Kulliyyah of Information Communication Technology (KICT)

Semester 1

Data Warehousing

Group project:

Title: MovieLens Data Warehouse

Members names	Matric No
SHAIKHAN HAMMAM SHAIKHAN AL-HEBSHI	1536947
ISSA MAHAMAT ISSA AHMAT	1526367
MUHAMMAD ASYRAF DANISH BIN AHMAD SUMARDIE SHAN	1621629
HASAN MD TANVIR	1716763
SAMIRA HASSAN MOHAMED AHMED	1520812

Introduction

Our group has obtained a dataset from the Kaggle.com website where datasets there are open source and available to use. The dataset we have is called MovieLens and its about movie ratings and Customer behavior, the dataset was obtained from the following link:

https://www.kaggle.com/snehal1409/movielens

MovieLens is a dataset which has around 100000 ratings and almost 1300 tag applications across more than 9100 movies. The data were created between January 09, 1995 and October 16, 2016. Our database contains four tables. The tables are Cinemas movies, Companies, and Ratings.

Cinemas is our first table, and the attributes are CinemaID which is the primary key, CinemaName, Location where the cinema is and hall represents the numbers of halls in each cinema. This table will record all the information related to each cinema. Our second table is Companies, these companies are the producers for the movies and it has two attributes, CompanyID and CompanyName.

The third table is Movies and it has four attributes MovieID, Title for the move, Genres which represent the style of the movie and CompanyName as the producing company. This table is going to record all the 9126 movie titles and genres alongside the producer company. All the data is obtained from Kaggle.com website.

Ratings is our forth table and it has seven attributes which are RatingID, MovieID, RatingScore, RatingDate, CustomerID, CinemaID and lastly Comment from the customer about the move. The table saves all the data for every movie's rating and stamp throughout the dataset.

The fifth table is Customers, which holds data of Customer who watch the movies, with a primary key of CustomerID. The attributes are CustomerID, FirstName, LastName and Date of birth of each customer.

Business Process

1. Since there are a lot of movies and each movie has a ratings left by the customers in order to evaluate the movie in terms of quality, excitement, performance and other considerations that the customers think are necessary to evaluate the movie. Customers after they finish watching a movie will leave their personal rating on the movie, these ratings will be stored in the table "Ratings". Our group decided to see what are the movies that are high in rating by the customers. From that point, a decision can be taken for the movies that are highly ranked in which day and time can be available for watching. Thus, we thought that, we could create a data warehouse, which calculates the Average rating of all the ratings left by customers.

Business Process: what movies that have high rating.

2. We also decided to create a data warehouse to know what is the preferred genres which the style of the movie for the customers according to their ages. So a decision can be made for the most beloved movie style by particular age group of customers. This will help to make decision to what kind of movie should be made available in the cinemas and what kind of audience we target for each movie style.

Business Process: what is the preferred genres for specific audience?

Dimensional modeling

After we are clear with the business process and it is ready, in order to have the information that can help us in the decision-making process. We proceed with the Data Warehouse stage where we create the dimensional modeling which has the dimension tables and fact table.

From the steps we made in the process above, we have three dimensions and one fact for the first business process which is about the highest movie in terms of raring from the customers, the dimensions are (DimMovie, DimDate, DimRating) and Fact table as Fact_MovieRating.

For the second business process we also have three dimensions and one fact. In the second business process, we aim to know what the most preferred genre is. In other words, what is the most style of movie that the customers like to watch in the cinemas? The dimensions are (DimCustomer, DimRating, DimGenre), and the fact table is Fact_GenrePrefered.

There are attributes from each dimension, which are relevant to business and create for every dimension a surrogate key. For the fact table, we choose only the key value of each dimension table to create the fact table.

Identification of the information needs for the data warehouse

The problem faced by us is the incomplete data related to movie ratings that are available to us to build the data warehouse and we also faced the problem of using the right data sets to be put into the fact tables and dimension tables.

The only way to gather the information of the movie ratings is through obtaining the available datasets on the web. Many organizations related to movie ratings have sources of information they use to rate the movies.

As stated in introduction, our dataset named MovieLens is about movie ratings. We get the data sets from Kaggle.com, the URL is: https://www.kaggle.com/snehal1409/movielens

Often, analysts create analytical and summary reports of the dataset. These reports can be simple correlations of existing reports, or they can include information that people overlook with the existing information stored in spreadsheets and memos of the dataset.

