Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 2(2014) pp. 359-365

Web Log Data Analysis Using a Data Warehouse and OLAP

Mohammed Hamed Ahmed Elhebir [1], Murtada Khalafallah Elbashir Elfaki [2] and Ajith Abraham [3]

[1] Faculty of Mathematical and Computer Sciences, University of Gezira,
P.O. Box 20, Wad Medani, Sudan
elhibr@uofg.edu.sd

[2] Faculty of Mathematical and Computer Sciences, University of Gezira,
P.O. Box 20, Wad Medani, Sudan
murtadabashir@uofg.edu.sd

[3] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, P.O. Box 2259, WA, USA
ajith.abraham@ieee.org

***Abstract*: Many Internet businesses usually collect hundreds of megabytes of click-streams data every day that need to be analyzed. Performing systematic analysis on such a huge amount of data is time-consuming. Online Analytical Processing (OLAP) can be used for this purpose. The primary requirement in the construction of multi-dimensional data cube is the identification of dimensions and measures. In this paper, the web usage mining is analyzed by applying the Pattern Analysis techniques on web log data. First, the dimensions and measures in web usage data warehouse are nominated and then a technique on how to apply (OLAP) on web usage data warehouse is proposed.**

***Keywords*:** Web Usage Mining, Web Log Data, Data Warehouse, Pattern Analysis, OLAP, Dimension and Measures**.**

I. Introduction

Nowadays, the Web has turned to be the largest information source available on the planet. It is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies an incredible amount of information, and also raises the complexity of how to deal with the information from different perspectives of users view. Users usually want to have an effective search tool for finding relevant information easily and precisely. Web Mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information from web documents and services. Web data mining can be divided into three different processes: "Web Content" mining, "Web Structure" mining and "Web Usage" mining [1, 2, 3]. Web Usage mining is a heavily researched area in the field of data mining. It can be described as the discovery and analysis of user access patterns through mining of log files and associated data from a particular website.

Generally, Web Usage mining consists of three processes: Data Pre-processing for the web log file, Pattern discovery and Pattern analysis [4].The data source affects the quality of the pre-processed data and in turn the pre-processed data influences the results of pattern discovery and pattern analysis directly [5]. Although many areas and applications can be cited where Web Usage mining is useful, it can be said that the main idea behind Web Usage mining is to let users of a website use it with ease and effectively predict and recommend parts of the website to them based on their previous actions on the web site. A server log file is a file that automatically creates and maintains the activities performed on the server. This file is used to record each and every hit to a web site [6]. It maintains a history of page requests, also it helps in understanding how and when a website pages and application are being accessed by the web browser. It contains information such as the host IP address, proprietor, username, date, time, request method, status code, byte size, and referrer and user agent [7]. In this paper one of the widely used analytical tools and techniques is used to analyze the access patterns of the University of Sudan Science and Technology ' website. In this tool, we used the data warehouse to the extracted information from web log file in terms of dimensions and facts. The dimensions were represented by time, Protocol type, Users, Agent, IP address while the number of the accesses and the document size represented facts. Then an Online Analytical Processing (OLAP) was used to analyze the data in the data warehouse.

II. Methodology and Tool

As shown in Figure1, the methodology contains four sub-steps: Data capture from log file, data pre-processing, data warehouse schema and OLAP data cube.

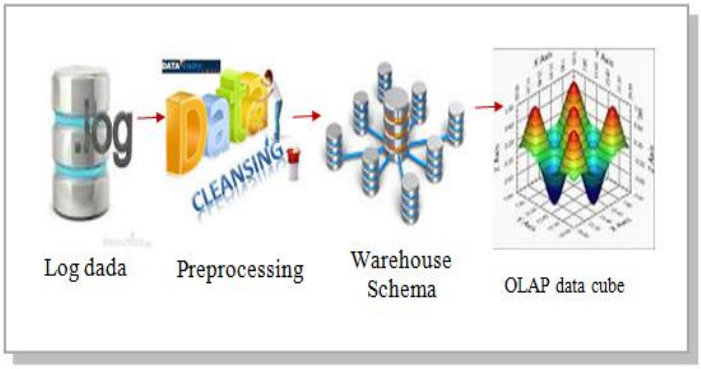


Figure 1. Methodology Steps

A. *Web Log Data*

The main source of the data for web usage mining was the Web server logs from Sudan University of Science and Technology. The Web server logs each visit to each web page with possibly IP address, refereed page, access time, browser type and version, and accessed page link. The period of the data source of our experiment was from 7/Nov/2008 to 10/Dec/2009. The size of the log file in this period was 567 MB containing 291642 cases.

