**Contents**

|  |  |  |
| --- | --- | --- |
| Sl. No | Title | Page No. |
|  | Abstract | 8 |
|  | * 1. Problem Definition   2. Project Aim | 9  9 |
|  | Introduction | 12 |
|  | * 1. Project Introduction | 13 |
|  | System Analysis | 14 |
|  | * 1. Analysis of Project      1. Existing System      2. Proposed System   2. Methodology of the Study   3. Movie-Predictor      1. Modules      2. Timeline Chart      3. Cost Estimation | 15  15  15  16  18  19  23  24 |
|  | System Requirements | 25 |
|  | * 1. S/w and h/w requirements   2. Operating System   3. Technology | 26  27  28 |

|  |  |  |
| --- | --- | --- |
| Sl. No | Title | Page No |
|  | Design and Development | 31 |
|  | * 1. System Definition      1. Structural Chart   2. Entity Relationship Diagram   3. Dataflow Diagrams   4. System Design and Development   5. Input and Output Design   6. Logical Design   7. Database Design      1. Documents | 32  32  34  37  44  45  46  46  47 |
|  | Testing | 50 |
|  | * 1. System Testing      1. Unit Testing      2. Module Testing      3. Integration Testing | 51  51  51  51 |
|  | Implementation | 52 |
|  | System Maintenance | 53 |
|  | Future Enhancement | 54 |
|  | Conclusion | 55 |
|  | Bibliography | 56 |
|  | Appendix | 57 |

# ABSTRACT

## - PROBLEM DEFINITION

We have been using a lot of Online Platforms to watch movies and TV Shows. Out of them, prediction of movies and tv shows that a user wants to watch is one of the prominent features which help users to watch movies of their personal preference.

Movie Predictor is an application designed specifically for the purpose of predicting movies to users according to their tastes, previously watched movies, and their previous ratings given to movies. To achieve this, we analyse the available datasets and perform a set of AI filtering techniques to get the most precise prediction. There will be lot of users who share similar tastes in watching movies. We use those users’ data to characterise other users as they use the application. This helps users to find the most interesting movies in their genres, which they would like to watch. We have tested this over 100k records of multiple users which was very precise in predicting the movies which a user wants to watch in comparison with users with similar watch pattern.

## - PROJECT AIM

The project aims to provide precise list of movies and TV Shows a user wants to watch with the highest accuracy possible. Also, the AI system will learn according to the users’ current watch patterns and start suggesting movies according to the changes in their patterns such as change of genres, ratings and watch lists. Movie-Predictor will let users to sign-in with their Google Accounts providing easy logging in. Later, the app will ask users to select their preferences on genres, actors, directors, etc which they would like to watch. With the current available data and as a new user, the app will provide a list of movies to the user and from there, user can rate the movies, add movies to their watch bucket. All these actions will be used by our machine learning algorithm to predict the movies and shows for the user next time the machine learning algorithm trains the model. All of the user actions on the movies and shows list will be recorded and send to the backend to store and process later whenever prediction algorithm trains the model again.

Users will be provided with a movie and TV shows forum section where they can discuss about the corresponding movies. Such a way, the behavioural pattern of users will be analysed in relation with how they respond in the forums. Also, they can find more movies to their tastes from the forum discussions. The user’s usage pattern on the app will be recorded in the backend to give more accuracy to the machine learning algorithm. Such that, if a user rates a movie far away from their usual watch pattern, next time when the model trains, the algorithm will predict movies for that user biased to that newly rated movie as well.

In a lot of ways, the Movie-Predictor app will help users to watch and track their favourite movies and TV shows.

# INTRODUCTION

## - PROJECT INTRODUCTION

Movie-Predictor is a Movie Recommendation System with a Machine Learning Filtration Algorithm to predict the movies and TV Shows according to the rating and preference of a user towards a domain specific movie or genre. So, when the user provides some data about self, the data is processed against a wide variety of Movie Sets where the ratings, genre details are already pre-recorded. For which, we use the largest Movie Dataset from MovieLens to do the prediction.

Almost everyone today uses technology to stream movies and television shows. While figuring out what to stream next can be daunting, recommendations are often made based on a viewer’s history and preferences. This is done through machine learning and AI predictions. This is implemented in Python with data from the MovieLens Dataset. Generated by more than 6,000 users, MovieLens currently includes more than 1 million movie ratings of 3,900 films.

**Advantages**

Mostly the Recommendation system will be predicting movies for the user as per the past preferences of the user. User will be able to select the movies of their preferences, choices and according to that, the app dashboard will prefer more such kind of movies for the user every time the machine learning model is trained again and again according to their behavioural and usage patterns.

But here, we will use a combined filtering algorithm with user-based collaborative filtering and content-based genre filtering where we will compare the user preferences with other users’ behaviour which are already recorded in the data set and is evolving as per their suggestions and usage of the app and the feature and characteristics of the movie and its genre.

# SYSTEM ANALYSIS

## – ANALYSIS OF PROJECT

### – Existing System

Currently in the industry, we do have a lot of OTT Platforms which provide us with the movies and tv shows that we would like to watch. Though, the algorithms used by such platforms collect a lot of user data to provide with best results. They are a completely biased to content-based algorithms. When we login to such websites and we start watching movies, they will collect data on the movies we watch, time spend on such apps, screenshots when paused, etc. And later the algorithms use these data to predict the movies that a user would like to watch. The algorithm uses content-based filtering as such, the movies to be shown are chosen in a way how a user watches other same kind of movies. For example, Netflix uses NRE (Netflix Recommendation Engine) which records user interactions, information about movie categories, genres, title, year of release, other viewers with similar preferences etc. Currently, we don’t have an existing movie prediction app which is dedicated for predicting movies with feature-based and user-based algorithms collaboratively.

### – Proposed System

The Movie-Predictor is a fully fledged web app which helps users to watch the movies of their tastes, add them to watchlists to watch later, and also to share their views and opinions on movies on a common forum. We use a predictive algorithm which make use of a movie feature to predict similar movies and then the user features to predict the movies similar users would like to watch. To achieve this, we collect only minimal data from user inputs. And later these data are processed against the dataset available in the AI System which is from MovieLens, one of the biggest movie datasets. We run two cosine similarity algorithms, one for collaborative filtering to figure out similar users, and the other is to predict movies according to the features of one movie, such as genres, tags, etc. Both algorithms are regression algorithms with an RMSE (Root Mean Square Error) value of .82 and .85 respectively. Hence, they provide best results with big datasets.