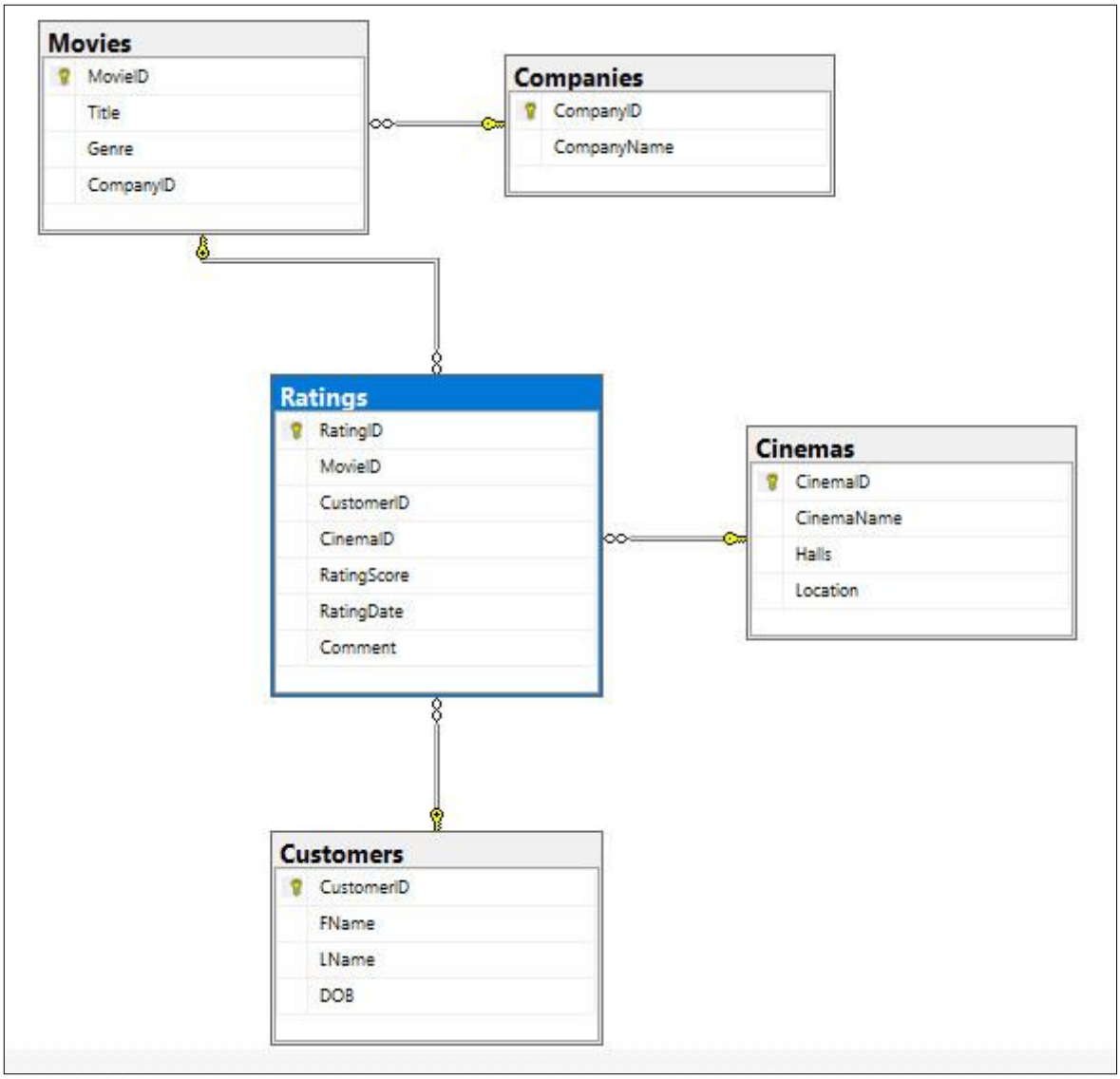
Another part of this collection and analysis phase is understanding how people gather and process the information because data warehouse can automate many reporting tasks, but we cannot automate what we have not identified and do not understand.

This process requires extensive interaction with various data sets. We need to go through these processes before we are ready to begin designing the warehouse.

