

## “Process Mining in Social Media: Applying Object-Centric Behavioral Constraint Models”

# A Project Report Submitted to

**Jawaharlal Nehru Technological University Anantapur, Ananthapuramu**

**in partial fulfillment of the requirements for the**

**Award of the degree of**

## BACHELOROF TECHNOLOGY

IN

**COMPUTER SCIENCE AND SYSTEMS ENGINEERING**

Submitted by

K.V.SAI HARINI 16121A1543

**B.PAVITHRA** **16121A1514**

**B.SWETHA REDDY** **16121A1513**

**B.KEERTHI CHANDANA** **16121A1512**

*Under the Guidance of*

**Dr.P.DHANALAKSHMI, M.Tech., Ph.D**

Associate Professor (SL)

Dept. of CSSE, SVEC



**Department of**

## Computer Science and Systems Engineering

Sree Vidyanikethan Engineering College (Autonomous)

Sree Sainath Nagar, Tirupati – 517 102 (2016-2020)

April, 2020



**SREEVIDYANIKETHAN ENGINEERING COLLEGE**

**(AUTONOMOUS)**

Sree Sainath Nagar, Tirupati

**DEPARTMENT OF COMPUTER SCIENCE AND SYSTEMS ENGINEERING**

## **CERTIFICATE**

This is to certify that the project report entitled

**“Process Mining in Social Media: Applying Object-Centric Behavioral  
Constraint Models”**

is the Bonafide work done by

<b>K.V.SAI HARINI</b>	<b>16121A1543</b>
<b>B.PAVITHRA</b>	<b>16121A1514</b>
<b>B.SWETHA REDDY</b>	<b>16121A1513</b>
<b>B.KEERTHI CHANADANA</b>	<b>16121A1512</b>

in the Department of **Computer Science and Systems Engineering**, and submitted to Jawaharlal Nehru Technological University Anantapur, Ananthapuramu in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Systems Engineering during the academic year 2019-2020. This work has been carried out under my supervision. The results of this project work have not been submitted to any university for the award of any degree or diploma.

***Guide:***

**Dr.P.DHANALAKSHMI**

Associate Professor

Dept. of CSSE

***Head:***

**Dr. C Madhusudhana Rao**

Professor & Head

Dept. of CSSE

**INTERNAL EXAMINER**

**EXTERNAL EXAMINER**

## **DECLARATION**

We hereby declare that the project report titled **“Process Mining in Social Media: Applying Object-Centric Behavioral Constraint Models”** is the genuine work carried out by us, in **B.Tech(Computer Science and Systems Engineering)** degree course of **JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR** and has not been submitted to any other college or University for the award of any degree or diploma.

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea / data / fact / source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature of the students

**K.V.SAI HARINI**  
**(16121A1543)**

**B.PAVITHRA**  
**(16121A1514)**

**B.SWETHA REDDY**  
**(16121A1513)**

**B.KEERTHI CHANDANA**  
**(16121A1512)**

## **ACKNOWLEDGEMENT**

I am extremely thankful to our beloved Chairman and Founder **Dr. M. Mohan Babu**, Padmasri awardee of SreeVidyanikethan Educational Institutions who took keen interest and encouraged me in every effort throughout this B. Tech Program.

I owe my gratitude to **Dr. P. C. Krishnamachary**, Principal, SreeVidyanikethan Engineering College for permitting me to use the facilities available to accomplish the Project work course successfully.

I express my heartfelt thanks to **Dr. C. Madhusudhanrao**, Professor and Head, Department of Computer Science and Systems Engineering, for his kind attention and valuable guidance to me throughout the Project work course.

I am thankful to our Seminar Coordinator **Dr. C. Sushama**, Associate Professor of CSSE for her valuable support and guidance throughout the Project work course.

I am extremely thankful to our Seminar Supervisor **Dr.P.Dhanalakshmi**, Associate Professor of CSSE department who took keen interest and encouraged me in every effort throughout the Project work course.

I am also thankful for all the teaching and non-teaching staff of Computer Science and Systems Engineering Department for their cooperation.

Signature of the students

**K.V.SAI HARINI**  
**(16121A1543)**

**B.PAVITHRA**  
**(16121A1514)**

**B.SWETHA REDDY**  
**(16121A1513)**

**B.KEERTHI CHANDANA**  
**(16121A1512)**

# **Abstract**

Internet is a network, which provides a variety of information and communication facilities, consisting of interconnected networks using standardized communication protocols. In internet social media plays an important role to ease the data transfer between people related to different fields and personal information. The pervasive use of social media (e.g., Facebook, Stack Exchange, and Wikipedia) is providing unprecedented amounts of social data. Data mining techniques have been widely used to extract knowledge from such data, e.g., community detection and sentiment analysis. However, there is still much space to explore in terms of the event data (i.e., events with timestamps), such as posting a question, commenting on a tweet, and editing a Wikipedia article. These events reflect users' behavior patterns and operational processes in the media sites. Classical process mining techniques support to discover insights from event data generated by structured business processes. However, they fail to deal with the social media data which are from more flexible “media” processes and contain one-to-many and many-to-many relations. Based on real-life data, process models are mined to describe users' behavior patterns. Object-centric behavioral constraint model helps to understand the user behavior. Conformance and performance are analyzed to detect the deviations and bottlenecks in the question and answer process in the Stack Exchange website.

# TABLE OF CONTENTS

<b><u>Title Page</u></b>	<b><u>No</u></b>
<b>Abstract</b>	<b>VI</b>
<b>Acknowledgements</b>	<b>VII</b>
<b>Table of contents</b>	<b>IX</b>
<b>List of figures</b>	<b>X</b>
<b>CHAPTER 1: Introduction</b>	
<b>1.1</b> Introduction to the topic	<b>1</b>
<b>1.2</b> Statement of the problem	<b>4</b>
<b>1.3</b> Scope	<b>4</b>
<b>1.4</b> Objectives	<b>4</b>
<b>1.5</b> Motivation	<b>5</b>
<b>1.6</b> Existing System	<b>6</b>
<b>1.7</b> Proposed System	<b>6</b>
<b>1.8</b> Limitations	<b>9</b>
<b>CHAPTER 2: Review on Literature</b>	<b>10</b>
<b>CHAPTER 3 Report on the present investigation</b>	
<b>3.1</b> Experimental Setup	<b>14</b>
<b>3.2</b> Procedure Adopted	<b>14</b>
<b>3.3</b> Methodologies Developed	<b>15</b>
<b>3.4</b> Physical Model	<b>18</b>
<b>3.5</b> Mathematical Model	<b>19</b>

<b>3.6 Simulation Model</b>	<b>20</b>
<b>3.7 Performance Evaluation</b>	<b>21</b>
<b>3.8 Analysis</b>	<b>22</b>
<b>3.9 Testing and Implementations</b>	<b>24</b>
<b>CHAPTER 4 : Results and Discussion</b>	
<b>4.1 Data Analysis</b>	<b>29</b>
<b>4.2 Results</b>	<b>31</b>
<b>4.3 Discussion</b>	<b>32</b>
<b>CHAPTER 5 : Conclusion and Future work</b>	
<b>5.1 Conclusion</b>	<b>34</b>
<b>5.2 Future work</b>	<b>34</b>
<b>5.3 Recommendation</b>	<b>36</b>
<b>CHAPTER 6 : Appendices and Screenshots</b>	
<b>6.1 Appendix 1</b>	<b>38</b>
<b>6.2 Appendix 2</b>	<b>54</b>
<b>6.3 Screenshots</b>	<b>48</b>

# List of Figures

<b>Figure no</b>	<b>Title</b>	<b>Page no</b>
Figure1.1	Procedure of Process Mining	2
Figure 1.2	Social Media Landscape	3
Figure 3.2.1	Physical model for process mining using OCBC model	18
Figure 3.6.1	Conformance checking	20
Figure 3.8.1	Two column charts based on window sizes	22
Figure 3.7.2	The distribution of the cardinalities	23
Figure 4.1.1	Engagement rate /Time graph for existing model	29
Figure 4.1.2	Conformance checking for a set of 10 questions	30



# Abbreviations

XES	eXtensible Event Stream
XOC	eXtensible Object Centric
OCBC	Object Centric Behavioral Model
AOC	Activity Over Class Relationship
HTML	Hyper Text Markup Language
CC	Conformance Checking

# Notations

$\mathbf{a}_i$	block address of reference $i$
$n$	length of vectors
$\text{cost}(\mathbf{v})$	accumulated cost for vector $\mathbf{v}$
$A$	artifacts
$f$	set of reference $id$
$D$	data
$I$	number of artifacts
$G$	graph
$T^m$	artifact table
$T_0$	base table
$T_{attr}$	set of attribute table
$C_{attr}$	set of attribute columns
$S$	artifact schema
$F_{link.}$	tables of the artifact in the form of foreign keys
$C_{ID}$	Column ID primary key

# CHAPTER 1

## Introduction

### 1.1 Introduction to the topic

In recent times, the world is changing into an internet dependent world where unlimited raw data is generated and stored in the unstructured format[1]. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use[9]. Data mining techniques have been widely used in social media to extract insights to improve business intelligence, provide better services and develop innovative opportunities[2][10]. Representative areas contain community detection, information diffusion, topic detection and monitoring, and sentiment analysis and opinion mining [10][4].

Data mining enables people to understand data from various aspects, there is still much space to explore in terms of the event data [3]. Events refer to user behavior in social media platforms and include a broad range of actions: joining a group, becoming friends with a person, posting a photo, sharing a link, updating a status, posting a question, commenting on a tweet, editing a wikipedia article, etc. Events play a significant role in any dynamic system and all these events have a timestamp associated with them[8]. They can be exploited using process mining techniques to derive process-related insights to reflect users behavior patterns and operational processes in social media platforms. Process mining is a relatively young research discipline that bridges the gap between machine learning and data mining on the one hand and process modeling and analysis on the other hand[7]. The aim of process mining is to automatically provide an accurate view on how the process is executed, by using historical facts as recorded by

the information system. The starting point of process mining is the observed behavior of process executions, stored in so-called event logs. Based on event logs, various process mining techniques can be employed to reveal insights. In general, these techniques can be organized into three categories: discovery, conformance and performance analysis, and enhancement which is shown in figure 1.1.

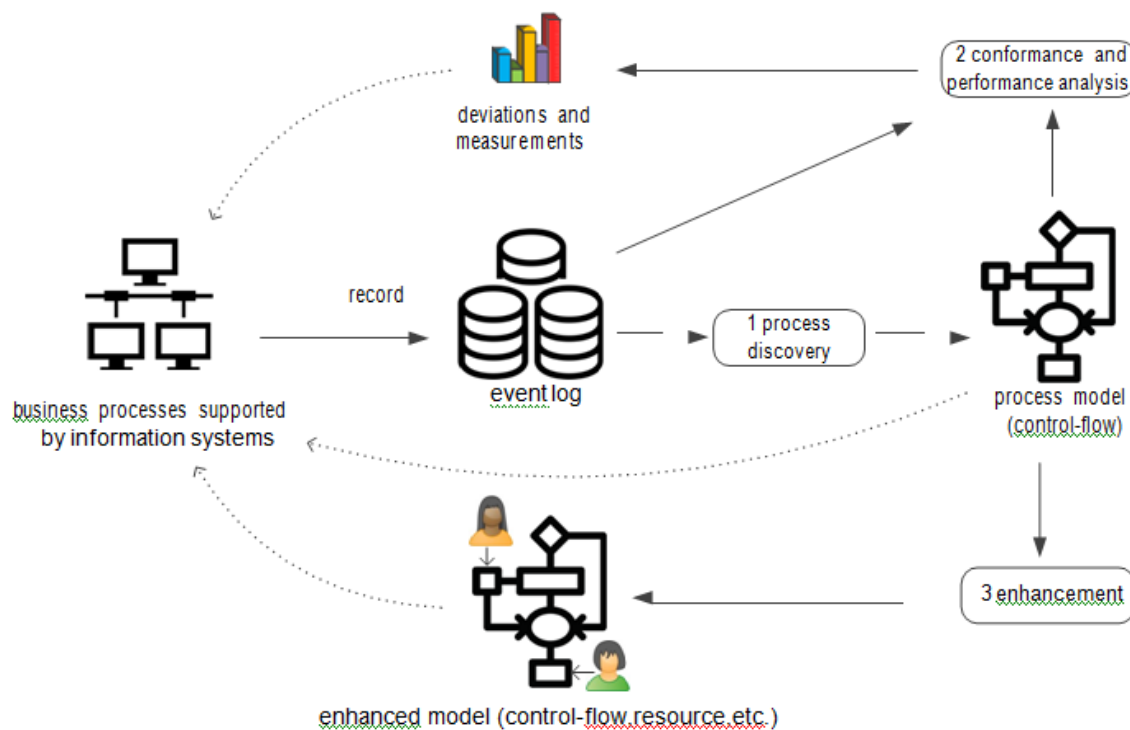


Figure1.1: Procedure of Process Mining.

Due to the large amount of data, social media has become an important source of information for understanding human behavior. It is critical for users, consumers, and service providers to mine social media to extract actionable patterns[8].

Social media is defined as a group of Internet-based applications that build on the ideological and technological foundations of the Web. It enables individuals to make, share, and additionally share data and thoughts[1]. Some outstanding social media websites are Facebook, Instagram, SnapChat, Stack Exchange, Quora, Wikipedia and Twitter. The ascent of web-based social networking has created extraordinary measures on social information[6]. Figure 1.2

explains how much information is made in one moment in various internet based life stages. For example, Facebook, which is the most dynamic of informal communities with over a billion months to month dynamic clients, makes the most measure of social information: clients post over 4.1 million for every moment. Instagram, with 300 million months to month clients in 2017, and secondly, photographs over 1.8 million likes every minute. It is critical for users, consumers, and service providers to mine social media to extract actionable patterns.