B. *Data Pre-processing*

Data pre-processing is particularly important for the whole Web Usage mining processes and the key of the Web Usage mining quality. This phase contains three sub-steps: Data Cleaning, User Identification, and Session Identification [8].The raw data should be cleaned to eliminate irrelevant information from the original log file and to make the Web log data convenient for pattern analysis. In order to remove useless requests from the log files, we needed to eliminate irrelevant entries [9]. The irrelevant entries in our log file were found to be: entries having suffixes like .jpg, .jpeg, .css, .map and entries having status code failure. We also needed to remove all records which do not contain method ” GET” and removed navigation sessions performed by Crawler ,Spider, and Robot. The data captured in the web logs were filtered to remove irrelevant information and a relational database was created to be loaded with the meaningful remaining data. For the mining part, we filtered out other unwanted entries like record accesses to image files that were embedded in the web pages whose ‘hit’ had already been logged. In the pre-processing phase, we created a file listing numbers of unique users, unique pages, and unique sessions in the log file. Unique users were the different users accessing the web site and that can be identified using IP address, User Agent and Referred URL field [10]. The users with the same IP address field were considered to be the same. If an IP address is the same but user agent is different, then each different agent will represent a different user. These rules were used to enable the program to identify users. After user identification, pages visited by each user were categorized into different unique sessions called session identification. A good time frame for each session was 30 minutes [11].The unique page was the distinct page without replication accessed by the users. After accomplishing this step of identification, it was found that there were 23,200 different users, 8861 unique pages, and 13869 sessions.

C. *Data Warehouse Construction*

To construct the data warehouse, first we nominated 6 dimensions. These were: Time, Protocol type, Users, Agent, IP address and Pages. Each dimension had primary key and other fields that can be used in the analysis. Second, we determined two facts. These facts were the number of visits and the document size. The facts were used as fields in the fact table. The other fields of the fact table were foreign keys that can be used in constructing the relations with the dimension tables to yield a schema called the star schema as shown in Figure 2. The time dimension was designed to contain the hierarchies (Year, Month, Day, Hour, Minute, Second). Once the data warehouse was constructed, we applied intelligent methods called OLAP or data mining techniques to extract data patterns. Generally OLAP is modeled by a multi-dimensional database structure called data cubes. Our constructed data warehouse was able to provide the data source for OLAP and also -if needed- for the data mining techniques.

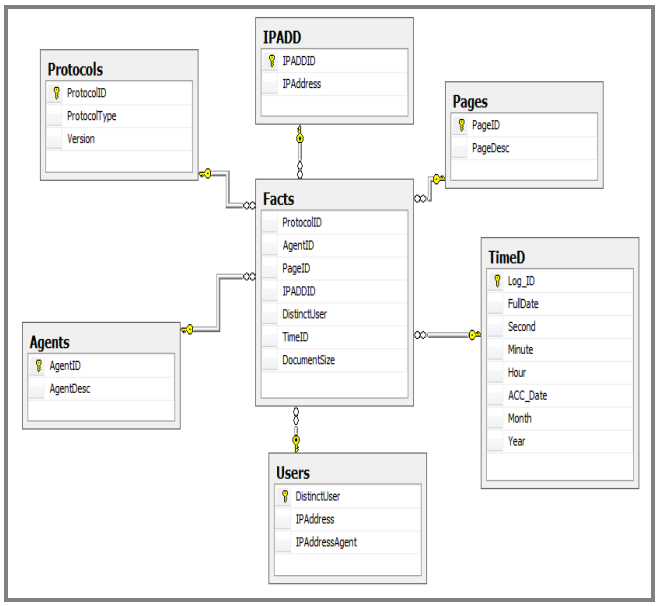


Figure 2. Web log Warehouse Schema

1) *Filling the Dimension Table*

Our dimension tables contained descriptive attributes, which were textual. These attributes were designed to form query constraint or filtering. Also they were used to label the results in the OLAP cube. They were filled directly from the attributes of the log file (i.e. the time dimension was filled from the time attribute in the log file). To accomplish the filling task, we used SQL statements within a VB.net program connected to both the log file and SQL Server.

The time attribute in the log file was divided into hours, minutes and seconds; years, months, and days. The attributes of the other dimension tables were taken, as found, from the log file (i.e. the IP Address attribute was used to fill the IP address field in the IP address dimension).

2) Filling the Fact Table

To fill the fact table, we needed to reference our dimension tables in the query. We used a temporary table as main driver of the query and then we looked up the resulting ID based on the primary keys of the dimension tables. The lookups were accomplished using the LEFT OUTER joins, which implies that the relationship may not exist in which case NULL value will go into the fact table.

D. OLAP

The most common form of pattern analysis consists of a knowledge query mechanism such as SQL (Structured Query Language), which needs end user access tools. OLAP (Online Analytical Processing) as an end-user access tool then can support advanced quarry by using a strong methodology called data cube [12]. OLAP can simplify the analysis of usage statistics of the server access logs. It pre-calculate summary information to enable roll-up or aggregation, which allows the user to move to the higher aggregation level, drilling, which is the reverse of a roll-up and represents the situation when the user moves down the hierarchy of aggregation, applying a more detailed grouping, pivoting, which changes the perspective in presenting the data to the user, slicing, which is based on selecting one dimension and focusing on a portion of a cube and dicing, which creates a sub-cube by focusing on two or more dimensions [13]. OLAP describes a set of technologies that allows analysts to quickly gain answers to the 'who' and 'what' questions premised on a, usually large, set of data. OLAP applications typically achieve this through multidimensional views of aggregate data derived from the data set. OLAP also answers tougher questions such as 'what if' and 'why' and this will be the emphasis of this paper. Some of the important questions are:

- Which are the top pages visited by user over the time?
- Which IP address accessed which site using which protocol and how many times?
- What is the distribution of network traffic over time (hour of the day, day of the week, month of the year etc.)?