## – METHODOLOGY OF THE STUDY

### Feasibility study

A feasibility study is made to see if the project on completion will serve the purpose of the organization for the amount of work, effort and the time that is spend on it. Feasibility study lets the developer foresee the future of the project and its usefulness. Feasibility study is a test of system proposed regarding its workability, impact on the organization, ability to meet the needs and effective use of resources

There are four aspects in the feasibility study portion of the preliminary investigation.

* Technical Feasibility
* Economic Feasibility
* Behavioural Feasibility
* Operational Feasibility

The proposed system must be evaluated from a technical point of view first, and if technically feasible their impact on the organization must be assessed. If compatible, the operational system can be devised.

### Technical feasibility

The system must be evaluated from the technical point of view first. In this we first check the availability of technical resources. Project uses technologies Python, JavaScript, MongoDB, and Google APIs which are free to use. As its open source and published in GitHub, anyone can simply fork the project and contribute. So, the project is technically feasible. The App is developed using NextJs, a production ready React Framework and FastAPI as the Asynchronous API framework based on Python. It is an SPA (Single Page Application) with responsive web designing on the frontend, hence it is compatible with mobile devices, and PCs.

### Economic feasibility

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return the cost it would require. In the development of this project there is no extra cost or efforts are required. In addition, we have created lot of reusable AI modules with can be used to reduce efforts in other Machine Learning Projects with similar application. So, the project is economically feasible. And it is open source and free for public.

### Behavioural feasibility

The present system is easily understandable. The maintenance and working of the new system need less human effort. The user interactions are as less as it can be and the features provided in app are for needful tasks only. All the behavioural aspects are considered carefully and have found the project is behaviourally feasible.

#### • Legal Feasibility study

In this privacy concern issues are considered. However, in this project we do not face any legal issues, because personal data like personal id proofs are not required or used in this system. We gather user data from Google APIs as a part of the Authentication which is already stated in Google Privacy Terms and as per user confirmation only.

#### • Scheduled Feasibility study

It is defined as the probability of a project to be completed within its scheduled time limits, by a planned due date. In this case the project is completed before the scheduled time.

### Operational feasibility

The project is completely operatable and implemented all sorts of Error Handlings inbuilt. All the features are working as expected, algorithms are training models with most accurate way and provides the user best predictions. Also, the machine will run the ML algorithm at an interval of 30 days. Considering all the above, the project is completely operational.

## – MOVIE-PREDICTOR

The Movie-Predictor aims at providing the best user experience to the user with collecting minimal app usage data. On the frontend, this is a Single Page Web Application which can be accessed over almost all IoT devices with a screen and web support. Movie-Predictor make use of Google Authentication APIs from Google Firebase for enhanced security. With the app, a user can login, view movies predicted by the AI System from the backend, read details about that movie, rate the movie, discuss about the movie and add reviews to the movie on the forums page, add movies to the watch list to watch them later.

We will use Firebase Authentication to enable access to the System. Firebase Authentication provides backend services, easy-to-use SDKs, and ready-made UI libraries to authenticate users to your app. It supports authentication using passwords, phone numbers, popular federated identity providers like Google, Facebook and Twitter, and more.

The Movie details which will be used for Prediction Algorithms is from the MovieLens Dataset. The module adds each movie to the database whenever a MovieLens adds a new movie to the dataset under supervision of the AI Systems Machine Learning Algorithm which tracks the Dataset changes.

Movie Prediction AI utilises MovieLens dataset to interpret and improvise the prediction using personalised algorithms which will provide most accurate predictions on which a user wants to or need to see next. According to that the system will suggest movies to the user, reviews and rating to user. The ML model will adapt itself as if we add more movies, reviews and ratings. We will use Collaborative Filtering and Content-based filtering technologies to implement the prediction system.

There is a forum for each movie predicted. Users will be able to add Reviews and ratings as per the movie they have watched. And it will provide a forum to discuss about the same. Users can personalize their preferences such as what kind of movies they are expecting, the movies they like to watch, genres, etc. Then we have a watchlist feature helps Users to add movies to watch later and they can remove once they have watched the movies.

### – MODULES

**TECHNICAL**

* Dataset preparation
* Read Data
* Apply ML
* Filter Recommendation
* Output generation
* Generate output

**USER**

* Sign Up
* Sign In
* List Movies
* Rate Movies
* Add Reviews/Comments
* Add/Remove movie to watchlist

#### Developing Solution Strategies

1. **Dataset Preparation**

We use python pandas data analysis library to read and perform preparation of our MovieLens Dataset. This step includes reading the dataset from our MongoDB Database and then converting it into ML identifiable data frames and then cleaning and processing the data.

* Understanding Parameters

We start off by understanding the required fields to build the ML training model. The required fields would be movieId, userId and rating for collaborative filtering and title and genres for content-based filtering algorithms respectively. We achieve this using pandas method *read\_json.*

* Cleaning

Since the database contains extra fields that are not required for the algorithms, we remove these fields and refine the received data using pandas dataframe cleaning methods such as *drop*.

1. **Approach**

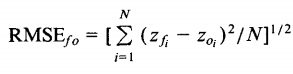
Machine learning algorithms in recommendation systems are typically classified into two categories — content based and collaborative filtering methods although modern recommenders combine both approaches, which can be seen in Movie-Predictor. Content based methods are based on similarity of item attributes and collaborative methods calculate similarity from interactions. From our requirement we chose to combine both collaborative method and Content-based method to achieve the best results. Collaborative methods work with the interaction matrix that can also be called rating matrix in our case since we are developing our first algorithm based on the ratings that are given by the user. The task of machine learning is to learn a function that predicts utility of items to each user. Content-based methods work with feature vectors and their similarity matrices which in turn used to train the model and perform filtering predictions on the selected features of a movie, such as its genre.

1. **Tool Selection**

For conversion of the BSON documents received from MongoDB into mathematical data frames we chose pandas method called *read\_json*. Pandas data frames are, arguably, the best way to easily carry out necessary refinement such as dropping columns, changing the datatype of values of particular columns etc. There are a lot of machine learning libraries available in python, and there are several recommendation-systems based libraries as well. We chose scikit-surprise as the best pick because of the simplicity, diversity and efficiency of the library. For content-based algorithm, we use the scikit-learn library to vectorize and perform similarity prediction of movies based on its features.

