Predictive Analysis and Warehousing of Web Log Data

Abdur Rahman Onik
Department of Information and Technology
Opsonin Pharma Limited
Dhaka, Bangladesh

Mashroor Zaman
Department of Computer Science and Engineering
Ahsanullah University
Dhaka, Bangladesh

Rasif Tahmid Islam
Department of Computer Science and Engineering
Ahsanullah University
Dhaka, Bangladesh

Farzana Yeasmeen Sabah
Department of Computer Science and Engineering
Ahsanullah University
Dhaka, Bangladesh

ABSTRACT

Data mining is a rapidly progressing and increasingly important sector in data science field. The core of data mining is none other than data warehousing, which is gradually becoming a self-sufficient technology for information, integration and data analysis. It is known that the decision support data model has the physical form of data warehouse and as such through storing a huge, thorough and systematic information decisions can be based upon which enterprises act. In this paper, we have analyzed different log file systems of a proxy server. The statistical representation of the log file data were our priority. Data are co-related with each other in every system. How those data are co-related is one of the object of our study. Here we studied the different processes how the data is managed actually. After that we had to consider closely on the types of data that are used so that valuable information and patterns could be discovered. The dynamic nature of modern distributed environments facilitates source data updates and schema changes, even concurrently, in different data sources. It should also be noted that volume of data increases very rapidly as a result. To address the challenge of analyzing data in an efficient way we developed a data warehouse by using multidimensional model. As a means of further analysis, we used predictive analysis on the data to empower appropriate authorities with making useful and accurate decisions.

Keywords

Data mining, Data warehouse, Predictive analysis.

1. INTRODUCTION

The present age is the age of technology, more specifically age of data. There has been a revolutionary change in the amount of data generated and analyzed today. The bulk of data is generated in cyber-space or in other words in the World Wide Web. Web mining is known as a specialized sector in data mining. Sometimes, in this process, data related to accessing web information is obtained from the web servers preserved in the form of log files. Our work is related to the representation and analysis of this log file data. Firstly, our effort was centered on analyzing the existing information of web server log files and also to find out co-relation between all these data. The log files of a server contain various information like access time, download size, duration, number of pages visited by users etc. We tried to analyze those types of data, and then build a warehouse. Secondly, we tried to analyze the data to generate predictions. Our work involved the study of operational database systems and ware housing the log file data, relevance analysis, and statistical analysis of the relevant data, predictive modeling and predictive model deployment.

2. INRODUCTION TO DATA MINING, LOG FILES AND PREDICTIVE ANALYSIS

2.1 Data Mining

Data mining is referred as knowledge discovery from data , is the programmed or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, the Web, other massive information repositories, or data streams.[1] In the scientific community, Data mining an inter disciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, logistic support and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations following some post-processing of discovered structure [2].

2.2 Web Server Logs

When someone visits a website the owner of the website can track information in order to administer the site and analyze its usage. [2].

2.3 Cookies

Cookies are small files which many web sites transfer to user's hard disk. They can inform the web site what pages the users visit, and their preferences, which enable web sites to provide user more personalized service. In particular, this is used to save user's preferences and login information, and provide personalized functionality.[2] A user can set their browser to refuse cookies, or to warn him\her before accepting them. However, a user may find there are parts of the site that he cannot access if the cookies are turned off (this particularly applies to users of Real Estate Defined site).They use the information to help them understand more about how their web site is used and to improve their site.

2.4 Predictive analysis

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. In other words, Predictive analytics is the process of extracting information from large data sets in order to make predictions and estimates about future outcomes.[new] Generally, the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it is in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. [12]The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions the level of data analysis and the quality of assumptions.[14].

3. DESIGN OF WAREHOUSING WEB SERVER LOG FILES

Warehouse of Web Server Log file Data is the repository of the information of a network server which can be used for decision making. Various parts of server record such as login time, duration time, downloading history etc. may be used for analysis and from this data warehouse valuable information can be found like which is the busiest time slot in the day, busiest day in the week etc. The whole process can be divided into three steps:

Step 1: Analysis of operational database system of the server.

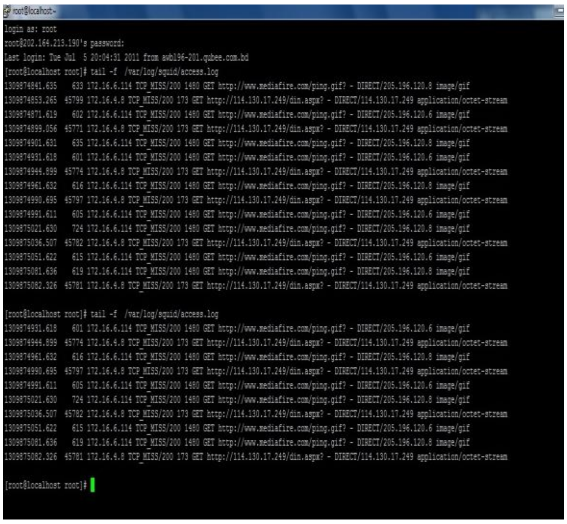
Step 2: Preprocessing and building a data warehouse.

