NEW YORK UNIVERSITY

School of Professional Studies

Data Warehousing and Data Mining

MASY 3510 Section 101 | Spring 2024

Amazon Prime Data Warehouse

SUBMITTED BY:

Saumay Killa – sk10882@nyu.edu

Charlotte Ding – yd2881@nyu.edu

Emily Song – js10574@nyu.edu

SUBMITTED ON:

May 4^{TH}, 2024

UNDER THE GUIDANCE OF:

Prof. Sam Sultan

Table of Contents

Cover Letter

May 4, 2024
Greetings Amazon Prime,

We are writing to introduce a proposal for a data warehouse project designed to enhance the company's efficiency and productivity. The project involves developing a data warehouse that would collect data from various sources, allowing for analysis of trends, customer behavior, and other important metrics.

The project team consists of three participating students: Saumay, Charlotte, and Emily. We are confident that our combined skills and expertise will be able to deliver a successful outcome for your business.

By examining user watching patterns and subscription trends, Amazon Prime can gain the following advantages:

- Optimization of contents by identifying the most popular shows across various groups. By doing this, it would be ensured that the library is customized to the interests of the user, increasing engagement and retention.
- Personalized Recommendations: By analyzing viewing habits, it would help recommend shows that users are likely to enjoy.
- Pricing Strategies: By analyzing subscription trends, pricing models based on user behavior and geography can be optimized to maximize income while preserving customer satisfaction.
- Marketing Efficiency: Identify user segments with specific viewing behaviors and design marketing campaigns to maximize ROI.

We are confident that our proposal will address the needs of Amazon Prime’s operations and offer valuable insight. We look forward to the opportunity to discuss this proposal with you further.

Thank you.

Sincerely,
Saumay, Charlotte, and Emily.

Executive Summary

The main goal of creating the Amazon Prime data warehouse is to centralize and analyze vital information about customer subscriptions and video content. By concentrating on these two crucial areas, we hope to accomplish several strategic objectives that are critical for the business's ongoing development and success in the fiercely competitive digital market.

The data warehouse wants to offer a thorough repository for subscriber data from customers. This involves collecting and analyzing a range of data, including demographics, subscription plans, and renewal dates. This would allow Amazon Prime to obtain an understanding about their subscriber behavior, preferences, and engagement patterns. This would help in enhancing overall client satisfaction and retention and assist with marketing campaigns and optimize subscription services.

A fact table for movies which are being watched by the customers on Amazon Prime would also be included in the data warehouse. Complete details about watch time, number of shows watched, and completion trends, will be kept in this fact table. This would enable the company to learn more about audience preferences, content acquisition tactics, and content performance by examining the data. Decisions about creation, licensing, and curation of material could be made with these insights in mind, which would improve the content collection and draw and keep customers.

Overall, the objective of the Amazon Prime data warehouse development is to empower data-driven decision-making processes across various aspects of the business. By analyzing customer subscription and movie content data, Amazon Prime could optimize operations, enhance customer experiences, and maintain its competitive edge in the dynamic digital entertainment landscape.

Business Need And Process

Goal 1. To identify the favorite movie / show of each customer.

- To understand customer preference to improve the content library and offerings.
- To better recommend shows according to customer preference thereby increasing customer satisfaction.

Goal 2. To identify subscription behavior and create targeted marketing campaigns.

- To understand which region is using the services more thereby creating marketing campaigns based on the region.
- To understand if user are joining using promotional offers thereby increasing number of promotion offer and adjusting subscription price

Amazon Prime needs to improve its content library to improve customer satisfaction and retention. To achieve this goal, we suggest that Amazon Prime create Session and Payment data marts to gather, organize, and analyze data from its business operations.

The Session data mart will provide insights into which shows are watched frequently by the customers. By analyzing Session data, Amazon Prime can identify popular shows and seasonal trends, which can help in improving its content library. For instance, if most customers are watching a particular show, we could infer that most of the customers like a particular genre or prefer shows involving a certain cast.

The Payment data mart will provide insights into the subscription behavior of the customers. By analyzing subscription behavior, Amazon Prime can determine their target market and develop marketing strategies around them by examining subscription behavior.

Proposal

To improve business performance and optimize decision-making, we propose the development of Session and Payment data marts. Each data mart will be designed with a star schema model, consisting of one fact table and multiple dimension tables.

The Session data mart will be sourced from multiple systems, which would provide information on customers' sessions. The Payment data mart will be sourced from a customer relationship management (CRM) system, which will provide information on customer demographics, subscription, and purchase history.

Each data mart will relate to two conformed dimensions, ensuring consistency across data marts and facilitating easy navigation and analysis. The Session data mart will utilize dimensions such as Users, Shows, Date, and Device. The Payment data mart will use dimensions such as User and Date.

By implementing these data marts, we will be able to extract valuable insights into show performance and customer behavior. This will enable us to make data-driven decisions and optimize pricing, marketing, and content library. The proposed data marts will provide Amazon Prime with a competitive advantage in the market.

Business Justification And Benefits

The implementation of a data mart for Amazon Prime is a significant step towards improving content library and increasing revenue by identifying customer behavior and preferences. With the help of these data mart, Amazon Prime will be able to gather, arrange, and analyze information about customer preferences, giving it useful insights that will assist in making important business decisions.