Answering the above questions requires the inclusion of the time, IP address, Page dimensions and it also requires the cube to render facts such as, the number of visitors, the document size by users or by IP address as shown in Figure 3.

A. Business Intelligence Development Studio

Business Intelligence Development Studio is Microsoft Visual Studio 2008 with additional project types that are specific to SQL Server business intelligence. Business Intelligence Development Studio is the primary environment that will be used to develop business solutions that include Analysis Services, Integration Services, and Reporting Services projects. Each project type supplies templates for creating the objects required for business intelligence solutions, and provides a variety of designers, tools, and wizards to work with the objects. We used visual studio to construct the data cubes. This was accomplished by: linking database, determined dimension, determined facts table and then running the cube.

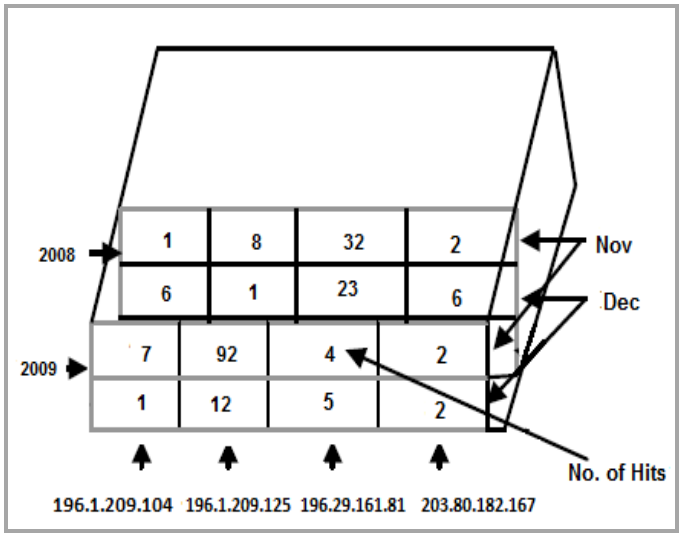


Figure 3. Data Cube

III. Experimental Results

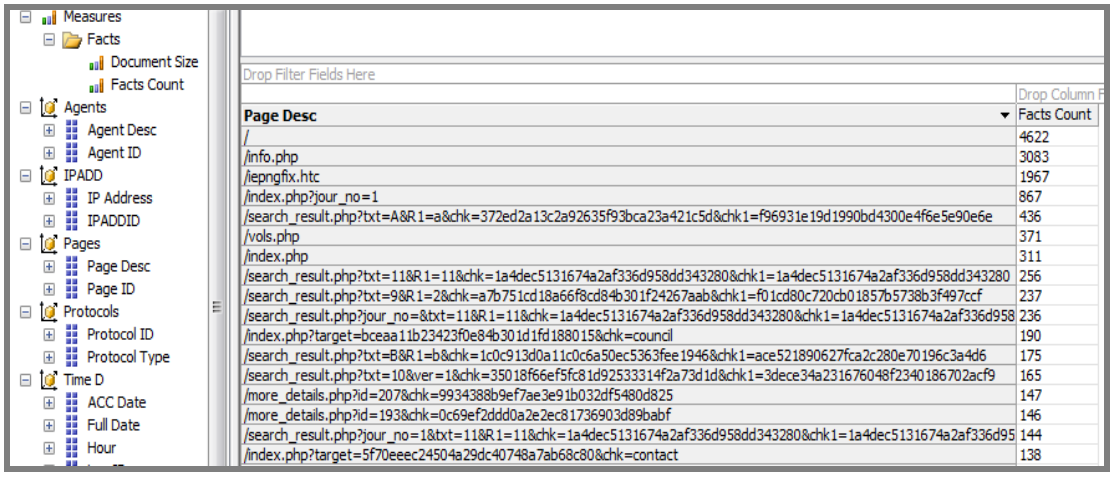
Our experiments were performed on a 2.8GHz Pentium CPU, 2GB of main memory, Windows 7 Ultimate, SQL Server 2008 and Microsoft Visual Studio 2010. During the data cleaning process, the number of the requests is reduced from 291642 to 122122. Table 1 shows the details of our data after the cleaning process.

Table 1. Statistical Summery

Number of records before cleaning	291642
Number of unique Users	23524
Number of Unique IP address	11030
Number of unique pages	8861
Number of sessions	13869

Users accessed each web page different number of times. Since each web page was not of the same interest. The top of the most frequently visited pages are illustrated in Figure 4 and the graphical representation of the top 7 visited pages are illustrated in Figure 5.

Figure 6 answers the question: Which IP address has accessed the website using which Protocol and how many times?. Figure 7 shows the number of bytes transferred on month 3 were greater than number of bytes transferred on month 10, although the number of users was equal. Also the number of bytes transferred slightly increased from month 4 through month 5 until month 6, although the number of users in these months decreased rapidly.