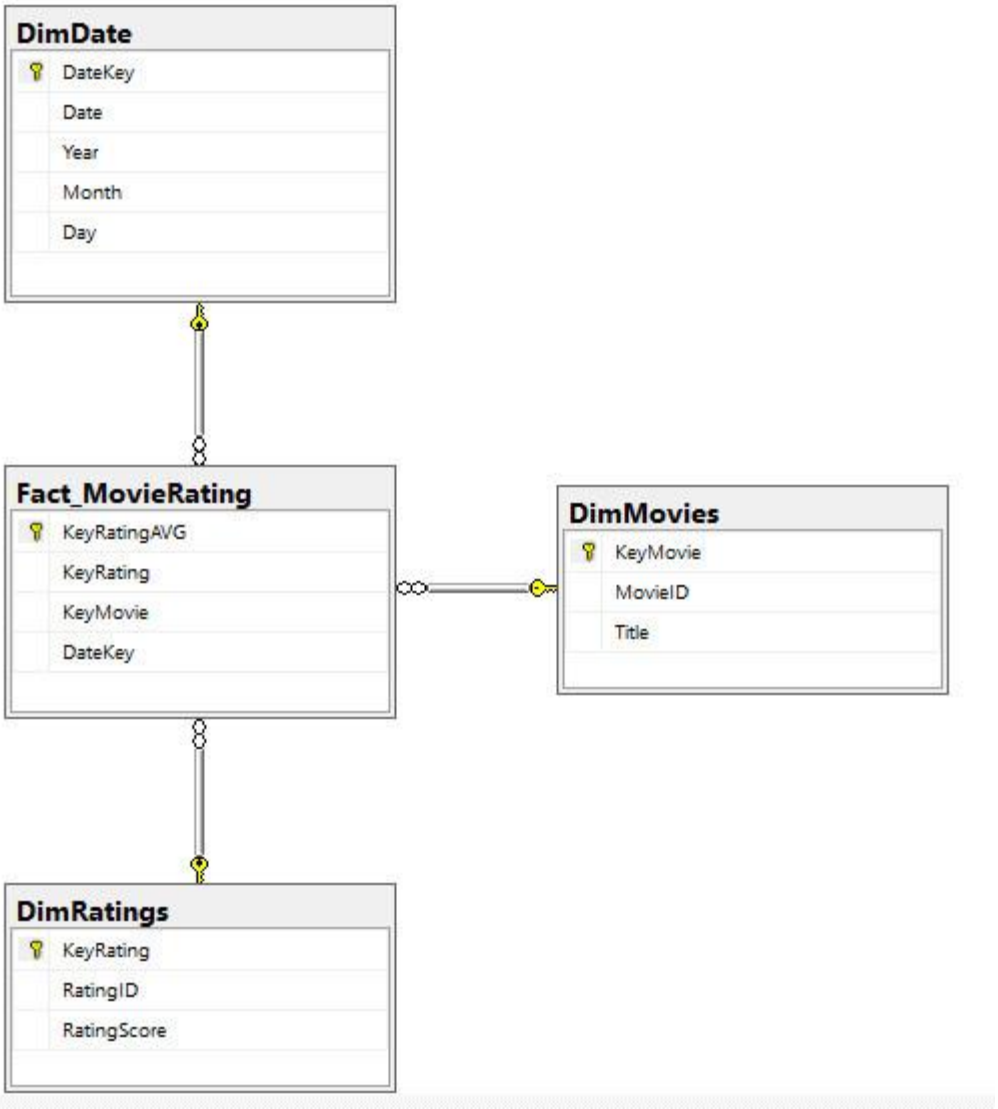
Interaction with various data sets helps us in completing the incomplete data and determining the right data to be put into our data warehouse. It helps us to design the data warehouse more accurately using the right data sets to be put into the fact tables and dimension tables.

After acquired all the data set for business process and getting the business process ready, we need to prepare our Data Warehouse by dimensional modelling that includes dimension tables and fact tables.
We defined five dimensions to make the fact table named Fact_MovieRating and Fact_GenrePrefered. The six dimensions are (DimCinema, DimMovie, DimCustomer, DimDate, DimRating). Each attribute for each dimension is relevant to our business process.