Step 3: Apply different Data mining techniques.

3.1 Source of data

- ✓ Data by accessing internet
- ✓ Data by accessing Server
- ✓ Accessing user activity

- ➢ Data by accessing internet

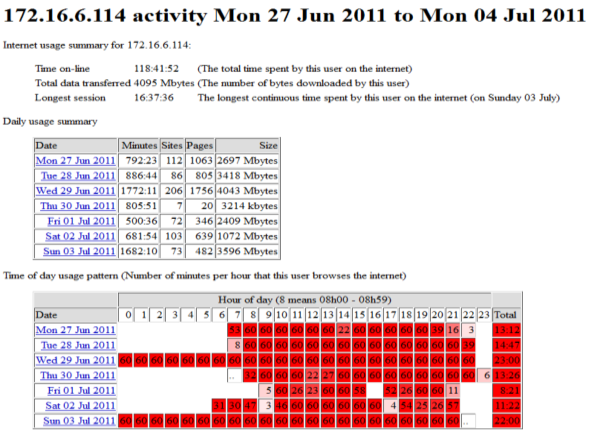


```
login as: root
root@202.164.213.190's password:
Last login: Tue Jul  5 20:04:31 2011 from awb196-201.qubee.com.bd
[root@localhost root]# tail -f  /var/log/squid/access.log
1309874841.635    633 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309874853.265  45799 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874871.619    602 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309874899.056  45771 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874901.631    635 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309874931.618    601 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309874944.899  45774 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874961.632    616 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309874990.695  45797 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874991.611    605 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309875021.630    724 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309875036.507  45782 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309875051.622    615 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309875081.636    619 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309875082.326  45781 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream

[root@localhost root]# tail -f  /var/log/squid/access.log
1309874931.618    601 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309874944.899  45774 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874961.632    616 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309874990.695  45797 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309874991.611    605 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309875021.630    724 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309875036.507  45782 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream
1309875051.622    615 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.6 image/gif
1309875081.636    619 172.16.6.114 TCP_MISS/200 1480 GET http://www.mediafire.com/ping.gif? - DIRECT/205.196.120.8 image/gif
1309875082.326  45781 172.16.4.8 TCP_MISS/200 173 GET http://114.130.17.249/din.aspx? - DIRECT/114.130.17.249 application/octet-stream

[root@localhost root]#
```

- ➢ Data by accessing server



172.16.6.114 activity Mon 27 Jun 2011 to Mon 04 Jul 2011

Internet usage summary for 172.16.6.114:

Time on-line	118:41:52	(The total time spent by this user on the internet)
Total data transferred	4095 Mbytes	(The number of bytes downloaded by this user)
Longest session	16:37:36	The longest continuous time spent by this user on the internet (on Sunday 03 July)

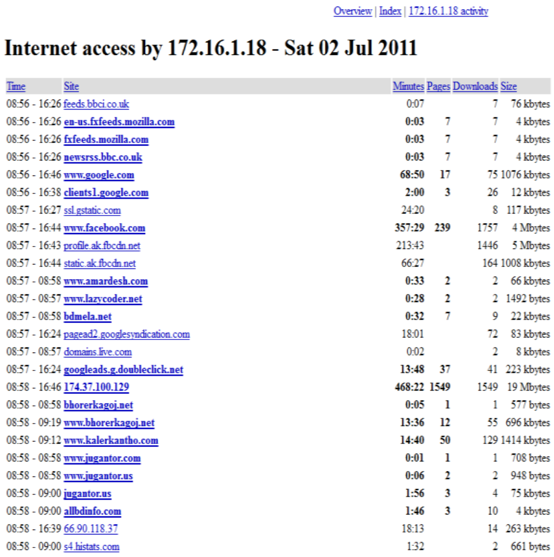
Daily usage summary

Date	Minutes	Sites	Pages	Size
Mon 27 Jun 2011	792:23	112	1063	2697 Mbytes
Tue 28 Jun 2011	886:44	86	805	3418 Mbytes
Wed 29 Jun 2011	1772:11	206	1756	4043 Mbytes
Thu 30 Jun 2011	805:51	7	20	3214 kbytes
Fri 01 Jul 2011	500:36	72	346	2409 Mbytes
Sat 02 Jul 2011	681:54	103	639	1072 Mbytes
Sun 03 Jul 2011	1682:10	73	482	3596 Mbytes

Time of day usage pattern (Number of minutes per hour that this user browses the internet)

