# Executive Summary

The entertainment industry is part of the technological advancements which has transformed the way business is being conducted. Data driven decisions give great insights into the business viability. Implementation of Data Warehouse can help various Production Houses utilize this data to make decisions about future movie productions.

Our Data Warehouse leverages the data residing in the IMDB databases and finds out the movie performance according to the various genres, so that this data can be used to drive decisions for any new projects that are undertaken. \*\* add for different fact tables

This data can also be used to find which is the most profitable genre, which Actor- Director pair works best for various genres etc. A data warehouse gets data from various sources and after the collation and analyses of such data, can be used by production houses to take advantage of the opportunities presented.

The Business Intelligence reports obtained from implementing this Data Warehouse would ultimately increase revenue by leveraging previous movie data. Using a consolidated Data Warehouse instead of separately housed data sources greatly improve the efficiency of creating Business Intelligence reports and aid the business in delivering the top performance.

# Metadata

|  |  |
| --- | --- |
| **Dimension meta data** | **Description** |
| Name of Dimension | Dim\_director |
| Business Definition | This dimension holds the data that contains the various attributes for the movie director. |
| Attributes | PK varchar |
| SCD | No Changes |
| Hierarchy | No Hierarchy |
| Load Frequency | Weekly |
| Source | Relational Database |
| Conformed | Yes. The event attributes remains same across all fact tables |
| Role Playing | No Roles |

# ETL Documentation

ETL refers to ‘Extract, Transform and load’. The various steps in the ETL process are as follows:

* Extracting data from databases
* Transforming the extracted data for storing it in relevant formats to enable query execution
* Loading the data into a final target database

The ETL process is conducted in the following manner:

* Data presented in the .csv files is extracted and loaded into the staging area by using Data flow tasks.
* Data present in the staging area is cleaned and transformed to cater to the Business Intelligence questions.
* The final tables that have been created are loaded into the Facts and Dimension tables.

Following is the layout of the **ETL plan** for Data Warehouse:

* Determine the target data needed in the Data Warehouse and the data sources used
* Preparation of Data mappings of the Data from sources in Excel to staging area and from the staging area to the data warehouse.
* Determine the Data extraction rules
* Determine the Data transformation and cleansing rules
* Plan and excute procedures for extraction and loading
* Load into Dimension tables
* Load into Fact tables.