1. **Algorithm**

After refining data, we start by choosing the best algorithm that is suitable for our dataset. We ran the GridSearchCV method of surprise to compare different algorithms with a suitable accuracy measure. The accuracy measurement we are using here is RMSE - Root Mean Squared Error. RMSE is a measure of how spread-out residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Residuals are a measure of how far from the regression line data points are. The formula is:



Where:

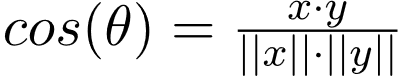
Σ = summation (“add up”)

(zfi – Zoi)2 = differences, squared

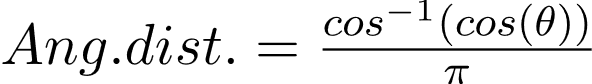
N = sample size.

The analysis led to two best suited algorithms namely, K-Nearest Neighbour (K-NN) and Matrix Factorization using Single Value Decomposition (SVD). Both these have an RMSE ranging from 0.8 to 0.9, which is very impressive. Hence, we consider the fit time of dataset which resulted in 32.0 for SVD and 0.07 for K-NN. Hence for better performance we chose K-NN.

The software creates a machine learning model using K-NN. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. The new case here is the new user and similarity is calculated between the movies rated by the using the ratings field as a deciding metrics. We are now creating a similarity matrix using the above algorithm. There are two choices to understand the distance (similarity metrics) between two datapoints in vector space namely Euclidean Distance and Cosine Similarity Matrix. We chose cosine similarity matrix since in vector space it gives more accurate measurements for our dataset. Cosine similarity between two vectors corresponds to their dot product divided by the product of their magnitudes. If x and y are vectors as defined above, their cosine similarity *cos(θ)* is:



The relationship between cosine similarity and the angular distance which we discussed above is fixed, and it’s possible to convert from one to the other with a formula:



After creating the similarity matrix, we will fit the dataset into the similarity matrix creating a similarity matrix with our dataset. This similarity matrix is the Machine Learning Model that we, in the future, use to predict movies for the user based on the user’s rating for related movies. In case of content-based algorithm, we do the same approach, but on the feature vector such that the algorithm generates the similarity based on the feature vector for movies, in our case, the movies with similar genres ad the best rating available will be returned.

1. **Filtering Outputs**

The model training will run every 30 days from the starting up of the server. So, each and every data from the database will evaluated for the changes and the training model and dataframes for the same will be generated using our machine learning algorithm and stored in the server using *pickle* methods provided by surprise library. Once these training models are created, we run our prediction algorithms over these training models to perform predictive filtering and return the outputs to the user on demand.

### – TIMELINE CHART

|  |  |  |
| --- | --- | --- |
| **System Analysis** | **Starting Date** | **Ending Date** |
| System Study | 20-04-2021 | 30-04-2021 |
| Requirement Analysis | 02-05-2021 | 09-05-2021 |

|  |  |  |
| --- | --- | --- |
| **Design Phase** | **Starting Date** | **Ending Date** |
| User interface Design | 16-06-2021 | 30-06-2021 |
| Coding | 02-07-2021 | 22-07-2021 |

|  |  |  |
| --- | --- | --- |
| **Database Design** | **Starting Date** | **Ending Date** |
| Physical database Design | 10-05-2021 | 05-05-2021 |
| DFD | 18-05-2021 | 20-05-2021 |
| Table Design | 30-05-2021 | 13-06-2021 |

|  |  |  |
| --- | --- | --- |
| **Testing and Implementation** | **Starting Date** | **Ending Date** |
| Unit Testing | 25-07-2021 | 30-07-2021 |
| Module Testing | 01-08-2021 | 08-08-2021 |
| Integration Testing | 09-08-2021 | 14-08-2021 |
| Implementation | 16-08-2021 | 24-08-2021 |

### – COST ESTIMATION

|  |  |
| --- | --- |
| **PURPOSE** | **COST** |
| Cost of the home page with whole design | **5000/-** |
| AI System | **20000/-** |
| Total Pages | **1** |
| Total Cost | **25000/-** |

# SYSTEM REQUIREMENTS

## – SOFTWARE AND HARDWARE REQUIREMENTS

### HARDWARE REQUIREMENTS

The selection of hardware is very important in the existence and proper working of any of the software. When selecting hardware, the size and capacity requirements are also important. The hardware must suit all application developments**.**

* Processor: i3 or above.
* System Bus: 32Bit or 64Bit
* RAM : 4 GB or Above
* HDD : 500 GB or Above
* Monitor: 14” LCD or Above
* Key Board: 108 Keys
* Mouse : Any Type of mouse
* Stable Network Connectivity

### SOFTWARE SPECIFICATION

One of the most difficult tasks is selecting software, once the system requirement is

find out then we have to determine whether a particular software package fits for those system requirements. This section summarizes the application requirement.

* Operating System: Windows 10 64-bit platform or above
* Front End: Next.js (JavaScript), FastAPI(Python)
* Back End: MongoDB
* IDE: VS Code, Python 3.10 or above, PyCharm
* Browser: Microsoft Edge, Google Chrome

## – OPERATING SYSTEM

### WINDOWS 10

Windows 10 is a series of personal computer operating systems produced by Microsoft as part of its Windows NT family of operating systems. It is the successor to Windows 8.1, and was released to manufacturing on July 15, 2015, and broadly released for retail sale on July 29, 2015. Windows 10 receives new builds on an ongoing basis, which are available at no additional cost to users, in addition to additional test builds of Windows 10 which are available to Windows Insiders. The latest stable build of Windows 10 is Version 1903 (May 2019 Update). Devices in enterprise environments can receive these updates at a slower pace, or use long-term support milestones that only receive critical updates, such as security patches, over their ten-year lifespan of extended support.

## 4.3 TECHNOLOGY

### Python

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

#### Pandas

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008. Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

#### Scikit-Surprise

Surprise is a Python scikit for building and analysing recommender systems that deal with explicit rating data. Surprise was designed with the following purposes in mind. Give users perfect control over their experiments. To this

end, a strong emphasis is laid on documentation, which we have tried to make as clear and precise as possible by pointing out every detail of the algorithms. Alleviate the pain of Dataset handling. Users can use both built-in datasets (MovieLens, Jester), and their own custom datasets. Provide various ready-to-use prediction algorithms such as baseline algorithms, neighbourhood methods, matrix factorization-based ( SVD, PMF, SVD++, NMF), and many others. Also, various similarity measures (cosine, MSD, pearson…) are built-in. Make it easy to implement new algorithm ideas. Provide tools to evaluate, analyse and compare the algorithms’ performance. Cross-validation procedures can be run very easily using powerful CV iterators (inspired by scikit-learn excellent tools), as well as exhaustive search over a set of parameters. The name SurPRISE stands for Simple Python RecommendatIon System Engine.