	Hour of day (8 means 08h00 - 08h59)																								
Date	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Total
Mon 27 Jun 2011								53	60	60	60	60	60	60	22	60	60	60	60	60	39	16	3		13:12
Tue 28 Jun 2011								8	60	60	60	60	60	60	60	60	60	60	60	60	60	60	39		14:47
Wed 29 Jun 2011	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60		23:00
Thu 30 Jun 2011								..	32	60	60	60	22	27	60	60	60	60	60	60	60	60	60	6	13:26
Fri 01 Jul 2011										5	60	26	23	60	60	58		52	26	60	60	11			8:21
Sat 02 Jul 2011							31	30	47	3	46	60	60	60	60	60	60	4	54	25	26	57			11:22
Sun 03 Jul 2011	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	..		22:00

- ➢ Data by accessing user activity



Overview | Index | 172.16.1.18 activity

Internet access by 172.16.1.18 - Sat 02 Jul 2011

Time	Site	Minutes	Pages	Downloads	Size
08:56 - 16:26	feeds.bbci.co.uk	0:07		7	76 kbytes
08:56 - 16:26	**en-us.fxfeeds.mozilla.com**	**0:03**	**7**	7	4 kbytes
08:56 - 16:26	**fxfeeds.mozilla.com**	**0:03**	**7**	7	4 kbytes
08:56 - 16:26	**newsrss.bbc.co.uk**	**0:03**	**7**	7	4 kbytes
08:56 - 16:46	**www.google.com**	**68:50**	**17**	75	1076 kbytes
08:56 - 16:38	**clients1.google.com**	**2:00**	**3**	26	12 kbytes
08:57 - 16:27	ssl.gstatic.com	24:20		8	117 kbytes
08:57 - 16:44	**www.facebook.com**	**357:29**	**239**	1757	4 Mbytes
08:57 - 16:43	profile.ak.fbcdn.net	213:43		1446	5 Mbytes
08:57 - 16:44	static.ak.fbcdn.net	66:27		164	1008 kbytes
08:57 - 08:58	**www.amardesh.com**	**0:33**	**2**	2	66 kbytes
08:57 - 08:57	**www.lazycoder.net**	**0:28**	**2**	2	1492 bytes
08:57 - 08:58	**bdmela.net**	**0:32**	**7**	9	22 kbytes
08:57 - 16:24	pagead2.googlesyndication.com	18:01		72	83 kbytes
08:57 - 08:57	domains.live.com	0:02		2	8 kbytes
08:57 - 16:24	**googleads.g.doubleclick.net**	**13:48**	**37**	41	223 kbytes
08:58 - 16:46	**174.37.100.129**	**468:22**	**1549**	1549	19 Mbytes
08:58 - 08:58	**bhorerkagoj.net**	**0:05**	**1**	1	577 bytes
08:58 - 09:19	**www.bhorerkagoj.net**	**13:36**	**12**	55	696 kbytes
08:58 - 09:12	**www.kalerkantho.com**	**14:40**	**50**	129	1414 kbytes
08:58 - 08:58	**www.jugantor.com**	**0:01**	**1**	1	708 bytes
08:58 - 08:58	**www.jugantor.us**	**0:06**	**2**	2	948 bytes
08:58 - 09:00	**jugantor.us**	**1:56**	**3**	4	75 kbytes
08:58 - 09:00	**allbdinfo.com**	**1:46**	**3**	10	4 kbytes
08:58 - 16:39	66.90.118.37	18:13		14	263 kbytes
08:58 - 09:00	s4.histats.com	1:32		2	661 bytes

3.2 Data preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. There are only one data table which stored data as raw data. This data is incomplete, inconsistent and lacking in certain behaviors. So to implement data warehouse, data should be preprocessed and should be relieved of any redundant data.The methods for preprocessing of data are organized into the following categories:

- ✓ Data selection
- ✓ Data cleaning
- ✓ Data integration
- ✓ Transformations
- ✓ Data reduction

3.3 Schema used in warehouse

Two types of schemas are usually used:

- ➢ Star schema
- ➢ Snowflack schema

3.4 Star and Snowflake Schema

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables.

3.5 Description of Warehousing Web Server Log File data

In the warehouse, there is one fact table and many dimension tables. List of fact table and dimension tables are given below:

Fact table :

- login_fact

Dimension tables :