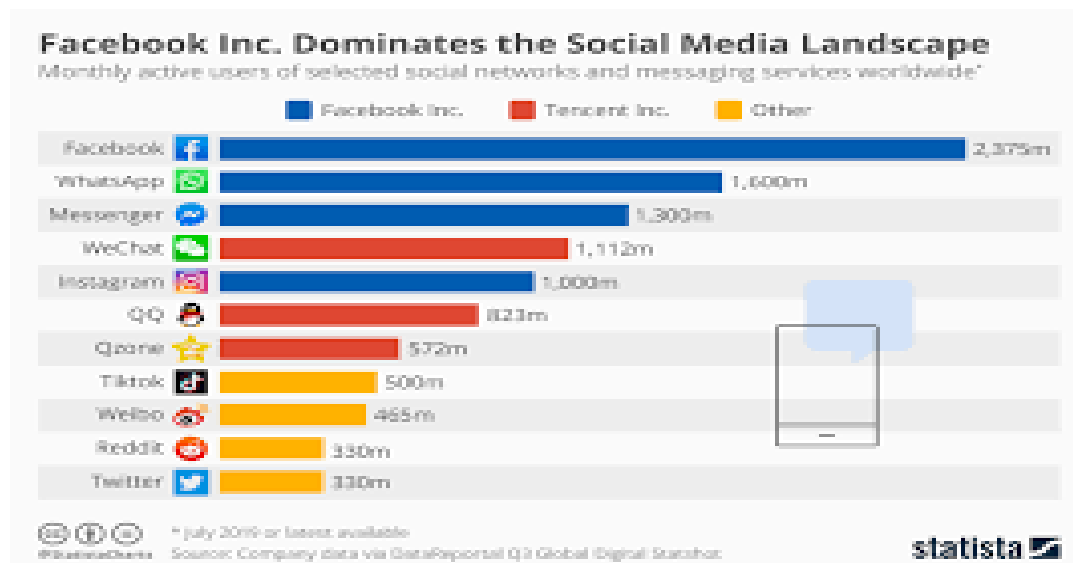


Figure 1.2: Social Media Landscape.

The data from social media is in different forms such as CSV files, database tables or XML files. Classical techniques fail to deal with the social media data which are from more flexible media processes and contain one-to-many and many-to-many relations [9]. The usage of process mining techniques, i.e. OCBC techniques, to real life data from social media, such as Facebook, Stack Exchange and Quora can discover the user's behavior patterns, e.g. Most liked answers given in one day after the corresponding questions are posted and a question is posted without any answer in Stack Exchange, and provide useful insights on the time perspective [3].

## **1.2 Statement of the problem:**

To provide the most desired and liked answers for a question posted in the web to make it easy for the users in the website like stack exchange. To improve the efficiency of discovering the user's behavior patterns, detect deviating and undesired behavior, provide useful insights on the time Perspective on social media data which reduces the ambiguity of providing correct answers to the users of platforms like Stack exchange.

## **1.3 Scope:**

We are using process mining and Object Centric Behavioral model together in order to extract the correct answers for the questions in websites like Stack Exchange which improves the time complexity. By combining these techniques will increase the automatic work flow model and to reconstruct the process as per the requirement according to the event logs. Using these we automated the process of extracting behavioral patterns and event logs accurately in the least possible time which leads to expressing answers according to the requirement. This helps in analyzing the data and providing an accurate and worst answer to the user to increase knowledge sharing.

## **1.4 Objectives:**

- To increase the usage of behavioral patterns in an effective way and to reduce data wastage and increase user readability.
- To improve the efficiency of platforms like Stack exchange which helps in sharing knowledge by developing this model we can provide the apt and most preferred solutions to the user.
- To decrease the time consumption while extracting the behavior patterns from the data.
- To give the most desired answers for the questions posted on the website in a very quick manner.

## **1.5 Motivation:**

- As there are more and more questions posted on social media, people would like to find more likeable answers in an effective way.

- The OCBC model is to give effective outputs or answers to the questions which will help the users like developers, researchers, customers and many others to get their results.
- Sharing and learning go hand in hand, sharing knowledge increases productivity and makes oneself work faster and smarter which helps in expertising and better problem solving.
- Worst answers have been taken into account which hinders knowledge sharing.
- Providing the most liked answers have become very important to reduce the contravencies.

### **1.6 Existing System:**

Due to the large amount of data, social media has become an important source of information for understanding human behavior. It is critical for users, consumers, and service providers to mine social media to extract actionable patterns. Data mining techniques have been widely used in social media to extract insights to improve business intelligence, provide better services and develop innovative opportunities. Representative areas contain community detection, information diffusion, topic detection and monitoring and sentiment analysis and opinion mining. But this process cannot be done automatically in the backend so the user should always collect data and replace it after following this technique in database.

### **Algorithm:**

- Collect data from social media website.
- The data should be parsed and converted into the relational databases after scoping the data based on a particular need.
- These databases were transferred into a XOC log files which can organize event data.
- Process mining techniques will be applied on XOC log file.
- The process will be discovered using object-centric behavioral constraint model.

**Merits:**

- Extraction of data becomes simpler while compared to usage of only data mining techniques.
- Generation of behavioral patterns became easy.
- Monitoring and tracking of process models is achieved.
- XOC log files extraction and event logs formation helped in sorting out the data properly.

**Demerits:**

- It is not automated method so user has to do it manually.
- Failed to notify deviations.
- Counting the activities (i.e., likes, answers, and comments) is inappropriate due to failure in extraction of required behavioral patterns.
- Conformance checking failed to provide appropriate results.
- Here they first parsed and imported into relational databases (after scoping the data based on a particular need) and then transformed into an XOC log that can better organize event data without a case notion.
- Since classical techniques fail to deal with the social media data which are from more flexible "media" processes and contain one-to-many and many-to-many relations.

**1.7 Proposed system:**

Events refer to user's behavior in social media platforms and include a broad range of actions: joining a group, becoming friends with a person, posting a photo, sharing a link, updating a status, posting a question, commenting on a tweet, editing a wikipedia article, etc. Events play a significant role in any dynamic system and all these events have a timestamp associated with them. They can be exploited using process mining techniques to derive process related insights to reflect users. This type of process mining techniques (based on *Object-Centric Behavioral*



*Constraint Models*) to analyze the data behavior patterns and operational processes in social media platforms. Compared with data mining techniques, our approaches can discover more complex user's behavior patterns, involving multiple activities, classes and interactions between them, rather than simple association rules. Besides, conformance checking can be applied to detect deviations or undesired behavior, e.g., a question without any answers in Stack Exchange. Moreover, performance can be analyzed to derive insights on the time perspective, e.g., most answers are given in one day after the corresponding questions are posted. Besides the described discovery and conformance checking techniques, we show how the performance analysis approach can be useful.

**Algorithm:**

1. Collect data from stack exchange website.
2. Include the Database.dmp file in Oracle database using SQL plus
3. Register user details in website
4. if detailsRemaining! = 0
5. Process mining starts Registration successful
6. end if
7. else
8. Registration failed
9. Go to step 3
10. if category = User
11. Access granted to View questions, comment and like
12. end if
13. if category = Admin
14. Access granted to Add and delete questions, View questions, View users
15. end if
16. if Question\_added >= 1 || Comment\_added >= 1 || like >= 1
17. Update database
18. end if
19. if like! = 0
20. Count ++

```
21. end if
22. else
23. Count --
24. while user! = 0
25. Number_of_users ++
26. user --
27. if user != 0 && like != 0
28. Engagement_Rate = Count/Number_of_users
29. end if
30. else
31. Engagement_Rate = 0
32. if Engagement_Rate > = Maximum_Threshold
33. Display green stars
34. end if
35. else
36. Display red stars
37. end
```

**Merits:**

- It is an automated working model which will not rely on human interaction.
- Conformance checking succeeded while compared to existing model.
- Discovery of more complex user's behavior patterns increased.
- No need to use the XOC log files as input direct database can be used.
- Reduced time complexity.

**Demerits:**

- Quality of an answer by using the process-related insights did not reach the required level.
- Natural language processing was not applied due to its complexity.

## 1.8 Limitations:

- Data from online social media platforms are dynamic, regular modifications and updates over a short period are not common but also a significant aspect to consider in dealing with social media data can be noisy.
- When you mine the social interaction, you try to get more information about the user than is visible at superficial layers.
- Incorrect information can also be the main drawback of data mining systems. When a user interacts on the social platform, they don't always assure us that they're being pristine in their thoughts.
- Even though there are many mining techniques it is impossible to provide 100% correct solution due to noise.
- Natural language processing is tougher.
- Due to enormous data the process of extracting behavior patterns leads to more complexity which makes human interaction necessary to solve them.
- Process mining and OCBC models alone cannot produce required data.

## CHAPTER: 2

### Review on Literature

W.M.P.vander Aalst, et.al, [1] has developed “**Declarativeworkflows**” for balancing flexibility and support in systems where workflow management system offer good process support as long as the processes are structured and do not require much flexibility. Information systems that allow for flexibility have a tendency to lack process-related support. Process-aware information systems (PAISs) support operational business processes by combining advances in information technology with recent insights from management science.

Workflow Management Systems (WFMSs) are such systems. WFMSs phases are Design time, Configuration time, Instantiation time, Run time, Auditing time. The declarative approach presented is based on constraints. Declareframework provides multiple constraint-based languages. Also it has different types of flexibility which became possible. A good example is an e-mail program like Outlook. The goal is to design processes that are correct.

However, the approach also has some short-comings. The limitation is that a constraint-based approach is not very suitable for processes that are of a strict procedural nature.K.Cai, et.al, [4] has developed a “**Sentiment Analysis for Topic Detection**” whereConsumers, non-profit organizations, and other forms of communities are extremely vocal about their opinions and perceptions on companies and their brands on the web. In particular, one important form can be derived from **sentiment analysis** on web content. Sentiment analysis traditionally emphasizes on classification of web comments into positive, neutral, and negative categories.

The sentiment classification is focusing on techniques that could detect the topics that are highly correlated with the positive and negative opinions. Such techniques, when coupled with sentiment classification, can help the business analysts to understand the overall sentiment scope. A detailed novel topic detection method used a point-wise mutual information and term frequency distribution. The effectiveness of all approaches known via several case studies on different social media data sets. D.Gruhl, et.al, [8] has developed “**Information Diffusion**” which is used for characterization and modeling topics.Using a large collection of weblogs for topic detection and tracking. Unlike traditional mechanisms for spreading information at the grassroots level, weblogs are open to frequent widespread observation, and thus offer an

inexpensive opportunity to capture large volumes of information flows at the individual level. The focus is on the propagation of topics from one blog to the next, based on the text of the weblog rather than its hyperlinks. Using this information, information diffusion can be done along with two dimensions: topics, individuals.

The topic model contributes to a solution for this problem by enabling us to identify subtopics that are experiencing spikes. Such an approach leverages the blogging community's reaction to external world events, as manifested by spikes in blog postings, to identify news events that are worthy of attention. But the process is limited to weblogs where some blog postings may be sufficiently important to merit notification, it can be difficult to identify the crucial posts in high-chatter topics as it can't handle large collection of data. G.Li, et.al, [11] has developed "**Object-Centric Behavioral Constraint Model**" to discover interacting instances and data dependencies between models where process discovery techniques have successfully been applied in a range of domains to automatically discover process models from event data. A novel modeling language which combines data models with declarative models. Moreover an algorithm been discovered to such models.

Object-Centric Behavioral Constraint(OCBC) modeling cardinality constraints to describe data and behavioral perspectives in a single diagram which overcomes the problems of existing data-aware approaches that separate the data and behavioral perspectives. An algorithm to discover OCBC models from object-centric event logs. Currently, the discovered models perfectly fit the source logs. But it failed to deal with larger scale logs in more complex scenarios, i.e., enabling the approach to discover compact models in a scalable manner (e.g., remove redundancies). Also using these OCBC models, many deviations cannot be detected. G. Li, et.al, [13] has developed "**Profiling**" which is a framework for detecting deviations in complex event logs. The Deviating behavior within an organization can lead to unexpected results. The effects of deviations are often negative, but sometimes also positive. However, existing model-based and cluster-based approaches are inaccurate or slow when dealing with complex event logs. A novel approach that is faster than cluster-based approaches because it creates a so-called profile which is less time-consuming than creating clusters. Deviation detection techniques can be divided into two categories, i.e., model-based approaches and cluster-based approaches.

Techniques in the first category basically employ conformance checking method on a discovered process model to detect deviations while the second one uses clustering to detect deviations. These techniques first mine an appropriate model as a reference model and then classify cases which do not fit the model as deviations.

A novel algorithm named profiling which is faster than cluster-based approaches and more accurate than model-based approaches. Clustering techniques are more suitable for complex processes. The profiling algorithm also has some limitations where the loops are not handled properly. X. Lu, et.al, [16] has developed “**Log Extraction**” for discovering interacting artifacts from ERP systems. The Enterprise Resource Planning (ERP) systems are widely used to manage business documents along a business processes and allow very detailed recording of event data of past process executions and involved documents. This recorded event data is the basis for auditing and detecting unusual flows. A semi-automatic end-end approach for analyzing event data in a database of an ERP system for unusual executions. This way, data divergence and convergence can be prevented. For the first time, a family of techniques to discover causal dependencies, between artifacts at the type level and at the event level. This information can be visualized as interactions between the extracted artifact life-cycle models.

For discovering artifacts still needs manual steps such as indicating a column for splitting the artifacts which is not handled using this automatic end-to-end approach. E.H.J. Nooijen, et.al, [18] has developed “**Extracting logs**” using mapping using data-centric and artifact-centric processes. Also there is a Process discovery technique that allows for automatically discovering a process model from recorded executions of a process as it happens in reality. This technique has successfully been applied for classical processes where one process execution is constituted by a single case with a unique case identifier. A structured data source R contains information about the events that have occurred in past process executions, usually it is in the form of timestamps written in the records of R. The complete approach combines a number of non-trivial, existing techniques for schema discovery, schema summarization, log extraction, and life-cycle discovery.

Technically, It is the first automatic discovery of schema-log mapping needed for log extraction. The technique currently focuses on discovering artifact life-cycles, but ignores artifact interactions which must be taken care to get accurate results.