### JavaScript

JavaScript (often shortened to JS) is a lightweight, interpreted, object-oriented language with first-class functions, and is best known as the scripting language for Web pages, but it's used in many non-browser environments as well. It is a prototype-based, multi-paradigm scripting language that is dynamic, and supports object-oriented, imperative, and functional programming styles.

JavaScript runs on the client side of the web, which can be used to design / program how the web pages behave on the occurrence of an event. JavaScript is an easy to learn and also powerful scripting language, widely used for controlling web page behaviour.

Contrary to popular misconception, JavaScript is not "Interpreted Java". In a nutshell, JavaScript is a dynamic scripting language supporting prototype-based object construction. The basic syntax is intentionally similar to both Java and C++ to reduce the number of new concepts required to learn the language. Language constructs, such as if statements, for and while loops, and switch and try ... catch blocks function the same as in these languages (or nearly so).

JavaScript can function as both a procedural and an object-oriented language. Objects are created programmatically in JavaScript, by attaching methods and properties to otherwise empty objects at run time, as opposed to the syntactic class definitions common in compiled languages like C++ and Java. Once an object has been constructed it can be used as a blueprint (or prototype) for creating similar objects.

* **NextJs**

Next.js is a React framework that gives you building blocks to create web applications. By framework, we mean Next.js handles the tooling and configuration needed for React, and provides additional structure, features, and optimizations for your application.

### MongoDB

MongoDB supports various forms of data. It is one of the many nonrelational database technologies that arose in the mid-2000s under the NoSQL banner -- normally, for use in big data applications and other processing jobs involving data that doesn't fit well in a rigid relational model. Instead of using tables and rows as in relational databases, the MongoDB architecture is made up of collections and documents.

Organizations can use Mongo DB for its ad-hoc queries, indexing, load balancing, aggregation, server-side JavaScript execution and other features.

### Visual Studio Code

Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux. It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages and runtimes (such as C++, C#, Java, Python, PHP, Go, .NET).

### PyCharm IDE

PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains (formerly known as IntelliJ). It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCS), and supports web development with Django as well as data science with Anaconda. PyCharm is cross platform, with Windows, macOS and Linux versions. The Community Edition is released under the Apache License, and there is also Professional Edition with extra features – released under a proprietary license.

# DESIGN AND DEVELOPMENT

## – SYSTEM DEFINITION

### Structural Chart

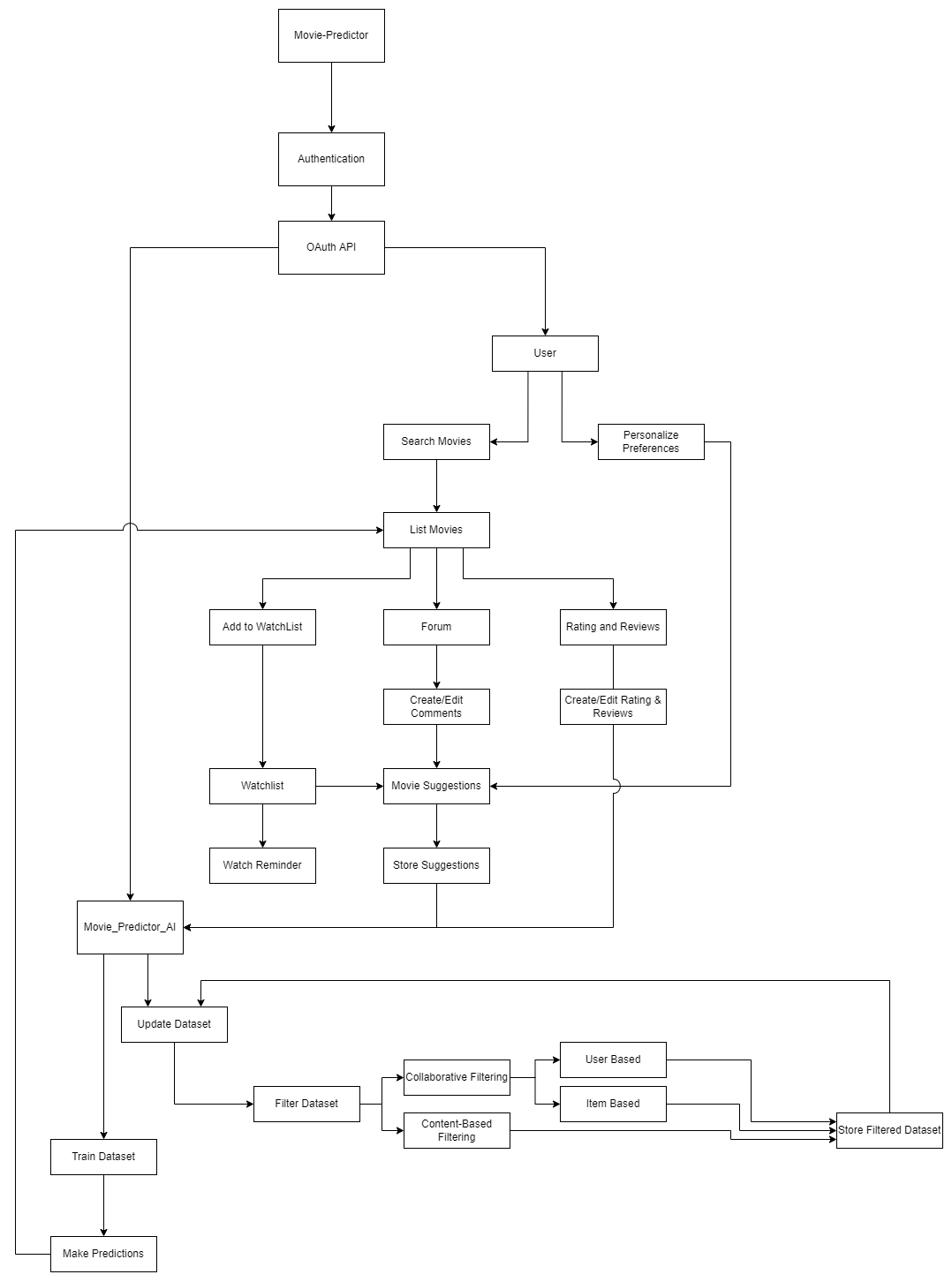
As well as a DFD, it is also useful to develop a structural system mode. This structural model shows how a function is realized by a number of other functions, which it calls. Structure charts are a graphical way to represent this decomposition hierarchy. Like DFD, they are dynamic rather than static system models. They show how one function calls others. They do not show a static bock structure of a function or procedure.