Page Desc	Facts Count
/	4622
/info.php	3083
/iepngfix.htc	1967
/index.php?jour_no=1	867
/search_result.php?txt=A&R1=a&chk=372ed2a13c2a92635f93bca23a421c5d&chk1=f96931e19d1990bd4300e4f6e5e90e6e	436
/vols.php	371
/index.php	311
/search_result.php?txt=11&R1=11&chk=1a4dec5131674a2af336d958dd343280&chk1=1a4dec5131674a2af336d958dd343280	256
/search_result.php?txt=9&R1=2&chk=a7b751cd18a66f8cd84b301f24267aab&chk1=f01cd80c720cb01857b5738b3f497ccf	237
/search_result.php?jour_no=&txt=11&R1=11&chk=1a4dec5131674a2af336d958dd343280&chk1=1a4dec5131674a2af336d958	236
/index.php?target=bceaa11b23423f0e84b301d1fd188015&chk=council	190
/search_result.php?txt=B&R1=b&chk=1c0c913d0a11c0c6a50ec5363fee1946&chk1=ace521890627fca2c280e70196c3a4d6	175
/search_result.php?txt=10&ver=1&chk=35018f66ef5fc81d92533314f2a73d1d&chk1=3dece34a231676048f2340186702acf9	165
/more_details.php?id=207&chk=9934388b9ef7ae3e91b032df5480d825	147
/more_details.php?id=193&chk=0c69ef2ddd0a2e2ec81736903d89babf	146
/search_result.php?jour_no=1&txt=11&R1=11&chk=1a4dec5131674a2af336d958dd343280&chk1=1a4dec5131674a2af336d95	144
/index.php?target=5f70eeec24504a29dc40748a7ab68c80&chk=contact	138

Figure 4. The top of the most frequently visited pages

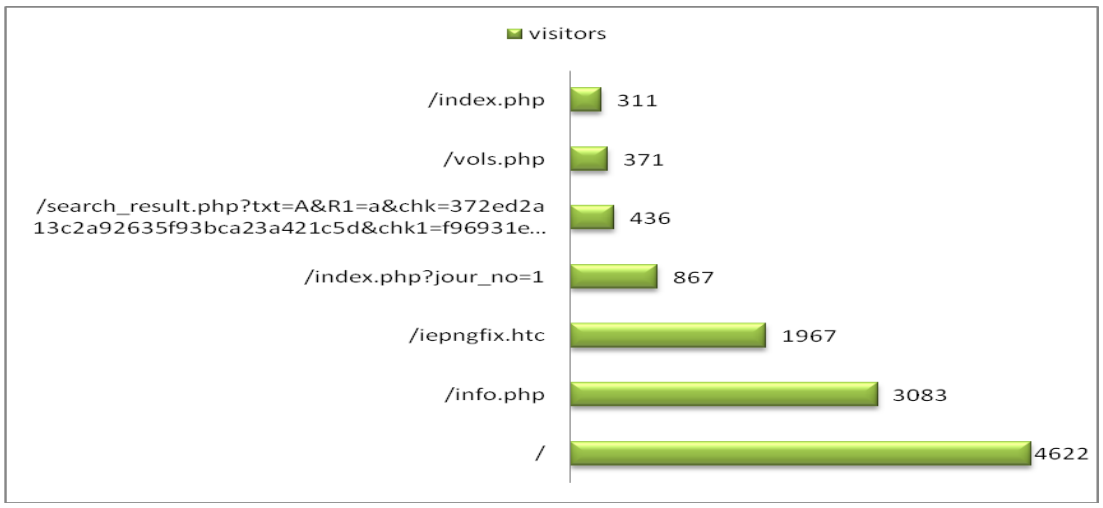
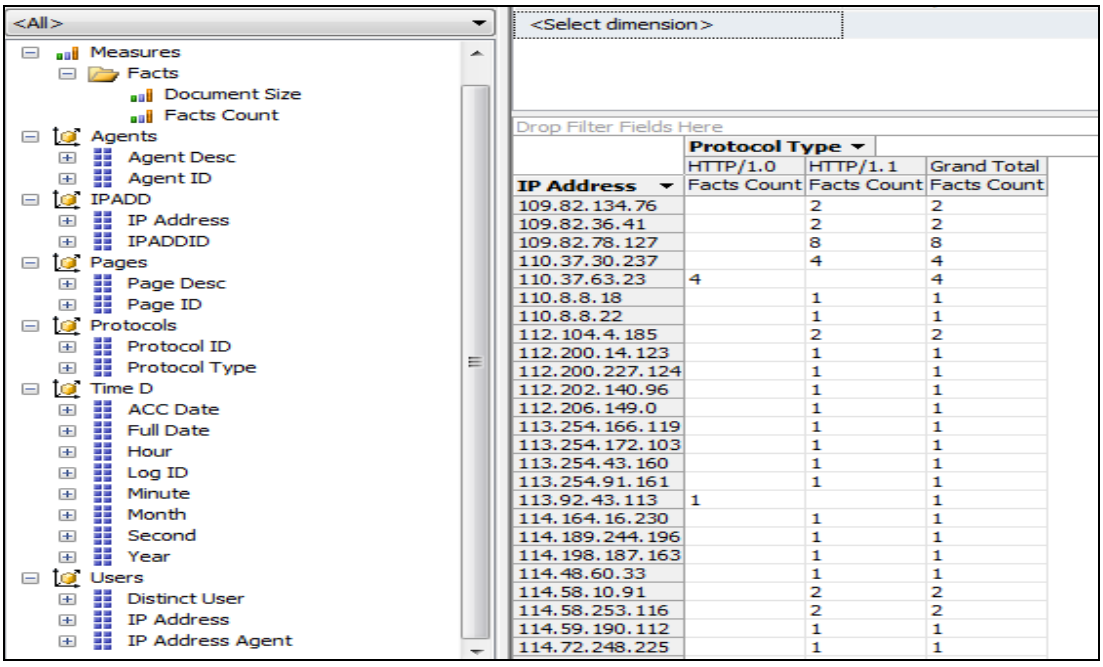