The organization will experience a number of advantages as a result of the data mart's adoption, including the ones listed below:

- **Better Content Library:** The data mart will give Amazon prime data on which shows are watched more and when. The company can identify these popular shows and seasonal trends by examining session data, which can help with decisions regarding content library, creating and licensing similar shows, and marketing tactics. This would also enable better recommendation to customers and optimize its pricing.
- **Increased Revenue:** The data mart would provide insights into consumer behavior and preferences, allowing it to customize marketing initiatives for certain clientele and create promotion offers to promote business. They could boost revenue by providing services that are tailored to consumer segments and modifying prices accordingly.
- **Increased Customer Satisfaction and Retention:** The data mart will help tailor their content library based on customer preferences. This would increase customer Satisfaction and Retention.

In conclusion, the development of a data mart for Amazon Prime is a strategic investment that will allow the business to improve customer satisfaction and generate more revenue and give them a competitive advantage in the digital streaming market.

Details of Dimension And Fact Tables

The data marts are connected through the shared conformed dimensions, which enable consistent and accurate reporting across the organization.

In the Session data mart, the fact table is the Session_Fact table, which has foreign keys (FK) to the following dimension tables: User_Dim table, Show_Dim table, Device_Dim table, and Date_Dim table. The User_Dim table and Date_Dim table are conformed dimensions shared with the Payment data marts.

In the Payment data mart, the fact table is the Payment_Fact table, which has foreign keys to the following dimension tables: User_Dim table and Date_Dim table. The User_Dim table and Date_Dim table are conformed dimensions shared with the Session data marts.

Session Data Mart Fact table: Session_Fact table

- **Session_ID (PK)**
- **User_ID (FK)**
- **Show_ID (FK)**
- **Device_ID (FK)**
- **Date_ID (FK)**
- Session_Start_Time
- Session_End_Time
- NUM_OF_EPISODE_WATCHED
- Watch_Time
- Ads_Shown
- Ads_Skipped

Dimensions table:

User_Dim Table - *Conformed dimension with Payment data mart*

- **User_ID (PK)**
- Name
- Email
- Gender
- DOB
- AGE
- Subscription_Plan
- Plan_Start_Date
- Plan_End_Date
- Renewal_Status
- Region_Name
- ZipCode
- Age_Group

Date_Dim Table - *Conformed dimension with Payment data mart*

- **Date_ID (PK)**
- Load_Date
- Day
- Month
- Year
- Quarter
- Is_Holiday
- Is_Weekend

Show_Dim Table

- **Show_ID (PK)**
- Title
- Type
- Runtime
- Release_Date
- Genre
- Cast
- Director
- Country
- Ratings
- Language
- Downloadable

Device_Dim Table

- **Device_ID (PK)**
- Device_Name
- Type
- OS

Payment Data Mart Fact table: Payment_Fact table

- **Payment_ID (PK)**
- **User_ID (FK)**
- **Date_ID (FK)**
- Payment_Method
- Status
- Price
- Promotional
- Tax
- Discount
- Payment_Method_Fees

User_ Dim Table ***- Conformed dimension with Payment data mart***

- **User_ID (PK)**
- Name
- Email
- Gender
- DOB
- AGE
- Subscription_Plan
- Plan_Start_Date
- Plan_End_Date
- Renewal_Status
- Region_Name
- ZipCode
- Age_Group

Date_Dim Table ***- Conformed dimension with Payment data mart***

- **Date_ID (PK)**
- Load_Date
- Day
- Month
- Year
- Quarter
- Is_Holiday
- Is_Weekend

