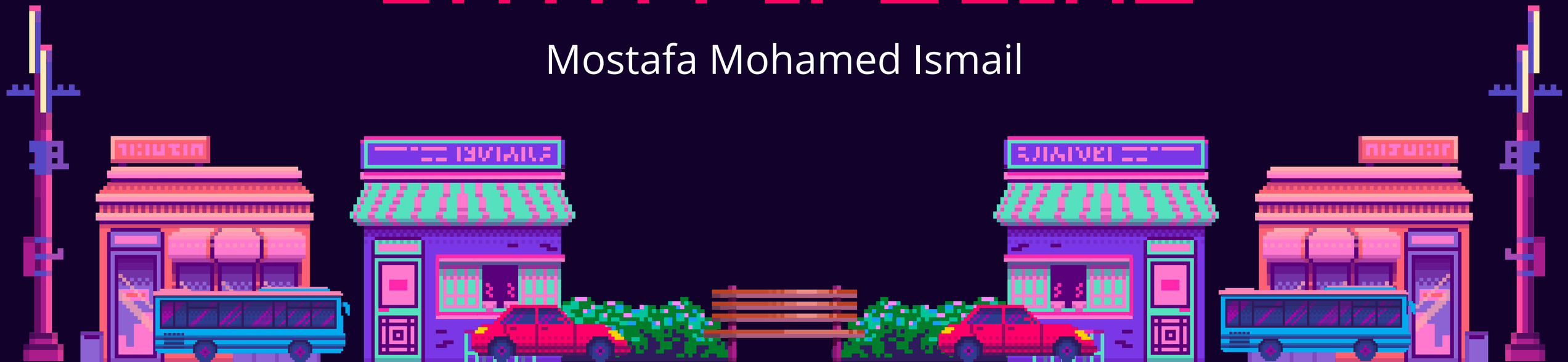
VIDEO GAMES SALES DATA PIPELINE

Mostafa Mohamed Ismail



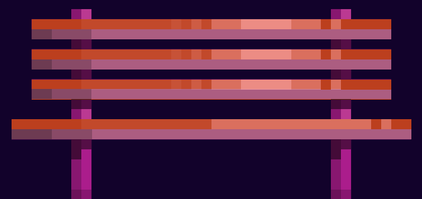
PROBLEM STATEMENT

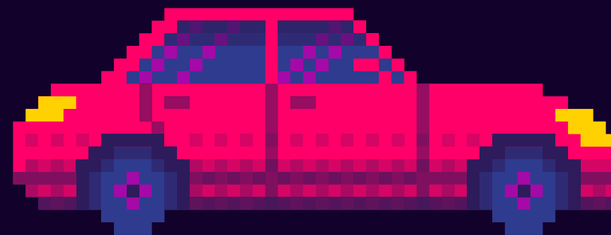
Raw data contains issues:

- "N/A" values instead of numbers
- Missing values
- Different data formats

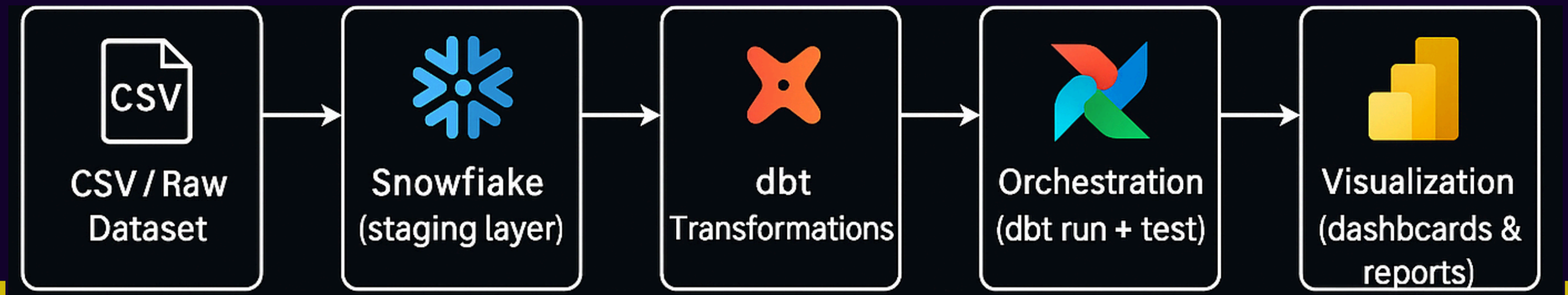
Business needs:

- Clean data
- Automated pipeline
- Easy visualization for insights





DATA PIPELINE ARCHITECTURE



DATA CLEANING

01 Removed incomplete rows
- Dropped records with missing values in critical fields (Name, Platform, Year, Genre, Publisher, Sales).

02 Enforced correct data types
- Converted Rank and Year into integers.
- Converted Sales values into floats for accurate calculations.

03 Handled missing values
- Replaced any remaining null sales values with 0.0 using COALESCE.

04 Standardized column names
- Renamed raw columns into meaningful business terms (e.g., na_sales → North America Sales (M)).

DATA MODELING

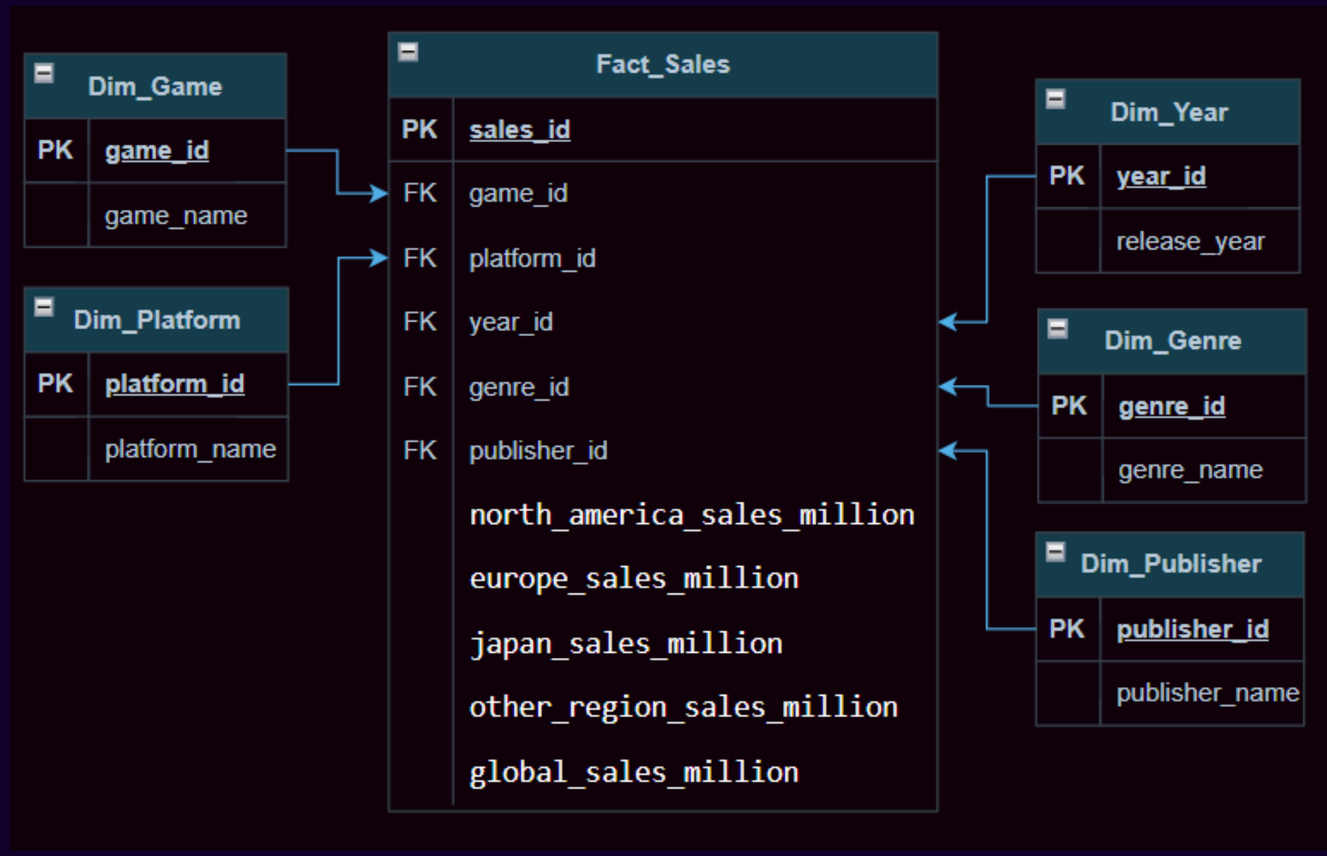
WHY STAR SCHEMA?