Figure 5. Graphical representation of the top 7 pages shown in Figure 4



	Protocol Type		
	HTTP/1.0	HTTP/1.1	Grand Total
IP Address	Facts Count	Facts Count	Facts Count
109.82.134.76		2	2
109.82.36.41		2	2
109.82.78.127		8	8
110.37.30.237		4	4
110.37.63.23	4		4
110.8.8.18		1	1
110.8.8.22		1	1
112.104.4.185		2	2
112.200.14.123		1	1
112.200.227.124		1	1
112.202.140.96		1	1
112.206.149.0		1	1
113.254.166.119		1	1
113.254.172.103		1	1
113.254.43.160		1	1
113.254.91.161		1	1
113.92.43.113	1		1
114.164.16.230		1	1
114.189.244.196		1	1
114.198.187.163		1	1
114.48.60.33		1	1
114.58.10.91		2	2
114.58.253.116		2	2
114.59.190.112		1	1
114.72.248.225		1	1

Figure 6. IP address accessed the web site using HTTP Protocol

Drop Filter Fields Here

	Drop Column Fields Here		
Year	**Month**	Facts Count	Document Size
2008	11	1520	32687965
	12	1780	35173103
	Total	3300	67861068
2009	1	1858	35407984
	10	2109	24277705
	11	1675	19895494
	12	557	6730123
	2	1836	35827284
	3	2109	32370546
	4	1790	25664183
	5	1619	27915992
	6	1491	36847786
	7	1799	21546580
	8	1669	19618279
	9	1712	20614214
	Total	20224	306716170
Grand Total		23524	374577238

Figure 7. Number of bytes transferred by Users

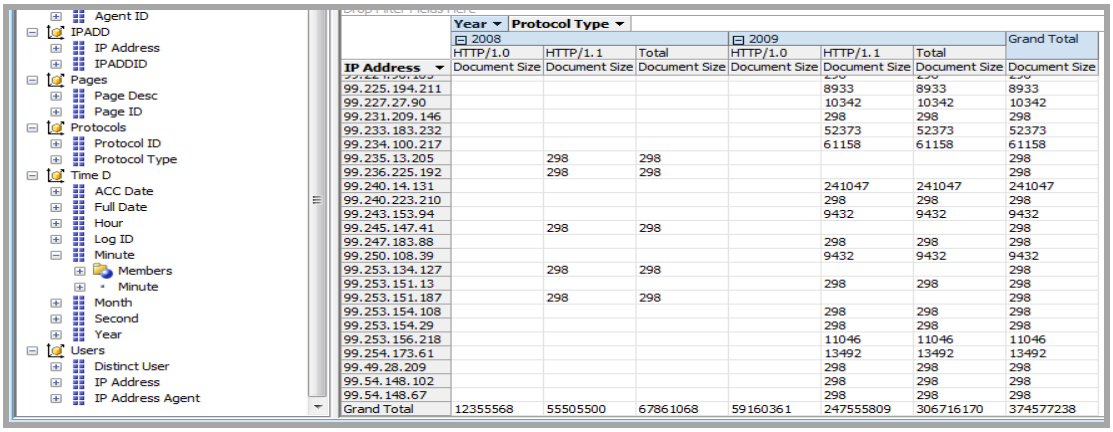
Figure 8 shows part of the Web log cube with 3 dimensions: Time, User IP Address and Protocol Type, where time was at level Year. Document size was numeric codes. For example in 2009, IP Address 99.243.153.94 used protocol HTTP/1.1 to download a document with a size 9432 MB.

Figure 9 shows the drill down operations. Here we drilled down the data cube shown in Figure 8 into months and access date in the time dimension.

Figure 10 illustrates the roll up operation over the data cube shown in Figure 8. The figure allows the user to move to month 2, which was a higher aggregation level through the minutes of the hour 17 on the days 3 and 5.

Figure 11 shows the dicing operation on the cube shown in Figure 8. This diced cube contains only two dimensions: Time (Year) and User IP Address.

Using slicing operation (as shown in the Figure 12), we were able to focus on the values of the specific cells. In the Figure 12 we sliced the data cube for day 1. We can easily see users who accessed the website on day 1/1/2009 of each hour and minute. Note the absence of the user in a few hours like 3, 4 and 5.