A function is represented on a structure chart as a rectangle. The hierarchy is displayed by linking rectangles with lines. Input and outputs are indicated with annotated arrows. An arrow entering a box implies input, a box implies output. Data stores are shown as rounded rectangles and user inputs as circles.

#### Rules to be applied

Many systems can be considered as three stages, input validation and output. If data validation is required, function to implement these should be subordinate to an input function. The role of function near the top of the structural hierarchy may be to control and coordinate a set of lower-level hierarchy. The objective of design process is to have loosely coupled highly cohesive components. Each node in the structure chart should have between two and seven subordinates.

**STRUCTURAL CHART**

****

## – ENTITY RELATIONSHIP DIAGRAMS (ERDs)

The E-R model was introduced by P.P Chen. Entity –Relationship modelling is a detailed, logical representation of the entities, associations and data elements for an organization or business area. This technique is used in database design, that helps to describe how entities in an enterprise are related to one another. The entity relationship model for data uses three features to describe data. These are the following

### Entities

An entity is a person, place, thing or event of interest to the organization and about which data are captured, stored or processed. For example, an Employee is an entity.

### Attributes

Various type of data items that describe an entity are known as attributes. For example, Name, address, DOB (Date of Birth) etc are attributes of the entity Employee.

### Relationship

An association of several entities in an Entity-Relation model is called relationship.

#### Name Symbol Meaning

Rectangle  Represents entity set

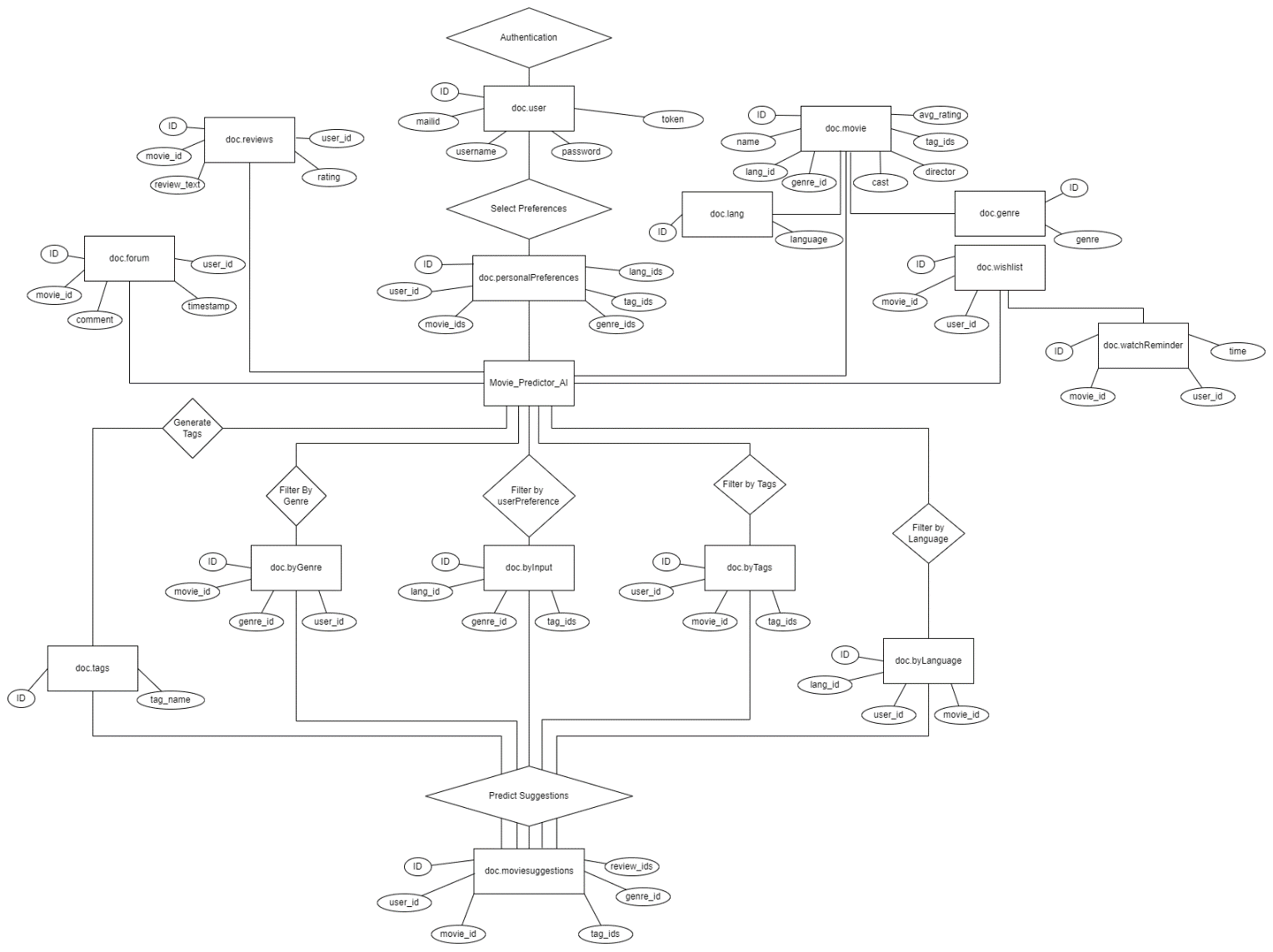
Oval Represents attributes

Diamond Represents among entity set.

Line Links attributes to entity set and entity set to Relationships.

While drawing the entity relationship (E-R) diagram, entity names are represented by a rectangle, relationships are represented by a diamond and oval shapes are used for representing attributes. Three types of relationships exist among entities. These are One to one (1:1) one to many (1: M) many to many (M: M). One to one relationship is an association only between two entities. One to many relationships exist when one entity is related to more than one entity.

### ER-DIAGRAM



## – DATAFLOW DIAGRAMS

Data Flow Diagram is a way of expressing system requirements in a graphical form. It has the purpose of identifying major transformation that will become programs in system design.

