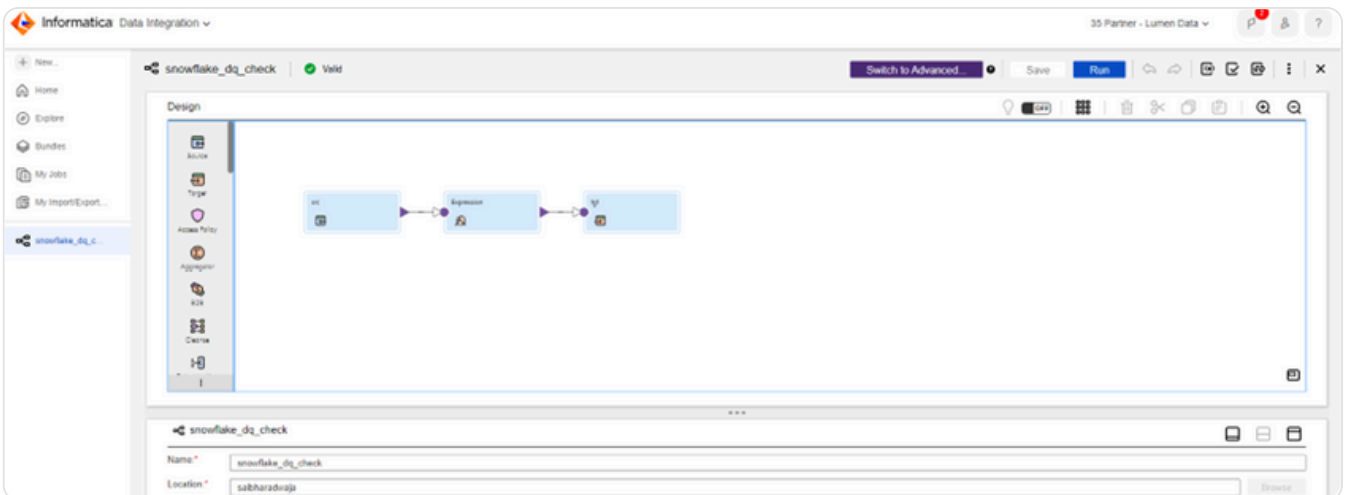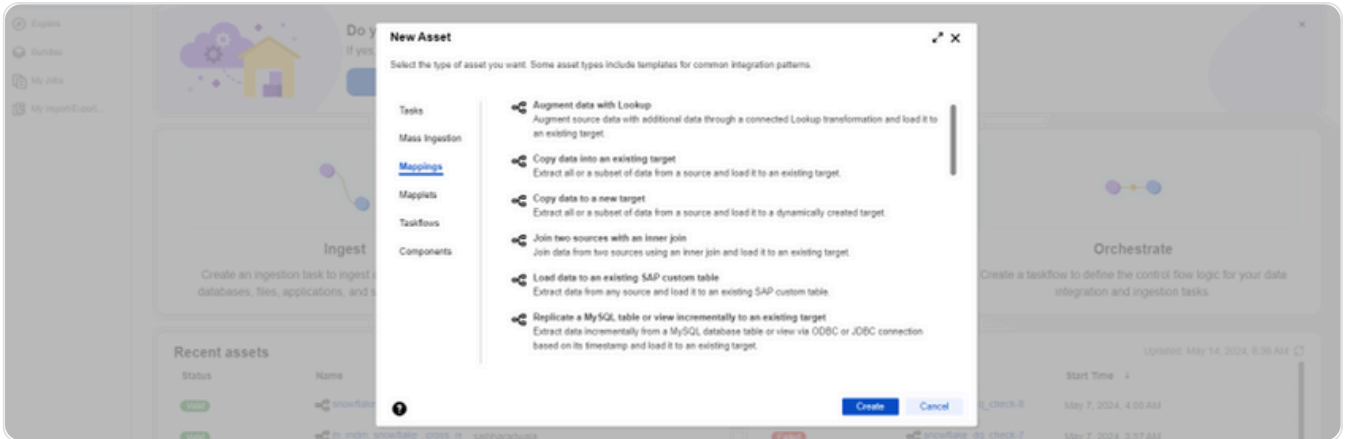
- **Now to view the data in Informatica, we need to create a connection under Administrator.**
