Frozen foods dataset

Business and data warehouse description

Contents

1 Business Description

1.1 Business background

The business sells frozen foods globally, in all corners of the globe, and also offers online ordering and in-person shopping at its regional stores.

1.2 Problems because of poor data management

The company's management would like to better understand the global distribution of sales and revenues by product, sales channel (online web sales or local sales), product category and territory.

1.3 Benefits from implementing a Data Warehouse

Using data warehouse can help with problems described above. Implementing a data warehouse can answer the following questions:

- Which product have the highest profit?
- Which ones have the widest distribution of globally?
- Is there a typical price distribution across products or within specific categories?

1.4 DATASETS DESCRIPTION

The first dataset includes online orders globally.

Product Information:

Product: Frozen food.

Category: The category of the frozen food (meat based, plant based, etc.)

Subcategory: Further classification of the food.

Manufacturers:

name: name of the manufacturer.

addresses: the address includes the exact street and house number, city, postcode, country, region.

Customer Information:

name, address (includes the exact street and house number, city, postcode, country, region.), tel number, e-mail

Other Attributes:

transaction time, delivery time (the time from order to delivery)

The second dataset records data on local, in store sales (offline). Compared to the first dataset, it does not include information about customers, but includes information about the location of the sale and about the employees who made the sale.

1.5 GRAIN / DIM / FACT

1.5.1 Business process selecting (1/4):

The business process consists of online delivery of frozen foods to order and local in-store sales, where geographic data, product and manufacturer details are recorded.

1.5.2 Declaring grain (2/4):

The grain of DWH is the sale of the product. In each product-sale the product, manufacturer, customer, employee (and therefore the store and place), saled quantity, procurement cost and paid amount is recorded.

1.5.3 Identifying dimensions (3/4):

The following dimensions can be associated to each online sales based on the data model:

- date
- product, which contains product categories in hierarchy
- customer, which contains geography in hierarchy
- manufacturer, which contains geography in hierarchy

The following dimensions can be associated to each online sales based on the data model:

- date
- product, which contains product categories in hierarchy
- employees
- manufacturers
- cities, which is the location, where a store is located. In one town there is only one store.

1.5.4 Identifying facts (4/4):

Each elementary sale is made up of the following:

- date
- product identifier
- customer identifier
- employee identifier
- city identifier
- manufacturer identifier
- sold product amount
- cost per sale
- paid amount per sale
- sale channel

1.5.5 Dimensional tables:

product

contains informations from products

Column name	Description	Data Type
product_id	PK of table	bigint
product_name	name of the product (frozen food)	varchar(100)
product_category_id	identifier of product category	bigint
product_category	name of the product category	varchar(100)
product_sub_category_id	identifier of product sub category	bigint
product_sub_category_name	name of the product sub category	varchar(100)
unit_gram_per_pack	the unit package weight in grams	int

Example with filled data

product_id	product_name	product_category_id	product_category	product_sub_category_id	product_sub_category_name	unit_gram_per_pack
30	Frozen Croquette	2	Plant based	3	Potatoes	1000
31	Frozen Spicy Potato Wedges	2	Plant based	3	Potatoes	1000
32	Frozen Sea Fish Fillet (Alaska Pollock with 20% glaze)	3	Meat based	4	Seafood	600

manufacturer

contains informations about manufacturers of products

Column name	Description	Data Type
manufacturer_id	PK of table	int
manufacturer_name	name of the manufacturer	varchar(100)
manufacturer_address	address of manufacturer (street name, number)	varchar(100)
manufacturer_city_id	city identifier of address	int

