Business Intelligence and Data Warehousing (ANL408)

- By Sabarish Nair

Recap from last week....

- Dimension Tables
- Date Dimensions
- NULLs in Dimensions
- Hierarchies in Dimensions
- Conformed Dimensions
- Degenerate Dimensions
- Junk Dimension
- Role Playing Dimension
- Practical: Create a staging schema
- Practical: Populate data into staging table from CSV file

Slowly Changing Dimensions (SCD)

- Data does change!
- Need to capture and track historical changes over time.
- Used for historical reporting and analysis
- Introduced by Kimball in 1995
- Example: Customer Data (changes in demographics)

Slowly Changing Dimensions

ASK PROACTIVELY ABOUT POTENTIAL CHANGES

CONNECT WITH BUSINESS AND TECHNICAL TEAMS

DEFINE A STRATEGY FOR EACH CHANGING ATTRIBUTE

Type 0 - Original

- No changes tracked
- No historical information is preserved
- Most Recent data is available
- Very simple and easy to maintain
- Example: Data Table (Except for holidays)

Type 1 - Overwrite

- Old Data is overwritten with the new data
- Only current state is reflected
- Most Recent data is available
- Historical Information is not preserved
- History is lost!
- Might break/affect existing queries

Example: Type 1

Product_ID	Name	Category
1	Bat	Toy
2	Milk	Beverage
3	Ball	Toy

Update

Product_ID	Name	Category
1	Bat	Toy
2	**Almond** Milk	**Liquid**
3	Ball	Toy

Not significant

Significant

Type 2 - New Row

- Introduce a new row for the change
- Historical Data is available
- Perfectly partitions history

Example: Type 2

- Fact Table starts pointing to the new product_ID (i.e. 4) as foreign key
- Dimension Table will have an additional entry
- COUNT (products) from dimension table will give accurate results?

Product_ID	Name	Category
1	Bat	Toy
2	Milk	Beverage
3	Ball	Toy
4	Almond Milk	Liquid

Administering Type 2 Dimensions

- Introduce 2 date columns, effective date and expiration date.
- Add a new column (IsActive)

Product_PK	Product_ID	Name	Category	Effective_Date	Expiry_Date	Is_Active
1	PR_001	Bat	Toy	2022-01-02	2100-01-01	Yes
2	PR_002	Milk	Beverage	2022-01-02	2022-06-01	No
3	PR_004	Ball	Toy	2022-01-02	2100-01-01	Yes
4	PR_004	Almond Milk	Liquid	2022-06-02	2100-01-01	Yes

Type 2 SCD Steps

- Add a new row in the dimensions
- Fact Table: Lookup in the dimension with the Natural Key + Ef/Ex Date
- Add Is_Current/Is_Active Flag

Type 1 + Type 2

- Can use Type 1 or 2 depending on the attributes
- Use Type 1 for low significant changes
- Use Type 2 for high significant changes
- Not a technical, but a business decision
- No set in stone rules

Type 3: Additional Attribute

Type 1- Static

Type 2: Default strategy to maintain history

Type 3: Switching back and forth between versions

Type 3 – Add a new attribute

- Typically used for significant changes at a time (e.g. restructurings in organizations)
- Switching between historic and current view
- Introduce additional columns to keep tracking multiple changes
- Not suitable for frequent or unpredictable changes

Product_ID	Name	Prev category	Category
1	Bat	Toy	Toy
2	Milk	Beverage	**Liquid**
3	Ball	Toy	Toy

Practical: Exploratory Data Analysis (EDA)

- Check the data quality
- Check for missing/ NULL values in the dataset
- Write queries for finding such records
- Get the overall record count
- Understand the data!!!

Get the record count

Query Query History

```
1  SELECT COUNT(1)
2  FROM "Staging"."tbl_ProductsData";
```

Query Query History

```
1  SELECT COUNT(*)
2  FROM "Staging"."tbl_ProductsData";
```

Data Output Messages Notifications

	count bigint
1	20

Query Query History

```
1  SELECT distinct customer_name
2  FROM "Staging"."tbl_ProductsData";
```

Data Output Messages Notifi

	customer_name character varying
1	[null]
2	Sophia Martinez
3	Sophia Brown
4	Jacob Thomas
5	Olivia Wilson
6	Alex Johnson
7	Michael Lee
8	John Doe
9	Ethan Anderson
10	Emily Wang
11	William Lewis

Get Distinct Records in a column

Query Query History

```sql
SELECT *
FROM "Staging"."tbl_ProductsData"
WHERE customer_name IS NULL;
```