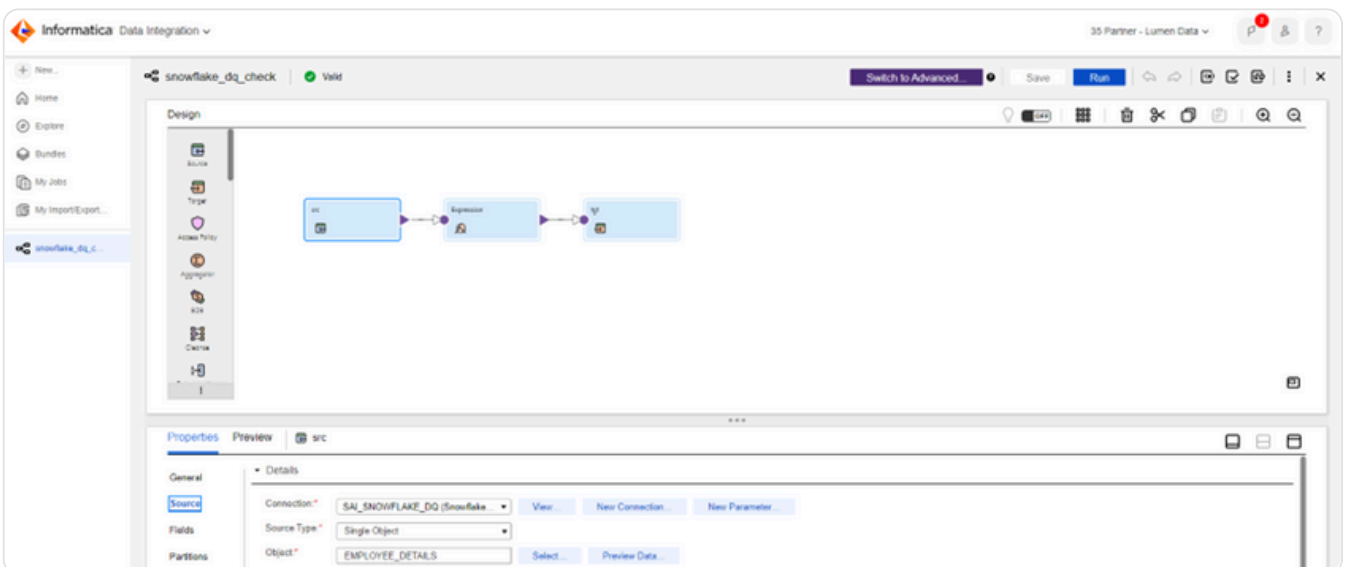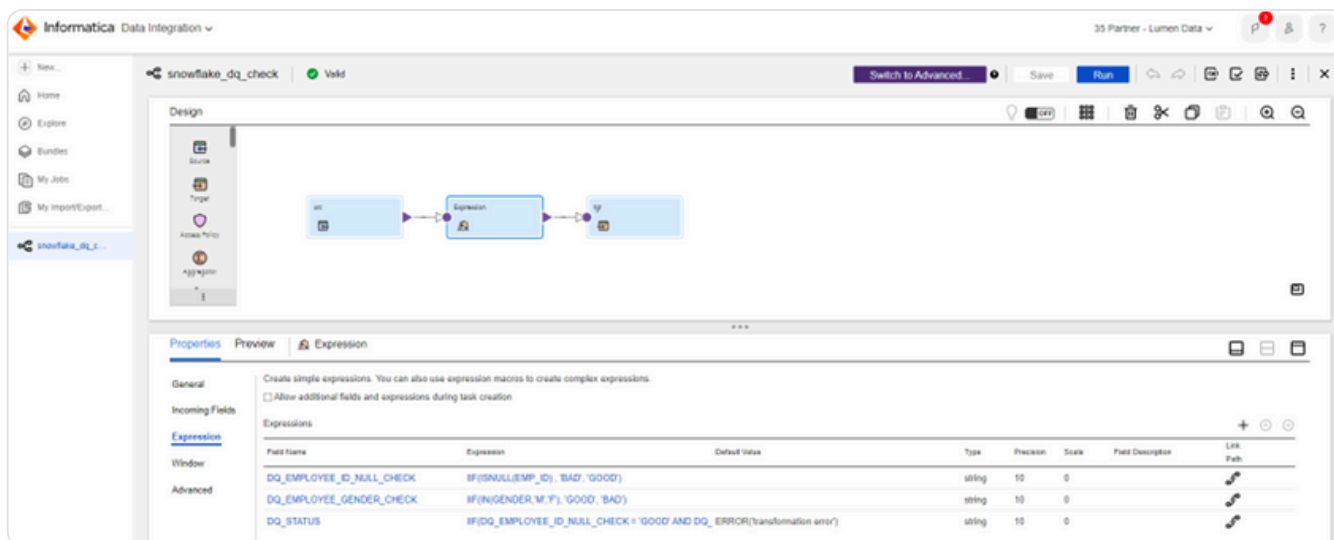
- **Create a mapping in Data Integration to view the data in Informatica.**