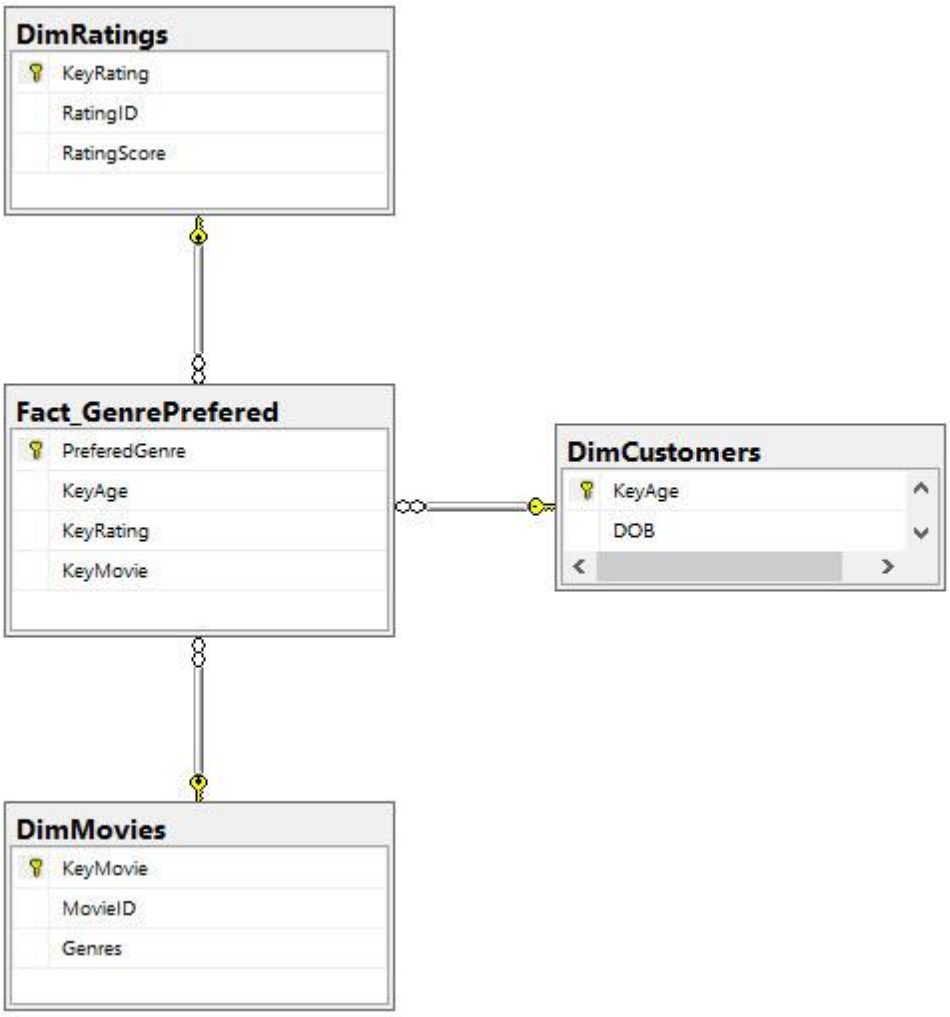
MoveLens star schema



Dimensional model diagram Business Process 1: Movie Rating



Dimensional model diagram Business Process 2: Genres Preferred:



Bus Matrix:

	Data Warehouse Bus Matrix for MovieLens								
Business Process	Fact Table	Granularity	Facts	DimMovie	DimDate	DimRating	DimCustomers	DimGenre
movie Rating	Fact_MovieRating	repreasnts the average rate of each movie based on CustomersRating	RatingAVG	x	x	x		
Genres Preferred	Fact_GenreAgePreferred	represents the genres preferred based on the frequency of Age per Genre	Age_per_genre			x	x	x

ETL plan and source-to-target mapping.

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyRating	Rating key to identify each rating	-1	1	int	4	PK, FK		N		DW	DimRating	KeyRating	int
KeyMovie	Movie Key to identify each movie	-1	7127 1900	int	4	FK		N		DW	DimMovie	KeyMovie	int
KeyDate	Date Key to identify each date	-1	3	date	4	FK		N		DW	DimDate	KeyDate	date
RatingAVG	Displays the rating average of a particular movie	-1	1771900	float				Y		DW			

Table Name : Fact_GenreAgePreferred
Table Type : Fact_GenreAgePreferred
Table Description : Each row will display the most frequent age for each Movie genre.