- duration
- downloadtime_slot
- time_slot
- page_visited

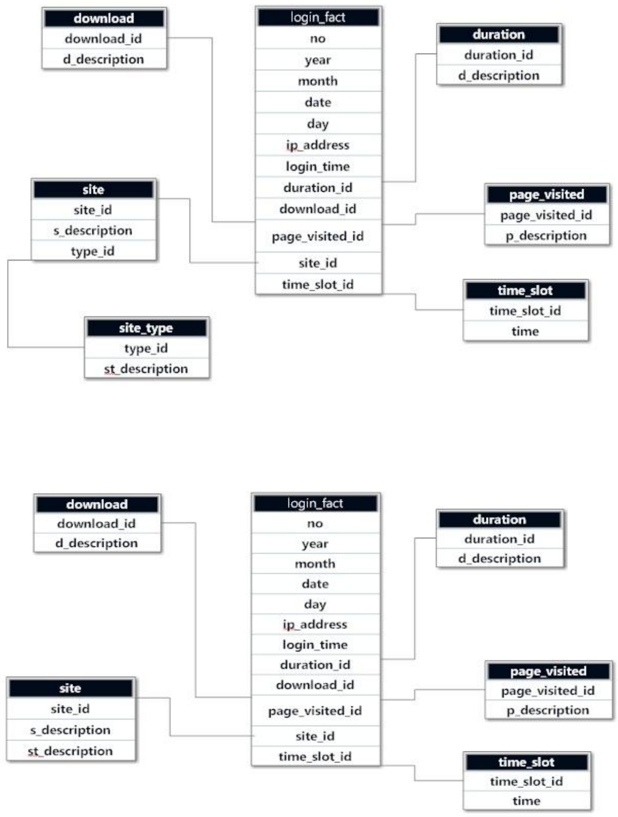
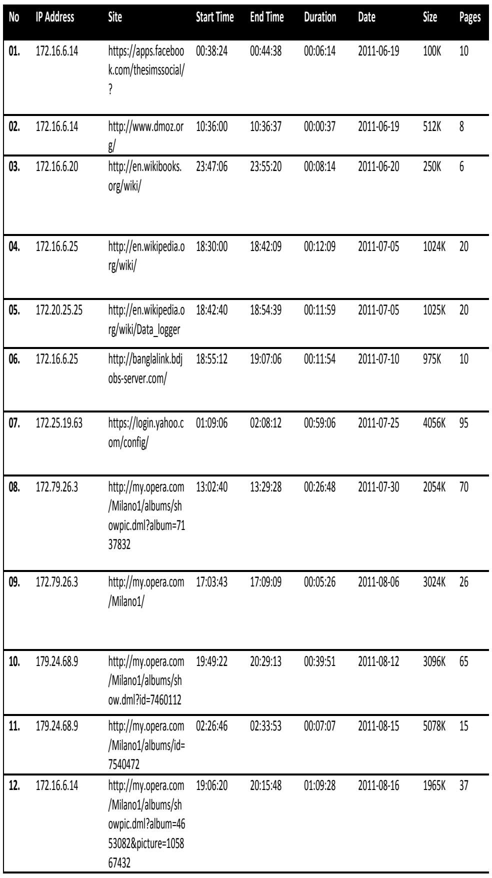


Fig 1: Star and Snowflake Schema

Table 1. Raw Table data



No	IP Address	Site	Start Time	End Time	Duration	Date	Size	Pages
01.	172.16.6.14	https://apps.facebook.com/thesimssocial/?	00:38:24	00:44:38	00:06:14	2011-06-19	100K	10
02.	172.16.6.14	http://www.dmoz.org/	10:36:00	10:36:37	00:00:37	2011-06-19	512K	8
03.	172.16.6.20	http://en.wikibooks.org/wiki/	23:47:06	23:55:20	00:08:14	2011-06-20	250K	6
04.	172.16.6.25	http://en.wikipedia.org/wiki/	18:30:00	18:42:09	00:12:09	2011-07-05	1024K	20
05.	172.20.25.25	http://en.wikipedia.org/wiki/Data_logger	18:42:40	18:54:39	00:11:59	2011-07-05	1025K	20
06.	172.16.6.25	http://banglalink.bdjobs-server.com/	18:55:12	19:07:06	00:11:54	2011-07-10	975K	10
07.	172.25.19.63	https://login.yahoo.com/config/	01:09:06	02:08:12	00:59:06	2011-07-25	4056K	95
08.	172.79.26.3	http://my.opera.com/Milano1/albums/showpic.dml?album=7137832	13:02:40	13:29:28	00:26:48	2011-07-30	2054K	70
09.	172.79.26.3	http://my.opera.com/Milano1/	17:03:43	17:09:09	00:05:26	2011-08-06	3024K	26
10.	179.24.68.9	http://my.opera.com/Milano1/albums/show.dml?id=7460112	19:49:22	20:29:13	00:39:51	2011-08-12	3096K	65
11.	179.24.68.9	http://my.opera.com/Milano1/albums/id=7540472	02:26:46	02:33:53	00:07:07	2011-08-15	5078K	15
12.	172.16.6.14	http://my.opera.com/Milano1/albums/showpic.dml?album=4653082&picture=105867432	19:06:20	20:15:48	01:09:28	2011-08-16	1965K	37

4. IMPLEMENTATION OF WAREHOUSING WEB SERVER LOG FILE DATA

A prerequisite of data warehousing is designing of data warehouse. Then follows the implementation process. This implementation may involve complexities in view of the facts that the technology used should support data warehousing, warehouse may involve huge number of schemas, there may be many constraints, rules, and it may need to maintain lots of relationships between fact and dimension tables.