## **CHAPTER 3**

### **Report on the present investigation**

The emergence of new social media such as blogs, message boards, news, and web content in general has dramatically changed the ecosystems of corporations. Consumers, non-profit organizations, and other forms of communities are extremely vocal about their opinions and perceptions on technologies, companies and their brands on the web. The ability to leverage such “voice of the web” to gain knowledge, consumer, brand, and market insights can be truly differentiating and valuable to today’s corporations. In particular, one important form of insights can be derived from sentiment analysis on web content. Sentiment analysis traditionally emphasizes on classification of web comments into positive, neutral, and negative categories.

It is critical for users, consumers, and service providers to mine social media to extract actionable patterns. Data mining techniques have been widely used in social media to extract insights to improve business intelligence, provide better services and develop innovative opportunities. Representative areas contain community detection, information diffusion, topic detection and monitoring and sentiment analysis and opinion mining.

Although data mining enables people to understand data from various aspects, there is still much space to explore in terms of the event data. In the approach used, they first parsed and imported into relational databases (after scoping the data based on a particular need) and then transformed into an XOC log that can better organize event data without a case notion. Since classical techniques fail to deal with the social media data which are from more flexible "media" processes and contain one-to-many and many-to-many relations.

The input data for OCBC techniques from databases (of artifact-centric information systems) can be very large. As a result, the current OCBC techniques still need to be improved to deal with the big data topic. In its current state, the performance analysis is time-consuming when the model is complex and there is no knowledge to identify some interesting correlation patterns, i.e., we have to compute the performance for all patterns.

### 3.1: Experimental Setup

#### Hardware Requirements:

Processor	:	Pentium IV
Hard Disk	:	500GB
RAM	:	2GB or more

#### Software Requirements:

Operating System	:	Windows XP/2003 or Linux
User Interface	:	HTML, CSS
Client-side Scripting	:	JavaScript
Programming Language	:	Java
Web Applications	:	JDBC, Servlet, JSP
IDE/Workbench	:	My Eclipse
Database	:	Oracle 10g
Server Deployment	:	Tomcat 7.X

### 3.2: Procedure Adopted

Process mining techniques can discover process-related insights, which can compensate data mining techniques on the time aspect. In comparison, our approaches can discover more complex users' behavior patterns involving multiple activities, classes and interactions between them rather than simple association rules. Besides, conformance and performance can be analyzed to detect deviations and bottlenecks. A process model can be viewed as a set of constraints.

Combining process mining and Object Centric Behavioral model techniques to solve problems which cannot be done only with either of them. For instance, in the Stack Exchange website, we can evaluate the quality of an answer by using the process-related and



time-related insights like considering the time after the corresponding question is posted and if the question receives “vote up” or “vote down” events before or after the answer discovered by process mining and content-related insights like to what extent the contents of the answers match the corresponding questions.

Using cardinality constraints restrictions can be provided for every iteration in posting question and answer in a website. Event logs will sort the data in database according to the time and date of event occurred. Using timestamps the data can be accessed and frequently updated in the database.

Deviation in activities can be tracked using conformance checking which can provide the result about the accuracy of process that occurs in particular time. It verifies the patterns and event logs that are being changed when the data is updated in the website. If there is a question without in answer is detected in the website it will automatically be deleted from website permanently which can remove the noise from the website. Following this whenever there is a deviation in event log it will be informed to administrator of website.

### **3.3: Methodologies Developed**

Mining in Social media is the process of obtaining big data from user-generated content on social media sites and mobile apps in order to extract patterns, form conclusions about users, and act upon the information. Stack Exchange is a great platform where user can share their knowledge about different fields in technology, sciences and industrial reviews. Using existing knowledge from the references the below methods are used:

**Artifact Type Level Interaction Discovery[13]**, is a semi-automatic end-to-end approach is used to identify artifacts in a relational data source and extract life-cycle event log for each identified artifact. A family of techniques been discovered for casual dependencies between artifacts at the type level and at event level. This can be viewed as interactions between extracted artifact life-cycle models (artifacts). This method helps for computing artifact type level interactions

### Artifact Type Level Interaction Discovery algorithm:

1. if  $D \leftarrow \emptyset$
2. call function ConstructInteractionGraph(A, F) /\* where A is artifacts and f is set of reference id \*/
3. **return**  $G = (A, D)$  /\* D refers to data \*/
4. **for**  $A > D - 1$
5. **do**  $I \leftarrow \emptyset$  /\* I refers to number of artifacts \*/
6. CalculatesInteraction(A, G) /\* G refers to graph \*/
7. **return** A

**Profiling method[16]**, which is faster than cluster-based approaches and more accurate than model-based approaches which are used to create a concrete approach for detecting deviations from specific perspectives that is control-flow perspective. Activities will be classified into normal cases from specific perspectives. In this method first event logs will be generated and then profile notions will be assigned to the cases.

### Profiling Algorithm:

- $a_i$ =block address of reference i
  - $n$ =length of vectors
  - $\text{cost}(\mathbf{v})$  = accumulated cost for vector  $\mathbf{v}$ , initially 0
1. **for each** reference I in program trace **do**
  2. **if**  $a_i$  is a compulsory miss **then**
  3.     push  $a_i$  to top of stack
  4. **end if**
  5. **else** /\*Count conflict vectors\*/
  6.     **for each**  $a_j$  on stack above  $a_i$  **do**
  7.          $\mathbf{V} = (a_i \oplus a_j)$  truncated to n bits increment  $\text{cost}(\mathbf{v})$
  8.     **end for**
  9.     Move  $a_i$  to top of stack

## 10. end for

**Process discovery**[18] is a technique that allows for automatically discovering a process model from recorded executions of a process as it happens in reality. Given a relational database that stores process execution information of a data-centric system, the technique extracts event information, case identifiers and their interrelations, discovers the central process data objects and their associated events, and decomposes the data source into multiple logs, each describing the cases of a separate data object. Then classical process discovery techniques can be applied to obtain a process model for each object.

### Process discovery Algorithm:

**Require:** A main artifact table  $T^m$ , base table  $T_0$  (with primary key  $C_{ID}$ ), a set of attribute table,  $T_{attr}$ , a set of attribute columns  $C_{attr}$  and an artifact schema  $S$ , links to other tables of the artifact in the form of foreign keys  $F_{link}$ .

1. collect data from table  $T^m$
2. **for all**  $T \in (T_{one2one} \setminus T_{attr})$  **do**
3.   **for all**  $C \in C_T \setminus C_{attr}$  **do**
4.      $AM \leftarrow \text{CreateAttributeMapping}(C)$
5.   **end for**
6. **end for**
7. **for all**  $T \in T_{one2many}$  **do**
8.    $\text{CreateMapping}(T^m, T, T_{attr}, C_{attr})$
9. **end for**
10. **return** general mapping item  $(C_{ID}, T_{from}, F_{link},)$

Besides, **conformance checking method** can be applied to detect deviations or undesired behavior, e.g., a question without any answers in Stack Exchange. Moreover, performance can be analyzed to derive insights on the time perspective, e.g., most answers are given in one day after the corresponding questions are posted.

By providing effective answers to the users can increase the popularity of website and knowledge sharing in wide range. Event logs normally record information about the users

executing the activities recorded in the log, so it is possible to extract social network from these logs for further analysis. The data will be mined using a sequence mining technique and stored in the Oracle database.

Now the key methods Process mining and Object Centric Behavioral Constraint model helps in extracting behavioral patterns and assigns event logs to the database. A Web User Interface is constructed which resembles the Stack Exchange website. Here the Users can create there account and access information. By using event logs behavioral patterns will be extracted and be uploaded to the database automatically in time. This will reduce the ambiguity in data and makes increases the accessibility.

### 3.4: Physical model:

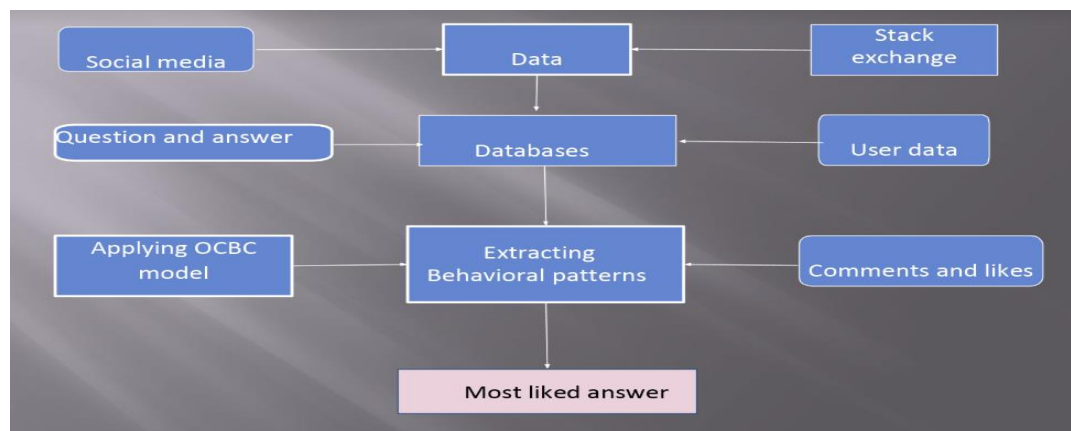


Fig 3.2.1:Physical model for process mining using OCBC model.

The required data will be collected from social media platform as we are considering Stack Exchange as the reference its data will b converted into a dump file. This data will be mined using a sequence mining technique and stored in the Oracle database. Now the key methods Process mining and Object Centric Behavioral Constraint model (OCBC) helps in extracting behavioral patterns and assigns event logs to the database. A Web User Interface is constructed which resembles the Stack Exchange website. Here the Users can create there account according to category like Admin and User where admin will have access to every detail of website liking posting and deleting questions updating the changes to website and user can comment and like

the question. The connection to database and Website will be provided using Servlet API. An algorithm is constructed to differentiate the data according to category in database like user name, admin name, question, answer, date and likes. Unlike other techniques this method can automatically calculate the number of likes and reduces time complexity by producing more number of behavioral patterns in the background. It can automatically delete deviated data like Question without answers according to a time period. These programs are constructed using java language in eclipse for better results.

### 3.5: Mathematical model

The engagement rate is a factor which is used to refer the level of interaction with followers that is generated from content created by a user. The engagement rate provides a more accurate representation of content performance than simply looking at absolute measures such as likes, shares, and comments. It is a popular equation used by most of the social media websites as it provides appropriate results while compared to other factors.

Formula for total engagement rate is as follows:

$$\text{Engagement rate} = \frac{\text{Total engagement}}{\text{Total followers}} * 100$$

#### Where:

- Total Engagement refers to the number of interactions (the measurement of which is dependent on the platform .For Stack exchange we are considering likes).
- Total Followers refers to the total amount of individuals that are following the page.

The engagement rate provides a more accurate representation of content performance than simply looking at individual absolute measures such as the number of likes, comments, shares, etc. It is a more comprehensive metric. It is a useful metric to use to

- Know the level of audience interaction.
- To gain insight into the quality of the content.

It is important to note that the makeup of “total engagement” can be altered in any way by the user who is using the metric. For example, the user may want the total engagement on Stack Exchange to include only the total amount of likes. Such alterations are valid if the user uses the same method of determining total engagement across all their calculations.

### 3.6: Simulation model

Simulation modeling is the process of creating and analyzing a digital prototype of a physical model to predict its performance in the real world.

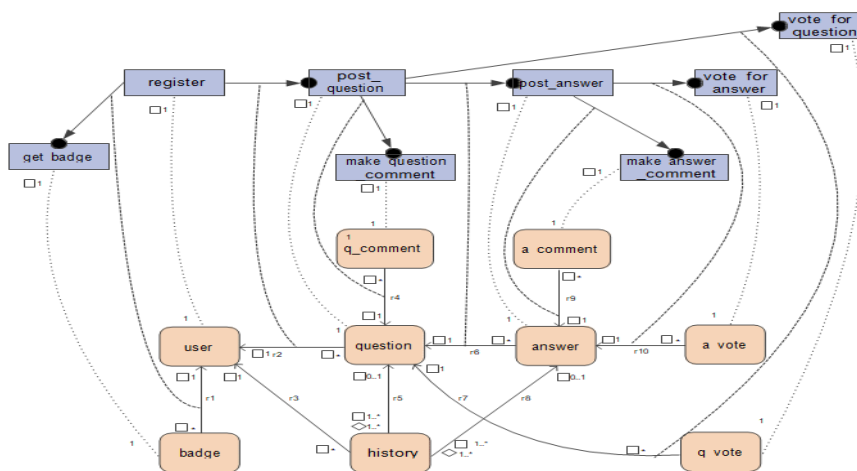


Fig 3.6.1: The discovered model is repaired to serve as the reference model for conformance checking. We add an eventually cardinality constraint from r1 to r6.

Conformance checking is a family of process mining techniques to compare a process model with an event log of the same process. Conformance checking techniques take as input a process model and event log and return a set of differences between the behavior captured in the process model and the behavior captured in the event log. We require the discovery technique to return a model where

- its fitness is 1,
- it only contains positive behavioral constraints and
- the discovered cardinalities are normalized

- ♦ represents Cardinality constraint
- □ Represents parent id
- Source: question
- Target: primary - answer, secondary - like

By taking r6 as an example the cardinality constraint 1 is discovered at its “question” side. It means that each answer always corresponds to precisely one question. Always, the cardinality constraint ♦1 is also discovered, which means that each answer eventually corresponds to precisely one question. It is omitted and are discovered at the “answer” side on r6 where is also omitted. They indicate that each question can correspond to any number of questions.

### 3.7: Performance Evaluation

Questions may have long life cycles. In other words, answers are still received even after a long time from the original question creation. This is indicated by the red dots which have long distances from their corresponding green dots.

The instances are sorted by timestamp. More precisely, the instances at the top of the dotted chart contain questions and answers posted at the moment.

There is an explosion of questions and answers just after the website was founded right after the website release, a quite steep green line curve and dense red dots can be noticed. The performance analysis on OCBC models can be split into independent analysis on so-called correlation patterns. Therefore, the most relevant pattern is the one which has “post question” as reference and “post answer” as target.

