"Movie Management System"

Ware House Design

1. Introduction

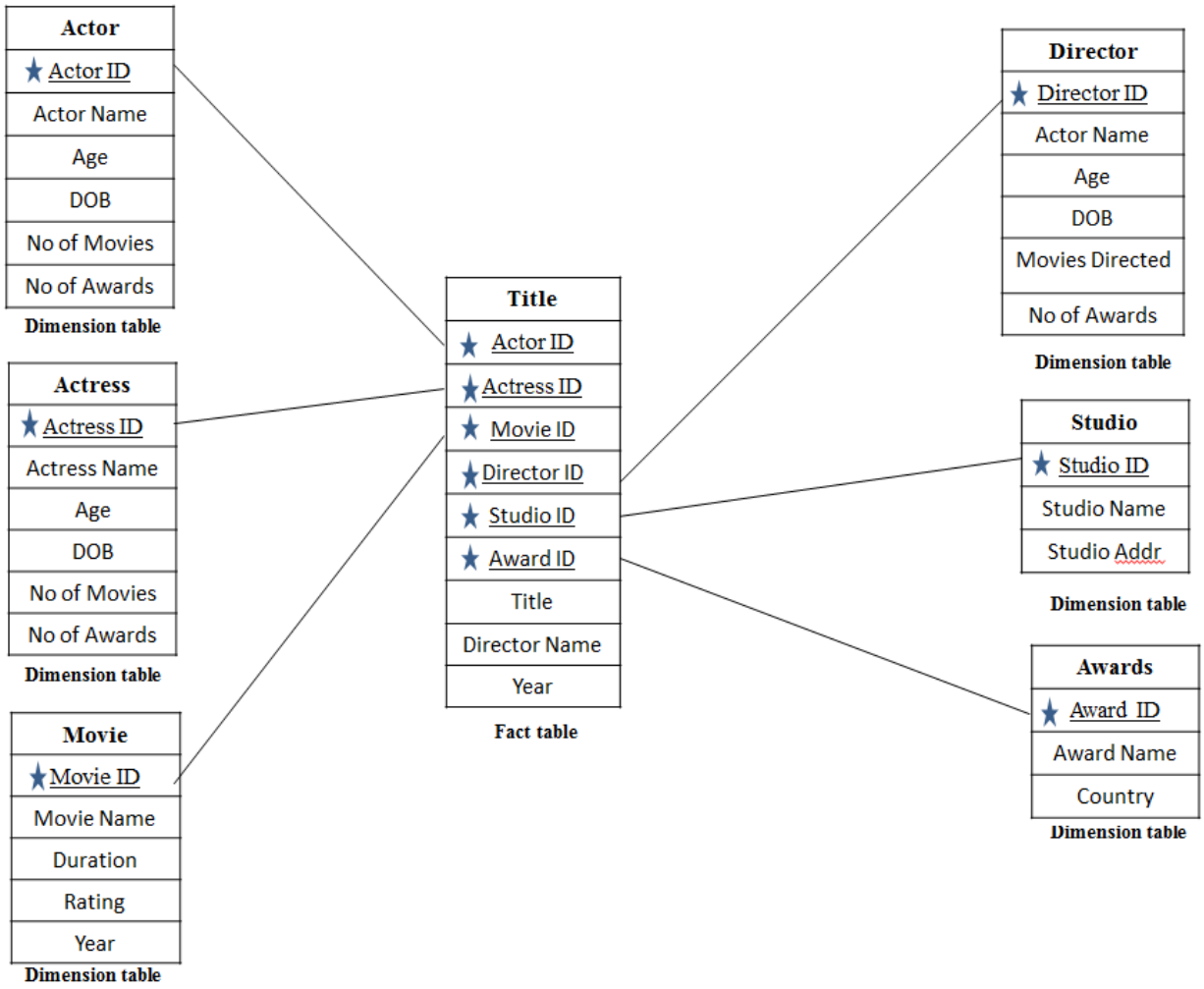
This Project involves in loading the Movie data's into the database by using the standard ETL process. The Process involves extraction of data from the source, cleansing and parsing the data and finally loading them into the database. The data's are further analyzed using OLAP functions. These functions help us to determine the data as a cube and find the modularity among them. Finally the data's are more scrutinized using few of the data mining algorithms.

