DATA MINNING AND BI DATA WAREHOUSING BASED IMPLEMENTATION FOR A RANDOM FILM STUDIO.

A Master's Project

Presented to

Department of Computer and Information Sciences

SUNY Polytechnic Institute

Utica, New York

In Partial Fulfilment of the requirements for the Master of Science Degree

By

Sneha Bonthi.

(U00281973)

December 2016

SUNY POLYTECHNIC INSTITUTE
DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES

Approved and recommended for acceptance as a project in partial fulfilment of the requirements for the degree of Master of Science in Computer and Information Sciences

January 26, 2016
Date

Bruno Andriamanalimanana
Dr. Bruno Andriamanalimanana, Ph. D. (Adviser)

Dr. Michael. J Reale, Ph. D.

M. Rezk
Dr. Mohamed Rezk, ~~Ph. D.~~ D. Eng

DECLARATION

I declare that this project is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

B. Sha

Sneha Bonthi.

ABSTRACT

The purpose of this report is to study a dataset of movies and analyse the possibility and feasibility of implementing a data warehousing or a data mining application to improve analytics and decision making.

The project report talks about the raw data originating from the data collection centres and box offices which can be modelled and transformed into a specific format and structure that would help the business analysts in identifying patterns and trends so as to take important business decisions. The report explores the benefits of extracting, transforming and loading this raw data into a dimensional model. According to the proposed implementation, one can create a reporting layer to perform aggregations and grouping them by various attributes like date, genre, actor and country and present them using dashboards and reports to enable better decision making.

This single point of data, which is the result of data mining activity, can be shared and brainstorming sessions can then be carried out to infer priceless market information and effectively utilize time and efforts to maximize profits.

TABLE OF CONTENTS

1. INTRODUCTION

Fox, Sony, Disney, Warner Bros are some typical studios in the entertainment business. The entertainment business is a strange place. Critically panned movies like transformers and twilight have been huge money-makers, while Oscar-winning movies generally do not fare as well at the box office. Coupled with the huge expenses involved in making and releasing films, the need to schedule and track releases and collections accurately became an urgent business imperative for any serious media and entertainment company. An exponential growth of online media users is challenging established business models for media and information services, companies and creating new revenue opportunities. The growth in the film industry resulted in the company's theatrical distribution systems becoming disparate, unconnected and dysfunctional. These companies recognized the importance to integrate its multiple theatrical distribution systems which were utilized in the different business operations within the theatrical distribution and hence provide accurate and timely business critical information on movie performance. [1]

One of these studios recently heard about the power of bi with data mining and is interested to implement business intelligence (BI), together with data mining based solution to handle huge sets of multi excel based data which is currently being handled manually, which incurs a lot of manual efforts.

1.1 Objective:

The objective here is to study the data and apply data mining techniques and based on data mining and analysis, designing an effective dimensional model which can provide help to business to fulfill their needs.

And also to know how business intelligence (BI), together with data warehousing and data mining can be utilized to help improve the operation and revenue of a studio with a single application, providing the company with much needed data available to perform movie release, planning, box-office reporting, contract management, and financial management and print control management.

2. BACKGROUND

Currently data management in a particular studio is completely based on Excel where multiple teams assigned with the responsibilities of collecting data and providing data to another team, who consolidate the data from different sources in excel and then perform analysis.

A data scientist is provided with BI resources with consolidated excels for multiple entities where the raw data is converted into meaningful data and then to propose a robust and effective ETL based solution to load excel based data into respective fact and dimension tables, fact and dimension tables will be the outcome of dimensional model design which also has to be completed before finalizing the ETL design.

2.1 Qlikview

Qlikview is a reporting tool developed by Qliktech to match the industry's reporting and analytical needs. Qlikview works on an in-memory principle i.e. every single record of the tables used is loaded onto the dashboard and the dashboard itself is loaded onto memory when opened. So each time, the user makes any selection or when he/she wishes to filter on a particular field, no query needs to be fired to the data base and all the charts seamlessly filter themselves based on the selection made. Qlikview integrates back-end scripting, data-modelling techniques and a wide array of front-end charts which are customizable in so many different ways that together it is argued to be one of the most widely used tools for building dashboards. [8]

2.2 Advantages of Qlikview over other reporting tools

- Qlikview is easy to access since it is an open source and doesn’t need any complex database connectivity.
- It is social, hence it is easy to collaborate with peers or team mates in an organization.
- It is mobile, hence it can generate business related decisions as data visualization quickly when compared to other tools.
- It is best for a small scale or a start-up organizations to use as a business intelligence tool which has a small amount of data base.[9]

2.3 Data sets used in the project

Raw data for below entities is collected from various sources, and is arranged as different data sets.

- Actor
- Franchise
- Genre
- Oscar
- Studio
- Top movies by country
- Top movies by year

Each dataset provides a range of parameters that can be used for many different purposes. Like, analyzing data thoroughly before designing and finalizing dimensional model. For example, concentrating on issues related to business competitiveness, revenue and margin growth for the studio. [2] [3] [4] [5]

2.4 Overview of the Project

So the steps involved to implement the project are:

- Thorough study of data for each entity and mine the data to be used for next phase.
- Using the mined data to come up with number of dimension and fact tables to be used further.
- Attributes/column for each of the dimension tables and fact table.
- Creating a dimensional model.
- Implementing the dimensional model in any of the database.
- Using ETL concept to load data.
- Design a visualization solution to show different patterns of dashboard which can be used by the studio in their important analysis.

3. DATA MINING

Archaically, mining refers to drilling down from the surface into the depths of the earth to extract and utilize the resource available. To be able to mine, we must have two per-requisites fulfilled.

1. We must have some raw material e.g. a fuel bed.
2. We must have the technology required to mine.

Similarly, data mining is the process of observing, analyzing and gathering information from raw data that is generated out of day-to-day or transactional processes. Data mining involves identifying and interpreting patterns and anomalies in the given data set so as to form meaningful conclusions on the data. [6]

As a part of a data mining tool, a developer has to design a method to prepare the existing raw data into a form of information that can be used further for analysis. A large part of data mining involves interpreting the data in such a way that it enables the processes downstream to:

- Identify patterns in the data.
- Identify relations between two or more fields in the data i.e. the association with each other.
- Group the data into sets and subsets which are logically related.

For the movies' data set provided, we have to develop a technology to extract, model and design an interpretation mechanism to identify patterns like those mentioned below:

1. Some movies are best-sellers due to the cast involved.
2. Some movies rope in high profits due to the period of the release.
3. Some genres and/or franchises generate more profit during festive seasons.
4. Certain countries prefer certain actors/actresses over other.

These are just a sample of the insight that can be gained by designing an effective tool and process to perform data mining. To achieve this, we have to:

- Use data modeling techniques to design and implement a dimensional model from the OLTP tables provided.
- Design a loading mechanism to generate the required structure and format of the data to populate the fact and dimension tables designed in step 1.
- Design and implement a front-end dashboard and/or reports to effectively highlight the patterns and associations observed in the data and present it to users.

4. DATABASE DIMENSION MODELLING AND IMPLEMENTATION

Each data point in a given data set can be split into two basic entities, namely attributes and measures. Attributes are those data points that answer to the question “what?” And measures are those that answer to the question “how much?” Attributes are also called dimensions while measures are called facts. Facts are the reporting unit of any data warehouse, i.e. those are the ones that are manipulated, measured and reported for analysis. Dimensions are those against which these facts are reported. So it can be said that dimensions give extra information regarding the facts. They also give us the ability to drill down into the measured entities.

The technique of restructuring the data set and grouping the data points into logical tables on the basis of facts and dimensions is called dimension modeling. The fact table consists of only keys from the dimension tables and the other measurable entities. The dimension tables each contain a primary key to identify and link to the fact table followed by the details or attributes from the raw data set. [6]

A dimensional model should be designed with the reporting layer in mind, i.e. facts and dimensions should be segregated so as to suit reporting and drill down purposes. There are two major types of dimensional models:

1. **Star schema:** A star schema consists of a central fact table which is liked directly to each of the dimension tables via a 1-1 or 1-many relationships.
2. **Snowflake schema:** A snow flake schema consists of some dimension tables that aren't linked directly to the fact table, but are linked indirectly via another dimension table, which in turn is linked directly to the fact table. These sorts of dimensions are usually known as details of the dimension that they are linked with.