1. Data Mappings for Data Warehouse (including sources, staging and target details and transformations)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimActor | DimActorID | int | Actor\_table.xlsx | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimActor | Actor\_Id | float | Actor\_table.xlsx | Primary Key |
| DimActor | First\_name | varchar | Actor\_table.xlsx |  |
| DimActor | Last\_name | varchar | Actor\_table.xlsx |  |
| DimActor | Gender | varchar | Actor\_table.xlsx |  |
| DimActor | Industry | varchar | Actor\_table.xlsx |  |
| DimActor | DOB | datetime | Actor\_table.xlsx |  |
| DimActor | Height\_in\_cm | float | Actor\_table.xlsx |  |
| DimActor | Weight\_in\_lb | float | Actor\_table.xlsx |  |
| DimActor | EyeColor | varchar | Actor\_table.xlsx |  |
| DimActor | Nationality | varchar | Actor\_table.xlsx |  |
| DimActor | PrimaryLanguage | varchar | Actor\_table.xlsx |  |
| DimActor | AgentName | varchar | Actor\_table.xlsx |  |
| DimActor | AgentContactNumber | varchar | Actor\_table.xlsx |  |
| DimActor | AgentContactEmail | varchar | Actor\_table.xlsx |  |
| DimActor | FavoriteGenre | varchar | Actor\_table.xlsx |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimAward | DimAwardID | int | award.csv | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimAward | Award\_Id | float | award.csv | Primary key |
| DimAward | Award\_name | varchar | award.csv |  |
| DimAward | Country | varchar | award.csv |  |
| DimAward | Venue | varchar | award.csv |  |
| DimAward | Host\_name | varchar | award.csv |  |
| DimAward | Month | varchar | award.csv |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimDate | DimdateID | int | Date.xlsx | Primary key |
| DimDate | Date | datetime | Date.xlsx |  |
| DimDate | Year | int | Date.xlsx |  |
| DimDate | MonthofYear | int | Date.xlsx |  |
| DimDate | MonthName | varchar | Date.xlsx |  |
| DimDate | DayofMonth | float | Date.xlsx |  |
| DimDate | WeekofYear | float | Date.xlsx |  |
| DimDate | DayofWeek | float | Date.xlsx |  |
| DimDate | WeekName | varchar | Date.xlsx |  |
| DimDate | DayofYear | float | Date.xlsx |  |
| DimDate | Fiscal Quarter | varchar | Date.xlsx |  |
| DimDate | Season | varchar | Date.xlsx |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimDirector | DimDirectorID | int | Director.xlsx | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimDirector | Director\_ID | float | Director.xlsx | Primary key |
| DimDirector | FirstName | varchar | Director.xlsx |  |
| DimDirector | LastName | varchar | Director.xlsx |  |
| DimDirector | DOB | datetime | Director.xlsx |  |
| DimDirector | Gender | varchar | Director.xlsx |  |
| DimDirector | MovieIndustry | float | Director.xlsx |  |
| DimDirector | Nationality | varchar | Director.xlsx |  |
| DimDirector | Language Spoken | varchar | Director.xlsx |  |
| DimDirector | MovieGenre | varchar | Director.xlsx |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimMovie | DimMovieID | int | Movie.xlsx | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimMovie | Movie\_ID | int | Movie.xlsx | Primary key |
| DimMovie | Num\_critic\_for\_reviews | float | Movie.xlsx |  |
| DimMovie | Duration\_in\_minutes | float | Movie.xlsx |  |
| DimMovie | GrossEarning | float | Movie.xlsx |  |
| DimMovie | Genres | varchar | Movie.xlsx |  |
| DimMovie | Movie\_title | varchar | Movie.xlsx |  |
| DimMovie | Num\_voted\_users | float | Movie.xlsx |  |
| DimMovie | Facenumber\_in\_poster | float | Movie.xlsx |  |
| DimMovie | Plot\_keywords | varchar | Movie.xlsx |  |
| DimMovie | Num\_user\_for\_reviews | float | Movie.xlsx |  |
| DimMovie | Language | varchar | Movie.xlsx |  |
| DimMovie | Country | varchar | Movie.xlsx |  |
| DimMovie | Content\_rating | varchar | Movie.xlsx |  |
| DimMovie | Budget | float | Movie.xlsx |  |
| DimMovie | Title\_year | float | Movie.xlsx |  |
| DimMovie | Imdb\_score | float | Movie.xlsx |  |
| DimMovie | Aspect\_ratio | float | Movie.xlsx |  |
| DimMovie | Is\_Multigenre | varchar | Movie.xlsx |  |
| DimMovie | 3D\_Released | varchar | Movie.xlsx |  |
| DimMovie | Is\_Rereleased | varchar | Movie.xlsx |  |
| DimMovie | Movie\_Start\_Date | datetime | Movie.xlsx |  |
| DimMovie | Movie\_End\_Date | datetime | Movie.xlsx |  |
| DimMovie | Movie\_Release\_Date | datetime | Movie.xlsx |  |
| DimMovie | DVD\_Release\_Date | datetime | Movie.xlsx |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimProdHouse | DimProdHouseID | int | production.xlsx | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimProdHouse | ProdHouse\_ID | float | production.xlsx | Primary key |
| DimProdHouse | CompanyName | varchar | production.xlsx |  |
| DimProdHouse | City | varchar | production.xlsx |  |
| DimProdHouse | Country | varchar | production.xlsx |  |
| DimProdHouse | Name | varchar | production.xlsx |  |
| DimProdHouse | ContactPersonEmail | varchar | production.xlsx |  |
| DimProdHouse | DateEstablished | datetime | production.xlsx |  |
| DimProdHouse | ContactPersonNumber | varchar | production.xlsx |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Source File | Transformation Rule |
| DimTheatre | DimTheatreID | int | Theatre.xlsx | Surrogate key of the dimension. Inserted as incremental key while loading data. |
| DimTheatre | theatre\_ID | Int | Theatre.xlsx | Primary key |
| DimTheatre | Theatre\_Name | varchar | Theatre.xlsx |  |
| DimTheatre | Total\_num\_of\_Screens\_2D | float | Theatre.xlsx |  |
| DimTheatre | Total\_num\_of\_Screens\_3D | float | Theatre.xlsx |  |
| DimTheatre | Seating\_Capacity | float | Theatre.xlsx |  |
| DimTheatre | IsMultiplex | bit | Theatre.xlsx |  |
| DimTheatre | Street\_Address | varchar | Theatre.xlsx |  |
| DimTheatre | City | varchar | Theatre.xlsx |  |
| DimTheatre | State | varchar | Theatre.xlsx |  |
| DimTheatre | Country | varchar | Theatre.xlsx |  |
| DimTheatre | ZipCode | float | Theatre.xlsx |  |
| DimTheatre | Ticket\_price\_2D | float | Theatre.xlsx |  |
| DimTheatre | Ticket\_price\_3D | float | Theatre.xlsx |  |

