

| Business Template  **Jackpot Insights** |
| --- |
|  |

**Contents**

[**1 Business Description**](#_heading=h.e19xqlcb2hpk) **3**

[1.1 Business background](#_heading=h.68azum271kq8) 3

[1.2 Problems because of poor data management](#_heading=h.ksu66fhjf57m) 3

[1.3 Benefits from implementing a Data Warehouse](#_heading=h.sshoqtp8hnlt) 3

[1.4 DATASETS DESCRIPTION](#_heading=h.5ga3teextrr6) 4

[1.5 GRAIN / DIM / FACT](#_heading=h.t6bgnnl825m1) 7

[1. Selecting the Business Process](#_heading=h.yj9yixg47vn8) 7

[2. Declaring the Grain](#_heading=h.8au0iu7fit0m) 7

[3. Identifying the Dimensions](#_heading=h.x5oq5snkl10t) 8

[4. Identifying the Facts](#_heading=h.r7b1vos2s14n) 9

[**2 Business Layer 3NF**](#_heading=h.lo4oygxeb6ul) **17**

[**3 Business Layer Dimensional Model**](#_heading=h.eixhmi6lspx0) **19**

[**4 Logical Scheme**](#_heading=h.ii9fn9hmpapn) **19**

[**5 Data Flow**](#_heading=h.sop8pmu8e784) **19**

[**6 Fact Table Partitioning Strategy**](#_heading=h.wfgkbvqc9afb) **19**

# 

# Business Description

## Business background

Lottery games, or luck-based games as we might call them, are part of a large and profitable

industry around the world. People keep playing because of the strong desire to win

big—sometimes just by scratching a ticket or picking a few lucky numbers. These games

continue to be popular in many places.

In the United States, the lottery market is very large, and Texas stands out as one of the

top states in terms of ticket sales. That’s why this project focuses specifically on the Texas

lottery. There are still many important questions to explore: Which games are the most or

least profitable? What types of games do people in Texas prefer? Are instant-win scratch

games more popular, or do players prefer waiting for draw results?

To find answers, it is important to understand how consumers behave, how well different

games perform, and how retailers contribute to sales. This can be done by collecting and

analyzing structured sales data from across Texas, which will help reveal useful insights and

support better business decisions in this growing industry.

## Problems because of poor data management

Poor data management can significantly hinder success in the lottery business. Without

proper data, it’s difficult to know which games are doing well and which ones are not. If

you don’t use tools that help you collect and analyze sales information, you won’t be able

to understand what players want or how retailers are performing.

In a competitive market like the Texas lottery, not having the right data can lead to

missed opportunities, poor planning, and less profit. To stay competitive and make smart

decisions, it’s important to manage data properly and use it to create effective strategies.

## Benefits from implementing a Data Warehouse

Using a data warehouse can help solve the problems mentioned above. Implementing a data

warehouse can answer important questions like:

• Which lottery games generate the most revenue?

• Which games have the widest range of sales across different retailers?

• Are there clear patterns in player preferences for instant-win versus draw-based games?

Further analysis of the data can also help to:

• Understand how sales vary by region or retailer type.

• Identify trends in customer behavior over time.

• Improve marketing strategies by targeting popular games.

• And many other useful insights.

By collecting and organizing data in one place, a data warehouse makes it easier to

analyze sales, support better decision-making, and improve overall business performance.

## DATASETS DESCRIPTION

In our project, we examine the performance of lottery games across different regions of the state of Texas, focusing on two distinct source systems and two key years: 2021 and 2022. These years are particularly significant as 2021 marked the early post-COVID period, still affected by restrictions and shifts in consumer behavior.

We have designed two separate source datasets:

* **SRC\_LOTTERY\_SCRATCH\_OFF**: Contains all sales transactions related to scratch-off lottery tickets.
* **SRC\_LOTTERY\_DRAW**: Contains all sales transactions related to draw-based lottery games.

This division reflects two fundamentally different player mindsets. Scratch-off games provide instant win opportunities, feature variable pricing across retailers, and generally offer better odds. In contrast, draw-based games have fixed prices, are standardized across vendors, and typically offer much lower odds of winning, but with higher potential jackpots.

By comparing sales trends between these two types of games, we aim to analyze which format appeals more to consumers and whether the promise of instant gratification with higher odds outweighs the allure of large payouts with slim chances.