The movies' data set can be modelled as the below star schema. [7]

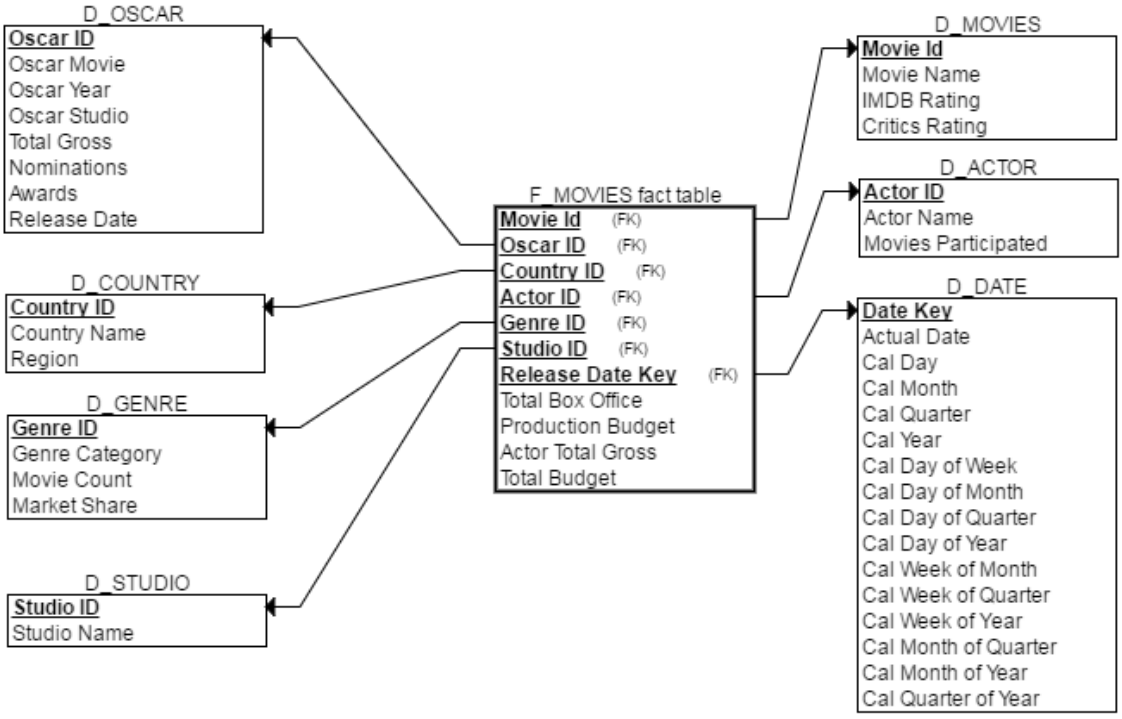


Fig 4.1 Star schema data model for given data set

The dimensional model consists of below tables:

1. D_movies: This is the dimension table that contains all the attributes of the movies like IMDB rating and critic rating along with the name. Movie ID is the primary key.
2. D_actor: This contains the names, IDs and the count of movies starred in for each actor/actress.
3. D_oscar: This is a collection of all the movies that have been nominated for and/or won an Oscar along with the nominations and wins for that particular movie along with other details.
4. D_country: This table contains the IDs and names of each country along with its region. This is a fairly static table and is rarely required to be refreshed.
5. D_genre: This table contains the genre IDs along with the names. Other details include market share and the movie count.

6. D_studio: This is a fairly static list of studios releasing the movies and is only updated in case of a new studio stepping into the market.
7. D_date: This is simply the calendar dimension for this data model with the respective system generated keys for each day along with other details.
8. F_movies (fact): This is the central fact table in the star schema. This table contains all the respective IDs of the dimensions and the reporting measures like production budget and total box-office collections.

The above data model has been designed after analyzing the data in the OLTP tables with a goal in mind to reduce complexity in the reporting layer.

5. ETL (Extraction, transformation and loading).

Once the dimensional model is created and the fact and dimension tables are designed, there is a need to design a method to extract the data from the OLTP files and insert/update into the dimension and fact tables. This concept of extract, transform and load is known as ETL and is the first step involved in a data mining or data warehousing application.

ETL is required in these particular scenarios since the data originating from the collection centres and the box offices has below anomalies.

- Some of the source files are from collection centres, while some are from box offices i.e. the sources of data are different.
- Some are extracts from several websites. This will give rise to a lot of cleansing activity that would be required.
- Besides that, this is a global implementation, i.e. there is data coming in from various countries, which means the same file can have a different format, the delimiters for numbers can be different (like decimals are represented by "," instead of "." in some European countries).

To join and relate these varied data sets effectively, it is imperative that we standardize the data first. Which is why using ETL becomes a must-have for this implementation.

These ETL workflows would have one or more of the OLTP data files as a source and the dimension and fact tables as the target. Also generated in the process are the fields that are derived or aggregated in some way from the given data set. Any cleansing or data quality checks that need to be performed are also carried out in this process flow.

Given below is the ETL flow for the movies' data mart.

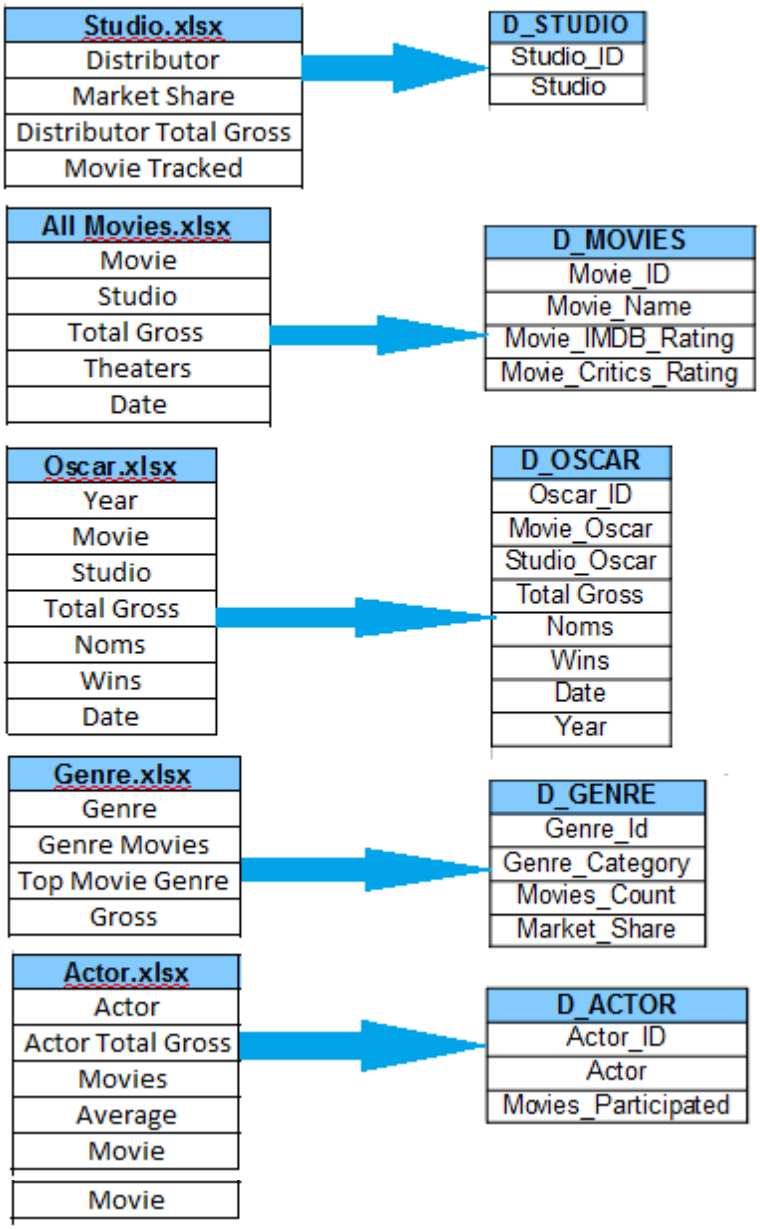


Fig 5.1 Dimension tables' data flow

Date and country dim are loaded from a static one-time file. Additionally, all IDs and keys are a system generated using a sequence generator.

The fact data flow is a relatively more complex as it consists of the base all movies file and several dimension tables. These dimension tables are referred to pass the respective attributes and return the corresponding attribute IDs. These attribute IDs are then processed onto the fact

table. This process is called a look-up in ETL terms and below flow illustrates the same. Please note that those in green are look-up tables.

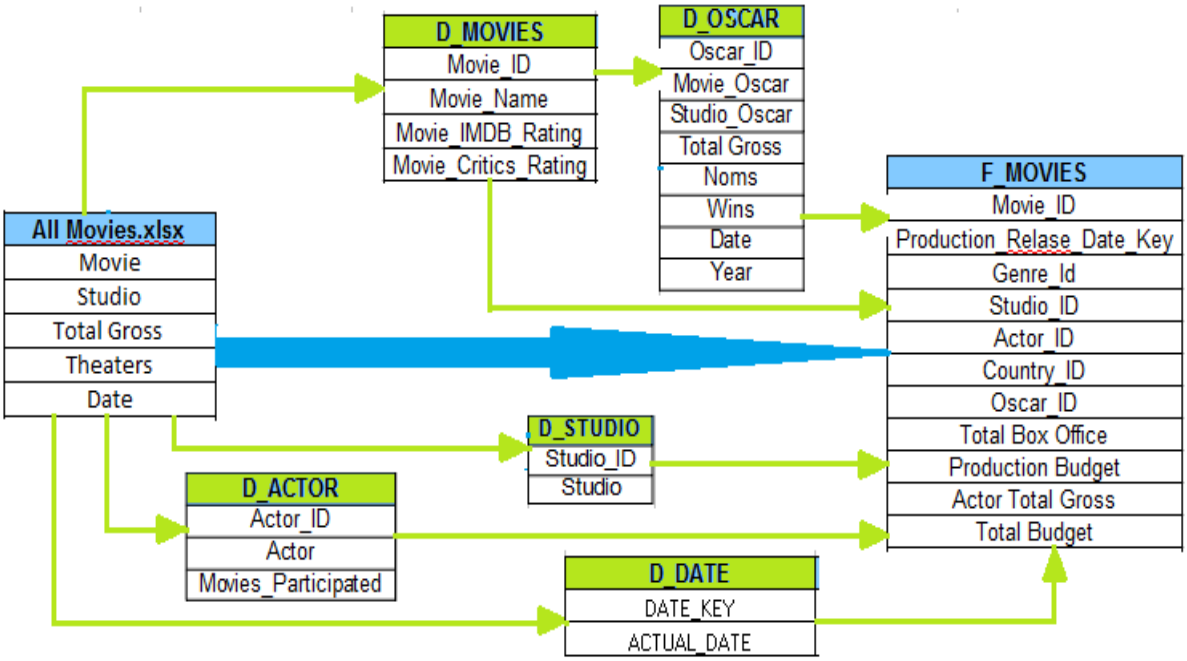


Fig 5.2 Fact table data flow

Care has to be taken while loading the fact table that all the dimension tables have been successfully loaded. If need be, and incremental logic can be implemented to only load the fresh data and not the historical one. This is particularly required in the entertainment business when the transactional data volume is too large.