Example with filled data

manufacturer_id	manufacturer_name	manufacturer_address	manufacturer_city_id
1	MeadowInnovate Inc.	23456 Maple Street 123	9264
2	yejmc InnovateHub Inc.	56789 Oak Avenue 456	2335
3	CalmHarbor Inc.	98765 Elm Lane 789	4340

customer

contains informations about customers

Column name	Description	Data Type
customer_id	PK of table	bigint
first_name	first name	varchar(100)
last_name	last name	varchar(100)
gender	gender of person	varchar(5)
date_of_birth	birth date of person	date
street_address	street name and number of address	varchar(100)
tel_number	telefon number of customer	varchar(100)
email	e-mail address of customer	varchar(100)
city_id	city identifier	bigint

Example with filled data

customer_id	first_name	last_name	gender	date_of_birth	street_address	tel_number	email	city_id
1	Brian	Clark	M	1958-03-14	36. Jeremy Villages	(518)319-6737	jeannemiddleton@gmail.com	1
2	Keith	Brown	M	1987-10-16	47. Lee Rapid	+1-468-281-2071x207	smithjennifer@yahoo.com	2
3	Raymond	Bryan	M	2001-01-11	113. Emily Stravenue	(606)522-4533x536	ccurry@yahoo.com	3

employee

contains informations about employees

Column name	Description	Data Type
employee_id	PK of table	bigint
employee_name	first name	varchar(100)
date_of_birth	birth date of person	date
employee_email	e-mail address of employee	varchar(100)
store_id	identifier of the store, where employee works	bigint
store_city_id	city identifier in where the store is located	bigint
store_address	street name and number of the store, in where the employee works	varchar(100)

Example with filled data

employee_id	employee_name	date_of_birth	employee_email	store_id	store_city_id	store_address
10	Marco Fitzpatrick	1989-12-03	marcofitzpatrick10@frozenretail.com	1	9956	59. Rodriguez Street
11	Michael Wilson	1995-12-26	michaelwilson11@frozenretail.com	2	9773	9. Robles Stravenue
12	Heidi Martinez	1989-11-25	heidimartinez12@frozenretail.com	2	9773	9. Robles Stravenue

city dimension

contains informations about cities. This dimension contains geographical hierography.

Column name	Description	Data Type
city_id	PK of table	bigint
city	name of the city	varchar(100)

postal_code	postal code	varchar(100)
country_id	identifier of country, in which is the city located	int
country_name	name of the country	varchar(100)
country_province_id	identifier of the province in which the country is located	bigint
country_province	name of the province	varchar(100)
continent_id	identifier of the continent	smallint
continent	name of the continent	varchar(100)

Example with filled data

city_id	city	postal_code	country_id	country_name	country_province_id	country_province	continent_id	continent
1	Amyfort	39051	89	Armenia	120	Kotayk	5	Asia
2	Dustinmouth	22450	155	Cuba	127	Guantanamo	4	North America
3	Hollyshire	89071	28	Kiribati	92	Phoenix Islands	6	Australia and Oceania

date dimension

contains informations about each date day

Column name	Description	Data Type
date_id	PK of table in date format	date
day_name	day name of the date	varchar(100)
day_number_in_week	day number of the week	int
day_number_in_month	day number of the month	int
calendar_week_number	week number of the year	int
calendar_month_number	month number of the year	int

Example with filled data

date_id	day_name	day_number_in_week	day_number_in_month	calendar_week_number	calendar_month_number

2023-01-20	Saturday	6	20	3	1
2023-01-21	Sunday	7	21	3	1
2023-01-22	Monday	1	22	4	1

DIM_channels dimension

contains sale channel informations

Column name	Description	Data Type
channel_id	PK of table	int
channel_name	name of the channel	varchar(100)