Model Diagram

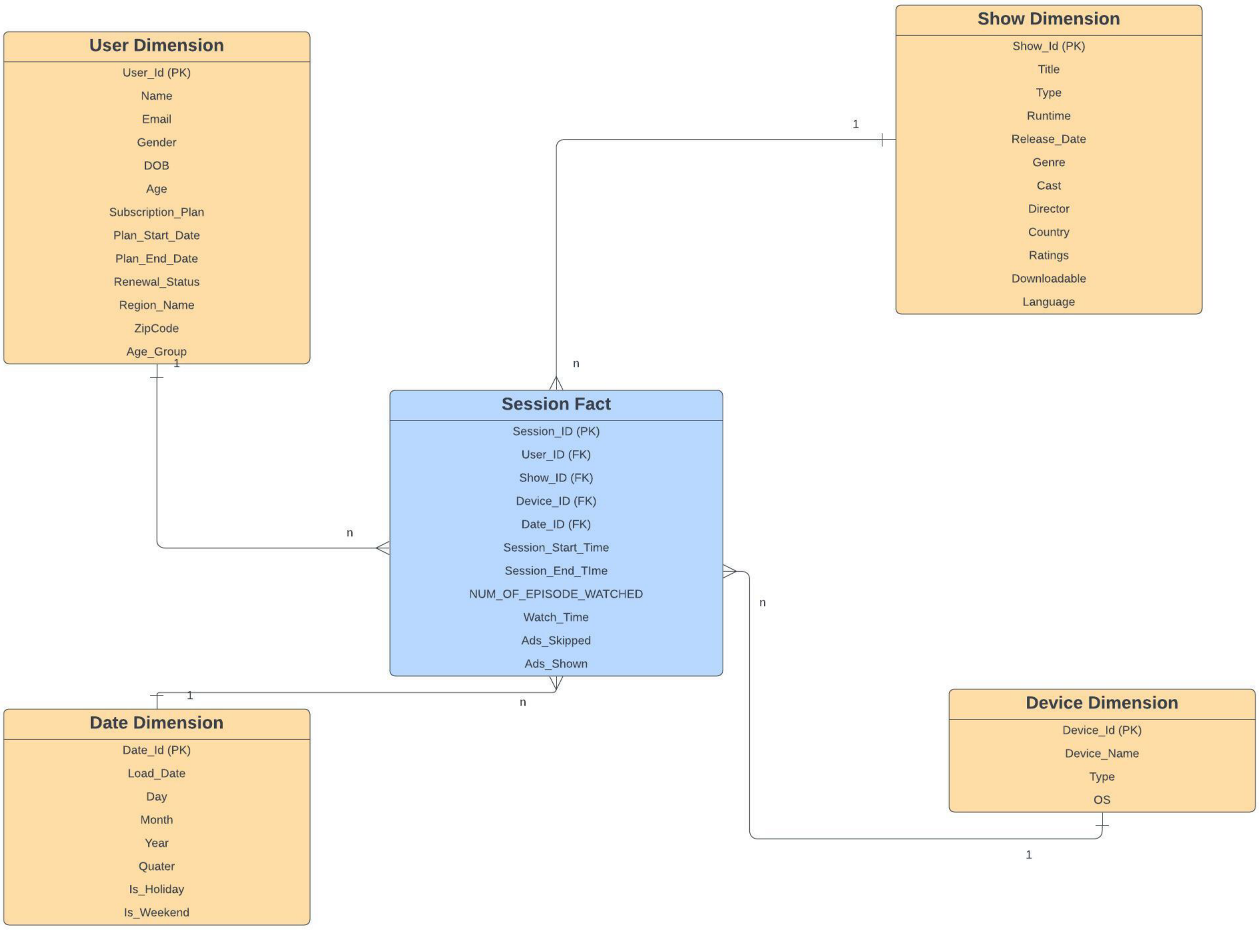


Fig 1: Session Data Mart

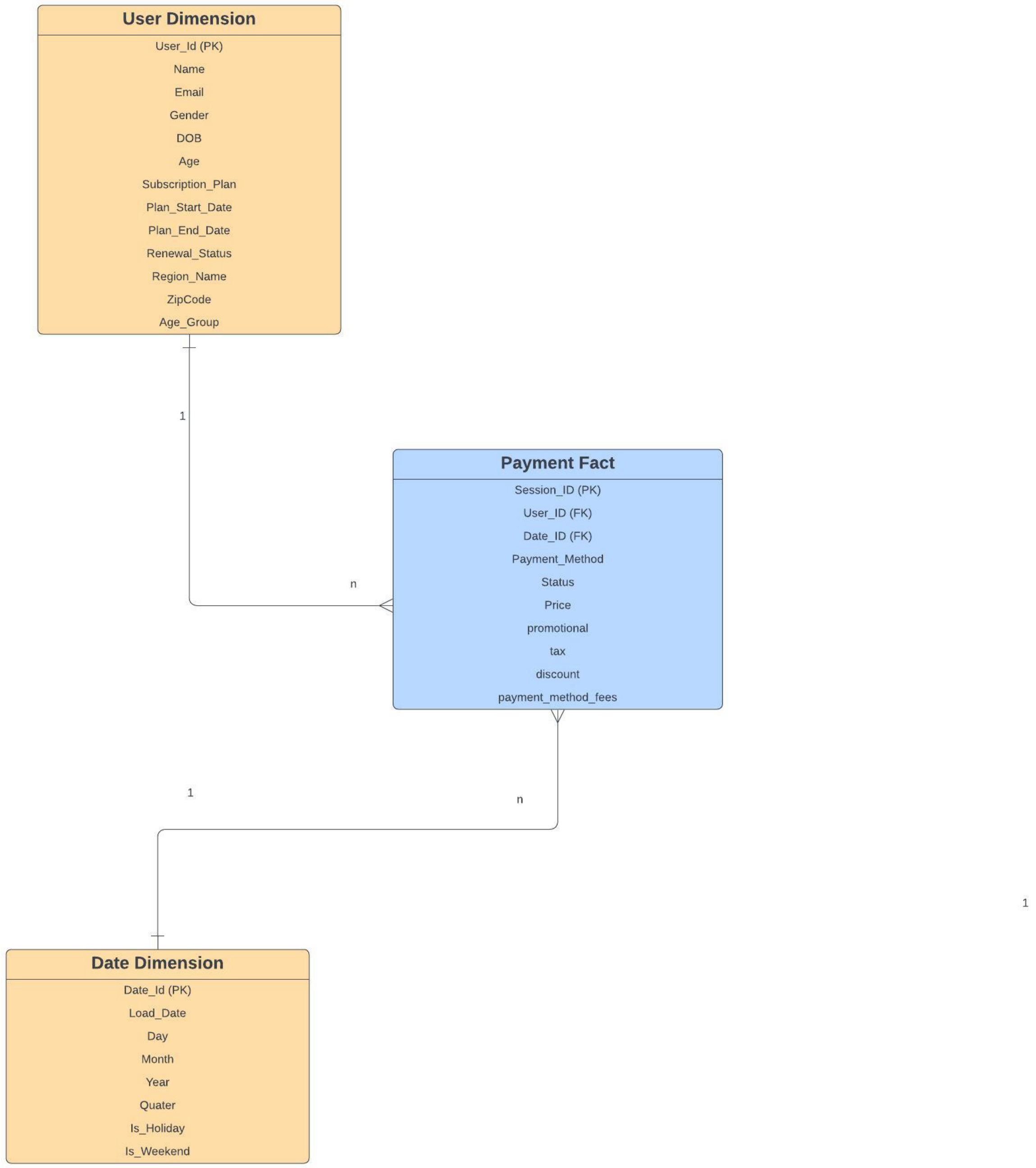


Fig 2: Payment Data Mart

Extract Transform And Load

For the ETL process, the first step is to identify the data sources. This includes Subscription Management System, Customer Relationship Management (CRM) System, Billing System, Content Management System (CMS), Streaming Analytics Platform, Market Research Reports and any other systems that store data relevant to the business.

Then, we would identify the types of files that contain the data and from where the data is extracted. The data files that we came across are mostly Excel files. Our goal is to extract the data from the already existing files and transform them to meet the data mart requirements.

After transforming the data, we will load the data in the database. If the data set is not too large, then we decide to query and code the data using SQL. If the data is huge and complex, we would like to read the file directly from SQL Developer. In this step, we would also ensure that the data is consistent and accurate.

The data mart will be kept up to date and checked upon from time to time by our team members.

Slowly Changing Dimensions

The Type 2 Slowly Changing Dimension (SCD) technique will be used for this project to manage changes in the Show dimension and User dimension.

The Type 2 technique creates new rows of records with updated attributes while keeping the previous record to keep track of past modifications and the evolution of a dimension over time.

The Show_Dim table in the Session data mart would use Type 2 technique to monitor changes in Ratings. Every time there is a change, a new record is made with a new surrogate key, a start date, and an end date. The end date is set to a high value, such as 9999-12-30, indicating that the record is still active, and the start date denotes the day the change took place. The original record would be closed by updating the end date to the date before the alteration took place. This enables historical customer demographic reporting.