6. DASHBOARD DESIGN, VISUALIZATION AND IMPLEMENTATION

Once the data is in place in the dimensional model that has been designed, the next and final step of data mining is to design and implement a reporting layer to be able to analyse and drill down into the data that is collected. The dashboard design is split into two categories:

6.1 Data modelling in Qlikview

In most cases, the existing data model is pulled as is onto the Qlikview dashboard since that seems to suffice. In some other cases, an additional data model is designed on top of the database model so as to simplify and optimize reporting. Either way, scripts have to be written to load the data from the database into the dashboard. These scripts also offer the ability to change column names, create reporting columns, cleanse the data and create grouped dimensions for reporting.

Below is the data-model created on Qlikview for the movies' data.

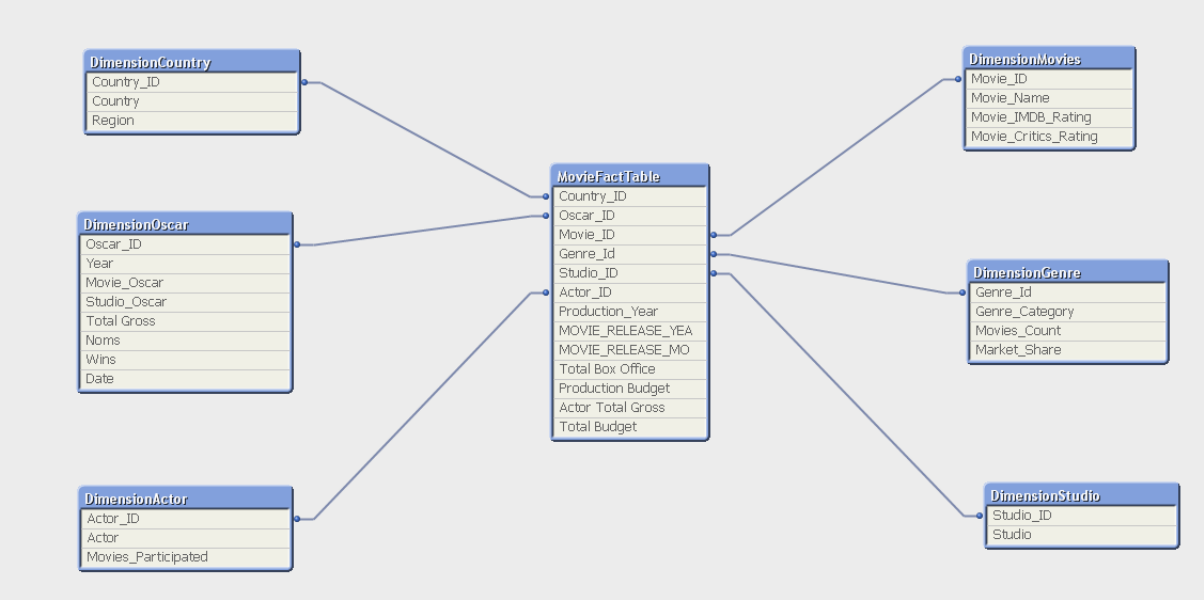


Fig 6.1 Data model in Qlikview

6.2 Data Visualization in Qlikview

Once the data is loaded onto the dashboard and modelled as per reporting requirements the most crucial stage is the data presentation on the front-end. On the Qlikview dashboard, charts

can be created after aggregating certain measures and display them in a variety of ways with various dimensions.

Front-end filters can be added for year, quarter, country, studio and genre, etc. To make the data responsive to customer inputs.

Below are some sample charts that have been created for the movies' data.

6.2.1 Top movies by a country:

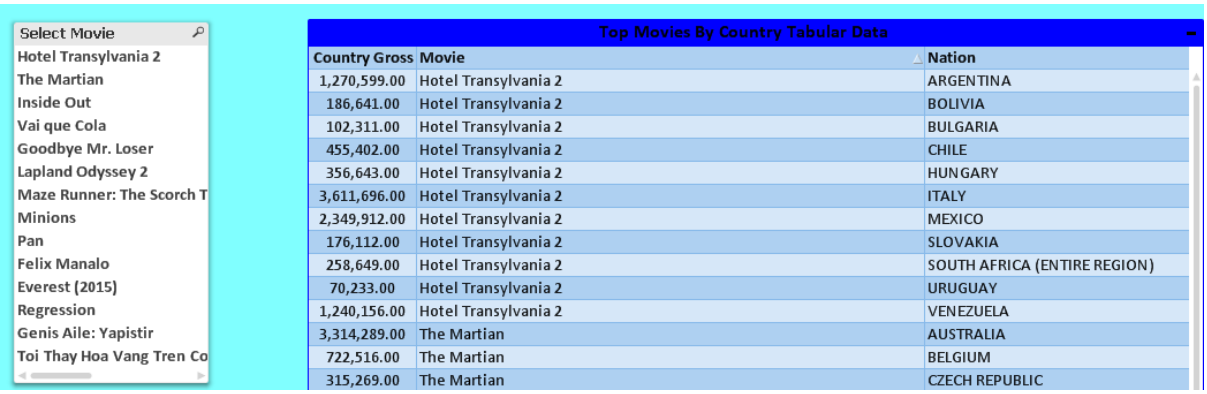
The below chart is a simple tabular representation of movies per country. These are ranked top to bottom by the gross earnings out of box-office for that particular movie in that particular country.

Chart information: this is a straight table, i.e. a simple tabular representation where the total gross amount is summed up and grouped by country and movie.

Type: A straight table.

KPI: The total gross.

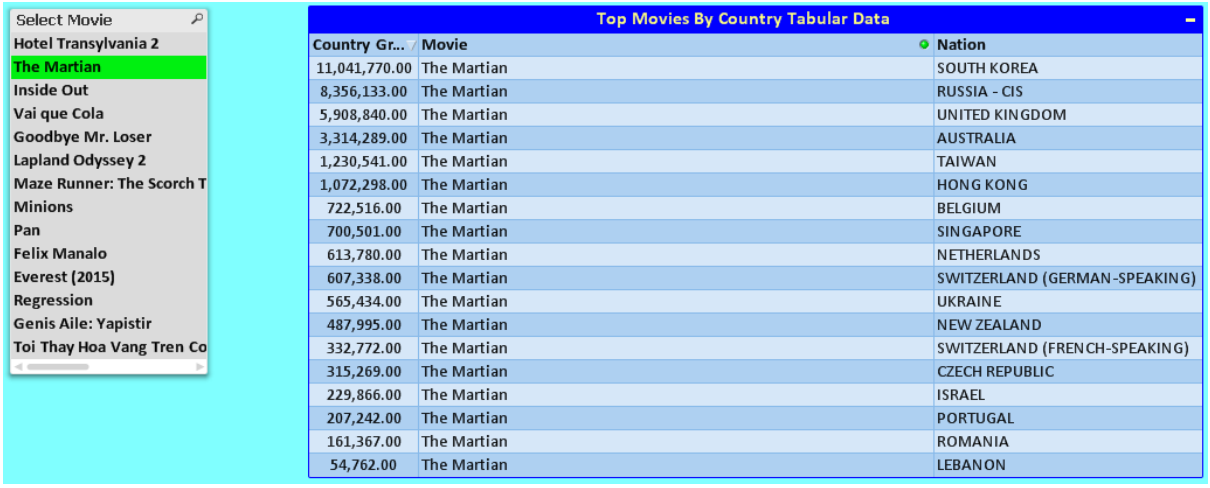
Dimension: Movie, country.



Country Gross	Movie	Nation
1,270,599.00	Hotel Transylvania 2	ARGENTINA
186,641.00	Hotel Transylvania 2	BOLIVIA
102,311.00	Hotel Transylvania 2	BULGARIA
455,402.00	Hotel Transylvania 2	CHILE
356,643.00	Hotel Transylvania 2	HUNGARY
3,611,696.00	Hotel Transylvania 2	ITALY
2,349,912.00	Hotel Transylvania 2	MEXICO
176,112.00	Hotel Transylvania 2	SLOVAKIA
258,649.00	Hotel Transylvania 2	SOUTH AFRICA (ENTIRE REGION)
70,233.00	Hotel Transylvania 2	URUGUAY
1,240,156.00	Hotel Transylvania 2	VENEZUELA
3,314,289.00	The Martian	AUSTRALIA
722,516.00	The Martian	BELGIUM
315,269.00	The Martian	CZECH REPUBLIC

Fig 6.2.1(a) Top movies by country

The filter on the left allows the business analysts or other users to select a particular movie if required. The below chart depicts that The Martian did very well in South Korea followed by Russia. Since the Martian is a sci-fi movie, it can be predicted that sci-fi movies do very well in the Eurasian continent as compared to the other regions.