A Data Flow Diagram (DFD) or a Bubble chart is a graphical tool for structured analysis. DFD models a system by using external entities from which data flow to process, which transforms the data and creates, output-data-flows which go to other processes or external entities or files. There are various symbols used in a DFD. Bubbles represent the processes; Named arrows indicate the data flow. External entities are representing by rectangles and are outside the system such as vendors or customers with whom the system interacts. That either supply or consume data. Entities supplying data are known as source and those that consume data are called sinks. Generally, DFDs are used as a design notation to represent architectural design (External design) and top-level design (Internal design) specification. DFD‟s represents the system in hierarchical manner with one top level and many lower-level diagrams with each representing separate parts of the system. A DFD shows what kind of information about will be input to and output from the system, where the data will come from and go to, and where the data will be stored.

**To construct a dataflow diagram, the following symbols are used:**

**Arrow**

**Circle**

**Open End Box**

**Squares**

Five rules for constructing a data flow diagram

1.Arrows should not cross each other.

2.Squares, circles and files must bear names.

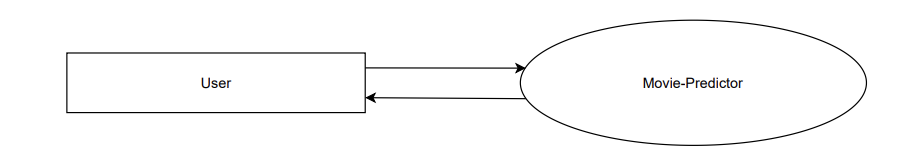
3.Decomposed data flow squares and circles can have same names.

4.Choose meaningful names for data flow.

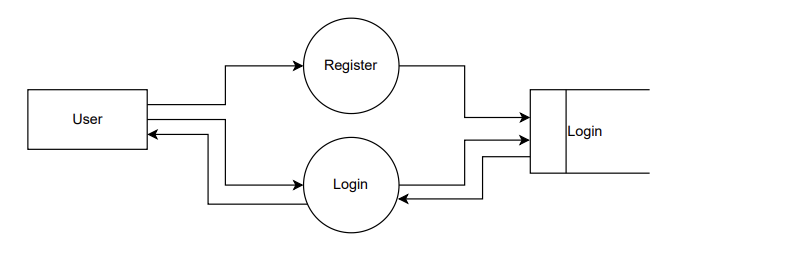
5.Draw all data flows around the outside of the diagram