Example with filled data

channel_id	channel_name
1	online
2	offline

1.5.6 Fact table:

dim_fct_sales

contains datas of the sales

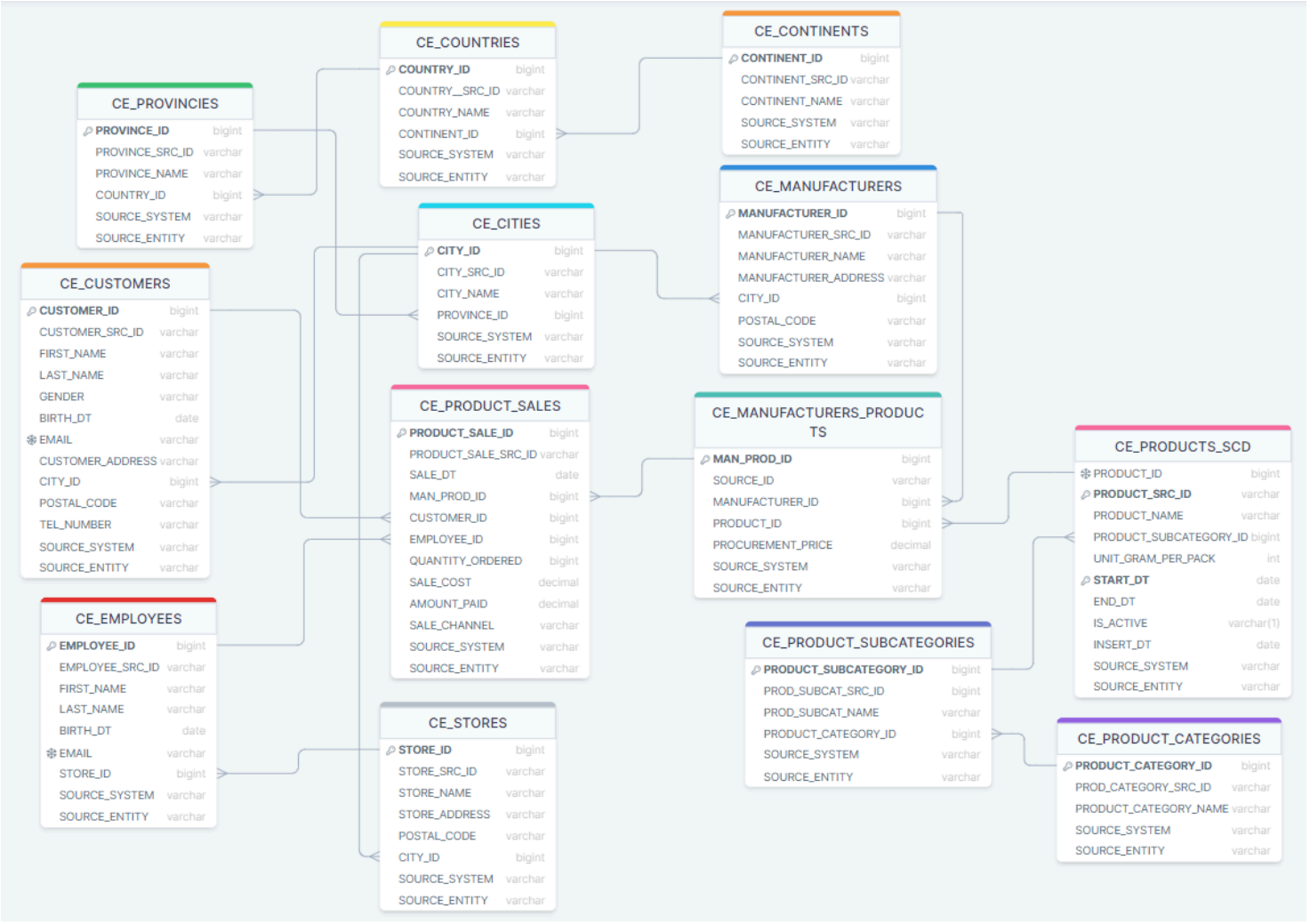
Column name	Description	Data Type
date_id	foreign key, PK of time dimension	date
product_id	foreign key, PK of product dimension	bigint
customer_id	foreign key, PK of customer dimension	bigint
manufacturer_id	foreign key, PK of manufacturer dimension	bigint
employee_id	foreign key, PK of employee dimension	intbigint
store_city_id	foreign key, PK of city dimension	bigint
quantity_ordered	quantity of the ordered product	int
order_cost	cost (including all summary costs) of the sale process	decimal

amount_paid	paid amount by the customer	decimal
DIM_channel	sale channel, on which transaction happened	varchar(100)