### 3.8: Analysis

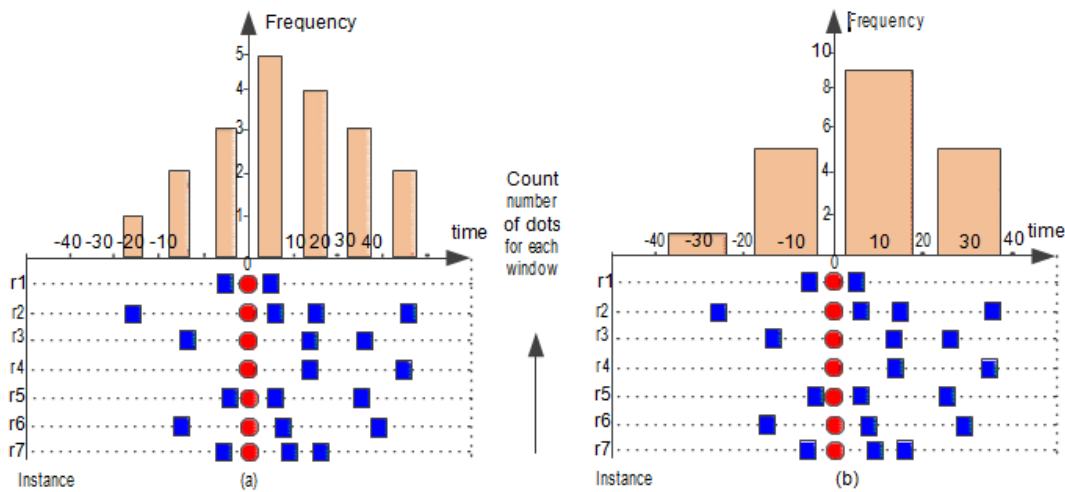


Fig 3.8.1: Two column charts based on window sizes of 10 (a) and 20 (b), respectively.

It shows how to derive a column chart, based on a dotted chart with relative time and a window size of 10 time units. More precisely, in the dotted chart each row depicts a pattern instance, where blue dots represent target events and red dots represent reference events. The horizontal (time) axis is divided into ten windows based on the given window size, and vertical axis has numbers to indicate the frequency of events in each window. In this example, we only focus on the distribution of target events, as reference events are always located at time zero in terms of relative time. Therefore we only count the blue dots (target events) for each window. For instance, there are five blue dots in the window (0, 10] and four blue dots in the window (10, 20].

Column charts can provide intuitive insights to users. The column chart in Fig 3.8.1 shows the frequency distribution of target events in terms of relative time. Apparently, there are more target events after reference events than those before reference events. Besides, target events most likely happen within 10 time units after the reference event. The probability decreases when the timestamps of events are getting far from the window (0, 10]. Based on the column chart, it is also possible to calculate the support of a customized window.

Where,



- Red dots indicate the reference events which are located at the time zero in terms of relative time.
- Blue dots are the target events for each set of questions.

Note that if the number of columns is very large, we can zoom out on the time and connect the top of each column, resulting in a line chart. The line chart lifts the information on a higher level, which can visually indicate insights, such as patterns and trends.

### Distribution of Cardinalities:

The discovered behavioral constraints between activities in Figure 9.6 specify the temporal restrictions on the behavioral perspective. For example, the “unary- precedence” constraint between “register” and “post\_question” activities indicates that someone is only able to post questions in the website after registering himself as a user. All these behavioral constraints describe the control-flow in a declarative manner. More precisely, after being registered, a user can post a question, and only then the question can be answered, commented or voted. Similarly, the provided answer can also be commented or voted. Besides, it is possible to get a badge based on one’s activities.

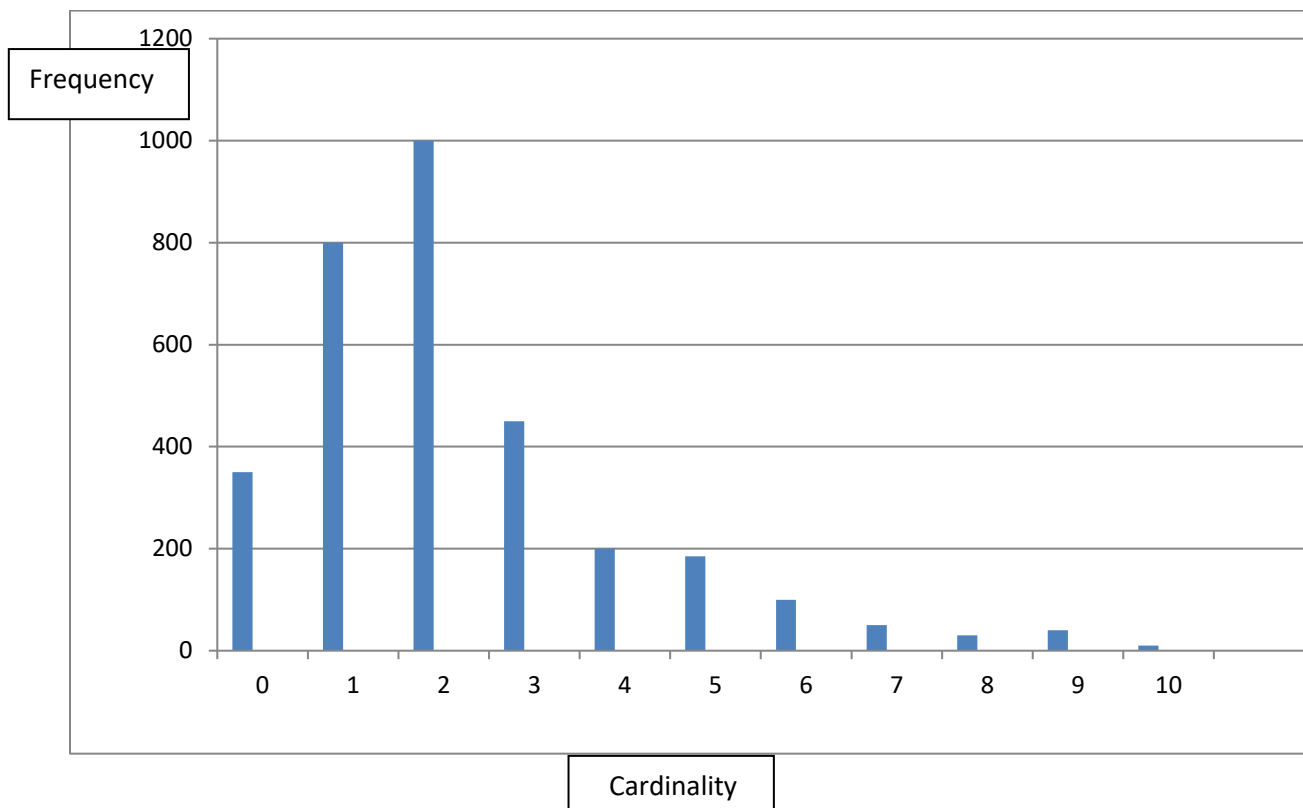


Fig 3.7.2: This diagram shows the distribution of the cardinalities of  $r_6$ , taking question as a reference. It indicates how many answers a question gets. The bar corresponding to the value “1” means that there are almost 1000 questions that received one answer.

In addition to discovering all behavioral constraints to describe the control-flow, our approach can also provide the cardinality numbers distribution of precedence/response target events in terms of a correlation pattern, by using the panel. The reference, target and intermediary drop-down menus in the panel indicate the reference activity, target activity and intermediary of the correlation pattern, respectively. The reference activity is “register”, the target activity is “post\_question” and the intermediary is “user-owneruserid- question” (corresponding to the class relation  $r_1$ ). By default, the distribution is for all reference events, in this case for all “register” events. It is possible to inspect the distribution for a particular reference event by setting specific instance values on the drop-down menu.

The discovered AOC relationships between activities and classes describe the constraints between events and objects. Here, all the discovered cardinalities on the relationships are 1 and 1, which indicate the one-to-one relation between events and objects. Thus, in the example, one “register” event corresponds to a “user” object and vice versa.

## 3.9 Testing and implementation

### Testing

White box testing is concerned with testing the implementation of the program. The intent of structural is not to exercise all the inputs or outputs but to exercise the different programming and data structure used in the program. Thus structural testing aims to achieve test cases that will force the desired coverage of different structures. Two types of path testing are statement testing coverage and branch testing coverage.

**Test Report:**

The module is working properly provided the user has to enter information. All data entry forms have tested with specified test cases and all data entry forms are working properly.

**Error Report:**

If the user does not enter data in specified order then the user will be prompted with error messages. Error handling was done to handle the expected and unexpected errors.

**TEST CASES:**

Test cases can be divided in to two types. First one is Positive test cases and second one is negative test cases. In positive test cases are conducted by the developer intention is to get the output. In negative test cases are conducted by the developer intention is to don't get the output.

**+VE TEST CASES**

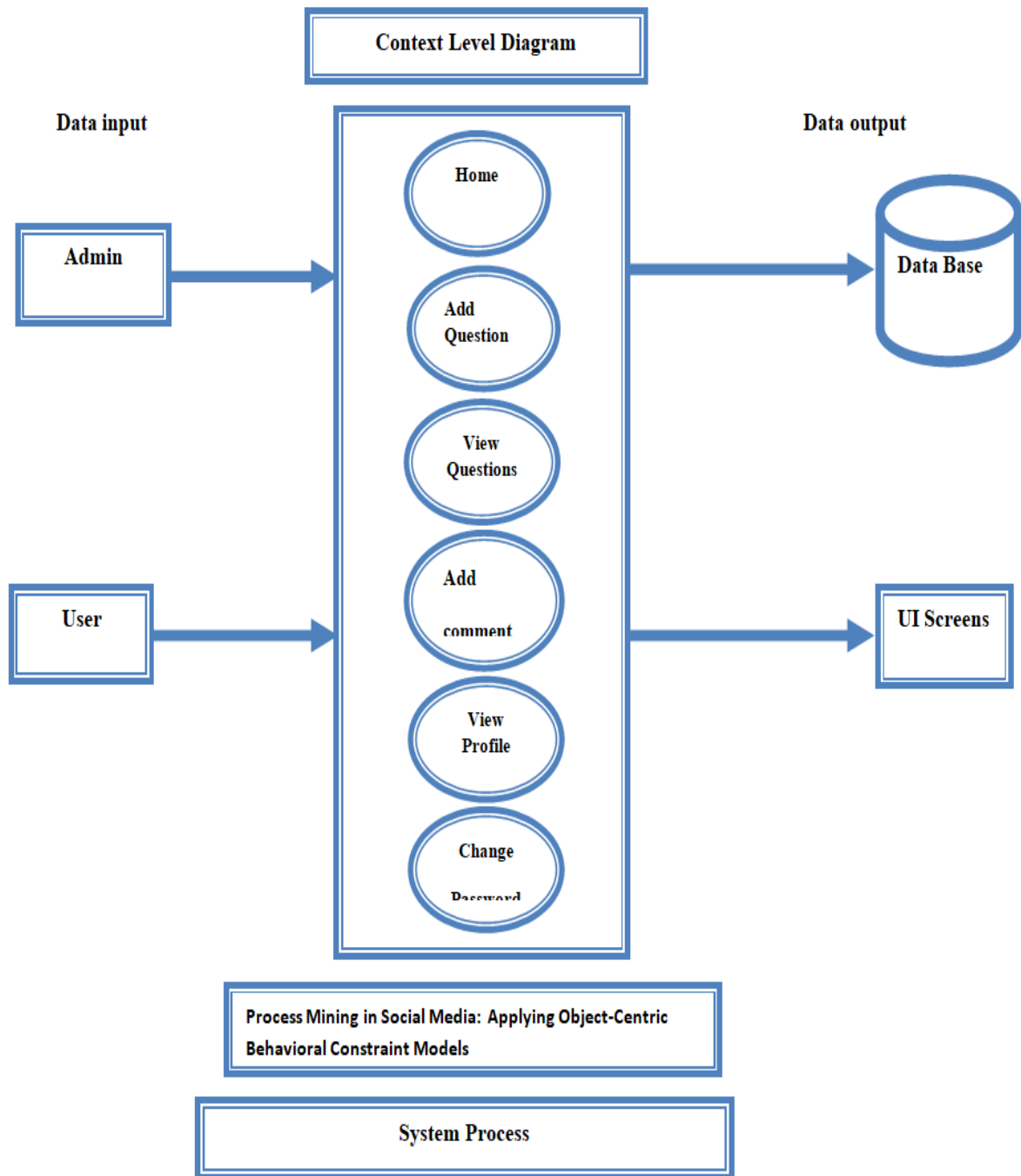
S .No	Test case Description	Actual value	Expected value	Result
1	Create new user registration process	Enter the personal info and address info.	Update personal info and address info in to oracle database successfully	True
2	Enter the username and password	Verification of login details.	Login Successfully	True
3	Add Questions	Enter all fields	Web data uploaded successfully	True
4	View All Questions.	Display All Data	Display All Added Questions	True

**-VE TEST CASES**

S .No	Test case Description	Actual value	Expected value	Result
-------	-----------------------	--------------	----------------	--------

<b>1</b>	Create the new user registration process	Enter the personal info and address info.	Personal info and address info its not update into database successfully.	False
<b>2</b>	Enter the username and password	Verification of login details.	Login failed	False
<b>3</b>	Add Questions	Enter all fields	Web data not added into Database	False
<b>4</b>	View All Questions.	Not Display added questions	No Data found	False

## Implementation:



Admin has to login with valid username and password. After login successful he can do some operations such as Add Questions, View Questions, view profile, and change password. Users will register before doing some operation. After registration successful he can login by using valid username and password. After login successful he can do some operations such as View Questions, Add Comments, view profile and change password.

The system is designed with completely automated process hence there is no or less user intervention. It is more reliable because of the qualities that are inherited from the chosen platform java. The code built by using java is more reliable.

This system is developing in the high level languages and using the advanced front-end and back-end technologies it will give response to the end user on client system with in very less time. It supports on a wide range of hardware and any software platform, which is having JVM, built into the system. It is implemented in web environment using struts framework.

The apache tomcat is used as the web server and windows xp professional is used as the platform. Interface the user interface is based on Struts provides HTML Tag. Inaccurate input data are the most common causes of errors in data processing. Errors entered by data entry operators can be controlled by the Input design. "Input design is the process of converting user originated inputs to computer based formats". It consists of developing specification and procedure for data preparation.