Country Gr...	Movie	Nation
11,041,770.00	The Martian	SOUTH KOREA
8,356,133.00	The Martian	RUSSIA - CIS
5,908,840.00	The Martian	UNITED KINGDOM
3,314,289.00	The Martian	AUSTRALIA
1,230,541.00	The Martian	TAIWAN
1,072,298.00	The Martian	HONG KONG
722,516.00	The Martian	BELGIUM
700,501.00	The Martian	SINGAPORE
613,780.00	The Martian	NETHERLANDS
607,338.00	The Martian	SWITZERLAND (GERMAN-SPEAKING)
565,434.00	The Martian	UKRAINE
487,995.00	The Martian	NEW ZEALAND
332,772.00	The Martian	SWITZERLAND (FRENCH-SPEAKING)
315,269.00	The Martian	CZECH REPUBLIC
229,866.00	The Martian	ISRAEL
207,242.00	The Martian	PORTUGAL
161,367.00	The Martian	ROMANIA
54,762.00	The Martian	LEBANON

Fig 6.2.1(b) Top movies by a country with a filter.

6.2.2 Top grossing movies all time:

The below chart depicts the top 10 grossing movies of all time in all countries.

Chart information: This is a bar chart displaying the total gross amounts for the top 10 movies. The dimension country is conditionally set to be enabled if a movie is selected on the left. Dimension limits are used to limit the representation to only the top 10 values.

Type: A bar chart.

KPI: The total gross.

Dimension: Movie, country (conditional).

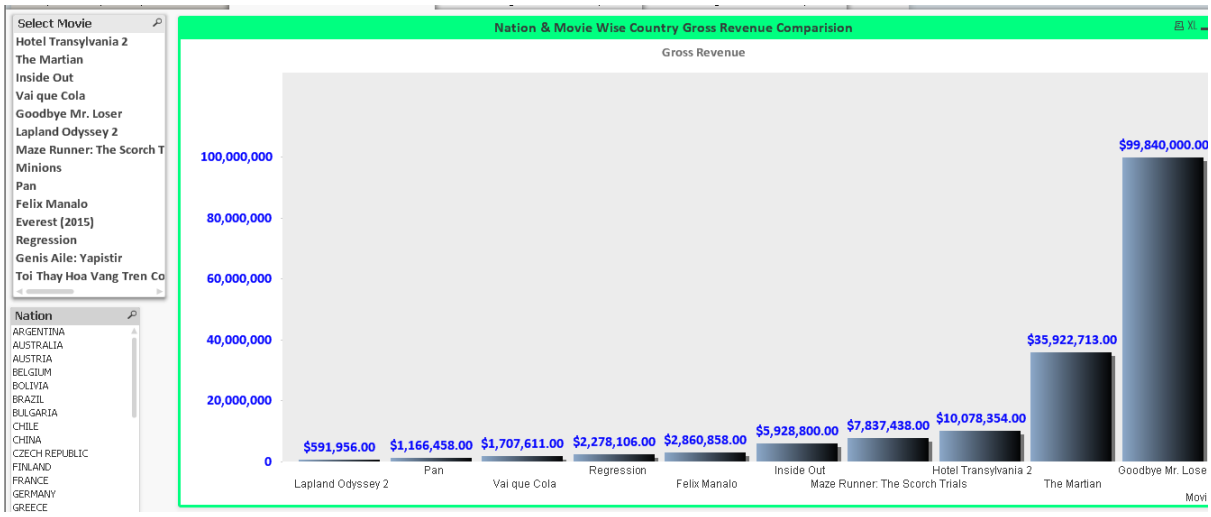


Fig 6.2.2(a) Nation and movie wise comparison of gross revenue.

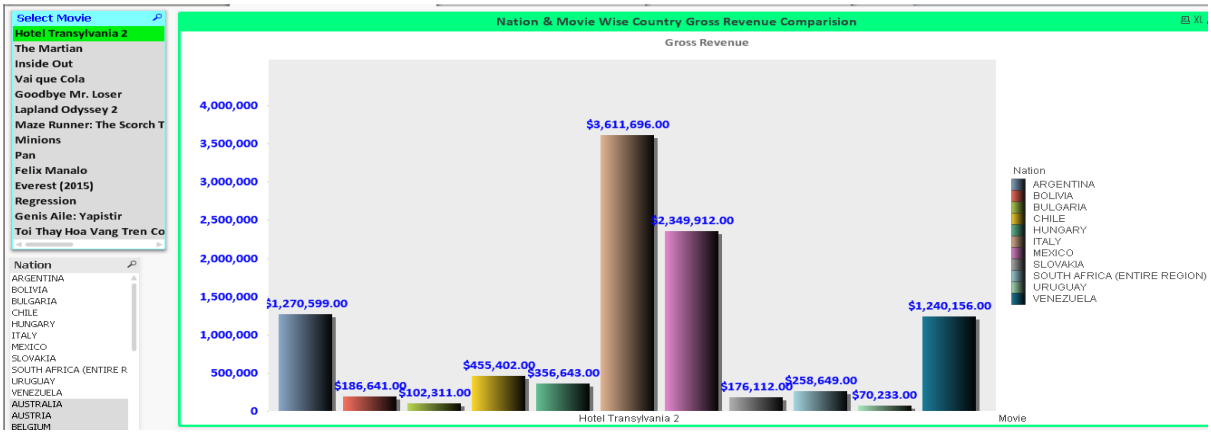


Fig 6.2.2(b) Nation and movie wise comparison of gross revenue with a filter.

6.2.3 Bar graph representation of Production budget v/s total box-office collections per country.

The below chart is a bar graph comparison of the production budget and the total box-office earnings collected in each country.

Chart information: This is a bar chart built for comparing the production budget v/s total box-office collections across countries. Dimension limits are used to limit the chart values to top 10 production budgets.

Type: A bar chart.

KPI: The total gross, production budget.

Dimension: Country.

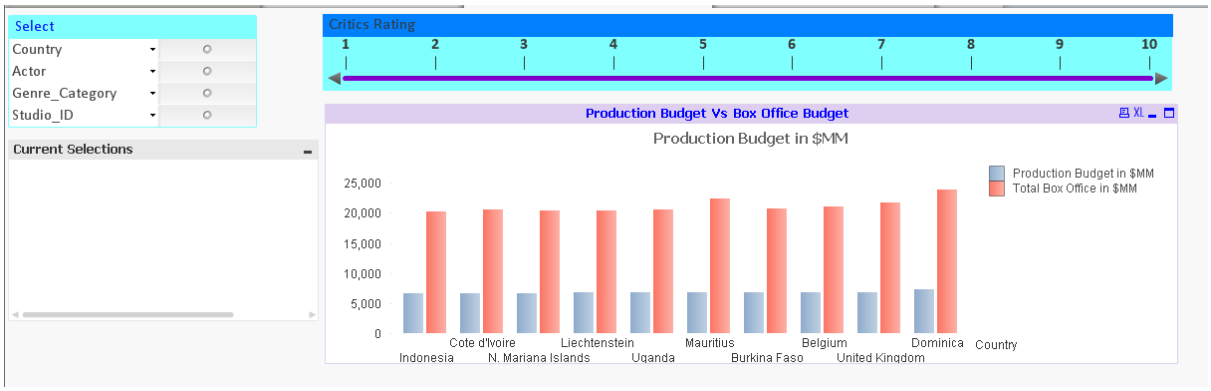


Fig 6.2.3(a) Production budget Vs Box-office budget

There is also the additional option of selecting either a country/actor/genre or studio.

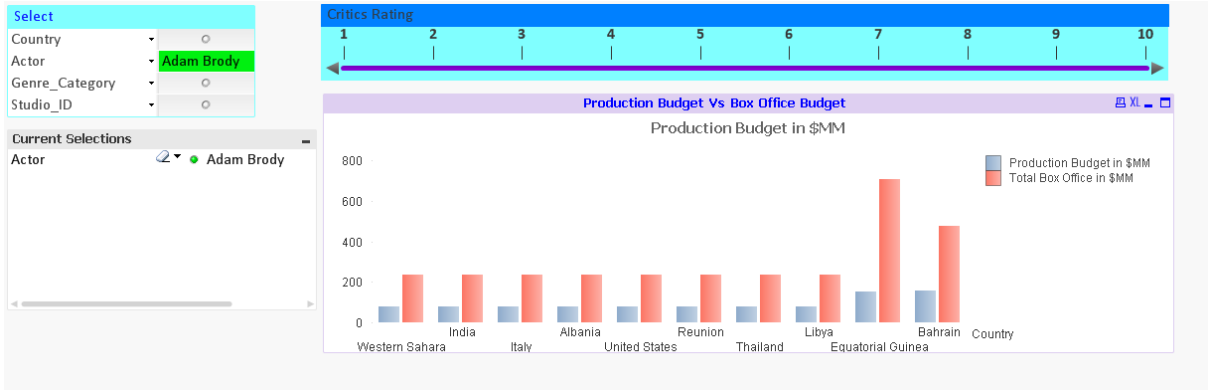


Fig 6.2.3(b) Production budget Vs Box-office budget with an actor as filter.