  Click on + new ⬚ mappings ⬚ - Copy data to a new target.

- **Click on src. Select the appropriate connection and table to perform data quality.**

- **The DQ rule needs to be added under the expression.**

We added two DQ rules. One for EMP_ID and another for GENDER.

- For EMP_ID, we check whether it has null values or not. If it has any null value, we mark it as a BAD RECORD. If it does not have a null value, we mark it as a GOOD RECORD.

    DQ RULE:

    IIF(ISNULL(EMP_ID) , 'BAD', 'GOOD')

- For GENDER, we check whether it has a valid string like 'M' or 'F' - If it has invalid strings other than 'M' or 'F' value, we mark it as a BAD RECORD. If it has valid strings 'M' or 'F' - it is a GOOD RECORD.
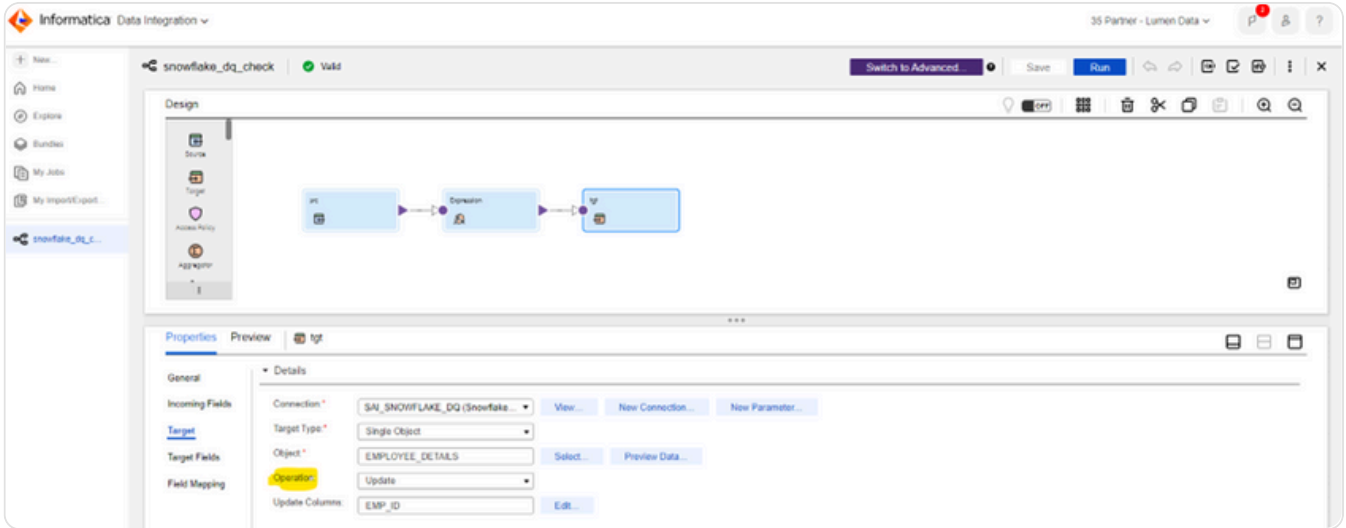
    DQ RULE:
    IIF(IN(GENDER,'M','F'), 'GOOD', 'BAD')