**DFDs – Movie-Predictor**

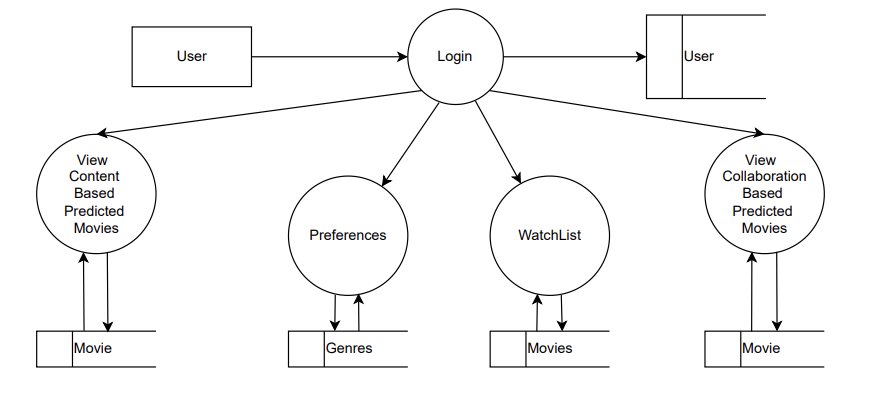
**Level 0**

****

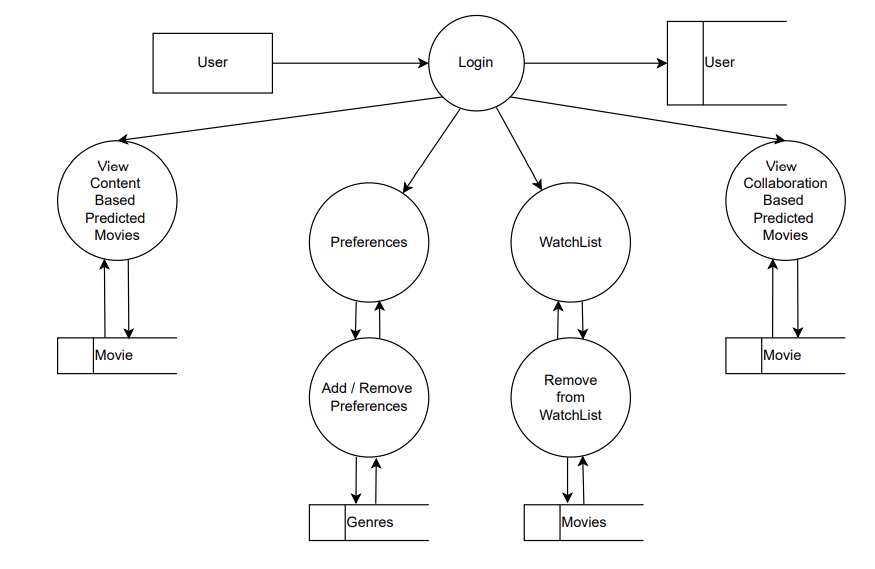
**Level 1**



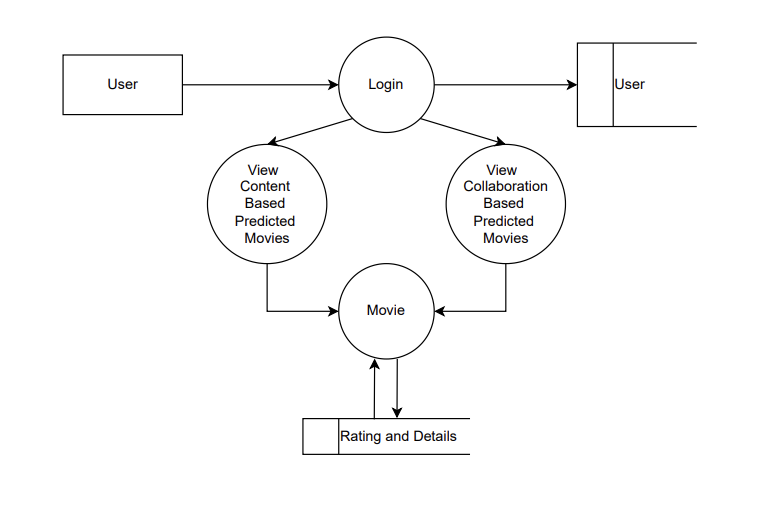
**Level 1.1**

****

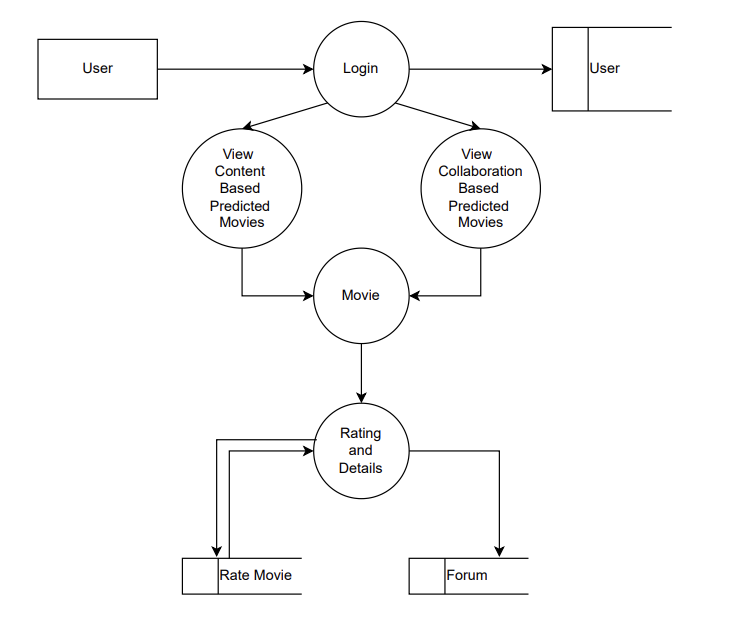
**Level 1.2**

****

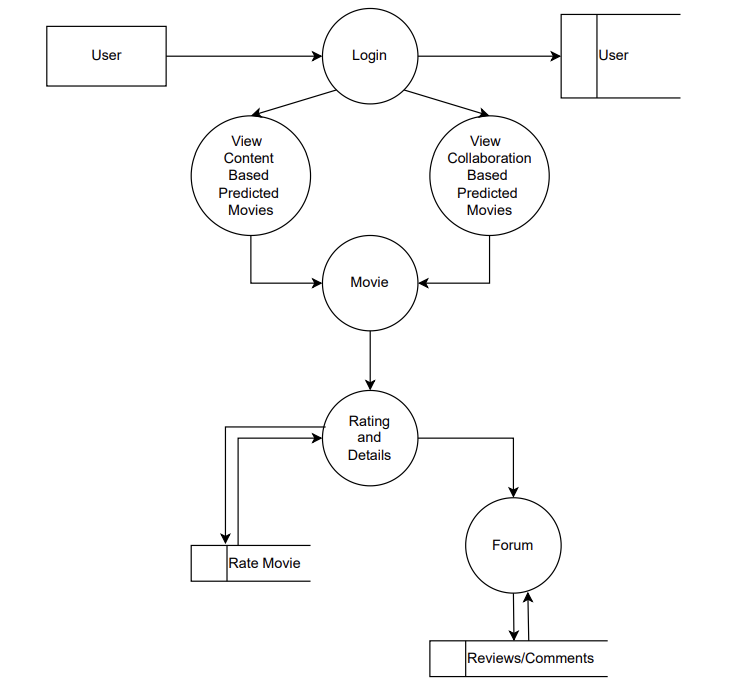
**Level 2**

****

**Level 2.1**



**Level 2.2**

****

## – SYSTEM DESIGN AND DEVELOPMENT

Design is the step in the development phase for every engineered product or system. Computer Software designing techniques like engineering design approach in the other, changes continuously as new method, better analysis and broader understanding evolve.

System design involves translating information requirement and conceptual design into technical specification and general flow of processing. After the user requirements are identified, related information is gathered to verify the problem and after evaluating the existing system, a new system is proposed. The proposed system consists of various table, their maintenance and report generation.

### SYSTEM DESIGN

#### Architectural Design

Architecture is an overall design of the system. Architecture takes into consideration the overall working of the system. Large system can be decomposed into sub-systems that provide some related set of services. The initial design process of identifying this sub-system and establishing a framework for sub-system control and communication is called architecture design. Architecture design usually comes before detailed system specification. Architecture decomposition is necessary to structure and organize the specification. There is no generally accepted process depends on application knowledge and on the skill and intuition of the system architect.

## – INPUT AND OUTPUT DESIGN

**User**

The User can login to the system through Login Window. Later, user can view the predicted movies, view the watch list and preferences. From preferences, user can change the genres they would like to watch. Accordingly, the content based predictive algorithm predicts different set of movies. Then, from the list of movies, user can view the movie details. From there, user can add movie to watch list, rate the movie, and go to Movie Forum. From watchlist, User can remove added movies. In forum page, user can add reviews for the movie.

### Input Design

Input design indicates the conversation of the user-originated inputs into the computer represent able form. The first step in the system design is to design input within predefine guidelines. In input design, User oriented data are converted to a computer-based format. Input design is the link that ties the information system into the worlds of its users.

The goal of input design is to make data entry as easy, logical and free from errors as possible. Input data are collected and organized into groups of similar data, appropriate input media we selected for processing. It consists of developing specifications and procedures for entering data into a system and must be in a simple format. A form can be used to enter these details using „VB‟ tools such as command boxes, text boxes etc.

### Output Design

Once the output requirements are determined, the system designer can decide what to include in the system and how to structure it so that the required output can be produced designing computer output should proceed in an organized, well throughout manner; the right output element is designed so that the people will find the system executed. The usefulness of the system is evaluated on the basis of their output.

## – Logical Design

In the logical design, represents the dataflow diagram of the proposed system. A data flow diagram is a graphical representation that depicts information flow and transforms that are applied as data to move from input to output.

A dataflow diagram may be used to represent a system or software at any level of abstraction. DFD‟s can be partitioned into levels that represent increasing information flow and functional details.

A level 0 DFD, also called fundamental system model or a context model, represents the entire software element as a single bubble with input and output data indicated by incoming and outgoing arrows, respectively. Each of the process represented at level 1 is a sub function of the overall system depicted in the context model.

## – Database Design

The objective in the development of the database technology has been to treat data as an organizational resource and make information access easy, inexpensive and flexible for the user in the database design. A database an integrated collection of data and provides centralized access to data. The organization of data in the database aims to achieve two major activities. They are data integrity and data independence. The organization of data in the database aims to achieve the following objectives. We use a NoSQL Database called MongoDB. The data is represented in MongoDB as Binary Document Objects.