4.1 Architecture used in Warehousing web server log file data

In warehousing web server log file data we have tried to implement a three-tier architecture. These three tier architecture is:

1. In warehousing web server log file data, first tier is a warehouse database server that is almost always a relational database system. We used MS SQL Server tools to store data in data warehouse from the operational database after performing some preprocessing tasks such as cleaning, transformation etc.
2. Second tier is an OLAP server which is implemented using multidimensional OLAP model. Here all the schemas of warehousing web server log file data are stored.
3. The top tier or third tier contains query and reporting tools, analysis tools and data mining tools.

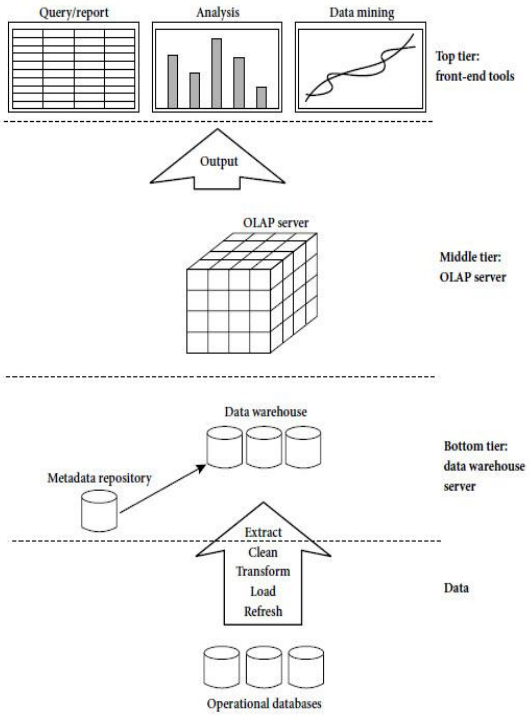
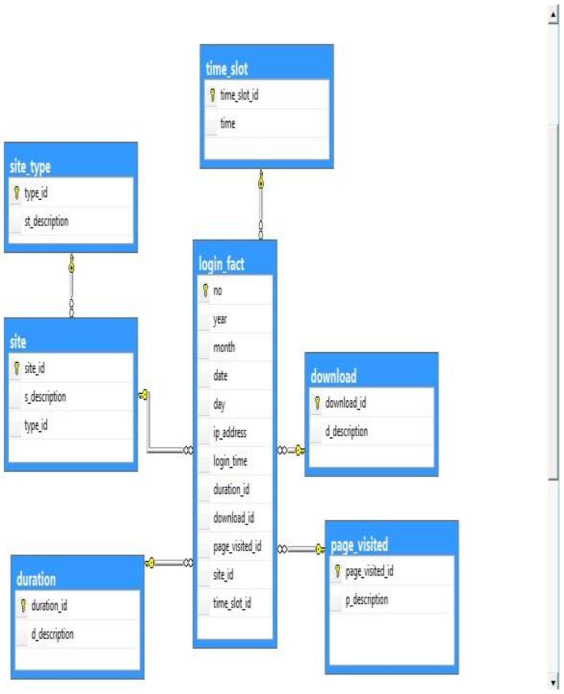


Fig: 2 tier Architecture

4.1.1 Technology used in warehousing web server log file data

Language: C#
Framework IDE: Visual Studio 2013
Architecture: 3-tier Layer Architecture (Presentation, Business, Data Access)
Database: SQL Server 2012
Operating System: Microsoft Windows 8.1

4.1.1.1 Implementation of fact constellation schema in MS SQL Server



4.1.1.2 Data Dictionary of Warehousing web server log file data

Raw data table:

Column Name	**Data Type**	**Detail Field Name**	**Comments**
no	int	Serial no	Primary key
ip_address	varchar(50)	IP address	Users ip address
site	varchar(50)	Site name	Which site visited by user
start_time	datetime	Login time	Users login time
duration	datetime	Duration time	Users duration time on server
date	datetime	date	Login date
download_size	varchar(50)	Download size	Download size of files
pages	int	Pages visited by user	Wvhich visited by users

Login fact table:

Column Name	Data Type	Detail Field Name	Comments
no	int	Serial no	Primary key
year	int	Year name	Year information
month	varchar(50)	Month name	Month information
date	int	Date	Date
day	varchar(50)	Day name	Day name
ip_address	varchar(50)	IP address	Users ip address
duration_id	int	Duration id	Foreign key of duration
download_id	int	Download id	Foreign key of download
page_visited_id	int	Page visited id	Foreign key of page_visited
site_id	int	Site id	Foreign key of site
time_slot_id	int	Time slot id	Foreign key of time_slot

Duration table:

Column Name	Data Type	Detail Field Name	Comments
duration_id	int	Duration id	Primary key
d_description	varchar(50)	Duration time	Description of duration time.

Download table:

Column Name	Data Type	Detail Field Name	Comments
download_id	int	Download id	Primary key
d_description	varchar(50)	Download time	Description of download time.

Page visited table:

Column Name	Data Type	Detail Field Name	Comments
page_visited_id	int	Page visited id	Primary key
p_description	varchar(50)	Page visited no	Page visited no which is vosoted by users.