Fact tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Staging table attributes | Source File | Transformation Rule |
| fact\_Awards | DimActorID | int | Actor\_Id |  | Foreign key of dimension table corresponding to Actor ID |
| fact\_Awards | DimDirectorID | Int | Director\_ID |  | Foreign key of dimension table corresponding to Director ID |
| fact\_Awards | DimMovieId | int | Movie\_Id |  | Foreign key of dimension table corresponding to movie ID |
| fact\_Awards | DimAwardID | int | Award\_Id |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Staging table attributes | Source File | Transformation Rule |
| fact\_Movie\_Performance | Movie\_Id | int | Movie\_Id |  | Foreign key of dimension table corresponding to movie ID |
| fact\_Movie\_Performance | LeadActorID | int | Actor\_ID |  | Foreign key of dimension table corresponding to actor ID |
| fact\_Movie\_Performance | SupportingActorID | int | Actor\_ID |  | Foreign key of dimension table corresponding to actor ID |
| fact\_Movie\_Performance | DirectorID | int | Director\_Id |  |  |
| fact\_Movie\_Performance | DateId | int | Date\_Id |  |  |
| fact\_Movie\_Performance | Number\_of\_tickets\_sold\_2D | float |  | ?? |  |
| fact\_Movie\_Performance | Number\_of\_tickets\_sold\_3D | float |  | ?? |  |
| fact\_Movie\_Performance | Gross\_Revenue\_Generated | float |  | ?? |  |
| fact\_Movie\_Performance | GrossProfit | float |  | ?? |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Staging table attributes | Source File | Transformation Rule |
| fact\_movie\_production | DimMovieID | int | Movie\_Id |  | Foreign key of dimension table corresponding to Movie ID |
| fact\_movie\_production | DimStartDateID | Int | Date\_ID |  | Foreign key of dimension table corresponding to Date ID |
| fact\_movie\_production | DimEndDateId | int | Date\_ID |  | Foreign key of dimension table corresponding to Date ID |
| fact\_movie\_production | DimReleaseDateId | int | Date\_ID |  | Foreign key of dimension table corresponding to Date ID |
| fact\_movie\_production | DimDVDReleaseDateId | int | Date\_ID |  | Foreign key of dimension table corresponding to Date ID |
| fact\_movie\_production | DimProdHouseID | int | ProdHouse\_ID |  | Foreign key of dimension table corresponding to ProdHouse\_ID |
| fact\_movie\_production | DimLeadActorID | int | Actor\_Id |  | Foreign key of dimension table corresponding to Actor\_ID |
| fact\_movie\_production | DimSupportingActorID | int | Actor\_id |  | Foreign key of dimension table corresponding to Actor\_ID |
| fact\_movie\_production | Cost\_of\_production | float |  |  |  |
| fact\_movie\_production | Number\_of\_Theatres\_Movie\_Released\_In | int |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Target table | Target table Attributes | Data Types | Staging table attributes | Source File | Transformation Rule |
| fact\_Theatre | DimMovieID | int | Movie\_Id |  | Foreign key of dimension table corresponding to Movie ID |
| fact\_Theatre | DimTheatreID | Int | Date\_ID |  | Foreign key of dimension table corresponding to Theatre ID |
| fact\_Theatre | DimDateId | int | Date\_ID |  | Foreign key of dimension table corresponding to Date ID |
| fact\_Theatre | Number\_of\_tickets\_sold\_2D | int |  |  |  |
| fact\_Theatre | Number\_of\_tickets\_sold\_3D | int |  |  |  |
| fact\_Theatre | Number\_of\_shows\_2D | int |  |  |  |
| fact\_Theatre | Number\_of\_shows\_3D | int |  |  |  |
| fact\_Theatre | Total\_number\_of\_Screens | int |  |  |  |
| fact\_Theatre | Released\_year | float |  |  |  |
| fact\_Theatre | Month\_of\_sale | int |  |  |  |
| fact\_Theatre | Price\_per\_ticket\_2D |  |  |  |  |
| fact\_Theatre | Price\_per\_ticket\_3D |  |  |  |  |

1. Data Extraction Rules

The process of retrieving data out from data sources for prcessing or storage is known as Data Extraction. Data present in data files is often poorly structured. The import to the staging system of such data is usually followed by data transformation before moving ahead.

To achieve data extraction, we employed the following steps:

* Source data that is present in the Comma Separated Value (.csv) format is extracted and imported into Microsoft SQL server as tables.

Put data that has been imported from .csv files ---one bullet point

* This data is used for Data transformation and further loading
* Once data has been put extracted into the staging area, it is cleaned and transformed to create Dimension and fact tables. Once these Dimension and fact tables have been verified with respect to the Business Intelligence needs, it is loaded into the Data Warehouse area.

1. Data transformation and cleansing rules

* Removal of dirty data:

ETL Documentation (screen shots with brief description of SSIS packages) Screenshots – Apurv – Formatting – Isha, Poonam, Adi

BI Reports- Screen shots of any dashboard/OLAP Cube a brief description of the types of report it can generate.  Screenshots –

Date Report Created - Poonam