In the first data source we have considered the following information about scratch-off ticket games sales in Texas:

**Sales Information:**

 - Transaction Date: The date when the draw tickets were sold

- Retailer License Number: The natural key of the retailer that performed the ticket sale

 - Customer ID: The unique ID of the customer that performed the ticket sale

 - Employee ID: The unique ID of the employee that performed the ticket sale

-Game Scratch Number:The Unique identifier for the scratch game

 - Sales Amount: The total revenue from the ticket sale

 - Number of Tickets Sold: The number of tickets sold for that particular draw game

 - Payout: the amount paid to the customer in case of winning(0 otherwise)

**Date Information:**

 - Date ID: Unique identifier for the date (e.g., YYYYMMDD)

 - Fiscal Year: Fiscal year number for the date

 - Fiscal Month: Fiscal month number

 - Fiscal Month Name and Number: Textual name and number of the fiscal month

**Customer Information:**

 - Customer ID: Unique identifier for the customer

 - Customer Name: Full name of the customer

 - Customer Gender: Gender of the customer

 - Customer Date of Birth: Birth date of the customer

**Employee Information:**

 - Employee ID: Unique identifier for the employee

 - Employee Name: Full name of the employee

 - Employee Department: Department where the employee works

 - Employee Status: Current employment status (e.g., active, terminated)

**Retailer Information:**

 - Retailer License Number: Unique license number for the retailer

 - Retailer Location Name: Name of the retailer’s physical location

 - Retailer ZIP Code: ZIP code of the retailer’s location

 - Retailer Location City: City where the retailer is located

**Game Information:**

 - Game Scratch Number: Identifier number of the scratch game

 - Ticket Price: Standard ticket price for the game

 - Game Category: Category of the scratch game (e.g., instant win)

 - Game Type: Type or theme of the game (scratch/non-scratch)

 - Average Odds: Average chance of winning

 - Average Odds Probabilities: Average probabilities of winning

 - Top Prize:Prize amount on the high tier

-Mid Pize: Prize on the middle tier

-Small Prize: Prize on the low tier

**Payment Information:**

 - Payment Method ID: Unique identifier for the payment method

 - Payment Method Name: Name of the payment method (e.g., Cash, Debit Card)

 - Is Allowed In Store: Indicates if the payment method is accepted at physical retail points

 - Is Allowed Online: Indicates if the payment method is accepted on online platforms

 - Is Cash Equivalent: Specifies if the payment method is treated like cash

 - Is Digital: Indicates whether the payment method is digital (e.g., mobile wallets)

 - Credit Based: Shows if the payment method involves borrowing or credit (e.g., credit cards)

 - Status: Current status of the payment method, such as active or inactive

 - Notes: Additional information or regulatory remarks about the payment method

In the second data source we have considered the following information about draw-based game tickets sales in Texas:

**Sales Information:**

  - Transaction Date: The date when the draw ticket was sold

- Retailer License Number: The natural key of the retailer that performed the ticket sale

  - Customer ID: The unique ID of the customer that performed the ticket sale

 - Employee ID: The unique ID of the employee that performed the ticket sale

-Game ID: Identifier of the draw-based game

 - Sales Amount: The total revenue from the ticket sale

 - Number of Tickets Sold: The number of tickets sold for that particular draw game

 - Payout: The amount claimed by a winning customers, 0 otherwise

 - Retailer License Number: The natural key of the retailer that performed the ticket sale

**Date Information:**

 - Date ID: Unique identifier for the date (e.g., YYYYMMDD)

 - Fiscal Year: Fiscal year number for the date

 - Fiscal Month: Fiscal month number

 - Fiscal Month Name and Number: Textual name and number of the fiscal month

**Customer Information:**

 - Customer ID: Unique identifier for the customer

 - Customer Email: Email address of the customer

 - Customer Phone: Phone number of the customer

 - Customer Registration Date: Date when the customer registered

 - Customer State: State of residence of the customer

**Employee Information:**

 - Employee ID: Unique identifier for the employee

 - Employee Email: Email address of the employee

 - Employee Phone: Phone number of the employee

 - Employee Hire Date: Date when the employee was hired

 - Employee Salary: Annual salary of the employee

**Retailer Information:**

 - Retailer License Number: Unique license number for the retailer

 - Retailer Location Name: Name of the retailer’s physical location

 - Retailer ZIP Code: ZIP code of the retailer’s location

 - Retailer Location State: State where the retailer is located

**Game Information:**

 - Game ID: Identifier of the draw-based game

 - Game Category: Category or type of the game (e.g., “million”)