Timeslot table

Column Name	Data Type	Detail Field Name	Comments
time_slot_id	int	Time slot id	Primary key
time	varchar(50)	Time slot	Time is divided in many slot.

Site table:

Column Name	Data Type	Detail Field Name	Comments
site_id	int	Site id	Primary key
s_description	varchar(50)	Site description	Which types of site is it.
type_id	int	Site type id	Foreign key

Site Type table:

Column Name	Data Type	Detail Field Name	Comments
type_id	int	Site type id	Primary key
st_description	varchar(50)	Site type description	Which types of site is it.

4.2 Feeding the warehouse from operational database

In data warehouse first we collect data from server which was in xx.txt file. From this xx.txt file we store data in data table. This data table will update day to day. This data table is a raw data table. This raw data is not suitable for work in data warehouse and analysis because here were some data which we do not use in data warehouse. So, from this raw data we

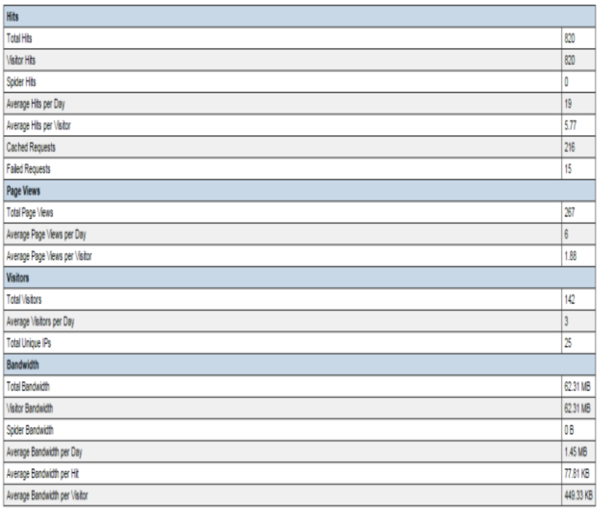
Transfer data in a fact table with some analysis and after removal of unnecessary data. This data is transferred automatically from table to table. After collecting this operational data we analyze the data.

5. APPROACHES FOR PREDICTIVE ANALYSIS OF WEB SERVER LOG FILE DATA

5.1 Steps of predictive analysis

The implementation approach of Predictive Analysis
is described in the Following way. The steps are described in detail below:
Step 1: Defining the Objective of Predictive Analysis.
Step 2: Data collection from different sources.
Step 3: Data Processing.
Step 4: Statistical Analysis.
Step 5: Modeling.

5.2 Statistical analysis



Hits	
Total Hits	820
Visitor Hits	820
Spider Hits	0
Average Hits per Day	19
Average Hits per Visitor	5.77
Cached Requests	216
Failed Requests	15
Page Views	
Total Page Views	267
Average Page Views per Day	6
Average Page Views per Visitor	1.88
Visitors	
Total Visitors	142
Average Visitors per Day	3
Total Unique IPs	25
Bandwidth	
Total Bandwidth	62.31 MB
Visitor Bandwidth	62.31 MB
Spider Bandwidth	0 B
Average Bandwidth per Day	1.45 MB
Average Bandwidth per Hit	77.81 KB
Average Bandwidth per Visitor	449.33 KB