01 Simplicity
- Easy to understand: one Fact table (sales) linked to multiple Dimension tables (games, publishers, platforms, time).

02 Query Performance
- Optimized for analytical queries in Snowflake & Power BI (fewer joins, faster aggregations).

03 Flexibility in Analysis
- Allows slicing and dicing data by Year, Platform, Genre, Publisher, etc.

04 Best Practice in Data Warehousing
- Star schema is the industry standard for BI tools like Power BI, Tableau.



AUTOMATION

01

DAG Definition

- Name: dbt_workflow
- Runs daily (@daily).

```
# Define the DAG
dag = DAG(
    'dbt_workflow',
    default_args=default_args,
    description='Run dbt models using dbt Core',
    schedule_interval='@daily',
    catchup=False,
)
```

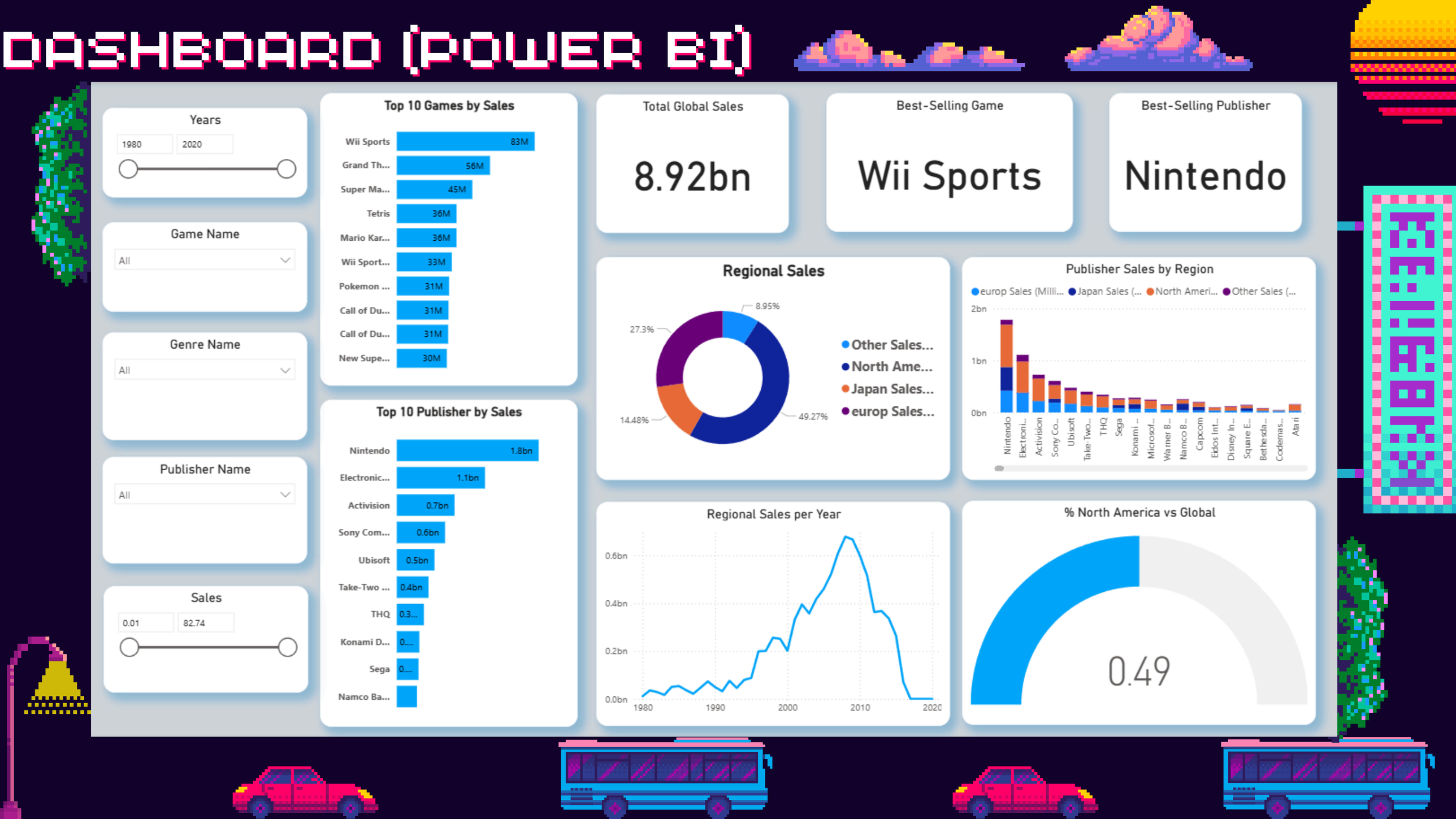
02

Task: Run dbt Models

- Executes dbt run command inside the project directory.
- Builds and refreshes data models in Snowflake.

```
# Task 1: Run dbt models
dbt_run = BashOperator(
    task_id='dbt_run',
    bash_command=f'cd {DBT_PROJECT_DIR} && dbt run',
    dag=dag,
)
# Define task dependencies
dbt_run
```

DASHBOARD (POWER BI)



DASHBOARD (POWER BI)