 - Ticket Price: Standard ticket price for the draw game

 - Draw Date: Date of the draw

 - Winning Jackpot: Big prize amount based on game played

 - Winning Chance: the probability of winning based on game category

**Payment Information:**

 - Payment Method ID: Unique identifier for the payment method

 - Payment Method Name: Name of the payment method (e.g., Cash, Debit Card)

 - Is Allowed In Store: Indicates if the payment method is accepted at physical retail points

 - Is Allowed Online: Indicates if the payment method is accepted on online platforms

 - Is Cash Equivalent: Specifies if the payment method is treated like cash

 - Is Digital: Indicates whether the payment method is digital (e.g., mobile wallets)

 - Credit Based: Shows if the payment method involves borrowing or credit (e.g., credit cards)

 - Status: Current status of the payment method, such as active or inactive

 - Notes: Additional information or regulatory remarks about the payment method

Further, in order to support analysis and ensure data is structured logically, we define the following hierarchies in our data warehouse:

**Location Hierarchy:** **LOCATIONS → ZIPS→ CITIES → STATES**

This hierarchy allows us to analyze sales performance from a broad regional level down to individual customers.

Time Hierarchy:

### DATES → CALENDAR\_MONTH\_NAME\_NUMBERS → CALENDAR\_MONTHS → CALENDAR\_YEARS

### Fiscal Time Hierarchy:

### DATES → FISCAL\_MONTH\_NAME\_NUMBERS → FISCAL\_MONTHS → FISCAL\_YEARS

This hierarchy enables temporal analysis of sales trends over time, allowing year-over-year, quarterly, and monthly comparisons.

**Game Hierarchy:** **GAMES → CATEGORIES→TYPES**

This hierarchy enables tracking the different categories and types of the lottery games played and the different attributes specific to the category and can be retrieved through this table

## GRAIN / DIM / FACT

In developing our data warehouse, we define the grain of our fact table using the standard four-step dimensional design process. While we are working with two separate source systems — one for scratch-off games and another for draw-based games — both sources relate to lottery ticket sales in the state of Texas. This shared geographic scope enables a unified modeling approach, even though each source has its own structure and game-specific attributes.

### 1. Selecting the Business Process

The business process we are analyzing is:

***"How many draw-based and scratch-off lottery tickets were sold in Texas, both in terms of units and dollar value?"***

Though the data originates from two distinct sources, both serve the same analytical purpose — evaluating lottery ticket sales across game types (scratch-off/draw-based) in a single market (Texas).

### 2. Declaring the Grain

The grain of the fact table is defined as a single ticket sales transaction. Each record in the fact table represents the sale of a certain number of tickets by an employee to a customer at a specific retailer on a particular date, for a specific round of a game (either draw-based or scratch-off), using a specific payment method.

This grain is applied consistently across both data sources. Despite differences in structure and metadata, both datasets can be normalized at this transaction level.

### 3. Identifying the Dimensions

To determine which dimensions are required, we apply the guiding questions:

***Where? Who buys? Who sells? What? When?***

We identify the following core dimensions, shared across both sources:

**Customer Dimension**

* Basic fields in both sources
* Draw-based games include extended attributes (e.g., email, phone, registration date, state)

**Employee Dimension**

* Includes ID and name
* Draw-based data also provides salary and hire date

**Retailer Dimension**

* Common across both: license number, name, ZIP code, and city
* State is constant (Texas)

**Date Dimension**

* Standard fields: date ID, fiscal month/year, month name/number, etc.

**Game Dimension**

* Differentiated by source:  
  + Scratch-off: Odds, prize structure, ticket price variability
  + Draw-based: Draw date, jackpot prize, fixed pricing

**Payment Dimension**

* Standard fields: payment id, payment type, boolean values, whether active or inactive as status

While the Game Dimension varies in content depending on the source, it can still be modeled as a unified dimension with optional or null attributes depending on the game type.

### 4. Identifying the Facts

To identify the facts, we ask:

***"What is the business process measuring?"***

**Additive Facts (aggregatable):**

* Units Sold – Number of tickets sold in the transaction
* Sales Amount – Ticket Price × Units Sold

**Non-Additive / Contextual Facts:**

* Ticket Price – May vary by source (dynamic in scratch-off, fixed in draw-based)
* Payout(if not win, then payout=0):  
  + Scratch-off: winning Jackpot
  + Draw-based: scaled prize

These facts enable flexible, multi-dimensional analysis, such as gross or net revenue by game type, sales by region or retailer, and trends across time.