The main objectives of input design are:

1. Controlling amount of input: Due to so many reasons, design should control the quantity of data for input. Reducing the data requirement can lower cost by reducing labour expenses. By reducing input requirement, the analyst can speed the entire process from data capture to providing results to the users.
2. Avoiding delay: A processing delay resulting from data preparation or data entry operator is called bottleneck. Avoiding bottleneck should always be one objective of the analyst while designing output.
3. Avoiding errors in data: The rate at which errors occurs depends on the quantity of data, i.e. smaller the amount of data to input the fewer the opportunities for errors.
4. Keeping the process simple: Simplicity works and is accepted by the users. Complexity should be avoided when there are simple alternatives.

## CHAPTER 4

### RESULTS AND DISSCUSSION

#### 4.1 Data Analysis:

Process mining deals with the discovery of process models (i.e., structures that model behavior) from event-based data. The goal is to construct a process model which reflects the behavior that has been observed in some kind of event log. An event log is a set of finite event sequences, whereas each event sequence corresponds to one particular materialization of the process.

The quality of process mining can be measured using some key factors like Engagement rate, Behavioral patterns Cardinality constraints, Conformance checking and many other factors.

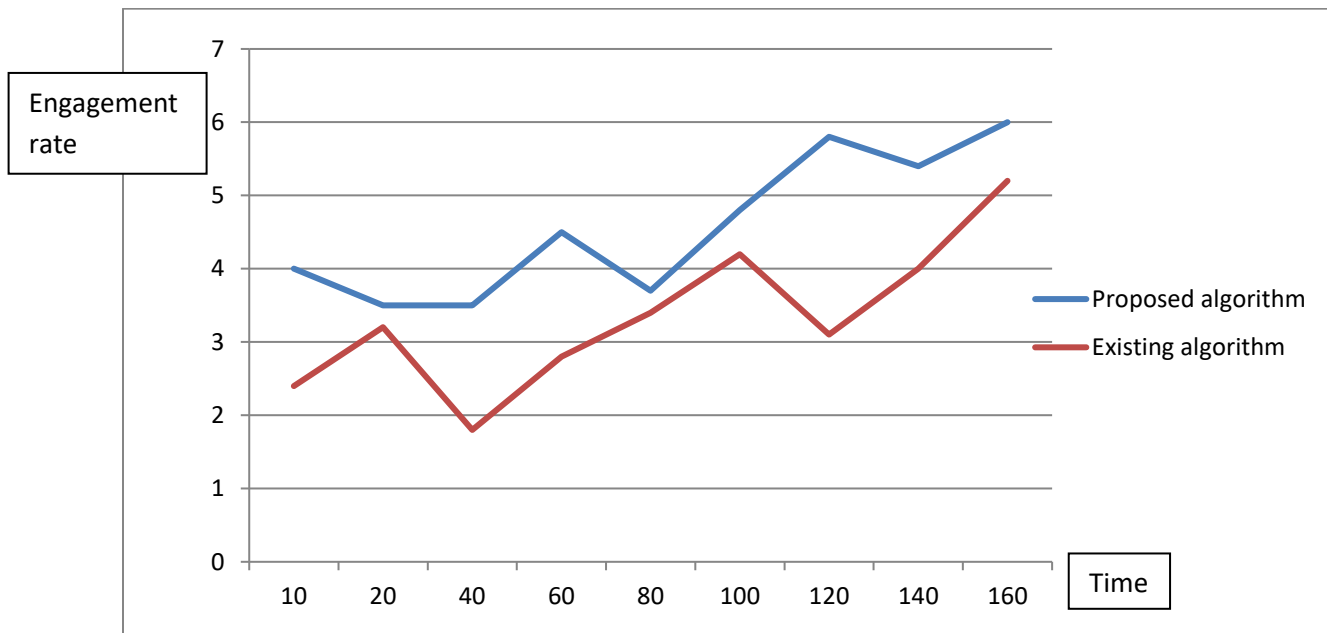


Fig 4.1.1: Engagement rate /Time graph for existing system and proposed System

The above graph shows the enhancement of high appropriate results obtaining from proposed algorithm than the existing algorithm. When process mining and Object Centric Behavioral Constraint Model is applied it increases the quality of extracting behavioral patterns.

An engagement rate is a metric that measures the level of engagement that a piece of created content is receiving from an audience. It shows how much people interact with the content. Factors that influence engagement include user's comments, shares, likes, and is a helpful metric to evaluate in a marketing competitive analysis.

The engagement rate was being calculated appropriately using proposed system in a more effective which reduced the time complexity.

### Conformance checking:

Conformance between a log and a model is difficult, since it is characterized by many dimensions. In traditional process mining, fitness, simplicity, precision, and generalization are used to evaluate the quality of a model discovered from a log. Actually, these criteria also reveal the conformance between the log and the discovered model, since a discovered model has good quality if it conforms to the log. Note that among these criteria, the simplicity dimension only indicates the complexity of the discovered model, and it is not related to behavior in the log. Therefore, in this section, we abstract from simplicity and use fitness, precision and generalization to quantify the conformance.

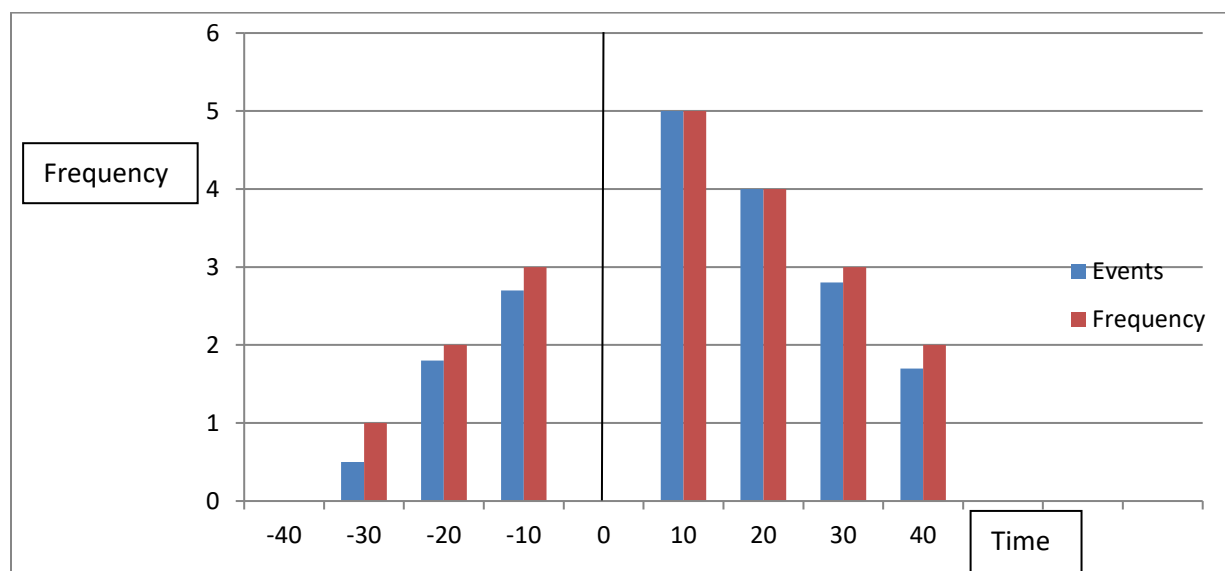


Fig 4.1.2: Conformance checking for a set of 10 questions.

Based on the idea of checking conformance checking using patterns as follows



- split the task of quantifying conformance on the whole process into several tasks of quantifying conformance on all correlation patterns in the process, and here we considered set of 10 questions as a quantity factor .
- Integrate the results of all correlation patterns into the whole criteria using number of events generated.
- In order to check the quality time frequency was considered.
- Identify the time window according to set of timestamps. Here a time window is a duration of time starting from the current moment and extending into the past.

After identifying time windows, we can build a column for each window. The height of the column indicates the number of events in the corresponding window. Note that filtering techniques can be employed when counting events in a window. For instance, we can set a restriction on an attribute of events such that only events that have required attribute values are counted. In this way, the column chart is more flexible to present the frequency distribution of different kinds of events.

## **4.2: Results**

Usage of process mining and Object Centric Behavioral model (OCBC) reduces the noise generation and increases the rate of quality in extracting behavioral patterns and events. They help in sorting the data in database according to the time. Events play a significant role in any dynamic system and all these events have a timestamp associated with them. They can be exploited using process mining techniques to derive process related insights to reflect type of process mining techniques based onto analyze the data' behavior patterns and operational processes in social media platforms. Compared with data mining techniques, our approaches can discover more complex users' behavior patterns, involving multiple activities, classes and interactions between them, rather than simple association rules. Besides, conformance checking can be applied to detect deviations or undesired behavior, e.g., a question without any answers in Stack Exchange. Moreover, performance can be analyzed to derive insights on the time perspective most answers are given in one day after the corresponding questions are posted. Besides the described discovery and conformance checking techniques, we show how the performance analysis approach can be useful. Time complexity is reduced due to usage of

Cardinality constraints as they restrict the path of execution when the process deviates from the original path.

### 4.3: Discussion

In general our work covers most types of process mining extracting event logs from execution data, discovering models from event logs, and checking conformance and analyzing performance based on logs and models. But there are some enhancements that cannot be achieved with this procedure.

- **EVENT LOGS AND EXTRACTION** - The notion of object- centric event data to abstract the data generated by artifact-centric information systems. Such data are different from case-centric event data and have their own features. In order to organize such data, we can use the event log format which does not require the case notion. Besides, an approach was proposed to automatically extract event logs from object-centric event data.
- **OCBC MODELS AND DISCOVERY** - We stated OCBC models by explaining all the involved elements: the data perspective, e.g., classes; the behavioral perspective, e.g., activities; and then the AOC relationships combining the first two. Approaches to automatically discover OCBC models from event logs are developed. First, a basic approach was illustrated to discover models from clean logs. Then, a more robust discovery approach was proposed to deal with noise in real life data.
- **OCBC CONFORMANCECHECKING**- Based on event log and manually designed or discovered OCBC model techniques to diagnose the conformance between them. By taking the data perspective into consideration, it is possible to detect and diagnose a range of conformance problems that would have remained undetected by conventional approaches. The diagnostic results (i.e., deviations related to the behavioral perspective, data perspective and inter- actions) were present in three different views: rule view, log view, and model view. Besides, metrics such as fitness, precision and generalization were defined to quantify the degree of conformance.
- **OCBC PERFORMANCE ANALYSIS**- Analyzed the performance in terms of frequencies and times by taking an event log and an OCBC model as input. In order to show performance results from different angles, (dotted and column) charts, and indicators

were used. With the obtained results in hands, it was possible to map them onto the model, and from this step important bottlenecks could be revealed.

## CHAPTER 5

# Conclusion and Future work

### 5.1: Conclusion

We applied a novel family of process mining techniques, i.e., OCBC techniques, to real life data from social media, such as Stack Exchange. The experiment shows that we can discover the users' behavior patterns, e.g., two patterns involving post, comment, answer activities in Stack exchange site, detect deviating and/or undesired behavior, e.g., a question is posted without any answer in Stack Exchange, and provide useful insights on the time perspective, e.g., most answers are given in one day after the corresponding questions are posted. The experiment shows that we can

- Discover the users' behavior patterns
- Detect deviating and/or undesired behavior, e.g., a question is posted without any answer in Stack Exchange, and
- Provide useful insights on the time perspective, e.g., most answers are given in one day after the corresponding questions are posted.

In order to provide the best knowledge to the users these techniques will help the website holders to automate the procedure of extracting data in background. Time complexity and human dependency will be reduced to great extent. By taking the data perspective into consideration, it is possible to detect and diagnose a range of conformance problems that would have remained undetected by conventional approaches. The diagnostic results were present in three different views: rule view, log view, and model view. Besides, metrics such as fitness, precision and generalization were defined to quantify the degree of conformance.

### 5.2: FUTURE WORK

We can combine process mining and data mining techniques to solve problems which cannot be done only with either of them. For instance, in the Stack Exchange website, we can evaluate the quality of an answer by using the process- related and time-related insights (e.g., considering the time after the corresponding question is posted and if the question receives “vote up” or “vote

down'' events before or after the answer) discovered by process mining and content-related insights (e.g., to what extent the contents of the answers match the corresponding questions) discovered by data(text) mining.

### **5.3: Scope for Future work**

**Semantics and reasoning on OCBC models:** Crucial reasoning tasks that exist for Declare and the like (such as consistency, dead activities, etc.) are not discussed in the context of OCBC models in this thesis, which may lead to potential problems. For instance, when dealing with noise, it is well known that as soon as constraints with support less than 100% are retained, the algorithm could produce a final inconsistent model. In future, these tasks should be taken into consideration to make our techniques more robust.

**Model patterns:** It is possible to identify typical behavioral PATTERNS that involve multiple instances or interaction between structure and behavior. Along this line, we plan to study the effect of introducing subtyping in the data model, a constraint present in all data modeling approaches. The interplay between behavioral constraints and subtyping gives rise to other interesting behavioral patterns.

**Distributed processing:** It is promising to relate process mining to Big Data technologies. In order to solve the limitations of OCBC techniques when dealing with large scale data, we can investigate how to incorporate the parallel computing approaches into these techniques. For instance, since OCBC models correlate events and analyze performance based on individual correlation patterns rather than the whole process, the performance analysis on the whole model can be split into independent smaller tasks for all correlation patterns. Similarly, the parallel manner can be also applied to log extraction and model discovery.