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyAge	Displays the age of a customer	-1	18	int	4	PK, FK	DimCustomers.KeyAge	N		DW	DimCustomers	KeyAge	int
KeyGenre	Movie Key to identify each movie	-1	Drama	int	44	FK	DimGenre.KeyGenre	N		DW	DimGenre	KeyGenre	int
KeyRating	Displays Rating of the movie	-1	2.3	float	4	FK	DimRating.KeyRating	N		DW	DimRating	KeyRating	int
Age_per_genre	Displays the most frequent age that occurred for a particular genre	-1	18	int	4			Y		MovieLens			int

Table Name : DimGenre
Table Type : Dimension
Table Description : Each row represents the Genre of a certain movie

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyGenre	Unique Key number that refers to each row of the genre dimention	-1	2	int	4	PK		N		DW			int
MovieID	Displalys the MovieID which identifies each movie in the source table	-1	204	int	4	FK	Movies.MovieID	N		MovieLens	Movies	MovieID	int
Genre	Shows the Genre of the movie	N/A	Drama	nvarchar	100	FK	Movies.Genre	N		MovieLens	Movies	Genre	nvarcha

Table Name : DimRating
Table Type : Dimension
Table Description : Each row displays the rating score left by one of the customers

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyRating	Display rating of the movie	-1	1	int		PK		N		DW			int
RatingID	Display RatingID which is unique to identify each rating row from source table.	-1	1001	int	4	FK	Ratings.RatingID	N		MovieLens	Ratings	RatingID	int
RatingScore	Shows the score which each movie got.	-1	2.3	float	120	FK	Ratings.RatingScore	N		MovieLens	Ratings	RatingScore	float

Table Description : Each row will display the age of the customers

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyAge	Display the key of the age in this dimension	-1	4	int	4	PK		N		DW			int
DOB	Shows the age of each customer	-1	4	int	4	FK	Customers.DOB	N		Movielens	Customers	DOB	int

Table Name : DimMovie
Table Type : Dimension
Table Description : Each row represents a movie and its title

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyMovie	Unique key to identify each row of the DimMovie Dimension.	-1	4	int	4	PK		N		DW			int
MovieID	Shows the unique Movie ID based on the source table	-1	250	int	4	FK	Movies.MovieID	N		Movielens	Movies	MovieID	int
Title	Displays the title name of each movie	N/A	The Usual suspect	nvarchar	100		Movies.Title	N		Movielens	Movies	Title	nvarchar

Table Name : DimDate
Table Type : Dimension
Table Description : Each row will display the age of the customers

				Target						Source			
Column Name	Description	Unknown Member	Example Values	Data Type	Size	Key	FK To	Null	Default Value	Source System	Source Table	Source Field Name	Source Data Type
KeyDate	Key of the Date in the dimension for each rating's date	-1	4	int	4	PK		N		DW			int
RatingDate	The rating date.	Unknown Attribute	7/7/2019	Date	4	FK	Rating.RatingDate	N		Movielens	Ratings	RatingDate	int

Data warehouse SQL Schema for business process 1.

```
CREATE TABLE  Fact_GenrePrefered(
   [PreferedGenre] int NOT NULL
,  [KeyAge] int NOT NULL
,  [KeyRating]  int NOT NULL
,  [KeyMovie] int NOT NULL
, CONSTRAINT [PK_PreferedGenre] PRIMARY KEY (PreferedGenre)
, CONSTRAINT [FK_KeyAge] FOREIGN KEY (KeyAge)
  REFERENCES DimCustomers (KeyAge)
, CONSTRAINT [FK_KeyRating] FOREIGN KEY (KeyRating)
  REFERENCES DimRatings (KeyRating)
, CONSTRAINT [FK_KeyMovie] FOREIGN KEY (KeyMovie)
  REFERENCES DimMovies (KeyMovie)
)

CREATE TABLE  DimRatings(
   [KeyRating]  int NOT NULL
,  [RatingID]  int NOT NULL
,  [RatingScore] [float](4)  NOT NULL
, CONSTRAINT [PK_KeyRating] PRIMARY KEY (KeyRating)
)



CREATE TABLE  DimCustomers(
   [KeyAge] int NOT NULL
,  [DOB] [date]  NOT NULL
, CONSTRAINT [PK_KeyAge] PRIMARY KEY (KeyAge)
)


CREATE TABLE  DimMovies(
   [KeyMovie]  int NOT NULL
,  [MovieID]  int NOT NULL
,  [Genres] [varchar](50)  NOT NULL
, CONSTRAINT [PK_KeyGeres] PRIMARY KEY (KeyMovie)
)
```