Although the data originates from two distinct systems — one for scratch-off games and another for draw-based games — both are part of the same business process and same geographic scope (Texas). Therefore, we design a unified dimensional model.

Most dimensions (Customer, Retailer, Employee, Date) are shared, and the differences in game types are isolated within the Game Dimension and source-specific fact attributes. This structure ensures high-quality integration, supports broad analytical queries, and maintains consistency across both data sources.

***FCT\_SALES***

All the sales transactions made

| Column name | Description | Data Type |
| --- | --- | --- |
| TRANSACTION\_DT\_ID(FK) | Natural primary key of the date dimension | BIGINT |
| GAME\_NUMBER(FK) | Natural primary key of the game dimension | BIGINT |
| CUSTOMER\_ID(FK) | Natural primary key of the customer dimension | VARCHAR(50) |
| EMPLOYEE\_ID(FK) | Natural primary key of the employee | VARCHAR(50) |
| RETAILER\_LICENSE\_NUMBER(FK) | Natural primary key of the retailer | INT |
| PAYMENT\_ID(FK) | The type of payment used during the transaction | VARCHAR(50) |
| TICKETS\_BOUGHT | (measurement) units boughts | INT |
| PAYOUT | (measurement)amount claimed by a winner customer, 0 otherwise | INT |
| SALES | (measurement)amount sold in dollars | INT |

Example with filled data

| TRANSACTION\_ID | SCRATCH\_GAME\_NUMBER | CUSTOMER\_ID | EMPLOYEE\_ID | RETAILER\_LICENSE\_NUMBER | TICKETS\_BOUGHT | PAYOUT | SALES |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 20210325 | scratch\_2053 | SAN ANTONIO\_137 | EMP\_100663\_3 | 100663 | 1 | 15000 | 30 |
| 20210112 | all\_or\_nothing\_05\_january\_100663 | SAN ANTONIO\_001 | EMP\_100663\_1 | 100663 | 3 | 0 | 3 |

***DIM\_CUSTOMERS***

In this dimension we include our customers along with their characteristics

| Column name | Description | Data Type |
| --- | --- | --- |
| CUSTOMER\_ID(PK) | Natural primary key of the employee | BIGINT |
| CUSTOMER\_NAME | Full name of the customer | Text |
| CUSTOMER\_GENDER | Gender (M/F) | CHAR(1) |
| CUSTOMER\_DOB | Date of birth | DATE |
| CUSTOMER\_EMAIL | Email of the customer | VARCHAR(100) |
| CUSTOMER\_PHONE | Phone Number of the customer | VARCHAR(20) |
| CUSTOMER\_REGISTRATION\_DT | Registration Date n the lottery commission database of the customer | DATE |
| CUSTOMER\_STATE | State of residence | VARCHAR(50) |
| CUSTOMER\_CITY | The city the customer lives in | VARCHAR(40) |
| CUSTOMER\_ZIP | The zip code of the customer | INT |

Example with filled data

| CUSTOMER\_ID | CUSTOMER\_NAME | CUSTOMER\_GENDER | CUSTOMER\_DOB | CUSTOMER\_EMAIL | CUSTOMER\_PHONE | CUSTOMER\_REGISTRATION\_DT | CUSTOMER\_STATE | CUSTOMER\_CITY | CUSTOMER\_ZIP |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SAN ANTONIO\_137 | Matthew Harris | M | 1990-05-15 | hmartinez@gmail.com | 887-564-9125x92300 | 2012-10-21 | TX | San Antonio | 24592 |

***DIM\_EMPLOYEES***

Here, we have information about employees.