6.2.4 Bar graph with multiple filter selection.

Chart information: A multi-box is used to enable 2 or more fields for user selection

Type: A multi box.

KPI: NA.

Dimension: Actor, country, genre and studio.

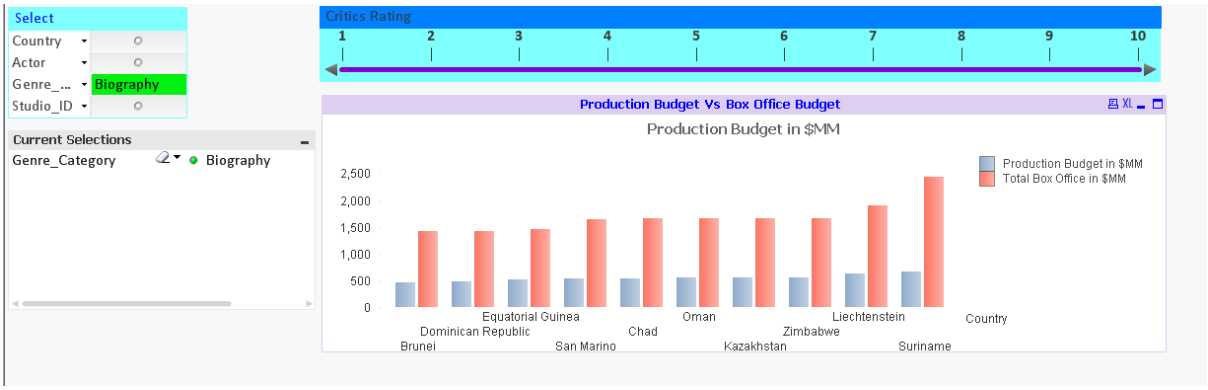


Fig 6.2.4 Production budget Vs Box-office budget with genre as filter.

6.2.5 Alert messages:

Text messages can be set-up to prompt the users to select specific fields e.g. Country to generate different charts. The below message prompts to select a country to generate charts to see genre and movie-wise total box-office collection.

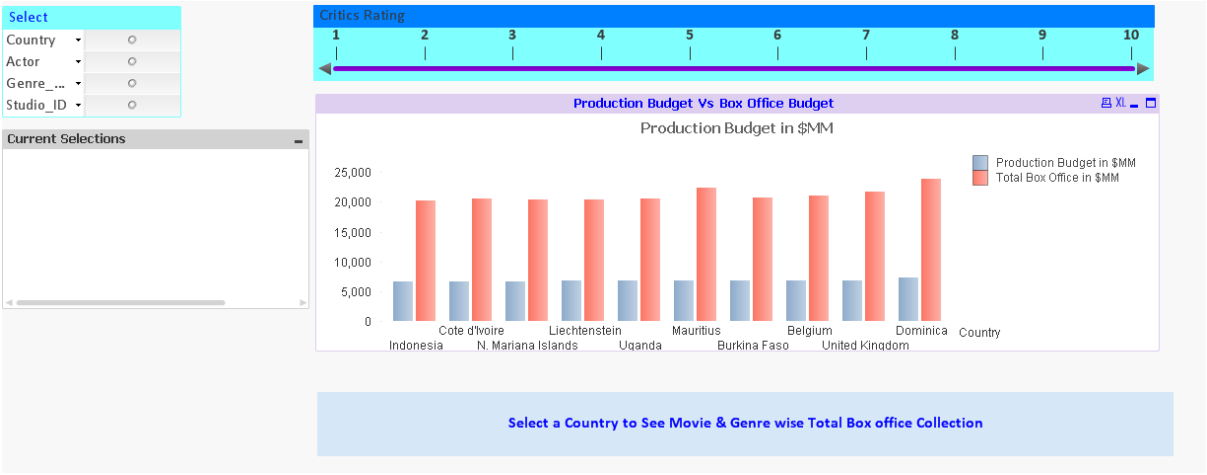


Fig 6.2.5 Alert message when a filter is not selected

Chart information: Alert messages are text objects with the desired message as the only attribute. These are conditionally hidden or displayed based on a certain condition viz. When a country is not selected in the multi-box.

Type: A text object.

6.2.6 Pie chart representation.

Upon the selection of a country two pie charts are triggered to appear at the bottom as visible in the next screen shot.

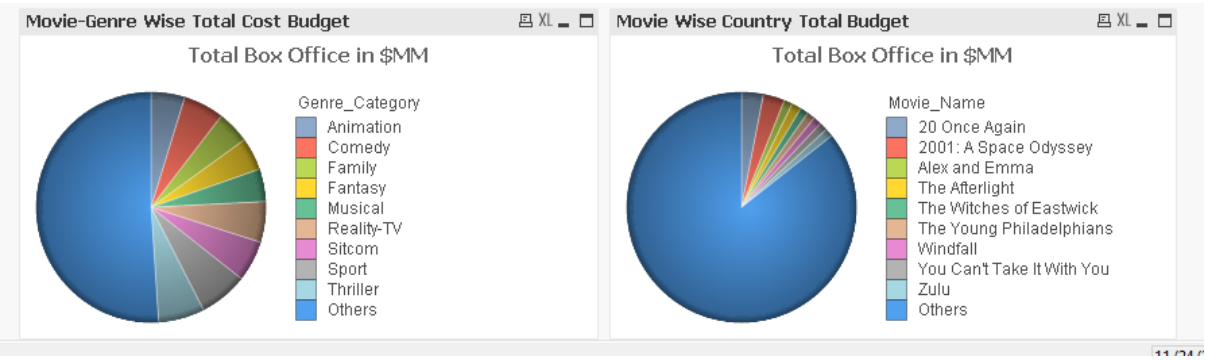


Fig 6.2.6 Pie charts each representing different attribute filters.

Chart information: These are pie charts set to hidden unless a particular variable is set. This variable is set only when a country is selected in the multibox.

Type: A pie chart.

KPI: The total gross in mm i.e. sum ([total box office])/1,000,000.

Dimension: Movie.

Chart information: these are pie charts set to hidden unless a particular variable is set. This variable is set only when a country is selected in the multibox.

Type: A pie chart.

KPI: The total gross in mm i.e. sum ([total box office])/1,000,000.

Dimension: Genre.

6.2.7 Critic rating wise comparison

A slider object can be provided to navigate between critic ratings for movies and display the production budget v/s box-office earnings for each country for movies of that particular critic rating.

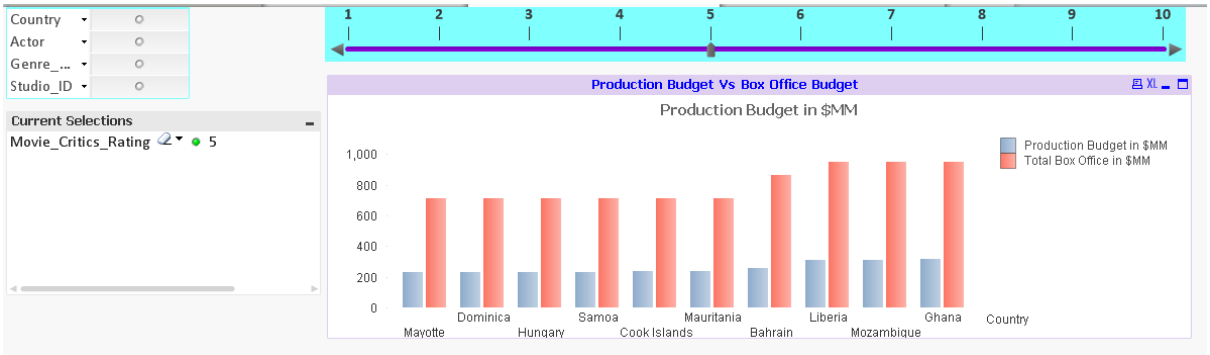


Fig 6.2.7 Critics' rating wise comparison.

Chart information: This is a slider object built using a single dimension, i.e. the critics' rating. The object allows the user to either select a range of values or a particular value for the field

Type: A slider.

KPI: NA.

Dimension: Critics rating.

6.2.8 IMDB rating wise comparison

The same can be implemented for the IMDB rating as well to understand if highly rated movies are bringing in the revenue expected. This is clear in the below screen shot.

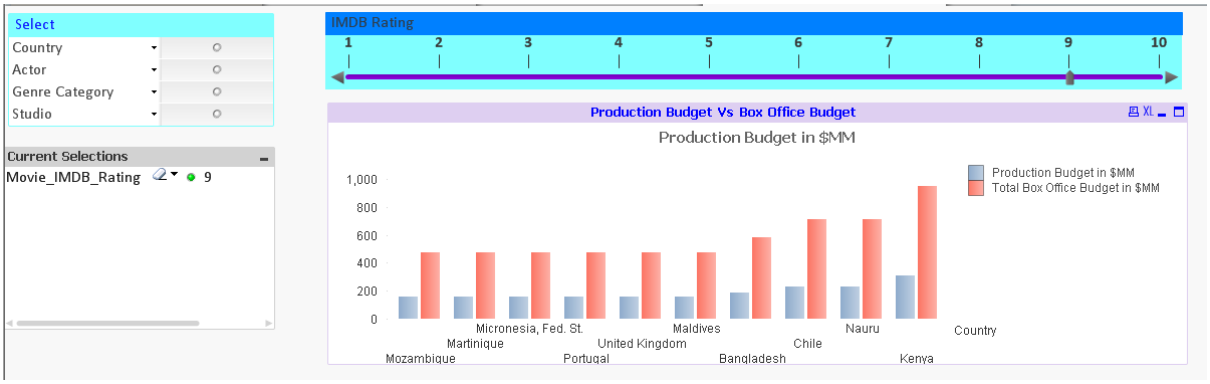


Fig 6.2.8 IMDB rating wise comparison.