The User_Dim table will use Type 2 technique in a similar manner to track modifications to Subscription Plan, Membership Start Date and Membership End Date.

While historical reporting would refer to the appropriate record depending on the relevant date range, the fact tables would refer to the most recent record in the relevant dimension table.

By employing this technique, we can ensure data accuracy and maintain historical reporting capabilities.

Aggregate Tables

We have decided not to use aggregate tables in our project for several reasons. First, the dataset is relatively small and easily manageable. This means that the data can be stored and processed using standard hardware and software without the need for specialized tools or resources.

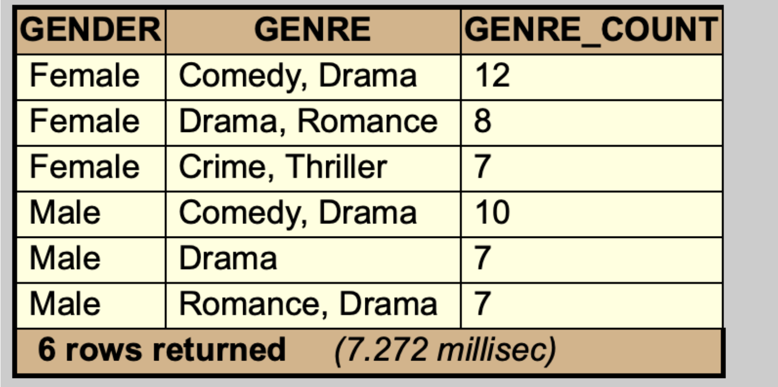
In addition, the data is not complex, suggesting no complicated data relationships or structures that would cause difficulties in summarizing or aggregating the data. Therefore, we believe that we can easily analyze the data without the need for additional manipulation or processing.

Given these factors, we have concluded that using aggregate tables is not necessary and cost-effective for the project. While aggregate tables could be useful in certain contexts, such as when dealing with large and complex datasets, it would not be necessary for the dataset currently on hand.

Analytical Queries

1. **Top 3 genres (most watched) by each gender**

```
select gender, genre, genre_count from (

select gender, genre,  count(genre) as genre_count, row_number() over (partition by gender order by count(genre) desc) as row_num

from user_dim

join session_fact using (user_id)

join show_dim using (show_id)

group by gender, genre)

where row_num <= 3
```

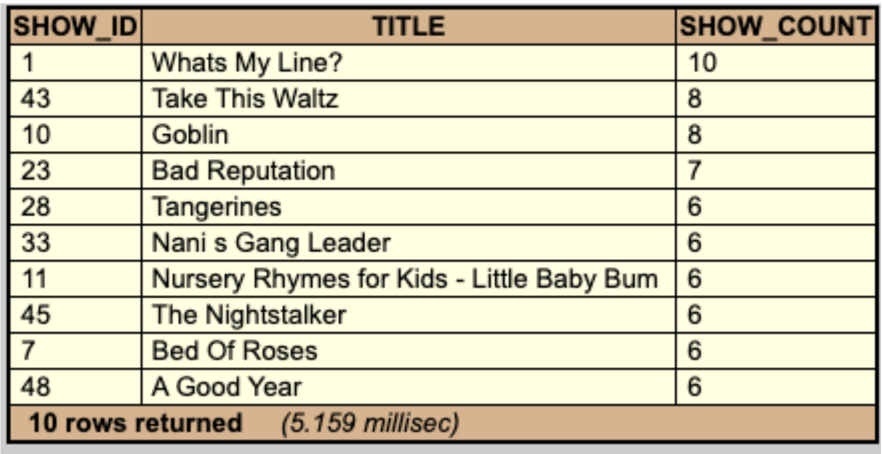


GENDER	GENRE	GENRE_COUNT
Female	Comedy, Drama	12
Female	Drama, Romance	8
Female	Crime, Thriller	7
Male	Comedy, Drama	10
Male	Drama	7
Male	Romance, Drama	7