## REFERENCES

- [1] W. M. P. van der Aalst, M. Pesic, and H. Schonenberg, “Declarative Workflows: Balancing between flexibility and support,” *Comput. Sci.-Res. Develop.*, vol. 23, no. 2, pp. 99–113, May 2009.
- [2] E. Baatarjav, S. Phithakkitnukoon, and R. Dantu, “Group recommendation system for Facebook,” in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.*, Nov. 2008, pp. 211–219.
- [3] G. Barbier and H. Liu, “Data mining in social media,” *Social Netw. Data Anal.*, vol. 17, pp. 327–352, Mar. 2011.
- [4] K. Cai, S. Spangler, Y. Chen, and L. Zhang, “Leveraging sentiment analysis for topic detection,” *Web Intell. Agent Syst., An Int. J.*, vol. 8, no. 3, pp. 291–302, Jan. 2010.
- [5] L. Cheng, B. F. van Dongen, and W. M. P. van der Aalst, “Efficient event correlation over distributed systems,” in *Proc. 17th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput.*, May 2017, pp. 1–10.
- [6] S. E. Community. (2019). Stack Exchange Data Dump. [Online]. Available: D. Fahland, *Artifact-Centric Process Mining*. Cham, Switzerland: Springer, 2018.
- [8] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” in *Proc. 13th Int. Conf. World Wide Web*, May 2004, pp. 491–501.
- [9] P. Gundecha and H. Liu, “Mining social media: A brief introduction,” in *Proc. New Directions Inform., Optim., Logistics, Prod.*, Sep. 2012, pp. 1–17.
- [10] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [11] G. Li, R. M. de Carvalho, and W. M. P. van der Aalst, “Automatic discovery of object-centric behavior and constraint models,” in *Proc. Int. Conf. Bus. Inf. Syst.*, May 2017, pp. 43–58.
- [12] G. Li, E. G. L. de Murillas, R. M. de Carvalho, and W. M. P. van der Aalst, “Extracting object-centric event logs to support process mining on databases,” in *Proc. Int. Conf. Adv. Inf. Syst. Eng.*, Berlin, Germany: Springer, Jun. 2018, pp. 182–199.
- [13] G. Li and W. M. P. van der Aalst, “A framework for detecting deviations in complex event logs,” *Intell. Data Anal.*, vol. 21, no. 4, pp. 759–779, Jan. 2017.

## CHAPTER 6

# Appendices and Screenshots

Appendices contain material that is too detailed to include in the main report, such as long mathematical derivations or calculations, detailed technical drawings, or tables of raw data. The content should be summarized and referred to at the appropriate point in the body of the report

My Eclipse is a commercially available Java EE IDE created and maintained by the company Genuitec, a founding member of the Eclipse Foundation. My Eclipse is built upon the Eclipse platform, and integrates both proprietary and open source code into the development environment.

We used Java programming language to construct the required code. In the Java programming language, all source code is first written in plain text files ending with the .java extension. Those source files are then compiled into .class files by the javac compiler. A .class file does not contain code that is native to your processor; it instead contains bytecodes the machine language of the Java Virtual Machine (Java VM). The java launcher tool then runs your application with an instance of the Java Virtual Machine.

JavaEE (Java Enterprise Edition) provides user an opensource platform where construction of web and java combined projects can work effectively due java API and Servlet API that are available in My Eclipse. Inbuilt libraries will handle all the requirements of java and HTML files included in the project.

Apache Tomcat server helps to run the project on online server using HTTP protocol to communicate with the internet and shows the result.

**Oracle database,** Oracle Database 10g Express Edition (Oracle Database XE) is a free version of the world's most capable relational database. Oracle Database XE is easy to install, easy to manage, and easy to develop with Administer the database. We used Oracle database to store data.

## APPENDIX 1:

### Coding

#### RegistrationDao.java:

```
package com.ProcessMining.dao;

import java.sql.Connection;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.util.ArrayList;

import com.ProcessMining.db.DbCon;
import com.ProcessMining.dto.Profilebean;

public class RegistrationDao extends DbCon {

    public int register(Profilebean pb) {

        int i=0;

        Connection con=null;

        String role=pb.getRole();

        con=getConnection();

        System.out.println("connection post*****"+con);

        try {

            PreparedStatement pstmt=con.prepareStatement("insert          into
USERDETAILS(user_id,PASSWORD,ROLE,USERNAME,MAIL,MOBILE,GENDER,
ADDRESS,STATUS) values(?,?,?,?,?,?,?,?,?)");

            pstmt.setString(1, pb.getLoginid());

            pstmt.setString(2, pb.getPassword());

            pstmt.setString(3, pb.getRole());

            pstmt.setString(4, pb.getUsername());
```



```

        pstmt.setString(5, pb.getEmail());
        pstmt.setString(6, pb.getMobilenno());
        pstmt.setString(7, pb.getGender());
        pstmt.setString(8, pb.getAddress());
        pstmt.setString(9, "Active");
        i=pstmt.executeUpdate();
        System.out.println(i+"Record is Inserted successfully");
        con.close();
    }
    catch (Exception e) {
        e.printStackTrace();
    }
    return i;
}

public ArrayList<Profilebean> viewprofile(String userid) {
    Connection con=getConnection();
    ArrayList<Profilebean> list=new ArrayList<Profilebean>();
    try {
        PreparedStatementps=con.prepareStatement("select
USER_ID,USERNAME,MAIL,MOBILE,GENDER,ADDRESS from USERDETAILS
where USER_ID=?");
        ps.setString(1, userid);
        ResultSet rs=ps.executeQuery();
        while (rs.next()) {
            Profilebean pb=new Profilebean();
            String loginid=rs.getString(1);
            String username=rs.getString(2);
            String mail=rs.getString(3);

```

```

        String mobile=rs.getString(4);
        String gender=rs.getString(5);
        String address=rs.getString(6);
        pb.setLoginid(loginid);
        pb.setUsername(username);
        pb.setEmail(mail);
        pb.setMobilenumber(mobile);
        pb.setGender(gender);
        pb.setAddress(address);
        list.add(pb);
    }
    con.close();
} catch (SQLException e) {
    e.printStackTrace();
}
return list;
}

public ArrayList<Profilebean> viewadminprofile(String userid) {
    Connection con=getConnection();
    ArrayList<Profilebean> list=new ArrayList<Profilebean>();
    try {
        PreparedStatement ps=con.prepareStatement("select
LOGINID,DOCTORNAME,MAIL,MOBILE,GENDER,ADDRESS from USERINFO
where LOGINID=?");
        ps.setString(1, userid);
        ResultSet rs=ps.executeQuery();
        while (rs.next()) {
            Profilebean pb=new Profilebean();

```

```

        String loginid=rs.getString(1);
        String username=rs.getString(2);
        String mail=rs.getString(3);
        String mobile=rs.getString(4);
        String gender=rs.getString(5);
        String address=rs.getString(6);
        pb.setLoginid(loginid);
        pb.setUsername(username);
        pb.setEmail(mail);
        pb.setMobilenumber(mobile);
        pb.setGender(gender);
        pb.setAddress(address);
        list.add(pb);
    }

```

```

        con.close();
    } catch (SQLException e) {
        e.printStackTrace();
    }
    return list;
}

public int changepassword(Profilebean dto) {
    int i=0;
    Connection con=null;
    con=getConnection();
    String userid=dto.getLoginid();
    String oldpassword=dto.getPassword();
    String password=dto.getNewpassword();

```

```

try
{
PreparedStatement ps=con.prepareStatement("select password from USERDETAILS
where USER_ID='"+userid+"' and PASSWORD='"+oldpassword+"'");

ResultSet rs=ps.executeQuery();
while(rs.next())
{
    if(rs.getString(1).equals(oldpassword)){
        try
        {
            PreparedStatement pstmt=con.prepareStatement("update USERDETAILS
set password=? where USER_ID=? ");
            pstmt.setString(1, password);
            pstmt.setString(2, userid);
            i=pstmt.executeUpdate();
            if(i!=0){
                return i;
            }
        }catch (Exception e) {
            e.printStackTrace();
        }
    }else{
        return i;
    }
}

con.close();
}catch(Exception e){
    e.printStackTrace();
}

```

```

    }
    return i;
}

public int adminchangepassword(Profilebean dto) {
    int i=0;
    Connection con=null;
    con=getConnection();
    String userid=dto.getLoginid();
    String oldpassword=dto.getPassword();
    String password=dto.getNewpassword();
    try
    {
        PreparedStatement ps=con.prepareStatement("select password from USERDETAILS
        where USER_ID='"+userid+"' and PASSWORD='"+oldpassword+"'");
        ResultSet rs=ps.executeQuery();
        while(rs.next())
        {
            if(rs.getString(1).equals(dto.getPassword())){
                try
                {
                    PreparedStatement pstmt=con.prepareStatement("update USERDETAILS
                    set password=? where USER_ID=? ");
                    pstmt.setString(1, dto.getNewpassword());
                    pstmt.setString(2, dto.getLoginid());
                    i=pstmt.executeUpdate();
                    if(i!=0){
                        return i;
                    }
                }
            }
        }
    }
}

```



```

        PreparedStatement pstmt=con.prepareStatement("insert into
QUESTION_TABLE(Q_ID,QUESTION,POSTED_DATE,POSTEDBY,POSTEDID)
values(?,?,?,?,?)");

        pstmt.setInt(1, qid);

        pstmt.setString(2, pb.getQuestion());

        pstmt.setString(3, pb.getPosteddate());

        pstmt.setString(4, pb.getPostedby());

        pstmt.setString(5, pb.getPostedid());


        i=pstmt.executeUpdate();

        con.close();

    }

    catch (Exception e) {

e.printStackTrace();

    }

    return i;

}

public ArrayList<Profilebean> viewAllQuestion() {

    Connection con=getConnection();

    ArrayList<Profilebean> list=new ArrayList<Profilebean>();

    try {

        PreparedStatement ps=con.prepareStatement("select Q_ID,QUESTION from
QUESTION_TABLE");

        ResultSet rs=ps.executeQuery();

        while (rs.next()) {

            Profilebean pb=new Profilebean();

            pb.setQ_id(rs.getInt(1));

            pb.setQuestion(rs.getString(2));

```

```

        //pb.setAnswer(rs.getString(3));
        list.add(pb);
    }

    con.close();
} catch (SQLException e) {
    e.printStackTrace();
}

return list;
}

public ArrayList<Profilebean> getdetails(String uid) {
    Connection con=getConnection();

    ArrayList<Profilebean> list=new ArrayList<Profilebean>();

    try {
        PreparedStatement ps=con.prepareStatement("select
USER_ID,USERNAME,MAIL,MOBILE,GENDER,ADDRESS from USERDETAILS
where USER_ID=?");

        ps.setString(1, uid);

        ResultSet rs=ps.executeQuery();

        while (rs.next()) {
            Profilebean pb=new Profilebean();

            pb.setLoginid(rs.getString(1));
            pb.setUsername(rs.getString(2));
            pb.setEmail(rs.getString(3));
            pb.setMobilenos(rs.getString(4));
            pb.setGender(rs.getString(5));
            pb.setAddress(rs.getString(6));

            list.add(pb);
        }
    }
}

```



```

        con.close();
    } catch (SQLException e) {
        e.printStackTrace();
    }
    return list;
}

public ArrayList<Profilebean> getanswer(Profilebean pb) {
    Connection con=getConnection();
    ArrayList<Profilebean> list=new ArrayList<Profilebean>();
    try {
        PreparedStatement ps=con.prepareStatement("select
q.Q_ID,q.QUESTION,c.COMMENTS,c.COMMENTBY from QUESTION_TABLE
q,COMMENT_TABLE c where q.Q_ID=c.QID and q.Q_ID=?");
        ps.setInt(1, pb.getQ_id());
        ResultSet rs=ps.executeQuery();
        while (rs.next()) {
            Profilebean pb1=new Profilebean();
            pb1.setQ_id(rs.getInt(1));
            pb1.setQuestion(rs.getString(2));
            //pb1.setAnswer(rs.getString(3));
            pb1.setComments(rs.getString(3));
            pb1.setLoginid(rs.getString(4));
            list.add(pb1);
        }
        con.close();
    } catch (SQLException e) {
        e.printStackTrace();
    }
}

```

```
        return list;
    }}
}
```

### Explanation:

Process mining techniques are applied on the database using prepared statements which allows java users to include the sql statements into the programs.