Chart information: This is a slider object built using a single dimension, i.e. IMDB rating. The object allows the user to either select a range of values or a particular value for the field.

Type: A slider.

KPI: NA.

Dimension: IMDB rating.

6.2.9 Horizontal Bar representation.

Reports can also be added to generate various representations of count of movies and total box-office earnings for a particular year and/or month.

Chart information: this is a horizontal bar chart to compare the earnings in billion USD with the number of movies for each genre.

Type: A bar chart.

KPI: Total box-office earnings in billions i.e. Sum ([total box office])/1,000,000,000.

Number of movies, i.e. count (distinct movie_name).

Dimension: Genre.

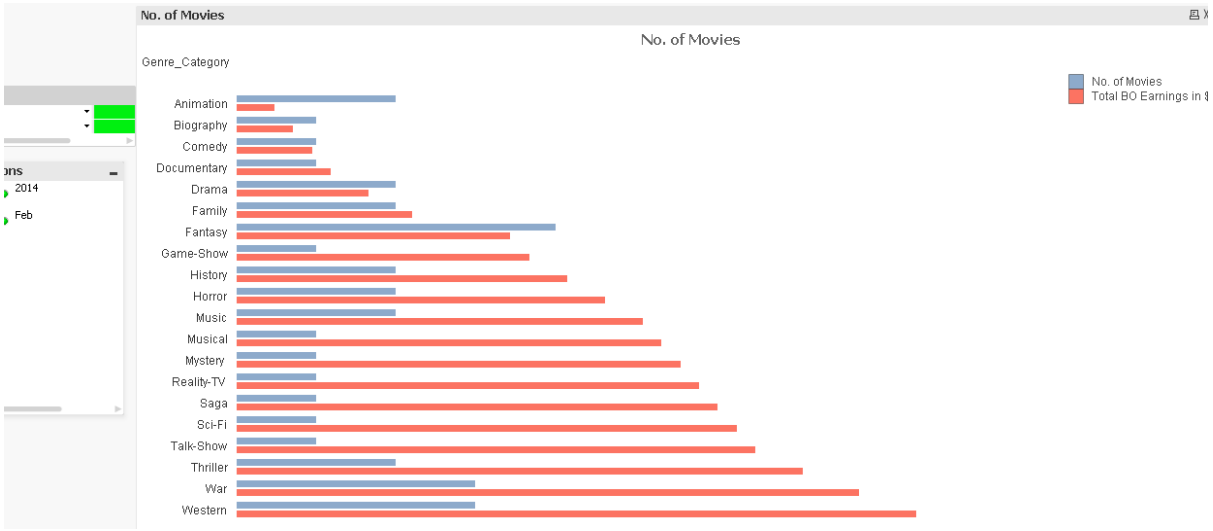


Fig 6.2.9 Bar chart to represent collection in billion USD in comparison with the number of movies.

6.2.10 Graphical representation

Fast type change mechanisms can be implemented to enable users to change the type of chart on the fly.

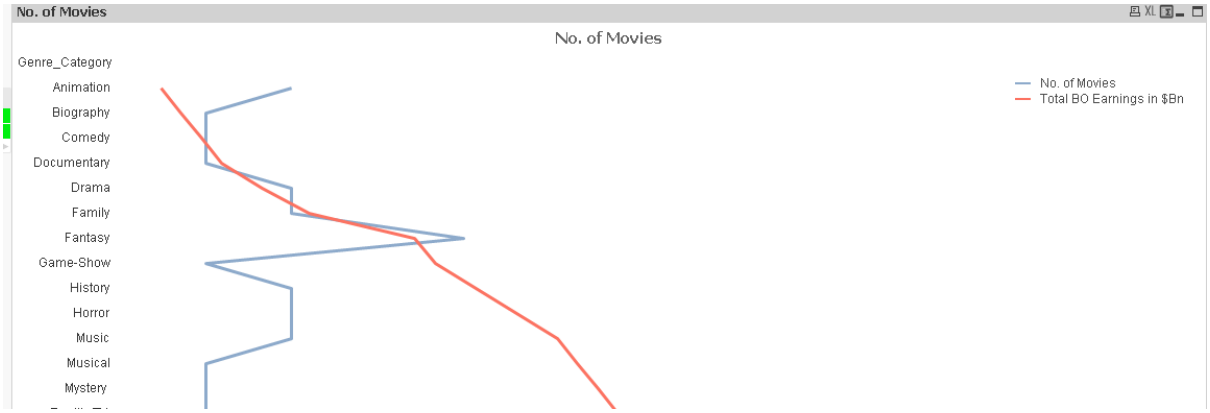


Fig 6.2.10 Graph representation of collection in billion USD in comparison with the number of movies.

The above charts are merely a sample of what can be done to identify patterns and trends among movies split over a variety of dimensions. There can be many more made and this particular dashboard can be scheduled to refresh every day with fresh data and distributed among all the business analysts so that they can maintain and work on a single version of truth.

In addition to the simple charts seen above, Qlikview allows us to create quite a lot of complex reports in a quite simple manner. This helps us avoid the complex queries that might be required in turn to fetch the same kind of data. Shown below are some examples.

6.2.11 Profitability percentage and movie count across genre for a selected country.

This chart allows the business analysts to answer questions like- What genres work better in a selected country? Which ones don't succeed at all despite repeated attempts?

There are several criteria that decide if a movie has done well. Here we have assumed that if the profitability is higher than average profitability then it falls into the "done well" category. The count of movie is represented as text on the bar charts with which the business analysts can gain associative insight into the profitability. For example, it can be inferred that in Australia, six movies of sport have a lower profitability than four of animation.

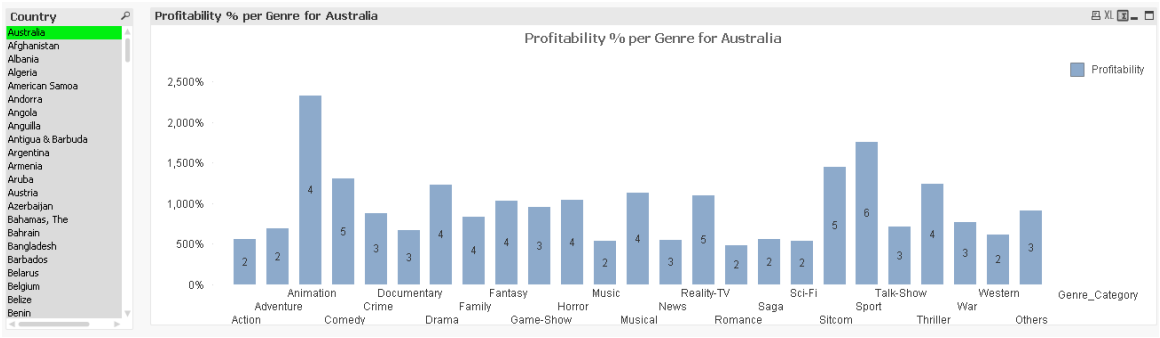


Fig.6.2.11 Profitability percentage and movie count across genre for a selected country.

Please refer GENRES_MORE_THAN_AVG part in appendix for SQL.

6.2.12 Critically acclaimed actors per country

This chart allows the business users to correlate Critics rating and profitability across actors for a selected country. This would serve to identify popularity (or lack of it) among the viewership for a particular country for the top 15 critically acclaimed actors, for example Anthony Hopkins, while clearly a critic's favorite, is certainly not a fan favorite in Belgium with barely a 124% average profitability %.

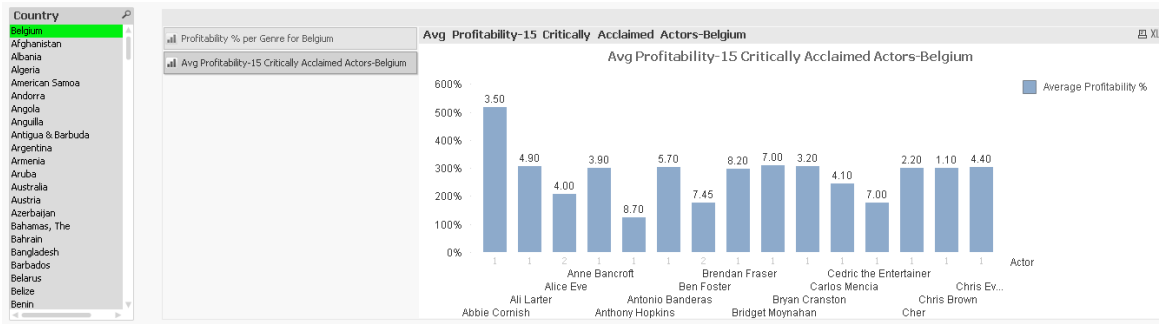


Fig.6.2.12 Critically acclaimed actors per country

Please refer ACTOR_CRITIC_RATING part in appendix for SQL.