6 rows returned *(7.272 millisec)*

2. **Show the 10 most watched shows of all time.**

```
select show_id, title, count(show_id) as show_count

from session_fact

join show_dim using(show_id)

group by show_id, title

order by show_count desc

fetch first 10 rows only;
```

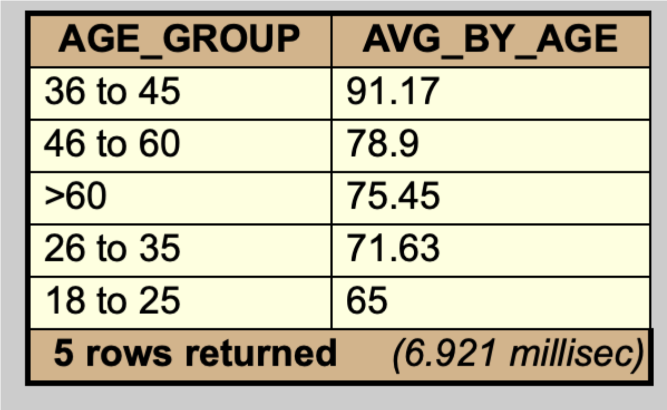


SHOW_ID	TITLE	SHOW_COUNT
1	Whats My Line?	10
43	Take This Waltz	8
10	Goblin	8
23	Bad Reputation	7
28	Tangerines	6
33	Nani s Gang Leader	6
11	Nursery Rhymes for Kids - Little Baby Bum	6
45	The Nightstalker	6
7	Bed Of Roses	6
48	A Good Year	6

10 rows returned *(5.159 millisec)*

3. **Display the average payment amount for each age group.**

```
select age_group, round(avg(price-discount+tax+payment_method_fees), 2)
avg_by_age

from user_dim

join payment_fact using (user_id)

group by age_group

order by avg_by_age desc
```



AGE_GROUP	AVG_BY_AGE
36 to 45	91.17
46 to 60	78.9
>60	75.45
26 to 35	71.63
18 to 25	65