### Screenshots:

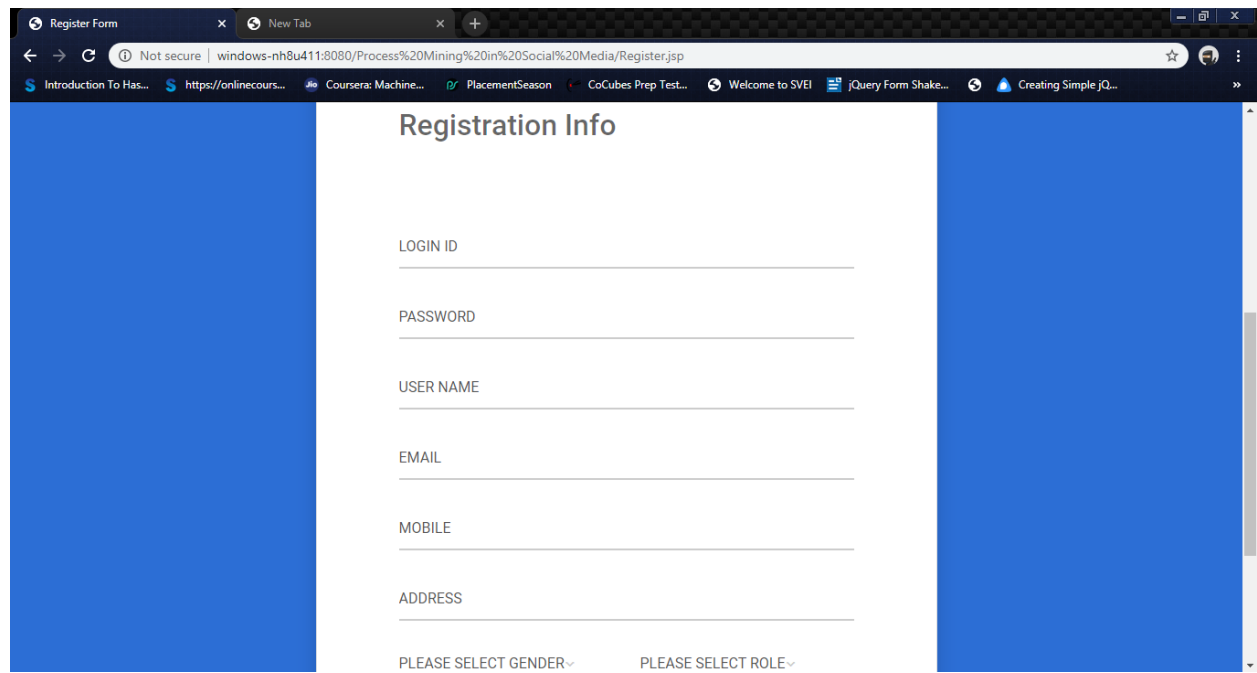
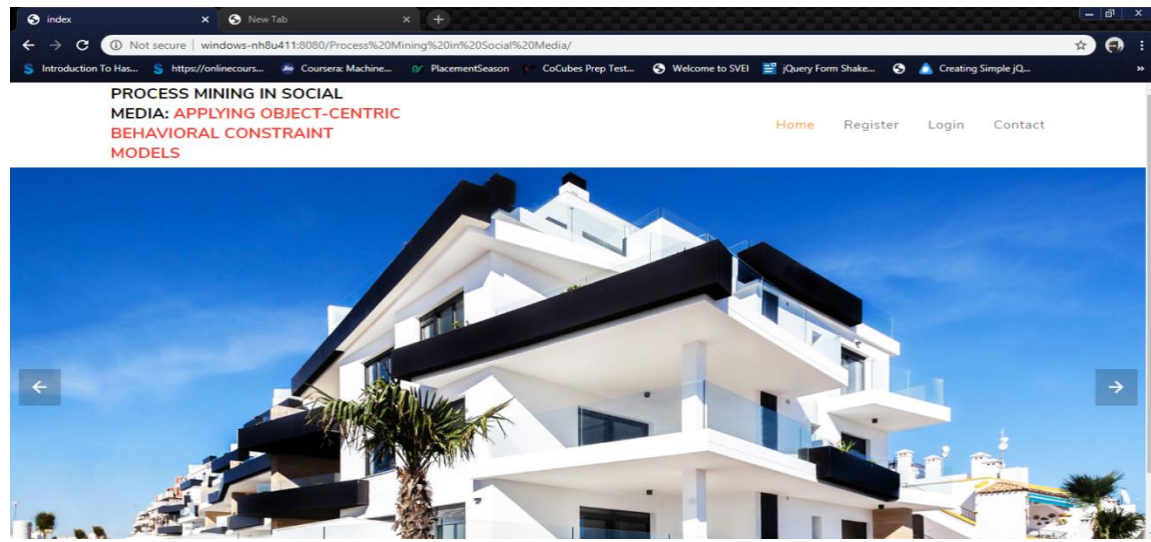
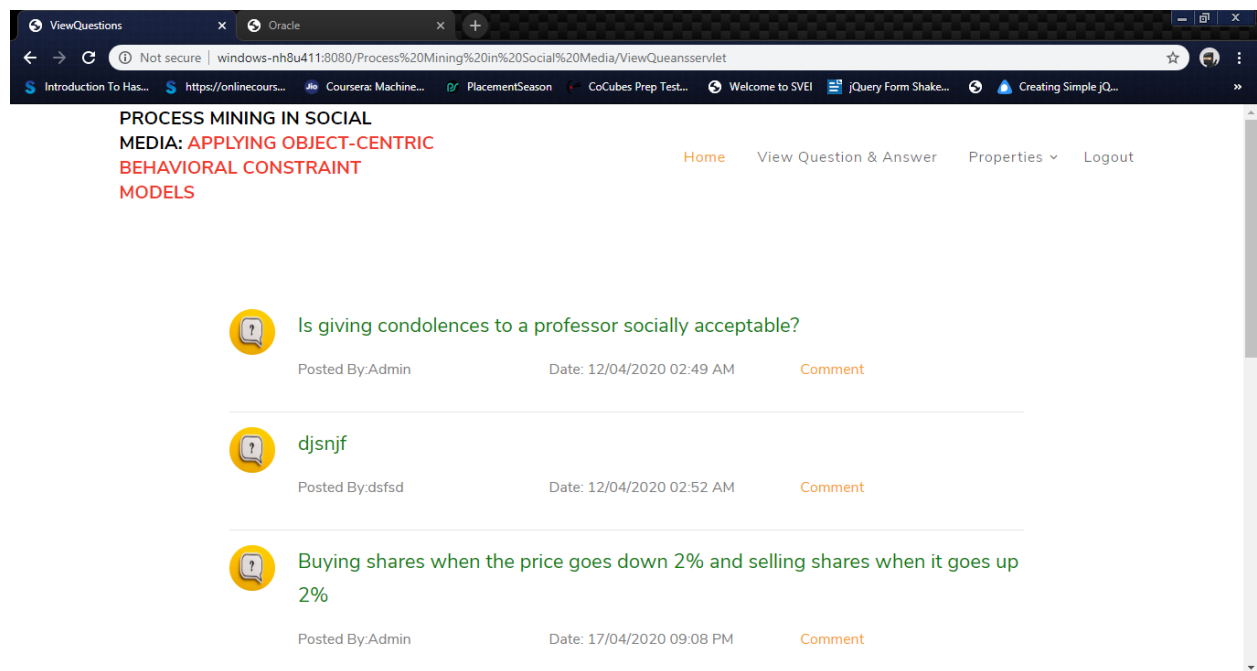
A screenshot of a web browser displaying a registration form. The browser's address bar shows a local file path. The form is titled "Registration Info" and is set against a blue background. It contains several text input fields for "LOGIN ID", "PASSWORD", "USER NAME", "EMAIL", "MOBILE", and "ADDRESS". At the bottom of the form, there are two dropdown menus with the text "PLEASE SELECT GENDER" and "PLEASE SELECT ROLE".

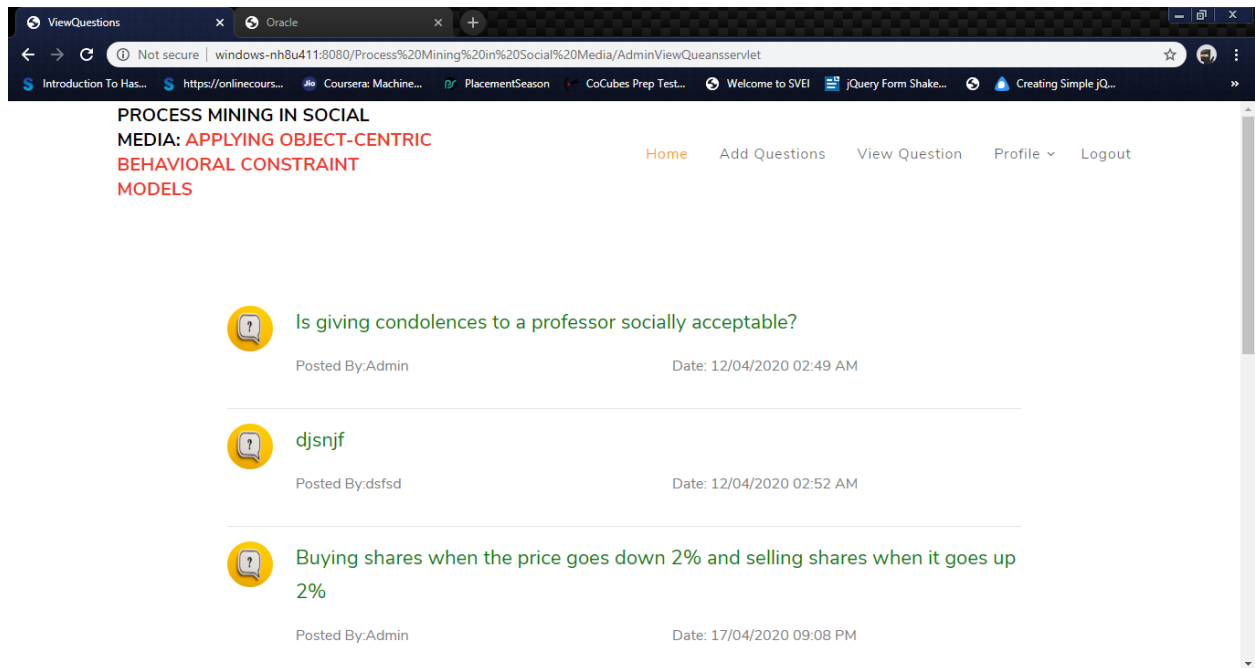
Fig 6.1: Registration form with required fields



**Fig 6.2:**Homepage

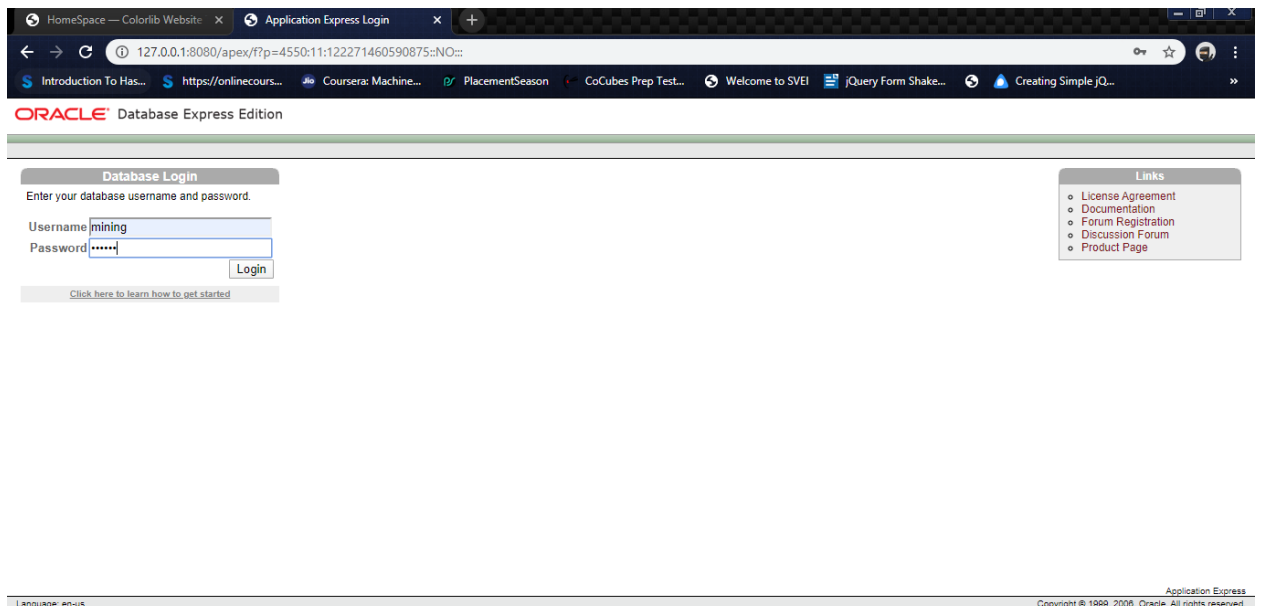


**Fig 6.3:**User home page



**Fig 6.4:**Admin home page

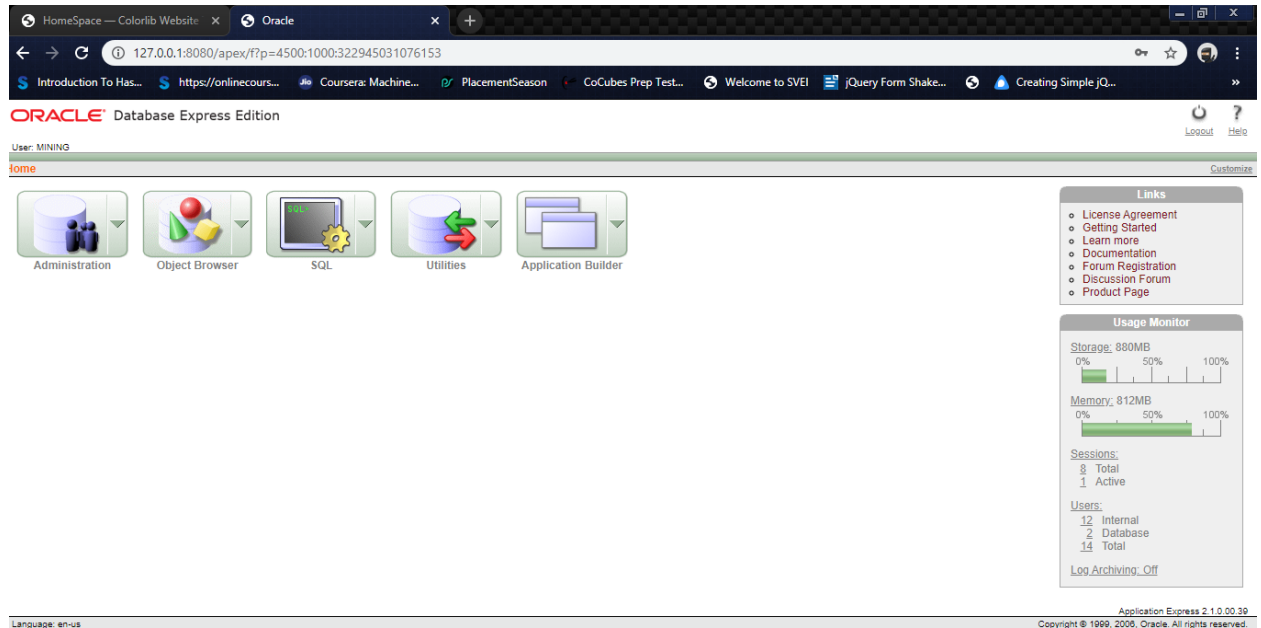
- Create an account in Oracle database and logon with credentials as required.