| Column name | Description | Data Type |
| --- | --- | --- |
| EMPLOYEE\_ID(PK) | The unique identifier of the employee based on retailer and hiring round | BIGINT |
| EMPLOYEE\_NAME | The name of the employee | VARCHAR(255) |
| EMPLOYEE\_DEPARTMENT | The department the employee works in | VARCHAR(255) |
| EMPLOYEE\_STATUS | Status(e.g. Active, Inactive) | VARCHAR(255) |
| EMPLOYEE\_EMAIL | Email of the employee | VARCHAR(255) |
| EMPLOYEE\_PHONE | The phone number of employee | VARCHAR(255) |
| EMPLOYEE\_HIRE\_DT | The hire date of the employee | DATE |
| EMPLOYEE\_SALARY | The salary of the employee | INT |

| EMPLOYEE\_ID | EMPLOYEE\_NAME | EMPLOYEE\_DEPARTMENT | EMPLOYEE\_STATUS | EMPLOYEE\_EMAIL | EMPLOYEE\_PHONE | EMPLOYEE\_HIRE\_DT | EMPLOYEE\_SALARY |
| --- | --- | --- | --- | --- | --- | --- | --- |
| EMP\_100663\_3 | John Waters | Sales Associate | Inactive | john.waters@retail.com | 887-564-9125x92300 | 2017-12-17 | 67182 |

***DIM\_RETAILERS***

In this dimension we define the retailers.

| Column name | Description | Data Type |
| --- | --- | --- |
| RETAILER\_LICENSE\_NUMBER(PK) | Natural key of the retailer | BIGINT |
| RETAILER\_LOCATION\_NAME | The location name of the retailer | VARCHAR(50) |
| RETAILER\_LOCATION\_ZIP\_CODE | The zip code of the retailer | TEXT |
| RETAILER\_LOCATION\_CITY | The City of the retailer | VARCHAR(100) |
| RETAILER\_LOCATION\_STATE | The state the retailer is in (e.g.TX) | VARCHAR(100) |

Example with filled data

| RETAILER\_LICENSE\_NUMBER | RETAILER\_LOCATION\_NAME | RETAILER\_LOCATION\_CITY | RETAILER\_LOCATION\_ZIP\_CODE | RETAILER\_LOCATION\_STATE |
| --- | --- | --- | --- | --- |
| 100663 | ISI-KAT | San Antonio | 78212 | TX |

***DIM\_GAME***

In this dimension, we are including all characteristics of the lottery games played

| Column name | Description | Data Type |
| --- | --- | --- |
| GAME\_NUMBER(PK) | The unique identifier of each scratch game round | BIGINT |
| TICKET\_PRICE | The ticket price of the scratch game round | INT |
| GAME\_TYPE | The type of game(i.e. Scratch/Non-Scratch) | VARCHAR(50) |
| GAME\_CATEGORY | The category of the game(MIllions, Lotto, Scratch Ticket) | VARCHAR(100) |
| AVERAGE\_ODDS | Random odds of winning in a scratch-off lottery game per round | VARCHAR(10) |
| AVERAGE\_ODDS\_PROB | The probability of winning in a scratch-off lottery game per round | FLOAT |
| TOP\_PRIZE | The value of the top tier prize in scratch-off games | INT |
| MID\_PRIZE | The value of the mid tier prize in scratch-off games | INT |
| SMALL\_PRIZE | The value of the small prize in scratch-off games | INT |
| WINNING\_CHANCE | Probability of winning in draw-based games | FLOAT |
| WINNING\_JACKPOT | Jackpot prize amount of draw based games | INT |
| DRAW\_DT | Date of draw event | DATE |

| GAME\_NUMBER | TICKET\_PRICE | GAME\_TYPE | GAME\_CATEGORY | AVERAGE\_ODDS | AVERAGE\_ODDS\_PROB | TOP\_PRIZE | MID\_PRIZE | SMALL\_PRIZE |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| all\_or\_nothing\_05\_january\_100663 | 1 | Non-scratch | All or Nothing | DEFAULT | DEFAULT | DEFAULT | DEFAULT | DEFAULT |
| scratch\_2053 | 30 | Scratch | Scratch Tickets | 1:3.44 | 0.2907 | 300 000 | 15000 | 300 |

| WINNING\_CHANCE | WINNING\_JACKPOT | DRAW\_DT |
| --- | --- | --- |
| 1.016590761223162e-06 | 250000 | 01/31/2021 |
| DEFAULT | DEFAULT | DEFAULT |

***DIM\_DATE***

In this dimension we are including all characteristics of the date id, in our csv file we only include DATE\_ID, but later we will code the rest of the fields, based on common knowledge for our database

| Column name | Description | Data Type |
| --- | --- | --- |
| DATE\_ID | The unique identifier of the date | BIGINT |
| EVENT\_DT(PK) | The unique natural identifier of the event(event date) | BIGINT |
| FISCAL\_YEAR | Fiscal year | INT |
| FISCAL MONTH | Fiscal month | INT |
| FISCAL\_MONTH\_NAME\_NUMBER | The start of the fiscal month | VARCHAR(50) |
| CALENDAR\_YEAR | The calendar year | INT |
| CALENDAR\_MONTH | The calendar month | INT |
| CALENDAR\_MONTH\_NAME\_NUMBER | The start of the calendar month | VARCHAR(50) |
| MONTH\_ENDING\_DATE | The end date of the month | DATE |