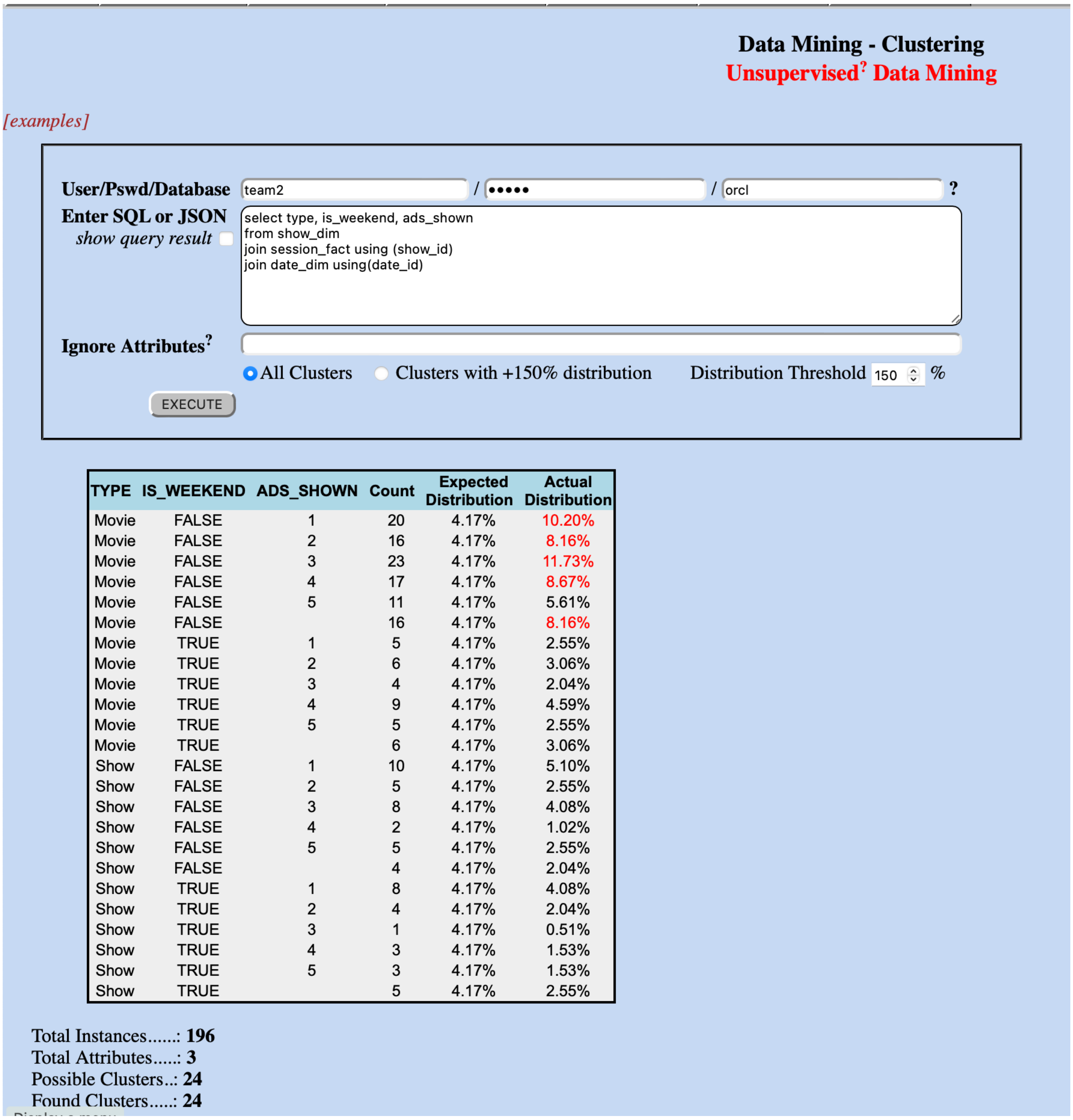
5 rows returned *(6.921 millisec)*

Data mining analysis

Clustering - Ads Selling

Purpose: To advise clients on advertising strategies based on show type, genre, and timing, and to set advertising rates according to viewership metrics.

For this purpose, we applied clustering to do the test. And the result is shown as below:



Data Mining - Clustering
Unsupervised? Data Mining

[examples]

User/Pswd/Database: team2 / ••••• / orcl ?

Enter SQL or JSON:

```
select type, is_weekend, ads_shown
from show_dim
join session_fact using (show_id)
join date_dim using(date_id)
```

show query result □

Ignore Attributes?:

● All Clusters ○ Clusters with +150% distribution Distribution Threshold 150 %

TYPE	IS_WEEKEND	ADS_SHOWN	Count	Expected Distribution	Actual Distribution