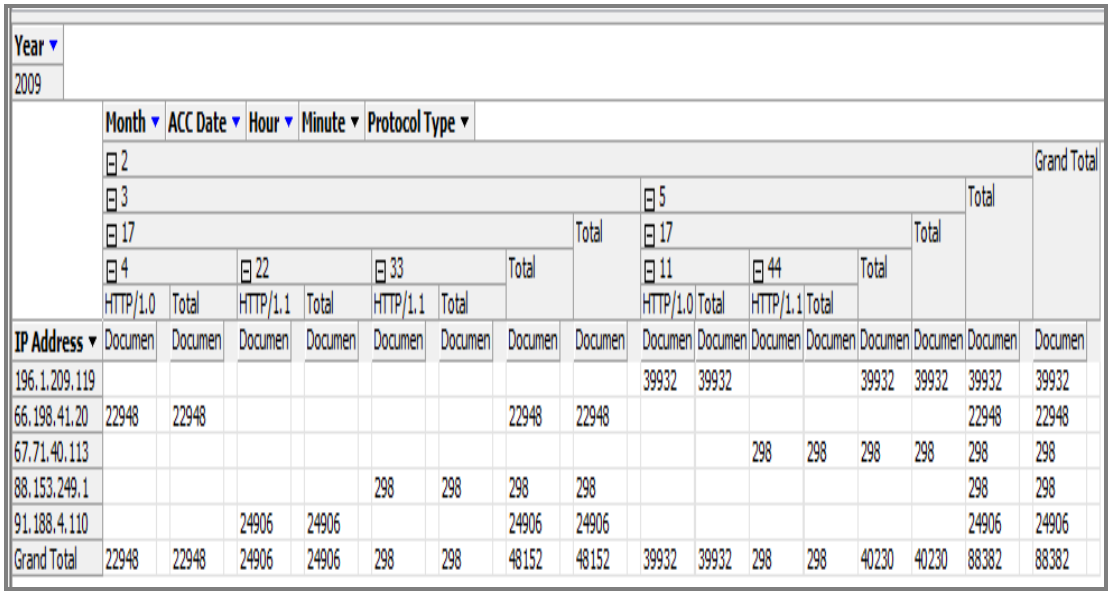
	Year / Protocol Type	2008			2009			Grand Total
		HTTP/1.0	HTTP/1.1	Total	HTTP/1.0	HTTP/1.1	Total	
IP Address		Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size
99.225.194.211						8933	8933	8933
99.227.27.90						10342	10342	10342
99.231.209.146						298	298	298
99.233.183.232						52373	52373	52373
99.234.100.217						61158	61158	61158
99.235.13.205			298	298				298
99.236.225.192			298	298				298
99.240.14.131						241047	241047	241047
99.240.223.210						298	298	298
99.243.153.94						9432	9432	9432
99.245.147.41			298	298				298
99.247.183.88						298	298	298
99.250.108.39						9432	9432	9432
99.253.134.127			298	298				298
99.253.151.13						298	298	298
99.253.151.187			298	298				298
99.253.154.108						298	298	298
99.253.154.29						298	298	298
99.253.156.218						11046	11046	11046
99.254.173.61						13492	13492	13492
99.49.28.209						298	298	298
99.54.148.102						298	298	298
99.54.148.67						298	298	298
Grand Total		12355568	55505500	67861068	59160361	247555809	306716170	374577238

Figure 8. Example data cube created having Time (Year), User IP Address, Protocol type as Dimensions and Document size transferred as a measure

Year: 2009

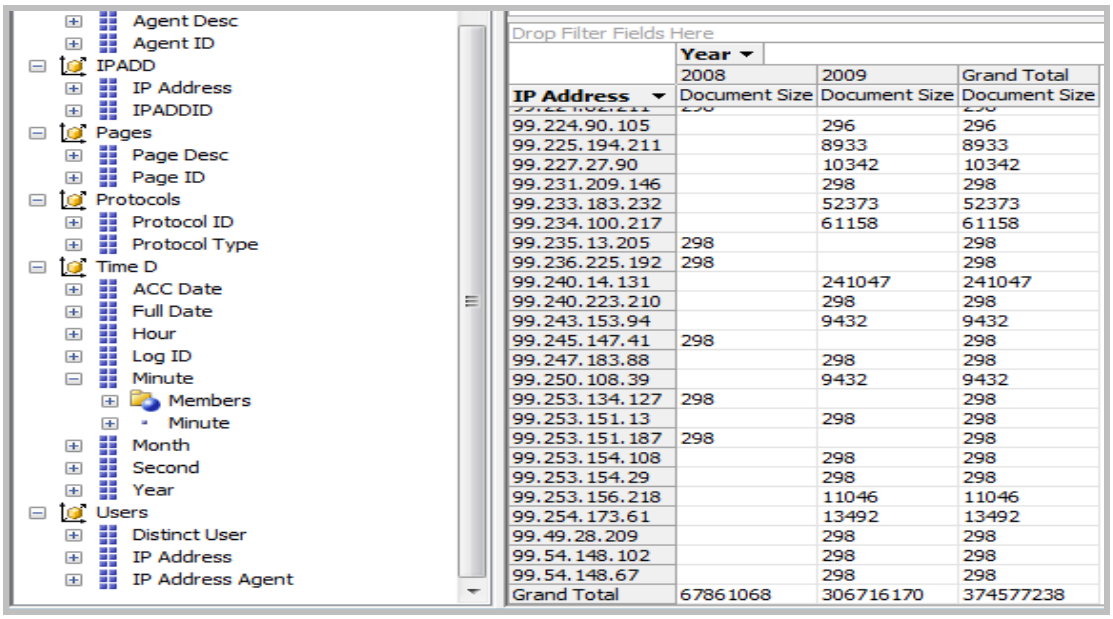
	Month / ACC Date / Protocol Type	2	3			4	5	6	7	8	9	Grand Total
			5									
			HTTP/1.0	HTTP/1.1	Total							
IP Address		Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size	Document Size
92.20.121.234				298	298	298						298
92.227.166.213		298										298
92.23.205.168		298										298
93.110.4.198		40230										40230
94.23.238.192												14003
94.27.64.86									10120			10120
94.96.55.116		61400										61400
94.97.37.61						22701						22701
94.97.93.69											50953	50953
94.98.108.201		17240										17240
94.98.32.79						7468						7468
95.170.210.4												16618
95.170.210.67									107029			107029
95.208.23.129				298	298	298						298
97.119.199.18							298					298
98.141.188.18												298
98.247.184.153									9740			9740
99.190.81.103		298										298
99.54.148.102				298	298	298						298
99.54.148.67				298	298	298						298
Grand Total		1864170	190973	714518	905491	905491	731010	755434	326866	581334	944264	242175	9866476