**Fig 6.5:** Oracle launch page login with credentials

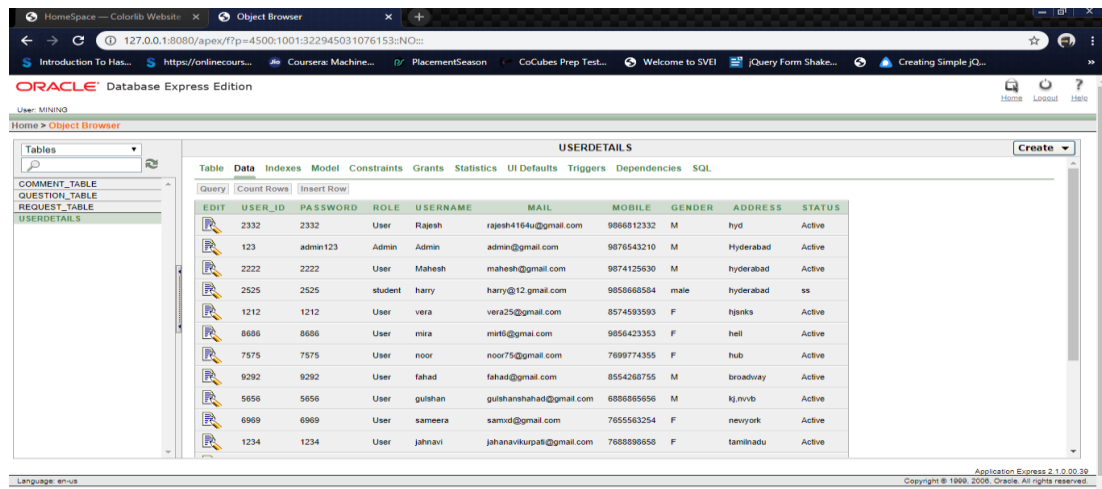
- Include dump file using SQL plus terminal with command by addressing file location  
SYNTAX: include c:/mypc/user/downloads/database.dmp

- Open Object browser to find the data that was imported into the oracle database in the form of a dump file.



**Fig 6.6:** Oracle database menu. Data will be stored in object browser

- Verify the tables to check whether the data is uploaded or not.



**Fig 6.7:** User table

Oracle Database Express Edition Object Browser showing the **REQUEST\_TABLE**.

EDIT	REQ_ID	FROMID	FROMNAME	TOID	TONAME	MAIL	MOBILE	GENDER	ADDRESS	STATUS
	1	2222	Maresh	2332	Rajesh	-	-	-	-	Friends

row(s) 1 - 1 of 1

Fig 6.8: Request table

Oracle Database Express Edition Object Browser showing the **QUESTION\_TABLE**.

EDIT	Q_ID	QUESTION	POSTED_DATE	POSTEDBY	POSTEDID
	1	Is giving condolences to a professor socially acceptable?	12/04/2020 02:49 AM	Admin	123
	2	djsnf	12/04/2020 02:52 AM	dlsfd	123
	3	Buying shares when the price goes down 2% and selling shares when it goes up 2%	17/04/2020 09:08 PM	Admin	123
	4	Guess a number between 1 and 16 with 7 attempts	17/04/2020 09:09 PM	Admin	123
	5	Why would an adventurer use a sword frog?	17/04/2020 09:09 PM	Admin	123
	6	Nodejs Passport Failure message not working	17/04/2020 10:18 PM	Admin	123
	7	Why is processing a sorted array faster than an unsorted array?	17/04/2020 10:20 PM	Admin	123
	8	Why is processing a sorted array faster than processing an unsorted array?	17/04/2020 10:22 PM	Admin	123
	9	helllofk	18/04/2020 02:31 PM	Admin	123

row(s) 1 - 9 of 9

Fig 6.9: Question table

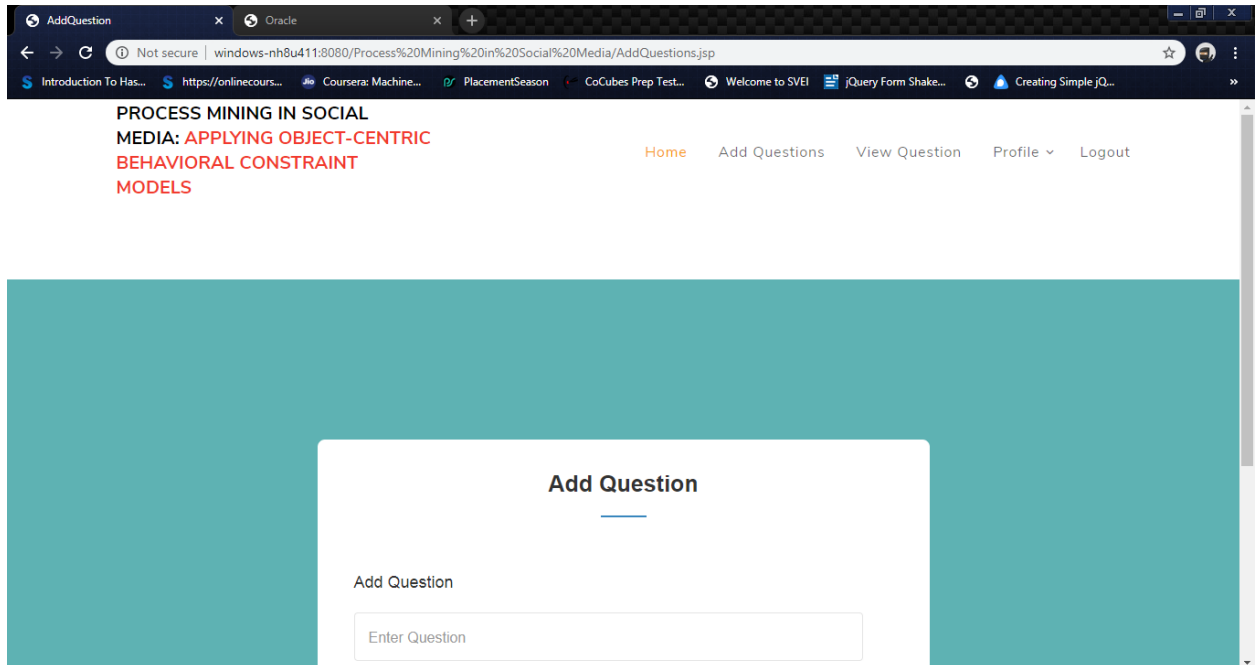
Oracle Database Express Edition Object Browser showing the **COMMENT\_TABLE**.

EDIT	CID	QID	COMMENTS	COMMENTBY	COMMENTDATE	LIKES	USERID
	1	1	My professor just told us via email that he will not be able to grade an exam until next week because his family member died from complications due to COVID-19. I don't care much about the delay, but I do like my professor, both as a person and an educator. Would it be out of line for me to send an email along the lines of "Hi professor, I'm sorry to hear your loved one passed."	Admin	12/04/2020 02:52 AM	23	123
	2	1	Gmail blocks 99.9% of dangerous emails before they reach you. If we think something seems phish-y, you'll get a warning.	Admin	14/04/2020 08:59 AM	6	123
	3	2	dldgdg	Admin	15/04/2020 01:50 AM	3	123
	4	1	no need	jahnvi	17/04/2020 09:04 PM	0	1234
	5	3	The problem is when the market goes up 10% and you cashed out at 2% because you thought it was going to go back down. You miss out on that 8% upswing waiting for that 2% dip that just isn't happening. On the other side, you buy in on a 2% downswing, but it keeps going down by 10%. Odds are good that the market will eventually come back up, and if you are patient, no loss. But this side works because you are in the market, rather than trying to time it.	svetha reddy	17/04/2020 09:41 PM	4	678
	6	5	Historically, swords were most often worn at the hip in either a sheath or a scabbard. The purpose of a sheath or scabbard was primarily to cover the blade to prevent the wearer from injuring themselves or people around them whilst walking around. It also serves a secondary function of covering the blade to protect it from the elements. A sword frog however, as far as I can tell, is more of a modern invention and is mainly used by the Live Action Roleplay (LARP) community. Swords used in LARP are typically made out of foam, meaning a wearer does not need to worry about accidentally cutting themselves or about their weapon rusting in the rain.	svetha reddy	17/04/2020 09:41 PM	2	678
	7	4	You should give more details on what kind of questions we can ask, and what kind of answer the answerer can give. For example, is it a yes/no question? Or is it we give a number then the answerer gives "larger, equal, or smaller?"	vera	17/04/2020 09:42 PM	6	1212
	8	4	the "Hamming(7,4)" code, a way of using 7 bits to transfer a 4-bit number and correct up to one error. Just take the below table: 1 0000000 2 1110000 3 1001100 4 0111100 5 0101010 6 1010101 7 1100110 8 0010110 9 1101001 10 0011001 11 0100101 12 1010101 13 1000011 14 0110011 15 0001111 16 1111111 and make your questions "Is the first bit in this number's entry 1? Is the second bit in this number's entry 1?", and so on. (If you don't want to reference a table, you can always just list numbers explicitly: your first question might be "Is the number 2, 3, 6, 7, 9, 12, 13, or 16?") This table has the special property that any two rows differ by at least three bit-flips. So if you only have one lie, you can't have two possible secret numbers: there is only one way to flip a bit and get to a row in the table.	Rajesh	17/04/2020 09:45 PM	26	2332

Fig 6.10: Comment table

## APPENDIX 2:

Admin page will handle every activity in the website. So that to improve the case study we used admin page to add the questions to website and handle the oracle database if there is any deviation in the original path the admin will be reminded immediately of the cause.



**Fig 6.11:**Adding a question to website using Admin login

### Results:


Counting number of likes for an answer cannot prove the proof that the answer is absolutely correct .So we are using an mathematical equation called Engagement rate which provides an average frequency of number of likes and the number of users who posted the question and comments for every single question. By providing a maximum threshold of engagement rate and minimum threshold of engagement rate ratings will be provided to the user like if the answer reaches max threshold then the answer was given five green stars if it fails the answer will be given 5 red stars. So that the user can analyze the pattern of the best answer and worst answer and increase their knowledge.


ViewAnswer Oracle

Not secure | windows-nh8u411:8080/Process%20Mining%20in%20Social%20Media/ViewAnswer.jsp?qid=7&q=Why%20is%20processing%20a%20sorted%20array%20faster%20... | Introduction To Has... | https://onlinecours... | Coursera: Machine... | PlacementSeason | CoCubes Prep Test... | Welcome to SVEI | jQuery Form Shake... | Creating Simple jQ...

BEHAVIORAL CONSTRAINT MODELS


Home View Question & Answer Properties Logout

 Why is processing a sorted array faster than an unsorted array?

2 Answers Asked Admin Date: 17/04/2020 10:20 PM 32 

---

Ans 1: My first thought was that sorting brings the data into the cache, but then I thought how silly that was because the array was just generated. What is going on? Why is processing a sorted array faster than processing an unsorted array? The code is summing up some independent terms, so the order should not matter.

Comment By: Rajesh Date: 17/04/2020 10:27 PM \*\*\*\*\*31 

---

Ans 2: any compiler that uses a cmov or other branchless implementation (like auto-vectorization with pcmptgtd) will have performance that's not data dependent on any CPU. But if it's branchy, it will be sort-dependent on any CPU with out-of-order speculative execution. (Even high-performance in-order CPUs use branch-prediction to avoid fetch/decode bubbles on taken branches; the miss penalty is smaller).


Comment By: jahnavi Date: 17/04/2020 10:32 PM \*\*\*\*\*1 

Fig 6.11: Representation of best and worst answer





SREE VIDYANIKETHAN ENGINEERING COLLEGE

(AUTONOMOUS)

Sree Sainath Nagar, Tirupati-517102

Department of Computer Science and Systems Engineering

## **PROGRAM OUTCOMES**

After the completion of the program, a successful student will be able to:

1. Acquire knowledge of mathematics, sciences and concepts of Computer Sciences and Engineering.
2. Ability to perform analysis of electronic systems, computer systems and software systems to meet the requirements.
3. Design and develop computer, software, mobile, embedded systems and high performance computing systems.
4. Skills to solve problems in hardware and software systems.
5. Use of computer science principles and modern tools to computing systems engineering practice.
6. Create solutions of social context the impact of Computer Science and Systems Engineering.
7. Practice computer sciences and engineering in compliance with environmental standards.
8. Follow ethical code of conduct in professional activities.
9. Achieve personal excellence and ability to work in groups.
10. Develop effective communication in professional transactions.
11. Life skills for effective project management.
12. Appreciate the significance and applications of computer science and engineering and to engage in lifelong learning for knowledge and skill upgradation.

## **PROGRAM SPECIFIC OUTCOMES**

On successful completion of the Program, the graduates will be able to

1. Acquire knowledge of mathematics, Computer Science and Systems Engineering to solve complex engineering problems.
2. Identify, Analyze, Design among alternatives and Develop software for applications and systems in the domain of Computers and its based Systems to meet the societal needs.
3. Use the research-based knowledge and methods to solve realworld problems in the fields of Computer Science and Systems Engineering.
4. Apply appropriate techniques, use modern programing languages, and packages to simulate and develop software by thoroughly understanding the requirements of the system and its constraints in Computer Science and Engineering.

## **PROGRAM EDUCATIONAL OBJECTIVES**

After few years of graduation:

1. Graduate will pursue advanced studies in Computer Science domain and Management.
2. Graduates will be employed in reputed Software Industries and develop Quality Software Systems.
3. Graduates will have career progression through professional skill development, continuing education with ethical attitude.



SREE VIDYANIKETHAN ENGINEERING COLLEGE

(AUTONOMOUS)

Sree Sainath Nagar, Tirupati-517102

Department of Computer Science and Systems Engineering

### **COURSE OUTCOMES**

**Completion of the project work enables a successful student to demonstrate:**

**CO1.** Knowledge on the project topic.

**CO2.** Analytical ability exercised in the project work.

**CO3.** Design skills applied on the project topic.

**CO4.** Ability to investigate and solve complex engineering problems faced during the project work.

**CO5.** Ability to apply tools and techniques to complex engineering activities with an understanding of limitations in the project work.

**CO6.** Ability to provide solutions as per societal needs with consideration to health, safety, legal and cultural issues considered in the project work.

**CO7.** Understanding of the impact of the professional engineering solutions in environmental context and need for sustainable development experienced during the project work.

**CO8.** Ability to apply ethics and norms of the engineering practice as applied in the project work.

**CO9.** Ability to function effectively as an individual as experienced during the project work.

**CO10.** Ability to present views cogently and precisely on the project work.

**CO11.** Project management skills as applied in the project work.

**CO12.** Ability to engage in life-long learning as experience during the project work