Movie	FALSE	1	20	4.17%	10.20%
Movie	FALSE	2	16	4.17%	8.16%
Movie	FALSE	3	23	4.17%	11.73%
Movie	FALSE	4	17	4.17%	8.67%
Movie	FALSE	5	11	4.17%	5.61%
Movie	FALSE		16	4.17%	8.16%
Movie	TRUE	1	5	4.17%	2.55%
Movie	TRUE	2	6	4.17%	3.06%
Movie	TRUE	3	4	4.17%	2.04%
Movie	TRUE	4	9	4.17%	4.59%
Movie	TRUE	5	5	4.17%	2.55%
Movie	TRUE		6	4.17%	3.06%
Show	FALSE	1	10	4.17%	5.10%
Show	FALSE	2	5	4.17%	2.55%
Show	FALSE	3	8	4.17%	4.08%
Show	FALSE	4	2	4.17%	1.02%
Show	FALSE	5	5	4.17%	2.55%
Show	FALSE		4	4.17%	2.04%
Show	TRUE	1	8	4.17%	4.08%
Show	TRUE	2	4	4.17%	2.04%
Show	TRUE	3	1	4.17%	0.51%
Show	TRUE	4	3	4.17%	1.53%
Show	TRUE	5	3	4.17%	1.53%
Show	TRUE		5	4.17%	2.55%

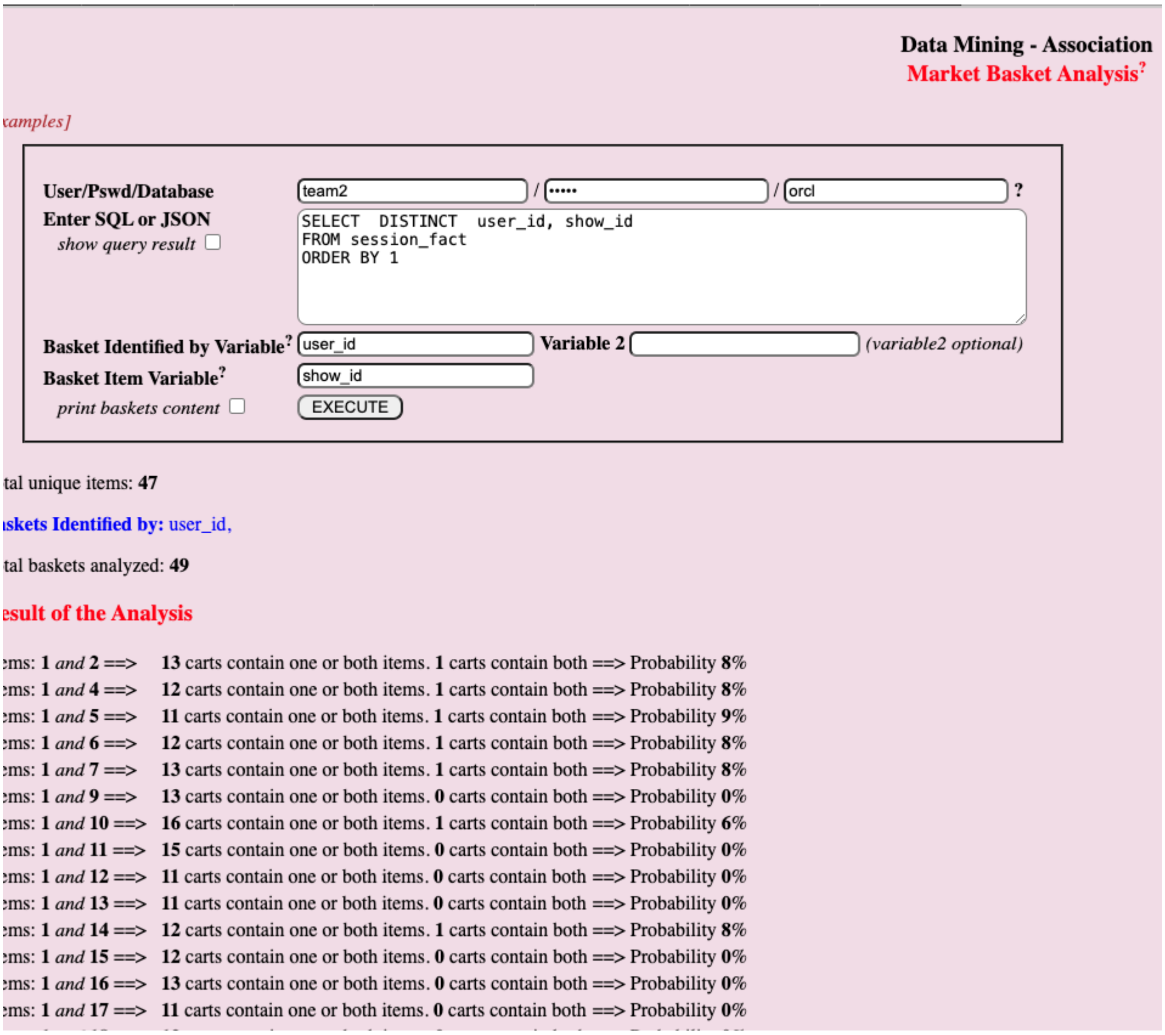
Total Instances......: **196**
Total Attributes.....: **3**
Possible Clusters..: **24**
Found Clusters.....: **24**