Years

1980 — 2020

Game Name

All

Genre Name

All

Publisher Name

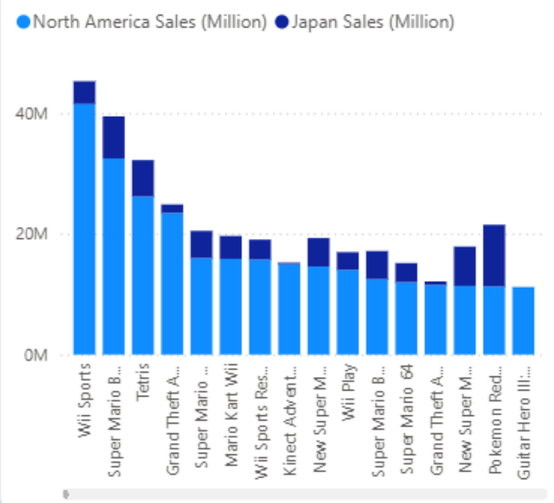
All

Sales

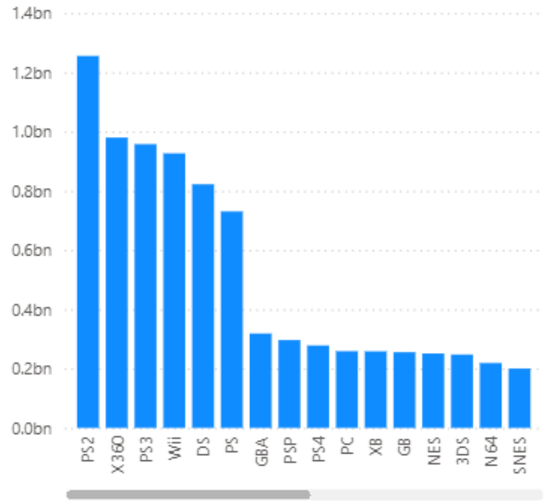
0.01 — 82.74

Sales by Genre

JP vs NA for Top Games

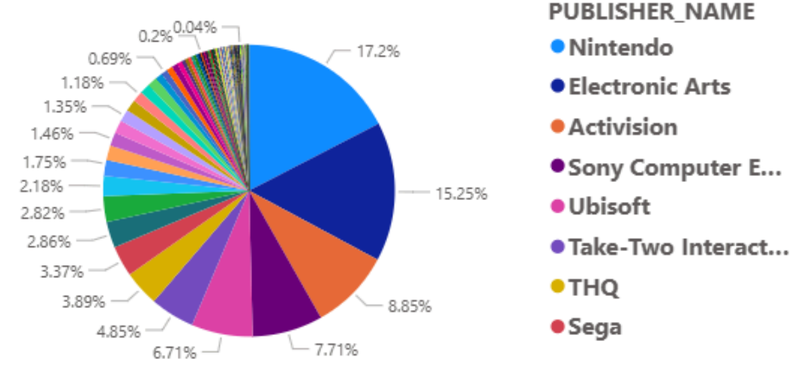


Platform Comparison



GENRE_NAME	2600	3DO	3DS	DC	DS	GB	GBA
Action	29,340,000.00		57,020,000.00	1,260,000.00	115,560,000.00	7,920,000.00	55,760,(
Adventure	1,700,000.00	60,000.00	4,810,000.00	2,500,000.00	47,290,000.00	17,160,000.00	14,680,(
Fighting	1,240,000.00		10,460,000.00	1,830,000.00	7,200,000.00		4,210,(
Misc	3,580,000.00		10,480,000.00		137,760,000.00	13,350,000.00	36,250,(
Platform	13,270,000.00		32,230,000.00	2,540,000.00	77,450,000.00	54,910,000.00	78,300,(
Puzzle	14,680,000.00	20,000.00	5,570,000.00		84,290,000.00	47,470,000.00	12,920,(
Racing	2,910,000.00		14,490,000.00	2,650,000.00	38,640,000.00	4,550,000.00	18,800,(
Role-Playing			75,740,000.00	680,000.00	126,850,000.00	88,240,000.00	64,210,(
