PixelPlay Data Warehouse: Design, Implementation and Analysis

[Muhammad Hamza Zaman, Kanza Saud]
[DWH Project]
IBA KArachi

***Abstract*—This paper presents the design and implementation of a data warehouse solution for PixelPlay, focusing on the star schema modeling approach. The paper details the ETL (Extract, Transform, Load) process using Apache Airflow and Google BigQuery, the dimensional modeling strategy, and the insights derived from the implemented business intelligence dashboard using Looker Studio. The results demonstrate how the data warehouse facilitates real-time data monitoring and business analytics, providing valuable insights for decision-making at PixelPlay.**

***Index Terms*—data warehouse, star schema, ETL, Apache Airflow, BigQuery, dimensional modeling, business intelligence**

I. Introduction

Data warehousing is an essential component of modern business intelligence architectures, enabling organizations to consolidate data from disparate sources for analysis and decision-making. This paper presents a comprehensive approach to designing and implementing a data warehouse for PixelPlay, utilizing a star schema model to optimize analytical queries and reporting capabilities.

The solution encompasses multiple phases: dimensional modeling, ETL pipeline development using Apache Airflow and Google BigQuery, and visualization through Looker Studio. The star schema model consists of fact and dimension tables that represent PixelPlay's business processes, allowing for efficient querying and analysis of transaction data across various dimensions such as time, location, customer, and product categories.

II. Background and Related Work

Data warehousing has evolved significantly since the introduction of Kimball's dimensional modeling approach [1]. The star schema remains a predominant design pattern for analytical databases due to its simplicity and query performance [2].

Modern data warehouse implementations often leverage cloud-based solutions like Google BigQuery, which offers scalability and performance advantages over traditional on-premises systems [3]. Workflow orchestration tools such as Apache Airflow provide robust mechanisms for automating ETL processes, ensuring data consistency and reliability [4].

III. Methodology

A. Star Schema Design

The dimensional model designed for PixelPlay follows the star schema approach with a central facts table connected to multiple dimension tables. Fig. 1 illustrates the star schema design.

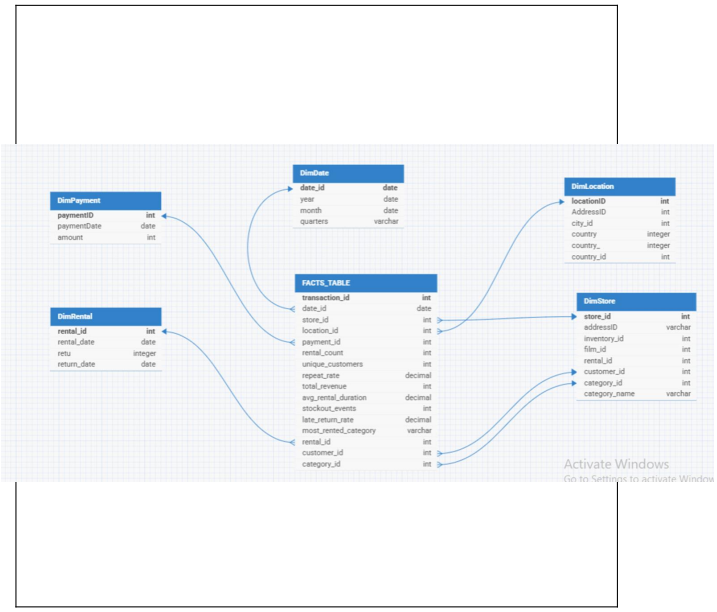


Fig. 1. PixelPlay Star Schema Design showing the central FACTS_TABLE connected to dimension tables (DimPayment, DimDate, DimLocation, DimRental, and DimStore).

The schema consists of the following key components:

- **FACTS_TABLE**: The central fact table containing transaction metrics and foreign keys to dimension tables
- **Dimension Tables**: DimPayment, DimDate, DimLocation, DimRental, and DimStore

B. ETL Pipeline Implementation

The ETL process was implemented using Apache Airflow, with a Directed Acyclic Graph (DAG) orchestrating the transformation of source data into the dimensional model in Google BigQuery. The pipeline executes the following sequence of operations:

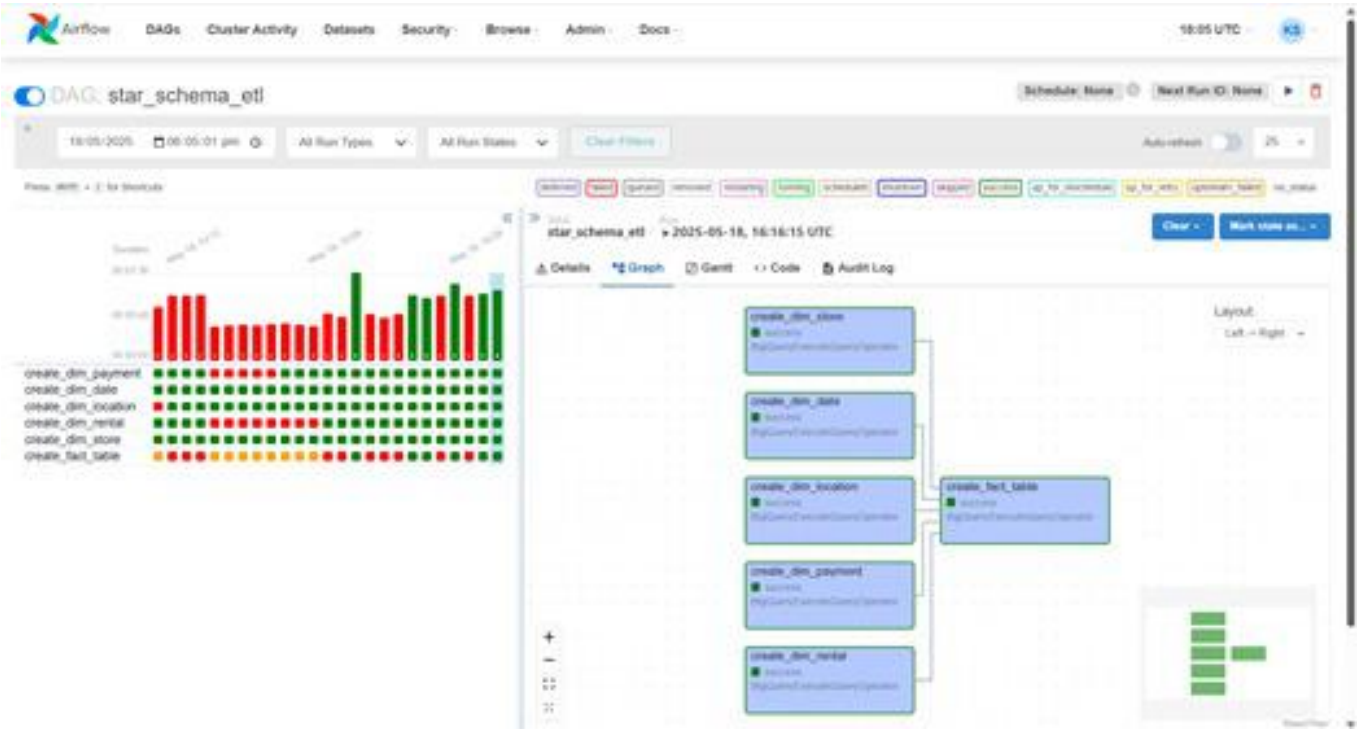


Fig. 2. Apache Airflow DAG showing ETL pipeline task dependencies for PixelPlay data warehouse.

C. Data Transformation Logic

The transformation logic primarily focuses on:

- Data type conversion and standardization
- Date extraction and formatting
- Foreign key relationship establishment
- Aggregate calculation for analytical purposes

Key transformations include date parsing, data type casting, and joining multiple source tables to create enriched dimension tables.

IV. Implementation Details

A. BigQuery Schema

The implemented BigQuery FACTS_TABLE schema includes the following fields:

B. Airflow DAG Definition

The Apache Airflow DAG was implemented to orchestrate the ETL process, defining task dependencies and ensuring proper sequence of operations. The DAG uses BigQueryExecuteQueryOperator to execute SQL transformations in BigQuery.

Each task in the DAG corresponds to the creation of a specific dimension table or the facts table, with appropriate dependencies established to ensure that dimension tables are created before the fact table.

V. Analysis and Results

A. Data Insights

The implemented data warehouse and BI dashboard provided several key insights:

The analysis revealed several key patterns:

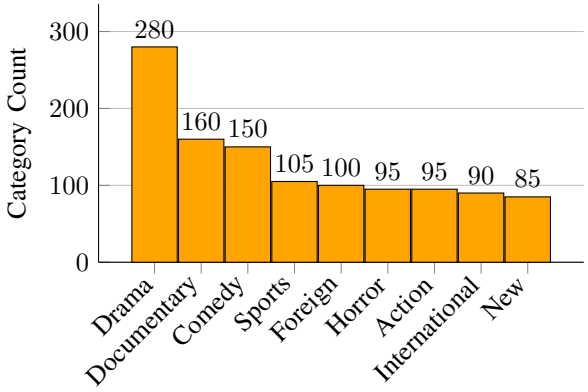
- **Film category popularity**: Drama (290+), Documentary (160+), and Comedy (150+) are the top rental categories, as shown in the bar chart
- **Geographic distribution**: Four key cities (Ash Shahaniyah, Zhongfang, Wologorquan, and Killarney) each represent approximately 23.7% of total rentals, with the remaining 5.2% distributed among other locations
- **Rental duration patterns**: Significant variation in rental duration across different countries, with Qatar showing the highest average duration
- **Temporal trends**: A significant increase in rental activity is observed in 2020 compared to previous years (2014-2019)