Example with filled data

| DATE\_ID | EVENT\_DT | FISCAL\_YEAR | FISCAL\_MONTH | FISCAL\_MONTH\_NAME\_NUMBER | CALENDAR\_YEAR | CALENDAR\_MONTH | CALENDAR\_MONTH\_NAME\_NUMBER | MONTH\_ENDING\_DATE |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20210112 | 2021-01-12 | 2021 | 5 | 05-January | 2021 | 1 | 01-January | 01/31/2021 |

***DIM\_PAYMENT***

In this dimension we are including all characteristics of the date id, in our csv file we only include DATE\_ID, but later we will code the rest of the fields, based on common knowledge for our database

| Column name | Description | Data Type |
| --- | --- | --- |
| PAYMENT\_METHOD\_ID (PK) | Unique identifier for the payment method | BIGINT |
| PAYMENT\_METHOD\_NAME | Name of the payment method | VARCHAR(100) |

Example with filled data

| **PAYMENT\_METHOD\_ID** | PAYMENT\_METHOD\_NAME |
| --- | --- |
| 1 | Cash |

# Business Layer 3NF

In the core concept of Slowly Changing Dimensions (SCD), there are three main types: Type 0, Type 1, and Type 2. Additionally, there is a hybrid type that combines Type 1 and Type 2 approaches.

**Type 0**: This dimension does not change over time. The data remains constant, such as a payment type dimension with fixed values like "prepaid" or "postpaid."

**Type 1**: In this type, changes overwrite the existing data without keeping any history. When an attribute (e.g., package type) is updated, the old value is simply replaced. This is used when history is not important.

**Type 2**: This type tracks the full history of changes. The dimension table includes additional columns such as START\_DT, END\_DT, and an IS\_CURRENT or IS\_ACTIVE flag. These columns allow us to keep historical versions of a row, marking which version was valid at any given time. Queries use date ranges like START\_DT <= SEARCH\_DATE < END\_DT to find the correct version. When a record expires, the END\_DT is set to the expiration date, and the current row’s END\_DT is set to a far future date (e.g., '31-Dec-9999'). The primary key may stay the same(e.g. the product as an entity stays the same only some descriptive attributes change) or change for each historical record(e.g. an individual moving through roles into a company), while the natural key remains the same.

**Type 3**: Instead of creating multiple rows for history, Type 3 stores previous attribute values in new columns within the same row. For example, you might have a package\_type column for the current value and an old\_package\_type column to store the previous value. This method captures only limited history (usually just one previous value) and is useful when tracking a small number of changes is sufficient.

**Hybrid Type**: This combines Type 1 and Type 2. Some attributes are tracked historically like in Type 2, while others are simply overwritten as in Type 1. For example, package information may have a full change history, while product volume might be updated directly without storing history.

Now that we have reviewed the concept of Slowly Changing Dimensions (SCD), we can apply it to modeling our Third Normal Form (3NF) schema. As we previously introduced the attributes, low-level dimensions, and fact table attributes, the next step is to design a 3NF model based on our source schema and tables, with the goal of normalizing our database.

To support data integration and tracking, we first introduce the concept of the **source triplet**, which consists of three key components: the source system, the source entity, and the source identifier. This triplet uniquely identifies each record in the source, enabling precise lineage and traceability.

Next, we define a <column\_name>\_surr\_id column as the primary key and surrogate key. This surrogate key is system-generated whenever a new record is integrated into the normalized schema (BL\_NF). Alongside it, we add a <column\_name>\_source\_id column, representing the natural key, which helps maintain connection to the original source data.

We also add SOURCE\_SYSTEM and SOURCE\_ENTITY columns to complete the source triplet, allowing backward tracing to the exact origin of each record.