Shooter	26,480,000.00		1,290,000.00	330,000.00	8,200,000.00	1,200,000.00	3,600,(
Simulation	450,000.00	20,000.00	27,080,000.00	520,000.00	132,030,000.00	3,550,000.00	5,910,(
Sports	3,430,000.00		6,200,000.00	3,660,000.00	31,830,000.00	9,050,000.00	16,410,(
Strategy			2,090,000.00		15,390,000.00	8,050,000.00	7,450,(
Total	**97,080,000.00**	**100,000.00**	**247,460,000.00**	**15,970,000.00**	**822,490,000.00**	**255,450,000.00**	**318,500,0**

EU Sales by Publisher



CHALLENGES & SOLUTIONS

01 Data quality issues → Solved by cleaning (null removal, type casting, handling N/A).

02 Large raw data → Automated with Airflow + dbt.

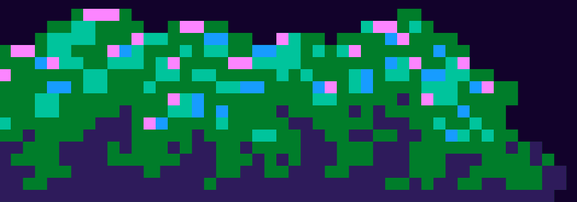
03 Business understanding → Visualized in Power BI.

CONCLUSION

01 Built an end-to-end data pipeline: raw → cleaned → Snowflake → Power BI.

02 Automated pipeline with Airflow.

03 Designed Star Schema for analytics.



THANK
YOU