Example with filled data

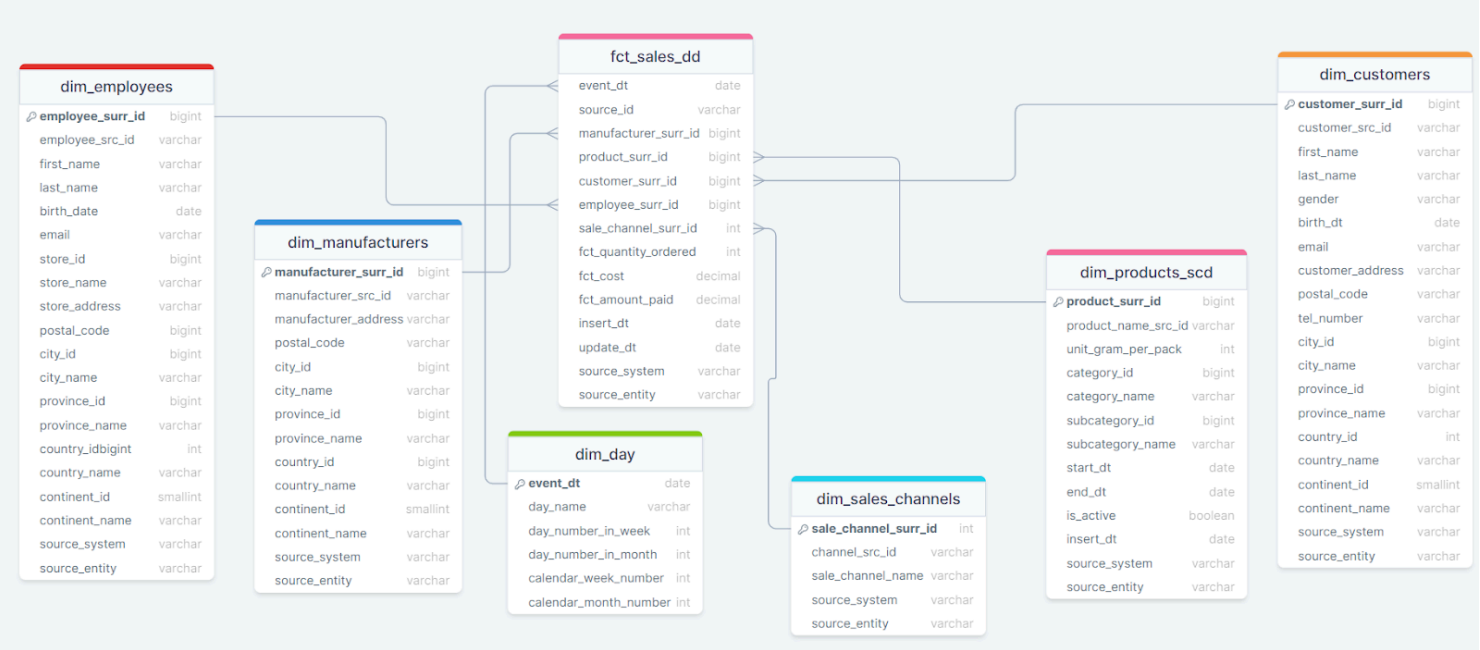
date_id	product_id	customer_id	manufacturer_id	employee_id	store_city_id	quantity_ordered	order_cost	amount_paid	sale_channel
2022-01-01	125	48772	43	15	32	2	6.72	8	1
2022-01-01	54	24818	12	43	6612	3	49.86	59.57	1
2022-01-02	95	56001	5	9	6541	4	23.5	26	2
2022-01-03	78	56001	6	1	216	1	15.35	20	2