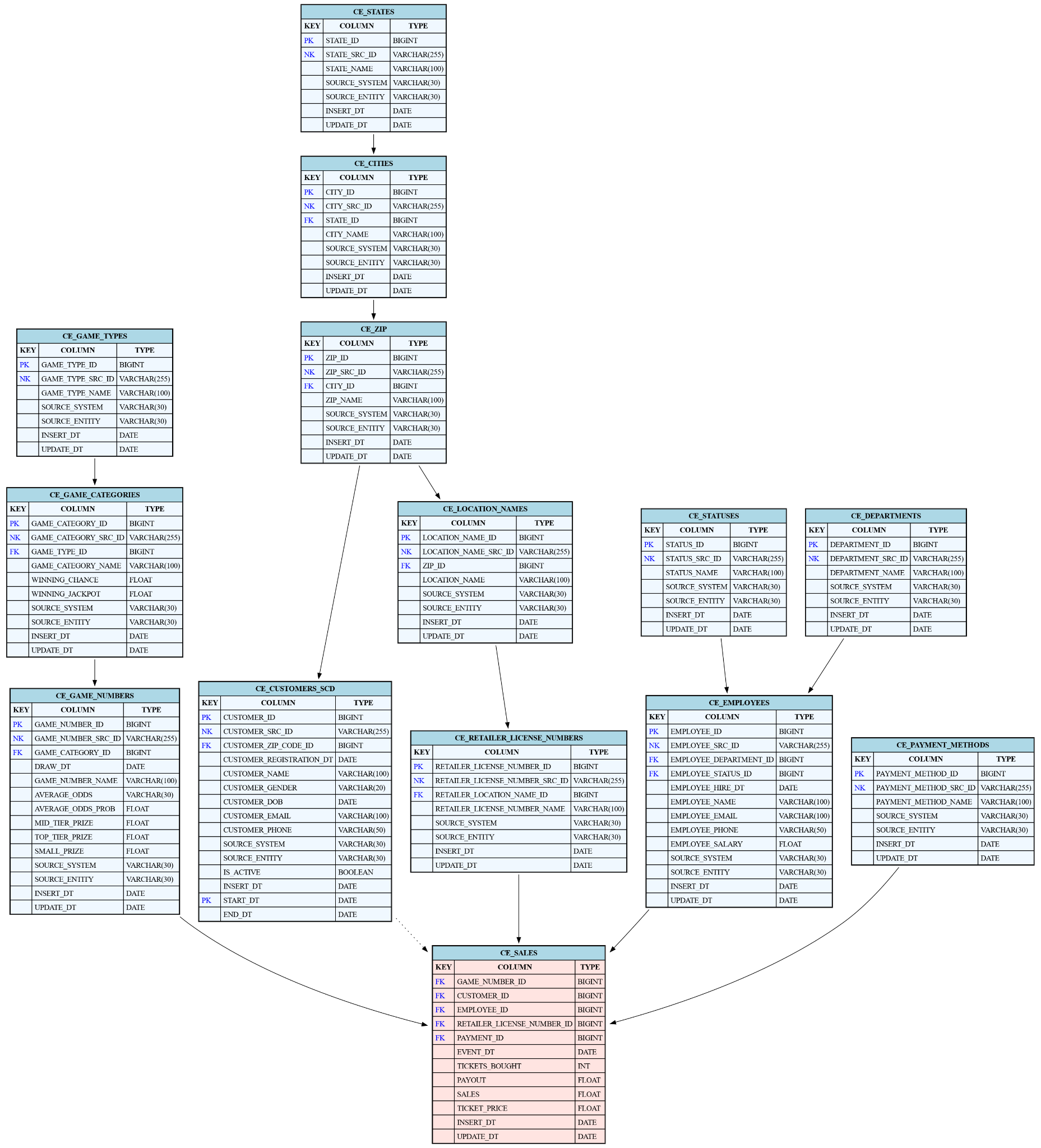
Finally, we apply the **Type 2 Slowly Changing Dimension** approach within our 3NF model by adding three additional columns: is\_active, start\_dt, and end\_dt. These columns enable us to preserve historical data — if the source system updates a record, rather than overwriting it, we insert a new record with the same surrogate and natural keys but with the updated attribute(s). The previous record remains in the table marked as inactive, ensuring the full history is retained.

In this modeling approach, when implementing **Slowly Changing Dimension Type 2 (SCD Type 2)** in a **Third Normal Form (3NF)** model, the relationship between the sales table and the customer (or employee) table is typically **logical** and is established through the **surrogate primary key combined with a condition on the start date**.