Data warehouse SQL Schema for business process 2.

```
CREATE TABLE  Fact_MovieRating(
   [KeyRatingAVG] int NOT NULL
,  [KeyRating]  int NOT NULL
,  [KeyMovie] int NOT NULL
,  [DateKey] int NOT NULL
, CONSTRAINT [PK_KeyRatingAVG] PRIMARY KEY (KeyRatingAVG)
, CONSTRAINT [FK_KeyRating] FOREIGN KEY (KeyRating)
  REFERENCES DimRatings(KeyRating)
, CONSTRAINT [FK_KeyMovie] FOREIGN KEY (KeyMovie)
  REFERENCES DimMovies (KeyMovie)
, CONSTRAINT [FK_DateKey] FOREIGN KEY (DateKey)
  REFERENCES DimDate (DateKey)
)

CREATE TABLE  DimRatings(
   [KeyRating]  int NOT NULL
,  [RatingID]  int NOT NULL
,  [RatingScore] [float](4)  NOT NULL
, CONSTRAINT [PK_KeyRating] PRIMARY KEY (KeyRating)
)



CREATE TABLE  DimMovies(
   [KeyMovie]  int NOT NULL
,  [MovieID]  int NOT NULL
,  [Title] [varchar](50)  NOT NULL
, CONSTRAINT [PK_KeyGeres] PRIMARY KEY (KeyMovie)
)



CREATE TABLE DimDate (
   [DateKey]  int NOT NULL
,  [Date] datetime   NOT NULL
,  [Year]  int   NOT NULL
,  [Month]  int   NOT NULL
,  [Day]  int   NOT NULL
, CONSTRAINT [PK_DateKey] PRIMARY KEY CLUSTERED( [DateKey] )
)
```

Screen Shots samples

CinemaID	CinemaName	Halls	Location
100	Wangsa Walk	8	Sri Rampai
101	SuriaKLCC	15	City Centre
102	Pavilion	10	Bukit Bintang
103	Pavilion	10	Bukit Bintang
104	Aion Big	4	Sri Rampai
105	Mid Valley	3	Bangsar
106	Starhill Gallery	6	Bukit Bintang
107	Sunway pyra...	4	Petaling Jaya
108	The Garden	7	Bangsar
109	Sungei Wang	9	Bukit Bintang
110	Plaza Idaman	2	Gombak

CustomerID	FName	LName	DOB
1	Adam	Ali	1993-12-17
2	Ahmed	Salem	1996-11-11
3	Noor Hafiza	Shafiq	2002-06-14
4	Siti Samia	Mohammed	2001-09-18
5	Adel	Ben Hakim	2000-07-24
6	Khaled	Nahin	1999-01-27
7	Aiman	Abdullah	1992-01-08
8	Ismit	Fitri	1997-11-10
9	Aniq	Najmi	2004-06-24
10	Zainatul Shimah	Ali	2001-09-05
11	Salem	Zulkifli	1999-03-23

RatingID	MovieID	CustomerID	CinemaID	RatingScore	RatingDate	Comment
401	201	3	100	4.5	2002-05-07	Very Good
402	202	2	105	4	1996-09-27	Good
403	203	2	111	2	1996-10-02	Very Poor
404	204	35	102	3	1999-07-09	Average
405	205	27	105	1	2000-07-16	Awful
406	206	4	107	2.5	2001-11-13	Poor
407	207	24	103	3	1998-12-27	Average
408	208	25	101	2	1997-04-05	Very Poor
409	209	50	110	3.5	1996-01-15	Okay
410	210	45	104	5	1995-11-11	Amazing

CompanyID	CompanyName
301	marvel
302	Disney
303	Pixer
304	Warner Bros
305	DC Comics
306	Dreamworks
307	Studio Ghibli

MovieID	Title	Genre	CompanyID
201	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	303
202	Jumanji (1995)	Adventure Children Fantasy	301
203	Grumpier Old Men (1995)	Comedy Romance	304
204	Waiting to Exhale (1995)	Comedy Drama Romance	305
205	Father of the Bride Part II (1995)	Comedy	301
206	Heat (1995)	Action Crime Thriller	306
207	Sabrina (1995)	Comedy Romance	303