2 Business Layer 3NF



- BL_3NF schema is made with drawSQL. Entities are made regarding to the business description. CE_PRODUCT_SALES contains datas to the fact tables. In BL_DIM fact table order_cost is calculated from CE_PRODUCT_SALES.SALE_COST and CE_MANUFACTURER_PRODUCTS.PROCUREMENT_PRICE.
- In each table (except CE_PRODUCTS, which is an SCD2 table with product_src_id - start_dt composite primary key) the primary keys are surrogate keys.
- On the left side of field names "snowflake" signs, that field contains unique values.

3 Business Layer Dimensional Model



SQL script to populate dim_time_day table:

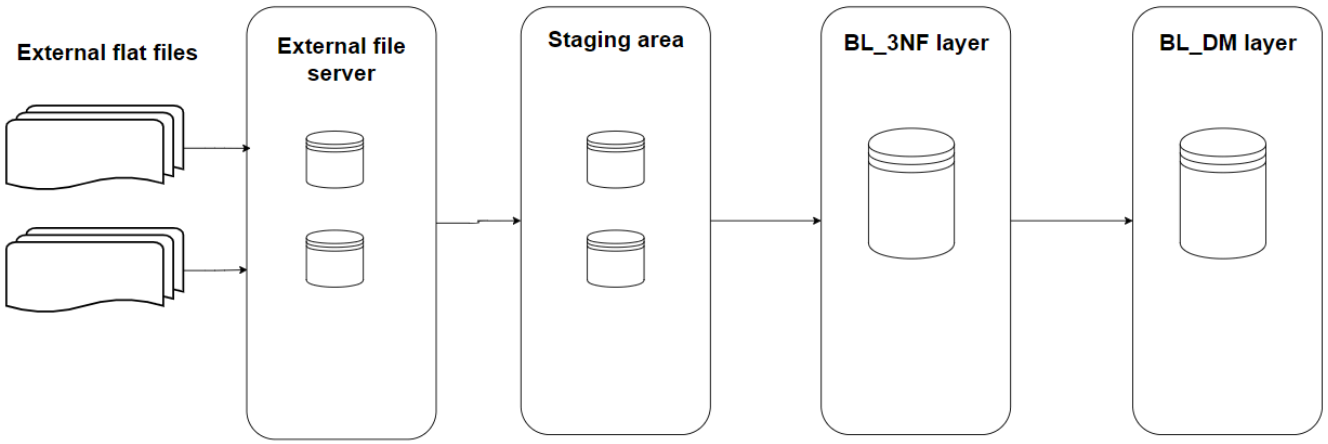
```
CREATE TABLE IF NOT EXISTS dim_day (
    event_dt date,
    day_name varchar(9),
    day_number_in_week varchar(1),
    day_number_in_month varchar(2),
    calendar_week_number varchar(2),
    calendar_month_number varchar(2));

INSERT INTO dim_day
      WITH days AS (SELECT generate_series('2022-01-01', '2023-12-31', INTERVAL '1
day')::date AS event_dt)
            SELECT      event_dt,
                        to_char(event_dt, 'day') AS day_name,
                        to_char(event_dt, 'ID')::int AS day_number_in_week,
                        to_char(event_dt, 'DD')::int AS day_number_in_month,
                        to_char(event_dt, 'WW')::int AS calendar_week_number,
                        to_char(event_dt, 'MM')::int AS calendar_month_number
            FROM        days d
            WHERE NOT EXISTS (SELECT * FROM dim_day dd WHERE dd.event_dt =
d.event_dt);
```