6.2.13 Region Wise Box Office Comparison

Qlikview allows a feature known as set analysis which allows a single chart to work on two different data sets. This can be harnessed into building a comparison report for the business analysts which displays the total box-office collections for Europe along with the same for Asia across several genres. This will help the analysts understand whether a particular genre can succeed in a particular region more than the rest.

For example, fantasy movies are a huge hit in Asian countries but not so much in European countries according to the below chart.

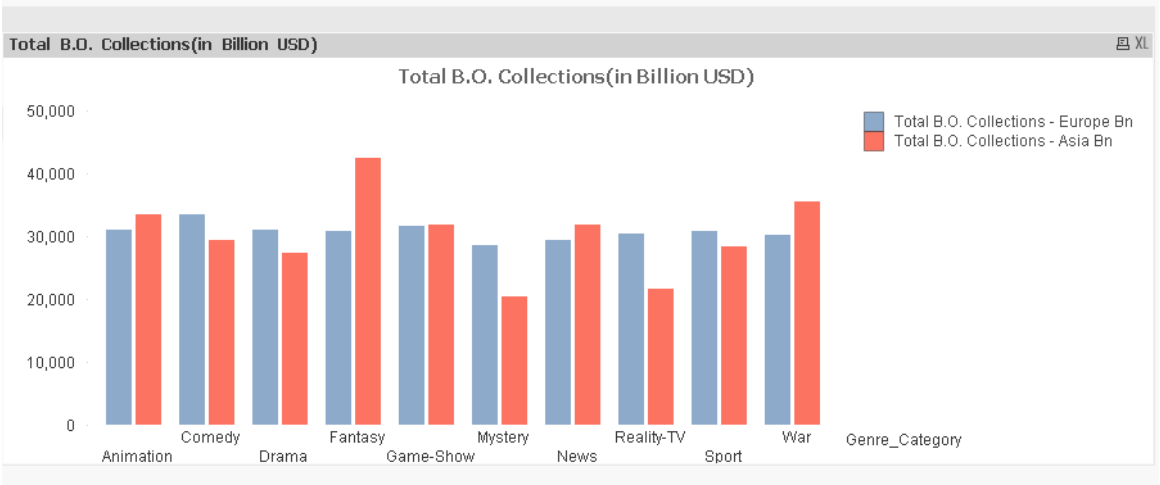
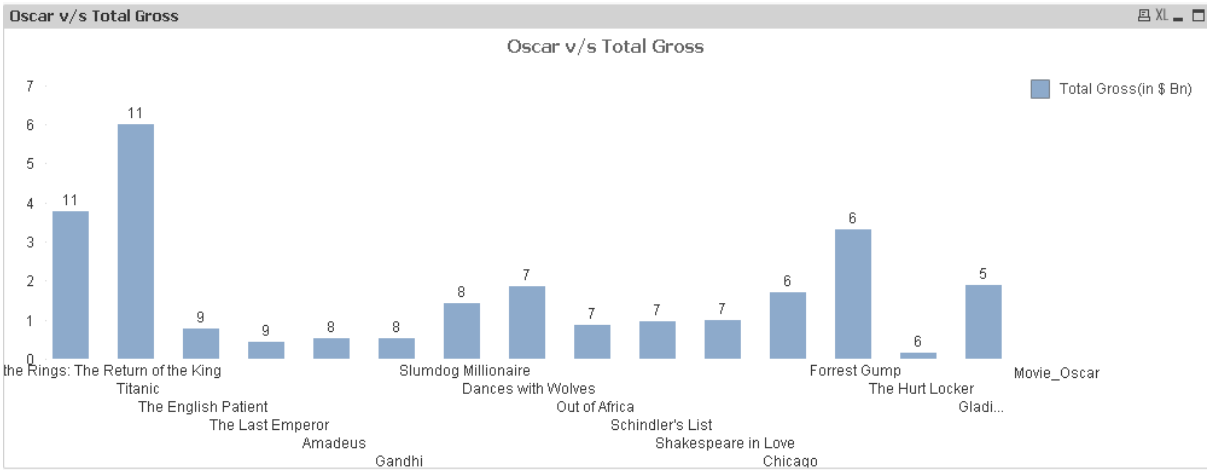


Fig 6.2.13 Region Wise Box Office Comparison

Please refer EURASIA_GENRE_COMPARISON part in appendix for SQL.

6.2.14 Oscar v/s Total Gross Chart

Many Oscar winning movies fail to do well in the box office and don't fetch as high rewards as their less popular peers. This chart portrays Gross earnings of top 15 Oscar winning movies in descending order.

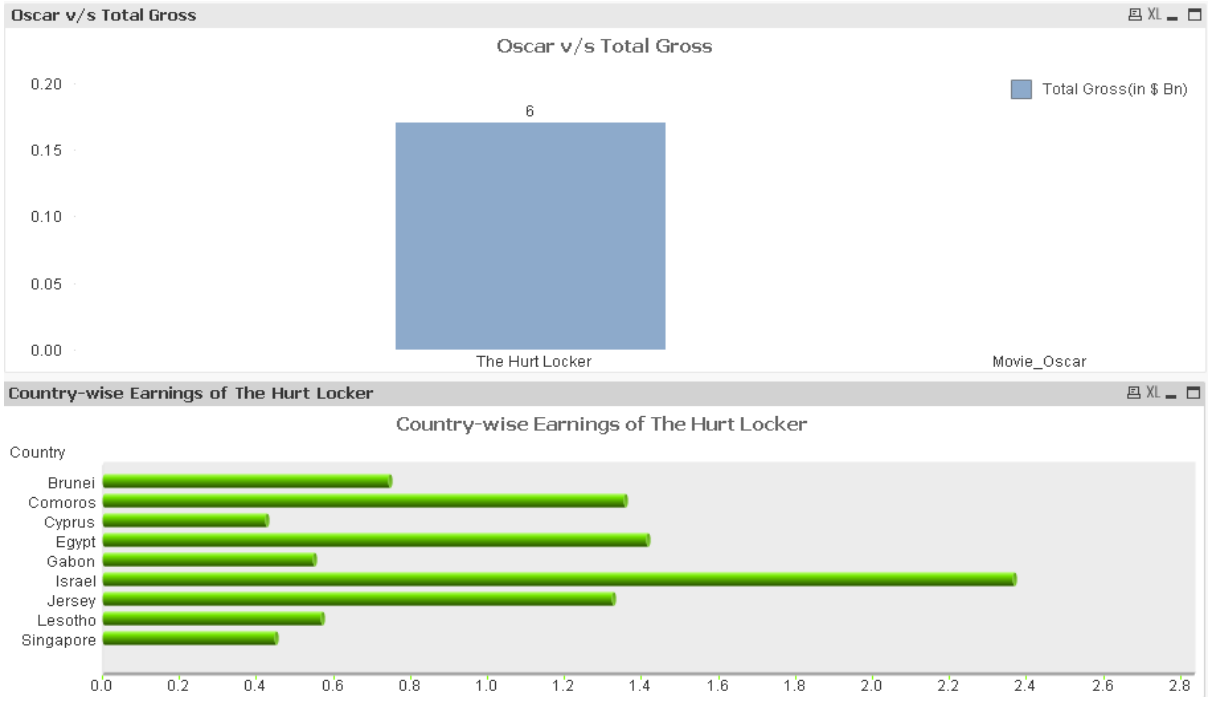


6.2.14 Oscar v/s Total Gross Chart

This chart helps the business analysts identify movies which have won several Oscars but haven't been very profitable. For example it can be inferred from the above chart that The Hurt Locker has fetched only 0.17 Billion USD in all even though it has won 6 Oscars.

6.2.15 Country wise earnings.

There is also an additional ability to drill down for the B.A.s if they select a particular Movie of interest say for example The Hurt Locker. This selection will trigger a chart displaying the earnings of that particular movie across the least 10 profitable countries.



6.2.15 Country wise earnings.

This will now enable the business analysts to understand which particular country is the cause of the problem.

Please refer **Oscar vs total gross** part in Appendix for SQL. The **country wise earnings** part covers the drill down feature. However, it isn't truly a drill down considering SQL limitations and we'd need to pass the selected movie name as a parameter to the query at run time.

7. CONCLUSIONS

Data mining has been proved over the years to be a very powerful tool to extract meaningful information out of raw data and present it in such a way so as to observe, identify and most importantly act on patterns. The entertainment business can benefit from this Qlikview tool, where we implement business intelligence considering the nature of the data and the complexity of the trends emerging out from them.

7.1 Accomplishments

This project involves business intelligence techniques to make sensible decisions based on the reports generated as the final outcome, though this method of decision making is familiar, I have tried implementing the project using some really quick techniques which is a part of my research on various business intelligence tools.

Where the following accomplishments are made in the project:

- A single version of truth is maintained i.e. a centralised data-store and a report is maintained.
- Easy distribution is possible via a URL as compared to the distribution of raw excels.
- Manual efforts spent at interpreting the data by complex querying is avoided with the use of Qlikview, the data instead talks to you.

Decision making is handed a major boost as business analysts are no longer forced to fly blind and can take informed decisions which would help maximize their business and improve the quality in management.