Figure 9. Resultant Data Cube after Drill down to Month and Access Date in the time dimension in the data cube given in Figure 8



Year	Month	ACC Date	Hour	Minute	Protocol Type	IP Address	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen	Documen
2009	2																			Grand Total
	3														5					Total
	17									Total					17					Total
	4		22		33			Total			11		44			Total				
	HTTP/1.0	Total	HTTP/1.1	Total	HTTP/1.1	Total					HTTP/1.0	Total	HTTP/1.1	Total						
						196.1.209.119									39932	39932			39932	39932	39932	39932
						66.198.41.20	22948	22948					22948	22948							22948	22948
						67.71.40.113											298	298	298	298	298	298
						88.153.249.1					298	298	298	298							298	298
						91.188.4.110			24906	24906			24906	24906							24906	24906
						Grand Total	22948	22948	24906	24906	298	298	48152	48152	39932	39932	298	298	40230	40230	88382	88382

Figure 10. Shows the rollup operation in the data cube shown in Figure 8

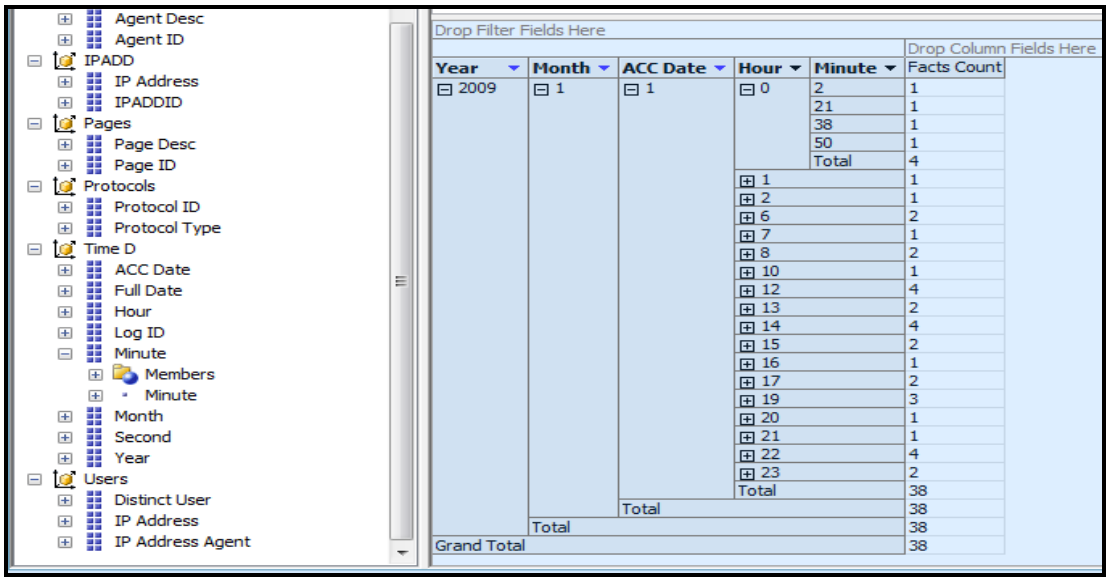


Agent Desc
Agent ID
IPADD
IP Address
IPADDID
Pages
Page Desc
Page ID
Protocols
Protocol ID
Protocol Type
Time D
ACC Date
Full Date
Hour
Log ID
Minute
Members
Minute
Month
Second
Year
Users
Distinct User
IP Address
IP Address Agent