Based on the both the rules, if the record satisfies the DQ RULES' condition, we mark it as a GOOD RECORD.
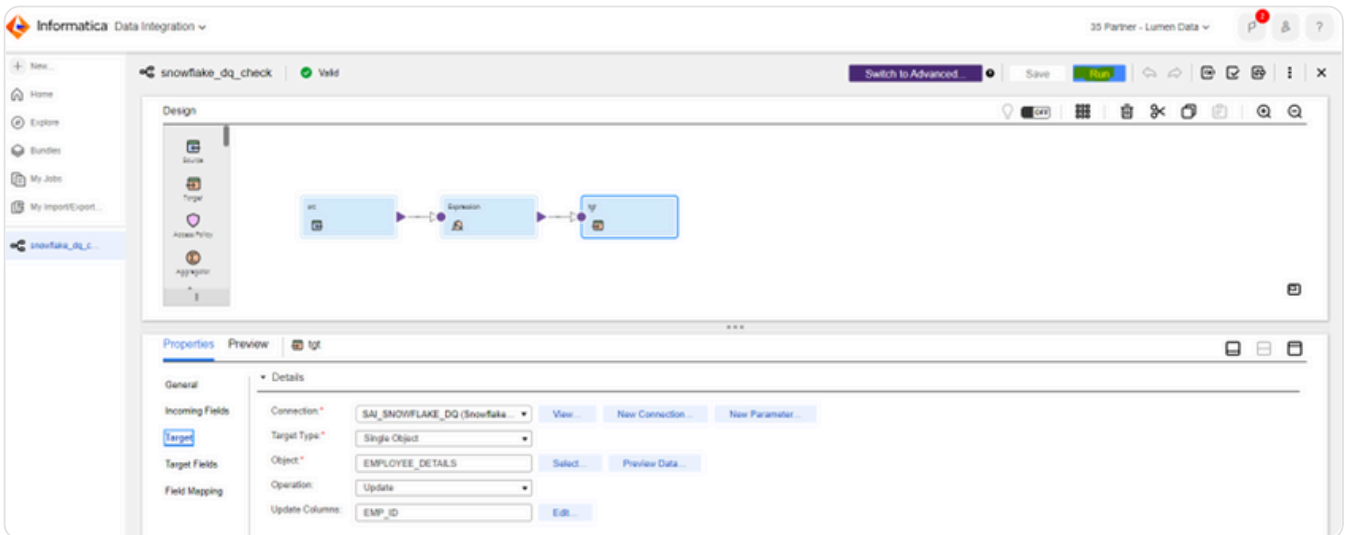
DQ RULE:

IIF(DQ_EMPLOYEE_ID_NULL_CHECK = 'GOOD' AND DQ_EMPLOYEE_GENDER_CHECK = 'GOOD', 'GOOD', 'BAD')

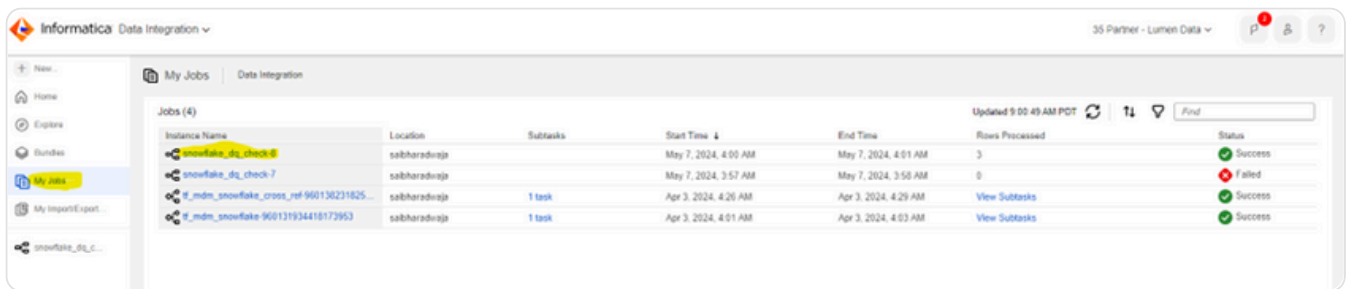- **Once we set the DQ RULES in expression, we add a condition under target to update the records in Snowflake.**