7.2 Future work

We have used only small databases to show the complete data mining process through thus project, the reason is it is time consuming to integrate huge data bases to work on the project, hence it is practical to implement such type of data sources to show a practical demo of what can actually be done in real time with vast data, which is the future of this project and my research.

REFERENCES

[1] "Competing on analytics: The new science of winning" by Thomas H. Davenport and Jeanne G. Harris.

[2] http://www.boxofficemojo.com/

[3] http://www.IMDB.com/

[4] www.opusdata.com

[5] http://www.the-numbers.com/

[6] "Data warehousing, data mining, and OLAP" by a Berson, SJ Smith

[7] https://erdplus.com/#/standalone

[8] http://www.qlik.com/us/products/why-qlik-is-different

[9] http://www.qlik.com/us/products/qlikview/getting-started

APPENDIX

SQL scripts to load the fact and the dimension tables:

MOVIEFACTTABLE:

```
LOAD [MOVIE_ID],
    [PRODUCTION_YEAR],
    YEAR(DATE#(PRODUCTION_YEAR,'MM/DD/YYYY')) AS
MOVIE_RELEASE_YEAR,
    MONTH(DATE#(PRODUCTION_YEAR,'MM/DD/YYYY')) AS
MOVIE_RELEASE_MONTH,
    [GENRE_ID],
    [STUDIO_ID],
    [ACTOR_ID],
    [COUNTRY_ID],
    [OSCAR_ID],
    [TOTAL BOX OFFICE],
    [PRODUCTION BUDGET],
    [ACTOR TOTAL GROSS],
    [TOTAL BUDGET];
SQL SELECT [MOVIE_ID],
    D_DATE.ACTUAL_DATE AS [PRODUCTION_YEAR],
    GENRE_ID,
    STUDIO_ID,
          ACTOR_ID,
    COUNTRY_ID,
    OSCAR_ID,
    TOTAL_BOX_OFFICE AS [TOTAL BOX OFFICE],
    PRODUCTION_BUDGET AS [PRODUCTION BUDGET],
    ACTOR_TOTAL_GROSS AS [ACTOR TOTAL GROSS],
    TOTAL_BUDGET AS [TOTAL BUDGET]
FROM F_MOVIES,D_DATE
WHERE F_MOVIES.PRODUCTION_RELEASE_DATE_KEY =
D_DATE.ACTUAL_DATE;
```

DIMENSION MOVIES:

LOAD [MOVIE_ID],
[MOVIE_NAME],
[MOVIE_IMDB_RATING],
[MOVIE_CRITICS_RATING];

SQL SELECT MOVIE_ID,
MOVIE_NAME,
MOVIE_IMDB_RATING,
MOVIE_CRITICS_RATING
FROM D_MOVIES;

DIMENSION GENRE:

LOAD [GENRE_ID],
[GENRE_CATEGORY],
MOVIES_COUNT,
[MARKET_SHARE];

SQL SELECT GENRE_ID,
GENRE_CATEGORY,
MOVIES_COUNT,
MARKET_SHARE
FROM D_GENRE;

DIMENSION STUDIO:

LOAD STUDIO,
[STUDIO_ID];

SQL SELECT STUDIO,
STUDIO_ID
FROM D_STUDIO;

DIMENSIONACTOR:

LOAD [ACTOR_ID],

```
    ACTOR,
    [MOVIES_PARTICIPATED];
SQL SELECT ACTOR_ID,
    ACTOR,
    MOVIES_PARTICIPATED
FROM D_ACTOR;

DIMENSION COUNTRY:
LOAD [COUNTRY_ID],
    COUNTRY,
    REGION;
SQL SELECT COUNTRY_ID,
    COUNTRY,
    REGION
FROM D_COUNTRY;

DIMENSION OSCAR:
LOAD YEAR,
    [OSCAR_ID],
    MOVIE_OSCAR,
        STUDIO_OSCAR,
    [TOTAL GROSS],
    NOMS,
    WINS,
    DATE;
SQL SELECT YEAR,
    OSCAR_ID,
    MOVIE_OSCAR,
    STUDIO_OSCAR,
    TOTAL_GROSS AS [TOTAL GROSS],
    NOMS,
    WINS,
    DATE
FROM D_OSCAR;
```

6.2.11

GENRES_MORE_THAN_AVG:

LOAD
GENRE_CATEGORY,
Profitability,
[NO OF MOVIES];
SQL

```
SELECT GENRE_CATEGORY,
     Sum(TOTAL_BOX_OFFICE / ( PRODUCTION_BUDGET +
ACTOR_TOTAL_GROSS ) * 100) AS Profitability,
          COUNT(DISTINCT MOVIE_ID) AS "NO OF MOVIES"
FROM   F_MOVIES,
     D_GENRE,
     D_MOVIES
WHERE  F_MOVIES.MOVIE_ID = D_MOVIES.MOVIE_ID
  AND F_MOVIES.COUNTRY_ID = D_GENRE.GENRE_ID
GROUP  BY GENRE_CATEGORY
HAVING Sum(TOTAL_BOX_OFFICE / ( PRODUCTION_BUDGET +
ACTOR_TOTAL_GROSS ) * 100) >
     (SELECT
Avg(TOTAL_BOX_OFFICE /( PRODUCTION_BUDGET + ACTOR_TOTAL_GROSS
)*100)
     FROM F_MOVIES);
```

6.2.12
ACTOR_CRITIC_RATING:
LOAD
ACTOR,
Avg_Critics_Rating,
Avg_Profitability
Where RowNo <16;
SQL

```
SELECT ACTOR,
               Avg(Movie_Critics_Rating) as Avg_Critics_Rating,
     Avg(TOTAL_BOX_OFFICE / ( PRODUCTION_BUDGET +
ACTOR_TOTAL_GROSS )) AS Avg_Profitability,
          COUNT(DISTINCT MOVIE_ID) AS "NO OF MOVIES"
FROM F_MOVIES,D_MOVIES,D_ACTOR
WHERE F_MOVIES.MOVIE_ID = D_MOVIES.MOVIE_ID
AND F_MOVIES.ACTOR_ID = D_ACTOR.ACTOR_ID
GROUP BY ACTOR
ORDER BY Avg(Movie_Critics_Rating) DESC;
```

6.2.13
EURASIA_GENRE_COMPARISON:
LOAD
GENRE_CATEGORY,
TOTAL_BO_COLLECTIONS_EUR,

```
TOTAL_BO_COLLECTIONS_ASIA;
SQL
SELECT EUR.GENRE_CATEGORY,
            TOTAL_BO_COLLECTIONS_EUR,
            TOTAL_BO_COLLECTIONS_ASIA
FROM
(SELECT
          GENRE_CATEGORY
    Sum(TOTAL_BOX_OFFICE) AS TOTAL_BO_COLLECTIONS_EUR
FROM F_MOVIES,D_GENRE, D_COUNTRY
WHERE F_MOVIES.COUNTRY_ID = D_GENRE.GENRE_ID
AND F_MOVIES.COUNTRY_ID = D_COUNTRY.COUNTRY_ID
AND REGION IN ('EASTERN EUROPE','WESTERN EUROPE')) EUR,
(SELECT
          GENRE_CATEGORY
    Sum(TOTAL_BOX_OFFICE) AS TOTAL_BO_COLLECTIONS_ASIA
FROM F_MOVIES,D_GENRE, D_COUNTRY
WHERE F_MOVIES.COUNTRY_ID = D_GENRE.GENRE_ID
AND F_MOVIES.COUNTRY_ID = D_COUNTRY.COUNTRY_ID
AND REGION IN ('NEAR EAST','ASIA (EX. NEAR EAST)')) ASIA
WHERE EUR.GENRE_CATEGORY = ASIA.GENRE_CATEGORY;

6.2.14
OSCAR_VS_TOTAL_GROSS:
LOAD
MOVIE_NAME,
TOTAL_GROSS,
OSCAR_WINS
Where RowNo <16;
SQL
SELECT
          MOVIE_NAME,
    Sum(TOTAL_BOX_OFFICE-TOTAL_BUDGET) AS TOTAL_GROSS,
    Sum(Wins) as OSCAR_WINS
FROM F_MOVIES,D_MOVIES,D_OSCAR
WHERE F_MOVIES.MOVIE_ID = D_MOVIES.MOVIE_ID
AND F_MOVIES.OSCAR_ID = D_OSCAR.OSCAR_ID
GROUP BY MOVIE_NAME
HAVING Sum(Wins)>0
ORDER BY Sum(Wins) DESC;

6.2.15
COUNTRYWISE_EARNINGS:
LOAD
COUNTRY,
TOTAL_GROSS_EARNINGS
Where RowNo <11;
SQL
SELECT
          COUNTRY,
```

```
    Sum(TOTAL_BOX_OFFICE-TOTAL_BUDGET) AS TOTAL_GROSS_EARNINGS
FROM F_MOVIES,D_COUNTRY, D_MOVIES
WHERE F_MOVIES.MOVIE_ID = D_MOVIES.MOVIE_ID
AND F_MOVIES.COUNTRY_ID = D_COUNTRY.COUNTRY_ID
AND MOVIE_NAME = &MOVIE_NAME   //A Variable/Prompt will need to be used to
pass the selected value if we're using SQL.
GROUP BY COUNTRY
ORDER BY Sum(TOTAL_BOX_OFFICE-TOTAL_BUDGET);
```