Fig 3: Statistical information

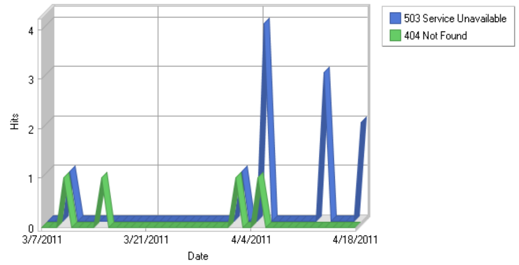


Fig 4: Error Types

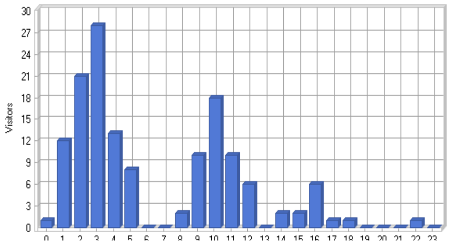


Fig 5: Activity by hour of day

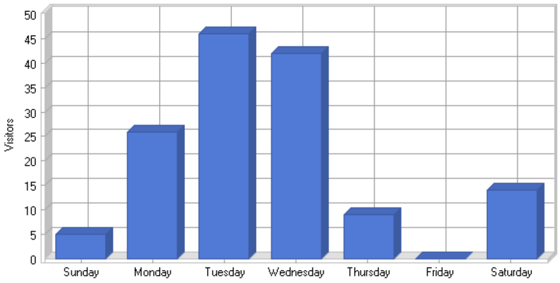


Fig 6: Activity by day of a week

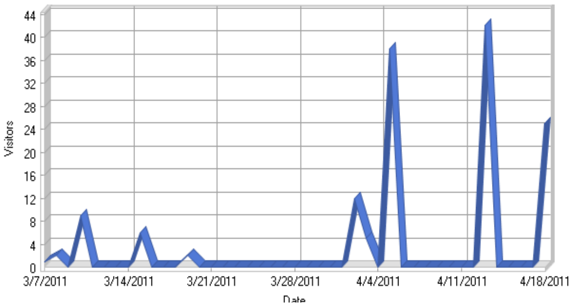


Fig 7: No of daily visitors

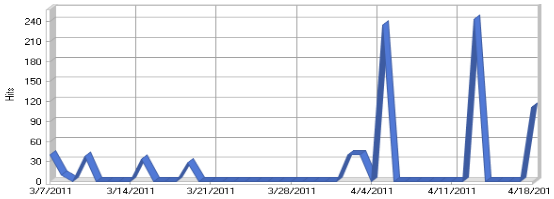


Fig 8: No of daily hits

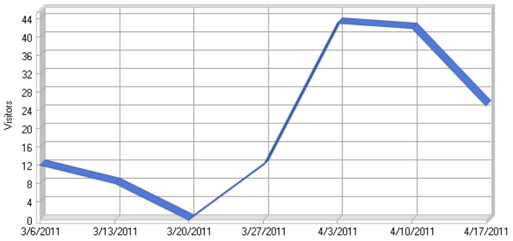


Fig 8: Activity by week

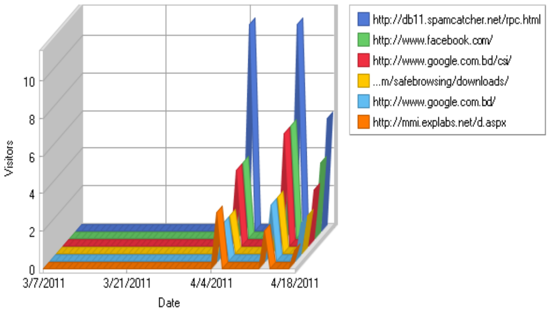


Fig 8: Daily page access

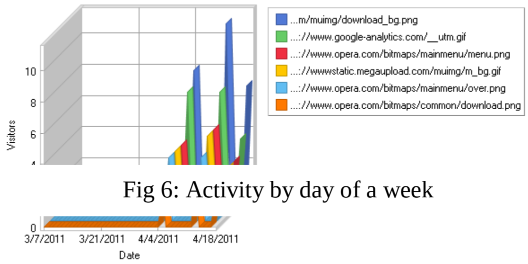


Fig 9: Daily image access

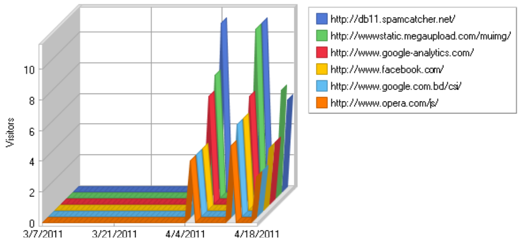


Fig 10: Daily directory access

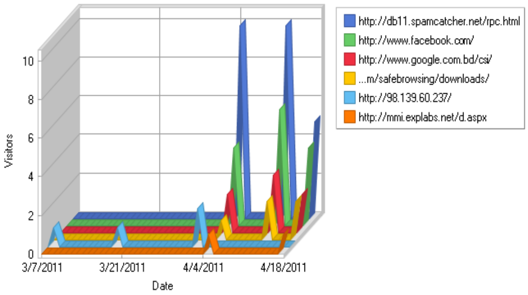


Fig 11: Daily entry pages

6. FUTURE WORK

Our work involved the basic design and implementation of data warehouse. After obtaining some basic knowledge we want to build a complete operational data warehouse system. Then we will try to implement more complex processing techniques, algorithm to store data in data warehouse. Here we worked on monthly data. In future we will try to work on yearly data which will help us to find out yearly information.

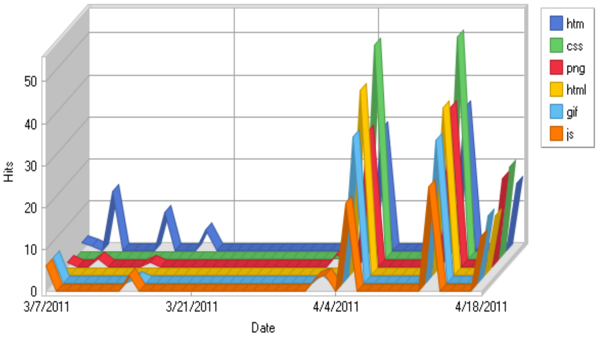


Fig 12: Daily file type access

7. REFERENCES

[1] Jian Pei, Jiawei Han, MichilineKamber. Data Mining concepts and Technique. Morgan Kaufmann, 3rd Edition, 2012.

[2] Bing Li. Web Data Mining: Exploring Hyperlinks, Contents and usage. Springer, 2nd Edition, 2011.