Based on the primary key "EMP_ID" I will update the records in snowflake either GOOD or BAD.
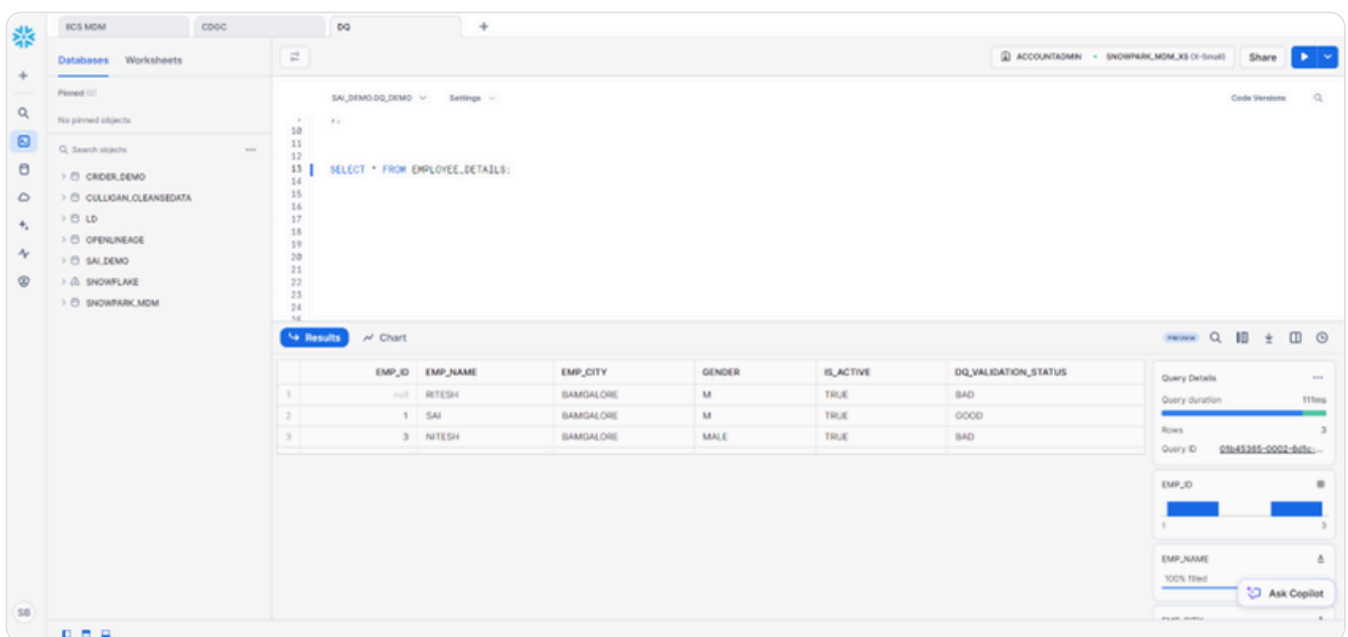
- **Once the mapping is done, click RUN.**

- **We can view the status of running jobs under My Jobs.**

LUMENDATA

- Once the Job status is "SUCCESS" - we can view the records updated with DQ_VALIDATION_STATUS.

As per the DQ RULE,
- The first record doesn't have EMP_ID. The record is marked as "BAD".
- The second record has EMP_ID and the GENDER has valid string either 'M' or 'F' – The record is marked as "GOOD".
- The third record has EMP_ID and the GENDER doesn't have valid string either 'M' or 'F' - Record is marked as "BAD".

About LumenData

LumenData is a leading provider of **Enterprise Data Management, Cloud & Analytics** solutions. We help businesses navigate their data visualization and analytics anxieties and enable them to accelerate their innovation journeys.

Founded in 2008, with locations in multiple countries, LumenData is privileged to serve over 100 leading companies. LumenData is **SOC2 certified** and has instituted extensive controls to protect client data, including adherence to GDPR and CCPA regulations.

LUMENDATA

Get in touch with us:
info@lumendata.com

Let us know what you need:
lumendata.com/contact-us