2. Schema Design

Table Name	***Table Type***	***Key Name***	***Key Time***
Actors	Dimension	Actor ID	Primary Key
Actress	Dimension	Actress ID	Primary Key
Directors	Dimension	Director ID	Primary Key
Movie	Dimension	Movie ID	Primary Key
Studio	Dimension	Studio ID	Primary Key
Awards	Dimension	Award ID	Primary Key
Title	Fact	Actor ID, Actress ID, Director ID, Movie ID, Studio ID, Award ID	Foreign Key

- Actors (**PK** Actor ID, Actor Name, Age, Date of Birth, No of Movies Acted, No of Awards)
- Actress (**PK** Actress ID, Actress Name, Age, Date of Birth, No of Movies Acted, No of Awards)
- Directors (**PK** Director ID, Director Name, Age, Date of Birth, No of Movies Directed, No of Awards)
- Movie (**PK** Movie ID, Movie Name, Duration, Rating, Year)
- Studio (**PK** Studio ID, Studio Name, Studio Address)
- Awards (**PK** Award ID, Awards Name, Country)
- Title (**FK** (Movie ID, Actor ID, Actress ID, Director ID, Studio ID, Award ID), Title, Director Name)
-

PK – Primary Key
FK – Foreign Key
Schema Type – Star Schema



3. ETL Design

The term ETL stands for Extract, Transform and Load. ETL process involves extracting the data from the source systems. In many cases this is the most challenging aspect of ETL, since extracting data correctly sets the stage for how subsequent processes go further.

Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization/format. Common data source formats are relational databases and flat files, but may include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even fetching from outside sources such as through screen-scraping. In general, the goal of the extraction phase is to convert the data into a single format which is appropriate for transformation processing. An intrinsic part of the extraction involves the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data may be rejected entirely or in part.

- ***Source*** : Amstat, Knomarix, autolab, amazon, myrrix.
- ***File type*** : Comma Separated files, Text files with Standard Delimiters
- ***Extraction*** : Files are extracted from the sources using a java batch files to a windows directory.
- ***Transformation***: Files are cleansed and parsed through a set of validation processes. On completion of the task a cleansed file is create and moved to the inbound directory.
- ***Loading*** : The Files are picked from the inbound directory and loaded into the Microsoft SQL database through the built in Sql services.

4. OLAP Design

In computing, online analytical processing, or OLAP an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management budgeting and forecasting, financial reporting and similar areas, with new applications coming up, such as agriculture. The term OLAP was created as a slight modification of the traditional database term OLTP (Online Transaction Processing).

OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time. They borrow aspects of navigational databases, hierarchical databases and relational databases.

- ***Roll-up*** **:** A roll-up involves summarizing the data along a dimension. The summarization rule might be computing totals along a hierarchy or applying a set of formulas such as movie rating = Total Rating/Number of Votes.
- ***Drill-Down*** **:** Drill Down allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down).
- ***Slicing*** **:** Slice is the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension. Finding out the movie that released in the year of 2004 are "sliced" out of the data cube.
- ***Dicing*** : The dice operation produces a sub cube by allowing the analyst to pick specific values of multiple dimensions. The new cube will show limited number of movies, the time and region dimensions cover the same range as before**.**

5. Mining Functions

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.
The mining model that an algorithm creates from your data can take various forms, including:

- A set of clusters that describe how the cases in a dataset are related.
- A decision tree that predicts an outcome, and describes how different criteria affect that outcome.
- A mathematical model that forecasts sales.
- A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together.

Microsoft SQL Server Analysis Services provides multiple algorithms for use in your data mining solutions. These algorithms are implementations of some of the most popular methodologies used in data mining. All of the Microsoft data mining algorithms can be customized and are fully programmable using the provided APIs, or by using the data mining components in SQL Server Integration Services.
You can also use third-party algorithms that comply with the OLE DB for Data Mining specification, or develop custom algorithms that can be registered as services and then used within the SQL Server Data Mining framework.

Analysis Services includes the following algorithm types:

- **Classification algorithms** predict one or more discrete variables, based on the other attributes in the dataset.
- **Regression algorithms** predict one or more continuous variables, such as profit or loss, based on other attributes in the dataset.
- **Segmentation algorithms** divide data into groups, or clusters, of items that have similar properties.
- **Association algorithms** find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- **Sequence analysis algorithms** summarize frequent sequences or episodes in data, such as a Web path flow.

6. Hardware and Software

- **Java** - Java is an object-oriented programming language developed by Sun Microsystems. Java is a platform-independent, multi-threaded programming environment designed for creating programs and applications for the Internet and Intranets.

- **JavaScript** - JavaScript is a scripting language developed by Netscape Communications designed for developing client and server Internet applications. Netscape Navigator is designed to interpret JavaScript embedded into Web pages. JavaScript is independent of Sun Microsystem's Java language.

- **Microsoft SQL Enterprise Edition**- The tool has inbuilt SQL server, Repository service, Integration Service, Analytical Services , SQL Database Management Studio. It's a tool which has both the ETL and the OLAP functionality with it.

- **Pl-Sql –** This is an Oracle commands which will be used to provide a flexible database activity in order to promote best achievements in obtaining a elegant working system.

- **Hardware's-** Windows 8 64 bit operating , 16 GB DDR3 Ram, i7 Intel Processor, 2 GB DDR5