[3] Nutan Farah Haq, Abdur Rahman Onik, (2015). "Application of Machine Learning Approaches in Intrusion Detection System: A Survey." (IJARAI) International.

[4] Onik A.R., Haq N.F. and Mustahin W. (2015). Cross-breed type Bayesian Network based Intrusion Detection System (CBNIDS). Published in the proceeding of the 18th International Conference on Computer and Information Technology (ICCIT 2015) Sponsored by IEEE, Dhaka, December 21-23, 2015

[5] Onik A. R., Haq, N. F., Alam, L., & Mamun, T. I. (2015). An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier. International Journal of Computer Applications, 124(13). DOI10.5120/ijca2015905706.

[6] Haq N.F., Onik A.R. and Shah F.M. (2015). An Ensemble Framework of Anomaly Detection Using Hybridized Feature Selection Approach (HFSA). SAI Intelligent Systems Conference (IntelliSys), 2015, IEEE. DOI 10.1109/IntelliSys.2015.7361264.

8. APPENDIX

```
1324170539 802 172.16.8.3 TCP_REFRESH_MISS/302 747 GET http://fxfeeds.mozilla.com/en-US/firefox/headlines.xml - DIRECT/93.184.216.119 text/html
1324170539 247 172.16.8.3 TCP_REFRESH_MISS/302 795 GET http://fxfeeds.mozilla.com/firefox/headlines.xml - DIRECT/93.184.216.119 text/html
1324170539 313 172.16.8.3 TCP_MISS/301 630 GET http://newsrss.bbc.co.uk/rss/newsonline_world_edition/front_page/rss.xml - DIRECT/125.252.225.167 text/html
1324170539 197 172.16.8.3 TCP_MISS/200 8497 GET http://feeds.bbci.co.uk/news/rss.xml? - DIRECT/125.252.225.166 text/xml
1324170557 1594 172.16.8.3 TCP_MISS/200 29423 GET http://www.google.com/search? - DIRECT/173.194.35.18 text/html
1324170558 783 172.16.8.3 TCP_MISS/200 529 GET http://id.google.com/verify/EAAAAMc3LCsyNXu8IpWPatEmQdk.gif - DIRECT/209.85.148.139 image/gif
1324170558 757 172.16.8.3 TCP_MISS/204 190 GET http://clients1.google.com/generate_204 - DIRECT/209.85.148.102 text/html
1324170559 580 172.16.8.3 TCP_MISS/204 280 GET http://www.google.com/csi? - DIRECT/173.194.35.17 image/gif
1324170560 607 172.16.8.3 TCP_MISS/200 631 GET http://www.google.com/url? - DIRECT/173.194.35.17 text/html
1324170562 614 172.16.8.3 TCP_MISS/200 808 POST http://ocsp.digicert.com/ - DIRECT/173.204.115.235 application/ocsp-response
1324170563 662 172.16.8.3 TCP_MISS/200 1438 POST http://ocsp.digicert.com/ - DIRECT/68.232.37.39 application/ocsp-response
1324170564 3946 172.16.8.3 TCP_MISS/200 33229 CONNECT login.yahoo.com:443 - DIRECT/98.139.241.94 text/html
1324170566 589 172.16.8.3 TCP_MISS/200 808 POST http://ocsp.digicert.com/ - DIRECT/64.151.73.102 application/ocsp-response
1324170566 1554 172.16.8.3 TCP_MISS/200 5004 CONNECT us.bc.yahoo.com:443 - DIRECT/111.67.226.185 text/html
1324170567 2033 172.16.6.44 TCP_MISS/200 5315 CONNECT login.yahoo.com:443 - DIRECT/98.139.241.94 text/html
1324170567 2129 172.16.6.44 TCP_MISS/200 4619 CONNECT login.yahoo.net:443 - DIRECT/98.139.241.93 text/html
1324170567 593 172.16.6.44 TCP_MISS/200 808 POST http://ocsp.digicert.com/ - DIRECT/173.204.115.235 application/ocsp-response
1324170568 1059 172.16.6.43 TCP_MISS/200 5004 CONNECT us.bc.yahoo.com:443 - DIRECT/111.67.226.184 text/html
1324170570 1486 172.16.6.43 TCP_MISS/200 5283 CONNECT login.yahoo.com:443 - DIRECT/98.139.241.94 text/html
1324170577 1608 172.16.8.3 TCP_MISS/200 6499 CONNECT login.yahoo.com:443 - DIRECT/98.139.241.94 text/html
```

Here, 1324170577 means date and duration time which is in decimal timestamp form. We transfer this decimal time stamp form in date and time. Code for transferring decimal time stamp form to date time in Microsoft SQL Server is given below.

```
          dateadd(s, timestamp, '19700101')
                    or
       SELECT DATEDEIFF(s, ‘1970-01-01 00:00:00’, GETUTDATE())
```

1608 means page no.
5283 means download size.
http://fxfeeds.mozilla.com/en-US/firefox/headlines.xml means a site name.

TCP REFRESH MISS/302, GET this are extra data.