Drop Filter Fields Here

IP Address	Year 2008 Document Size	2009 Document Size	Grand Total Document Size
99.224.90.105		296	296
99.225.194.211		8933	8933
99.227.27.90		10342	10342
99.231.209.146		298	298
99.233.183.232		52373	52373
99.234.100.217		61158	61158
99.235.13.205	298		298
99.236.225.192	298		298
99.240.14.131		241047	241047
99.240.223.210		298	298
99.243.153.94		9432	9432
99.245.147.41	298		298
99.247.183.88		298	298
99.250.108.39		9432	9432
99.253.134.127	298		298
99.253.151.13		298	298
99.253.151.187	298		298
99.253.154.108		298	298
99.253.154.29		298	298
99.253.156.218		11046	11046
99.254.173.61		13492	13492
99.49.28.209		298	298
99.54.148.102		298	298
99.54.148.67		298	298
Grand Total	67861068	306716170	374577238

Figure 11. Dicing Data cube shown in Figure 8 contains two dimensions Time and IP Address



Agent Desc
Agent ID
IPADD
IP Address
IPADDID
Pages
Page Desc
Page ID
Protocols
Protocol ID
Protocol Type
Time D
ACC Date
Full Date
Hour
Log ID
Minute
Members
Minute
Month
Second
Year
Users
Distinct User
IP Address
IP Address Agent

Drop Filter Fields Here

Drop Column Fields Here

Year	Month	ACC Date	Hour	Minute	Facts Count
2009	1	1	0	2	1
				21	1
				38	1
				50	1
				Total	4
			1		1
			2		1
			6		2
			7		1
			8		2
			10		1
			12		4
			13		2
			14		4
			15		2
			16		1
			17		2
			19		3
			20		1
			21		1
			22		4
			23		2
			Total		38
		Total			38
	Total				38
Grand Total					38

Figure 12. Slicing Data cube on the Time dimension for the day 1/1/ 2009

IV. Conclusions

In this paper, the Sudan University of Science and Technology' web log files were used to analyze the access patterns of the web server which contains a huge amount of information that if mined properly can help in taking the right decision to improve system designed. The log file normally contains a huge amount of information that needs to be organized and cleaned. The cleaning process was achieved by removing irrelevant data. Then the data was uploaded in a data warehouse in form of dimension tables and fact table. The organization of the log file was achieved by grouping the data into unique users, unique IP address, protocol, pages, and agent. Cleaned and organized data were presented in the form of a cube, the basic structure that can be used by the Online Analytical Process (OLAP). The results achieved have proved that the data warehouse can be implemented successfully to analyze the log files to make appropriate decisions.

References

1. V. Singhal, and G. Pandey, "A Web Based Recommendation Using Association Rule and Clustering," *Int. J. Comput. Commun. Eng. Res.*, vol. 1, no. 1, pp. 1–5, 2013.
2. F. Y. Tani, D. Farid, and M. Rahman, "Ensemble of Decision Tree Classifiers for Mining Web Data Streams," *Int. J. Appl. Inf. Syst.*, vol. 1, no. 2, pp. 30–36, 2012.
3. H. Rana and M. Patel, "A Study of Web Log Analysis Using Clustering Techniques," *Int. J. Innov. Res. Comput. Commun.Eng.*, vol. 1, no.1, pp. 925–929, 2013.
4. G. Katkar and A. Kasliwal, "Use of Log Data for Predictive Analytics through Data Mining," *Curr.Trends Technol. Sci.*, vol. 3, no. 3, pp. 217–222, 2014.
5. M. Helmy, A. Wahab, M. Norzali, H. Mohd, H. F. Thanasi, M. Farhan, and a Background, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm," *Proc. World Acad. Sci. Eng. Technol.*, vol. 36, no.1, pp. 970–977, 2008.
6. Om Kumar C.U. & P. Bhargavi, "ANALYSIS OF WEB SERVER LOG BY WEB USAGE MINING FOR EXTRACTING USERS PATTERNS," *Int. J. Comput. Sci. Eng. Inf. Technol. Res.*, vol. 3, no. 2, pp. 123–136, 2013.
7. R. Gupta and P. Gupta, "Application specific web log pre-processing, *Int.J.Computer Technology & Applications*, vol. 3, no. 1, pp. 160–162, 2012.
8. P. Nithya and P. Sumathi, "An Effective Web Usage Analysis using Fuzzy Clustering," *ARPN J. Sci. Technol., vol.* 3, no. 7, pp. 693–698, 2013.
9. K. B. Patel, "Process of Web Usage Mining to find Interesting Patterns from Web Usage Data," *International Journal of Computer Applications & Technology*, vol. 3, no. 1, pp. 144–148, 2012.
10. I. Technologies, "Identifying the User Access Pattern in Web Log Data," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 2, pp. 3536–3539, 2012.
11. S. Asadianfam and M. Mohammadi, "Identify navigational patterns of Web," *Int. J. Comput.Technol*., vol. 1, no.1, pp. 1–8, 2014.
12. K. J. Grace and D. Nagamalai, "Analysis of Web Loga and Web User," *Int. J. Netw. Secur.Its Appl.*, vol. 3, no. January 2011, pp. 99–110, 2011.
13. R. K. Jain, S. Jain, and R. S. Kasana, "On Line Analytical Mining of Web Usage Data Warehouse," *Int. J. Comput. Sci. Emerg. Technol.*, vol. 1, no. 1, pp. 15–24, 2010.