Movies show a higher engagement with ads on weekdays, particularly noticeable with 1 to 3 ads shown, exceeding expected distributions. Conversely, during weekends, there's a noticeable dip in ad engagement for movies. Shows display more consistent ad viewing patterns regardless of the day, with ad counts generally remaining low.

Movies have higher ad engagement on weekdays, making it advantageous to place premium ad slots during these times to maximize viewership and revenue. Therefore, ads shown during weekday movies should be priced the highest. On weekends, reducing the number of ads during movies can help keep viewers more engaged and happy. Shows display consistent viewing patterns, indicating that a uniform ad strategy is effective. Consequently, ads shown during shows on weekends should be priced the lowest.

Association - Market Basket Analysis

Purpose: Determine the shows that tend to be watched together to provide show recommendations to users.



Data Mining - Association
Market Basket Analysis?

xamples]

User/Pswd/Database: team2 / ••••• / orcl ?

Enter SQL or JSON

show query result □

```
SELECT  DISTINCT  user_id, show_id
FROM session_fact
ORDER BY 1
```

Basket Identified by Variable? user_id Variable 2 *(variable2 optional)*

Basket Item Variable? show_id

print baskets content □

tal unique items: **47**

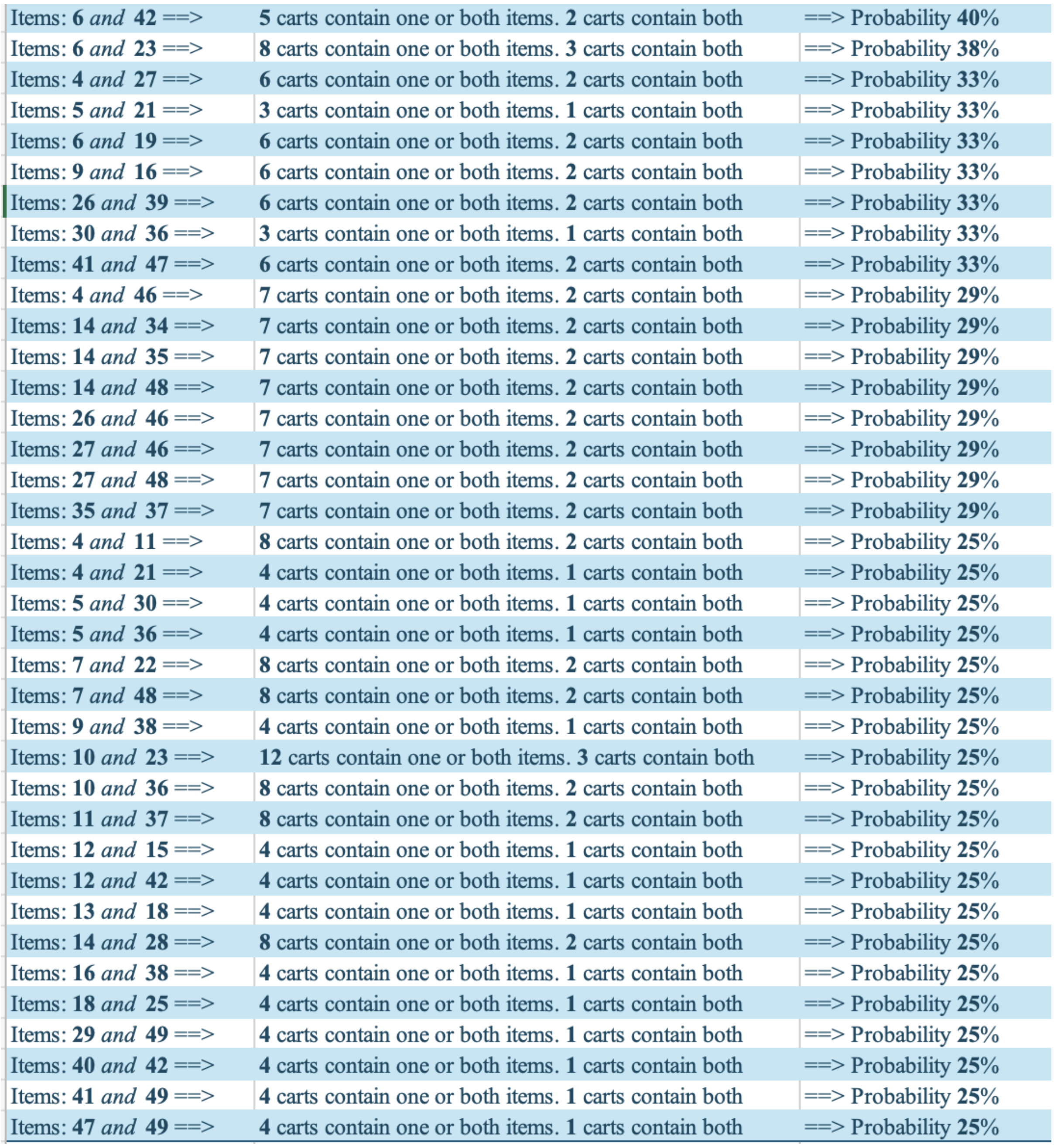
iskets Identified by: user_id,

tal baskets analyzed: **49**

esult of the Analysis

```
ems: 1 and 2 ==>    13 carts contain one or both items. 1 carts contain both ==> Probability 8%
ems: 1 and 4 ==>    12 carts contain one or both items. 1 carts contain both ==> Probability 8%
ems: 1 and 5 ==>    11 carts contain one or both items. 1 carts contain both ==> Probability 9%
ems: 1 and 6 ==>    12 carts contain one or both items. 1 carts contain both ==> Probability 8%
ems: 1 and 7 ==>    13 carts contain one or both items. 1 carts contain both ==> Probability 8%
ems: 1 and 9 ==>    13 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 10 ==>   16 carts contain one or both items. 1 carts contain both ==> Probability 6%
ems: 1 and 11 ==>   15 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 12 ==>   11 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 13 ==>   11 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 14 ==>   12 carts contain one or both items. 1 carts contain both ==> Probability 8%
ems: 1 and 15 ==>   12 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 16 ==>   13 carts contain one or both items. 0 carts contain both ==> Probability 0%
ems: 1 and 17 ==>   11 carts contain one or both items. 0 carts contain both ==> Probability 0%
```

The number of combinations analyzed here is 1081. However, many of them have 0% or very low probability. After eliminating the ones with low probability, I kept a table showing baskets with a probability of 25% or higher.



Items: 6 *and* 42 ==>	5 carts contain one or both items. 2 carts contain both	==> Probability 40%
Items: 6 *and* 23 ==>	8 carts contain one or both items. 3 carts contain both	==> Probability 38%
Items: 4 *and* 27 ==>	6 carts contain one or both items. 2 carts contain both	==> Probability 33%
Items: 5 *and* 21 ==>	3 carts contain one or both items. 1 carts contain both	==> Probability 33%
Items: 6 *and* 19 ==>	6 carts contain one or both items. 2 carts contain both	==> Probability 33%
Items: 9 *and* 16 ==>	6 carts contain one or both items. 2 carts contain both	==> Probability 33%
Items: 26 *and* 39 ==>	6 carts contain one or both items. 2 carts contain both	==> Probability 33%
Items: 30 *and* 36 ==>	3 carts contain one or both items. 1 carts contain both	==> Probability 33%
Items: 41 *and* 47 ==>	6 carts contain one or both items. 2 carts contain both	==> Probability 33%
Items: 4 *and* 46 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 14 *and* 34 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 14 *and* 35 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 14 *and* 48 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 26 *and* 46 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 27 *and* 46 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 27 *and* 48 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 35 *and* 37 ==>	7 carts contain one or both items. 2 carts contain both	==> Probability 29%
Items: 4 *and* 11 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 4 *and* 21 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 5 *and* 30 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 5 *and* 36 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 7 *and* 22 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 7 *and* 48 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 9 *and* 38 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 10 *and* 23 ==>	12 carts contain one or both items. 3 carts contain both	==> Probability 25%
Items: 10 *and* 36 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 11 *and* 37 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 12 *and* 15 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 12 *and* 42 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 13 *and* 18 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 14 *and* 28 ==>	8 carts contain one or both items. 2 carts contain both	==> Probability 25%
Items: 16 *and* 38 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 18 *and* 25 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 29 *and* 49 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 40 *and* 42 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 41 *and* 49 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%
Items: 47 *and* 49 ==>	4 carts contain one or both items. 1 carts contain both	==> Probability 25%

These refined results provide Amazon with valuable insights for enhancing show recommendations to users, potentially leading them to discover content aligned with their interests. This tailored approach holds the promise of boosting customer satisfaction as it increases the likelihood of users finding shows they genuinely enjoy.