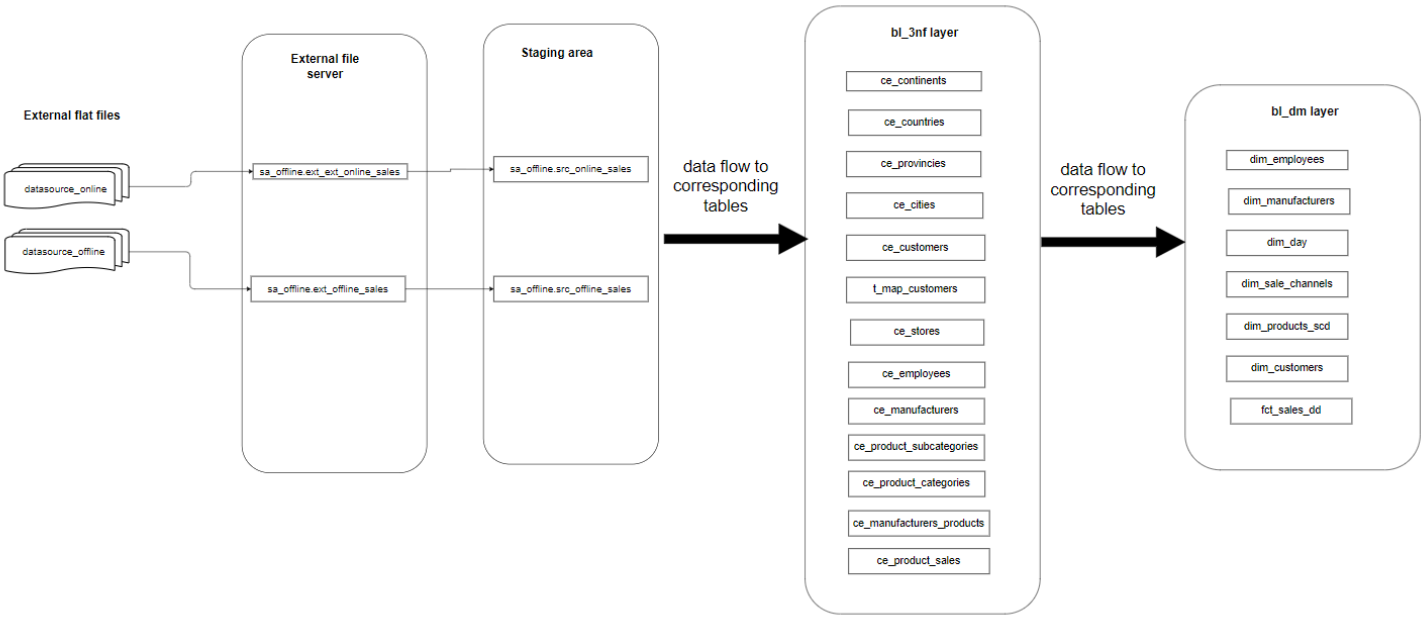
Metric calculation:

In fact table fct_cost is calculated from values from BL_3NF schema ce_product_sales.sale_cost values and ce_manufacturers_products.procurement_price value, they are added together calculating the ordered quantities: (ordered quantities) * (procurement prices) + sale_cost.

4 Logical Scheme



5 Data Flow



6 Fact Table Partitioning Strategy

dim_fct_sales_dd fact table is partitioned by range of date in yearly interval (dim_fct_sales_dd_2021, dim_fct_sales_dd_2022, dim_fct_sales_dd_2023, dim_fct_sales_dd_2024). New partitions are added yearly to the main table.

7 Incremental Load

During incremental load new records are filtered by source identifiers in all tables of all schemes, except in case of the fact table (dim_fct_sales_dd), where new records are filtered by dates. This incremental load strategy enables at least daily upload (but only one time in one day) of the data warehouse.

8 SCD 2 (Slowly Changing Dimension)

SCD 2 strategy is applied on tables of products (bl_3nf.ce_products_scd and bl_dm.dim_products_scd) with "MERGE INTO" postgres SQL command.