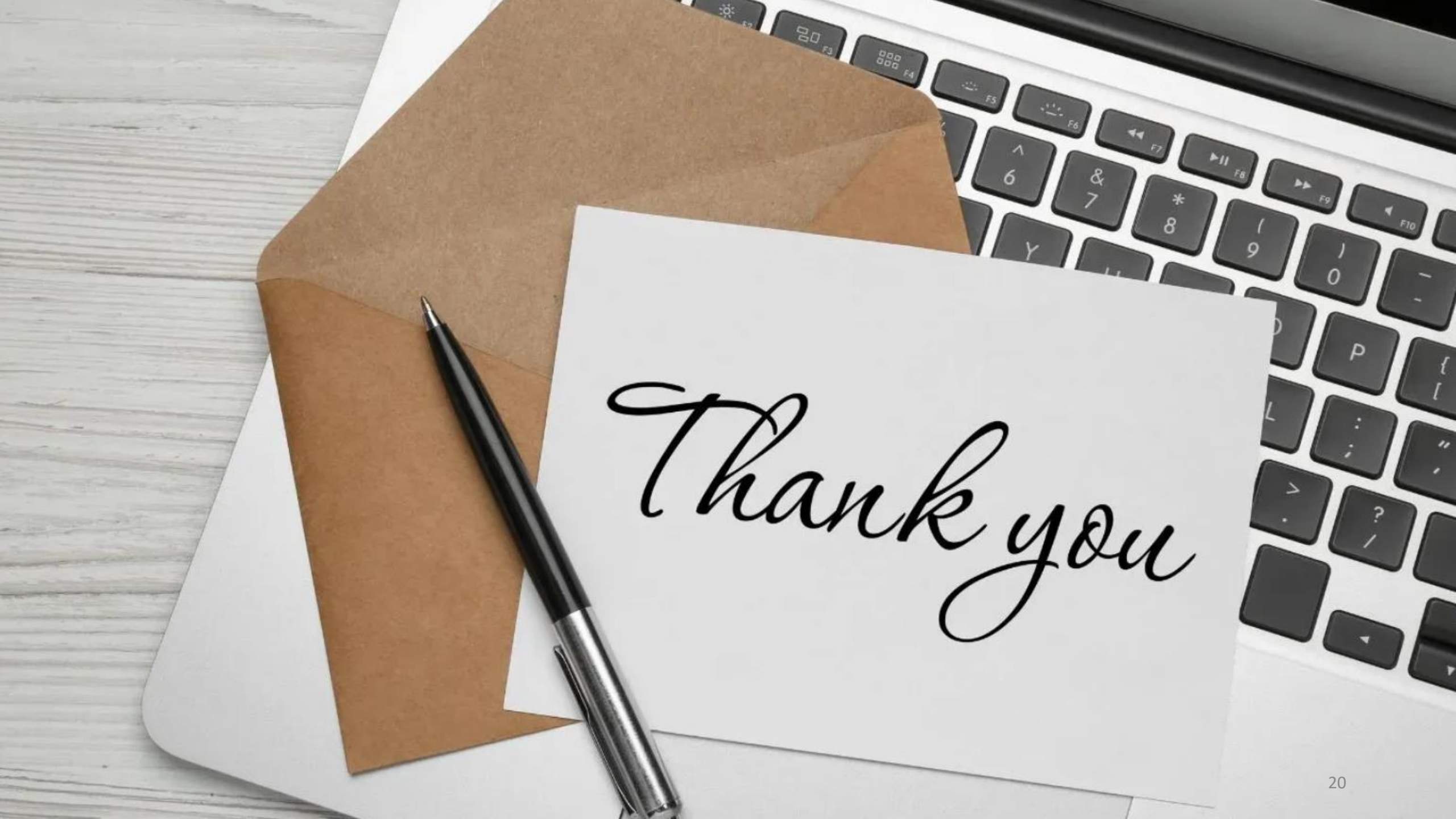
Data Output Messages Notifications

	sales_id [PK] integer	date_sales date	product_id integer	product_name character varying	category character varying	price numeric	customer_id integer	customer_name character varying	city character varying	country character varying
1	3	2022-01-03	3	Speaker	Electronics	[null]	3	[null]	Paris	France
2	4	2022-01-04	4	TV	Electronics	1500	[null]	[null]	Berlin	Germany
3	6	2022-01-06	[null]	Mouse	Electronics	25	6	[null]	Sydney	Australia
4	10	2022-01-10	10	Printer	Electronics	200	10	[null]	Paris	France

Check for NULL values in columns

Questions

- Total Rows?
- Total Distinct Products?
- Total Distinct Category?
- Maximum Price?
- Total Distinct Cities?
- Count of countries with blank values?


Thank you