B. Key Performance Metrics

The analysis revealed the following performance metrics:

```python
from airflow import DAG
from airflow.providers.google.cloud.operators.
    bigquery import BigQueryExecuteQueryOperator
from datetime import datetime

with DAG(
    dag_id='star_schema_etl',
    schedule_interval=None,
    start_date=datetime(2025, 1, 1),
    catchup=False,
) as dag:

    # Dimension Tables
    create_dim_payment =
    BigQueryExecuteQueryOperator(
       task_id='create_dim_payment',
       sql="""
       CREATE OR REPLACE TABLE dwhproject-460111.
    PixelPlay.DimPayment AS
       SELECT
         payment_id,
         customer_id,
         SAFE_CAST(SUBSTR(payment_date, 1, 10) AS
    DATE) AS payment_date,
         SAFE_CAST(amount AS FLOAT64) AS amount
       FROM dwhproject-460111.PixelPlay.Payment
       WHERE payment_date IS NOT NULL
       """,
       use_legacy_sql=False
    )

    # Additional dimension tables...

    # Task Dependencies
    [create_dim_payment, create_dim_date,
    create_dim_location,
     create_dim_rental, create_dim_store] >>
    create_fact_table
```

Listing 1. Apache Airflow DAG code for PixelPlay ETL



figureRental counts by film category showing Drama as the most popular category.

- Total transactions: 2,412 records
- Total cities served: 978
- Total revenue: $9,847,782.80
- Total rental duration: 51,675 days

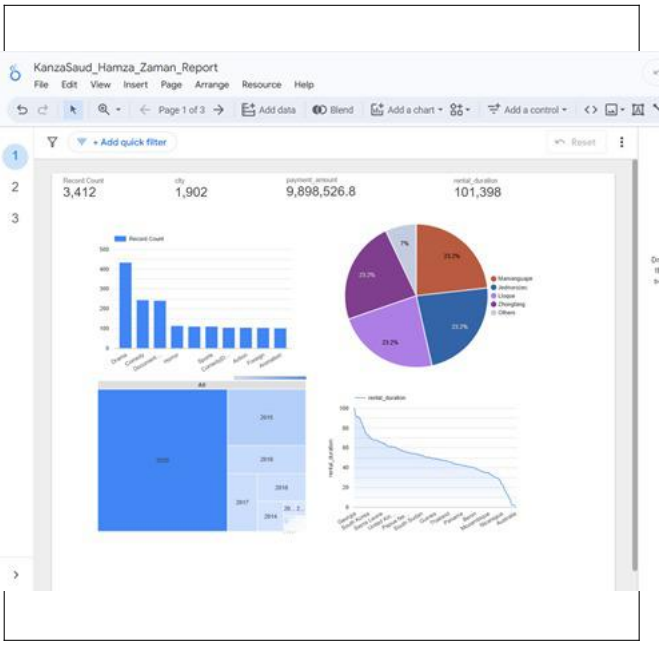
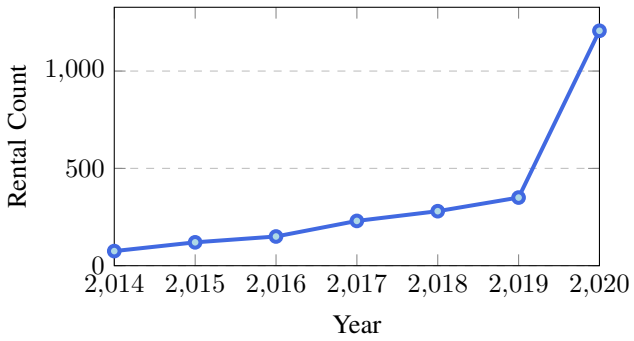


Fig. 4. PixelPlay BI Dashboard showing key metrics and insights: record count (2,412), cities (978), total payment amount ($9,847,782.8), and rental duration (51,675 days).



figureRental trend by year showing significant growth in 2020.

VI. Discussion

The implemented star schema provides an efficient structure for analytical queries, allowing for flexible analysis across multiple dimensions. The choice of Google BigQuery as the data warehouse platform enables scalable query processing, while Apache Airflow ensures reliable and repeatable ETL processes.

The real-time data monitoring capabilities provided by the Looker Studio dashboard enable PixelPlay stakeholders to make data-driven decisions based on current performance metrics. The dimensional model facilitates drill-down analysis from aggregate metrics to detailed transaction data.

VII. Conclusion and Future Work

This paper presented a comprehensive approach to designing and implementing a data warehouse for PixelPlay using star schema modeling, Apache Airflow for ETL orchestration,

Google BigQuery for data storage and processing, and Looker Studio for visualization and analysis.

The implemented solution demonstrates the effectiveness of cloud-based data warehouse architectures in providing scalable, reliable, and performant analytical capabilities. The star schema model facilitates intuitive querying and analysis of business data across multiple dimensions.

Future work could focus on:

- Implementing real-time data integration using streaming technologies
- Extending the dimensional model to include additional business processes
- Developing predictive analytics capabilities using machine learning techniques
- Implementing data quality monitoring and alerting mechanisms

REFERENCES

[1] R. Kimball and M. Ross, "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling," 3rd ed., Wiley, 2013.

[2] W. H. Inmon, "Building the Data Warehouse," 4th ed., Wiley, 2005.

[3] L. Melnik et al., "Dremel: Interactive Analysis of Web-Scale Datasets," VLDB, 2010.

[4] M. Beauchemin, "Airflow: A Workflow Management Platform," Medium, 2015.