In the 3NF model, the **surrogate key remains unchanged** for a customer even as their attributes change over time. History is tracked by maintaining multiple records with the same surrogate key but different attribute values and effective date ranges (such as start\_date and end\_date). To link facts to the correct customer record, the join between the sales and customer tables uses the surrogate key **along with a condition ensuring the sales date falls within the customer’s effective date range**.

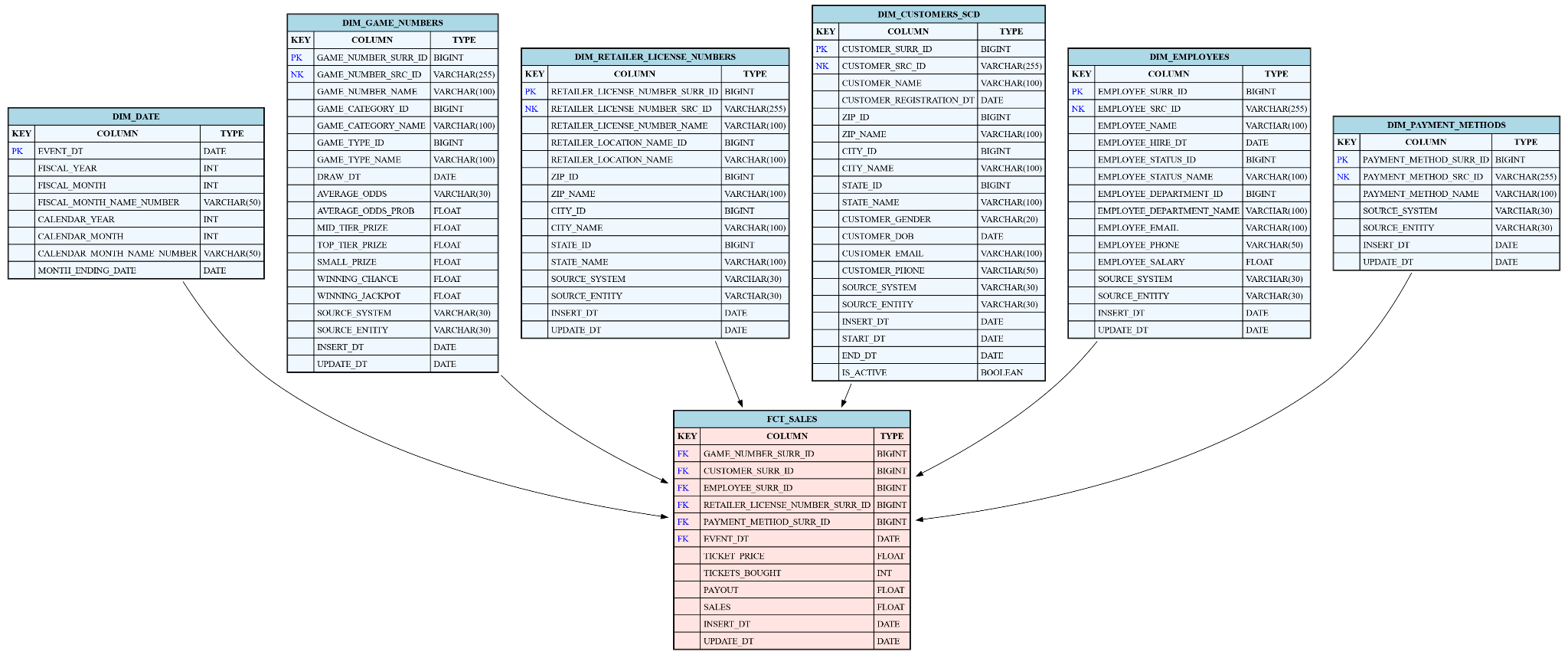
This means the relationship is logical because it depends not just on the key but also on the time context provided by the date fields.

In contrast, in a **dimensional star schema** implementing SCD Type 2, each attribute change results in a **new record with a new surrogate key** in the dimension table. The fact table references this changing surrogate key directly, creating a **physical relationship** that ties each fact to the exact version of the customer dimension valid at the time of the transaction. This simplifies querying but requires new keys for each change.



# 

# Business Layer Dimensional Model



**Additional metric**

A metric to consider in our fact table is (payout minus sales) divided by tickets\_bought. This gives the ***average net return per ticket***, showing how much, on average, each ticket contributes to the player's overall profit or loss.

# Logical Scheme

# Data Flow

# Fact Table Partitioning Strategy