* Controlled redundancy
* Data independence
* More information in low cost
* Accuracy and integrity
* Recovery from failures
* Privacy and security
* Performance

### – Documents

1. Col\_mp\_user

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| username | string | Unique | Username |
| email | string | Unique | User email |
| password | hashed string | Not Null | Password |
| userId | int | Unique | User ID |

1. Col\_mp\_movie

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| movieId | string | Unique | Movie ID |
| title | string | Not Null | Movie name |
| genres | List(string) | No Constraints | Movie Genres |
| tmdbId | string | Unique | Movie TMDB ID |

1. Col\_mp\_forum

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| movieId | string | Unique | Movie ID |
| comments | List (Object) | No Constraints | Reviews and Comments |

1. Obj\_mp\_comments

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| userId | string | No Constraints | User ID |
| timestamp | string | Not Null | Timestamp of comment recorded |
| comment | string | Not Null | Comment |

1. Col\_mp\_preferences

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| userId | string | Unique | User ID |
| preferences | List(string) | No Constraints | Genre Preferences |

1. Col\_mp\_rating

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| userId | string | No Constraints | User ID |
| movieId | string | Not Null | Movie ID |
| rating | string | Not Null | Movie Rating |
| timestamp | string | Not Null | Timestamp of rating recorded |

1. Col\_mp\_watchlists

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Datatype | Constraints | Description |
| \_id | BinaryJSON | Unique Identifier | Document ID |
| userId | string | No Contraints | User ID |
| movieIds | List(String) | Not Null | Movie IDs |

# TESTING

## – SYSTEM TESTING

### System Testing

System testing is actually a series of different tests whose primary purpose is to fully exercise the computer-based system. Although each test has a different purpose, all work to verify that all system elements have been properly integrated and perform allocated functions.

#### Testing Methodologies

The following are testing methodologies used here:

* Unit testing
* Module testing
* Integration testing

#### – UNIT TESTING

Unit testing enables a programmer to detect error in coding. A unit test focuses verification of the smallest unit of software design. In this project each form’s code is checked uniquely, to check the functioning of both input and output forms.in each form an input is inserted then check if it’s stored in correct location and original output is produced.

#### – MODULE TESTING

Instead of testing whole software program at once, module testing recommends testing the smaller building blocks of the program. In this project we can perform module testing as, when User changes the preferences information, the AI system responses with different set of movies as per the currently selected preferences. Multiple processes are connected here, so each modules are checked to ensure the other modules are correct.

#### – INTEGRATION TESTING

In this testing, units or individual components of the software are tested in a group. The focus of the integration testing level is to expose defects at the time of interaction between integrated components or units. All modules are combined and run as whole project. It checks the project works properly.

# IMPLEMENTATION

Implementation is the stage of the project where the theoretical design is turned into a working system. At this Stage the main work load, the greatest upheaval and the major impaction the existing system shifts to the user department. If the implementation is not carefully planned or controlled, it can cause chaos and confusion.

Implementation includes all those activities that take place to convert from the old system to new one. The new system may be totally new, replacing an existing manual or automated system or it may be a major modification to an existing system. Proper implementation is essential to provide a reliable system to meet the organization requirements. Successful implementation may not guarantee improvement in the organization using the new system, but improper installation will prevent it.

The process of putting the developed system in actual system in actual use is called system implementation. This includes all those activities that take place to convert from the old system to the new system. The system can be implemented only after through testing is done and if it is found to be working according to the specification. The system personnel check the feasibility of the system.

The most crucial stage is achieving a new successful system and giving confidence on the new system for the user that it will work efficiently and effectively. It involves careful planning, investigation of the current system and its constraints on implementation, design of methods to achieve the changeover.

# SYSTEM MAINTENANCE

The maintenance phase of the software cycle is the time in which a software product performs useful work.

After a system successfully implemented, it should be maintained in a proper manner. System maintenance is an important aspect in the software development life cycle. The need for the system maintenance is for it to make adaptable to the changes in the system environment. There may be social, technical and other environmental changes, which affect a system, which is being implemented software product enhancement may involve providing new functional capabilities, improving user displays and mode of iteration, upgrading the performance characteristics of the system. So only through proper system maintenance procedures, the system may adapt to cope up with these changes.

Software maintenance is of course, far more than “finding mistakes “we may define maintenance by describing four activities that are undertaken after a program is released for use. The first maintenance activity occurs because it is unreasonable to assume that software testing will uncover all latent errors in a large software system. During the use of any large program, errors will occur and are handled automatically by the Error Handler in the system.

The second activity that contributes to a definition of maintenance occurs because of the rapid change that is encountered in every aspect of computing. Therefore, adaptive maintenance an activity that modifies software to properly interface with changing environment is both necessary and commonplace.

The third activity that may be applied to a definition of maintenance occurs when a software package is successful. As the software is used, recommendation for new capabilities, modification to existing function, and general enhancement are received from users. To satisfy request in this category, perfective maintenance is performed. This activity for the majority of all effort expanded on software maintenance.

The fourth maintenance activity occurs when software is changed to improve future maintainability or reliability or to provide a better basis for future enhancement. Often called preventive maintenance, this activity is characterized by reverse engineering and reengineering technique.

# FUTURE ENHANCEMENT

The application developed is designed in such a way that any further enhancement can be done with ease. The system has the capability for easy integration with other system. New modules can be added to the existing system with less effort.

* Facility for modifying user detail.
* More interactive user interface.
* Facilities for Backup creation.

# CONCLUSION

Movie-Predictor is a single page web application from the frontend while being and AI and ML API System running on the backend to provide the most accurate results by analysing the datasets available to it. The AI system is very efficient to provide the best predictive results as we have combined collaborative filtering algorithm and content-based filtering algorithm altogether. Also, we perform the Machine Learning and Model training to run every 30 days from the backed server start-up. The Movie-Predictor as an SPA, becomes very user-friendly as it provides Google Authentication, say no to usernames and passwords. On the other hand, from the backend, it becomes a powerful AI API which responds as fast to provide the best prediction with available datasets. Also, the regression algorithms are very efficient to provide a margin of 0.9 RMSE metric.

# BIBLIOGRAPHY

ONLINE REFERENCES

1. <https://pandas.pydata.org/>
2. <https://nextjs.org/learn/basics/>
3. [FastAPI (tiangolo.com)](https://fastapi.tiangolo.com/)
4. [Surprise · A Python scikit for recommender systems. (surpriselib.com)](http://surpriselib.com/)
5. [Conda — conda documentation](https://docs.conda.io/en/latest/)

# 12.APPENDIX

## 12.1 - FINAL RESULT

**Screenshots**