ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Big Data Analytics

Rohit Kapdoskar[1], Sanket Gaonkar[2], Nihar Shelar[3], Akshaya Surve[4], Prof.Sachin Gavhane[5]

Student, Information Technology, Atharva College Of Engineering, Mumbai, India [1,2,3,4]

Professor, Information Technology, Atharva College Of Engineering, Mumbai, India [5]

Abstract: Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. The main goal of this project is to understand and implement the entire process of data mining and analytics. We will be extracting the information from data sources by implementing a web crawler. To remove the inconsistencies in the extracted data we will be cleaning it. The cleaned data will be migrated to database, analyzed and visualized.

Keywords: web crawler, open refine, visualization, analytics.

I. INTRODUCTION

Modern day systems produce tons of data. The volume, velocity, variety and veracity of data coming into the organizations has reached to a ground breaking level. This data contains a lot of information stored in the form of hidden patterns, unknown correlations which can be used to make better decisions. Analytics is making the businesses smarter and more productive by making better predictions by analyzing the trends in the data.

This paper is an attempt made to deep dive into analytics domain by performing some analysis over the car sale dataset. The analysis is made by keeping in mind information which can be extracted from the inventory car sale data that can prove to be useful to sales people and managers to improve the sales and overall profit. Questions for analysis can be like car sale across the various geographical locations every year, car sale by its type etc.

A lot of similar researches have been done is this field and many applications are there in the market for the same. But our project is a sincere effort made to learn step by step how analytics is actually performed.

II. LITERATUR SURVEY

A. Web Crawler

A web crawler is a mechanism used in search engine which helps search engines in finding and exploring the web. It is a algorithm for downloading various web pages automatically. It is an important software for compilation of data. It is also called as web spider. There are multiple types of web crawler for eg. Incremental web crawler.

Incremental web crawler [5] updates a current set of download pages rather than reinstating the crawling process from the start.

B. Data Cleaning

A lot of data is created everyday by various organization to get the best business decisions and profit it is necessary to observe the generated data. To observe the data, a data warehouse is the only solution. For the future aim the faultlessness of the data is very important. Thus, this quality data can be generated only after cleaning the data before adding it in the data warehouse.

Openrefine is an open source tool for data cleaning. It is a powerful tool for working on noisy data and cleaning it. Openrefine accepts data from various sources, analyzes different datasets quickly and apply various cell transformations on the data.

C. Data Integration

Data integration is basically linking of data from different sources and it provides a collective view of the data.[6,7,8]

D. Data Visuaization

While building visualizations, graphics developer mostly use multiple tools concurrently. This can be seen mostly in many websites, were collective visualizations combine different technologies. But sadly, this euphoric interoperability is mostly lost with visualization toolkits due to encapsulation of DOM with more functional forms.

Data Driven Documents (D3) is used for visualization process. D3 enables direct analysis and handling of a native delegation for HTML, SVG, CSS but D3.js is implemented on all the above standards. It has a great control over the ultimate visual outcome.

III. AIM AND OBJECTIVE

Aim is to develop a web crawler to extract data from websites and using data preprocessing techniques such as cleaning, integration and visualization.

IV. PROPOSED SYSTEM

- Information Gathering or Data collection
- Data Cleaning
- Data Integration
- Data Visualization

Above steps can be shown as follows:

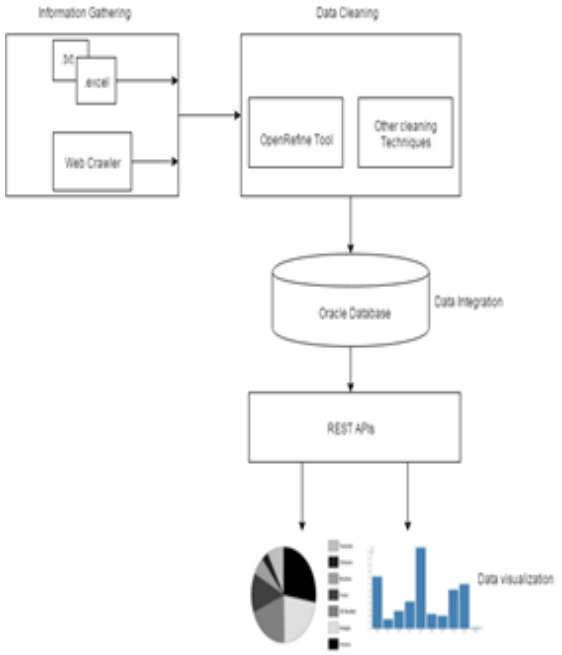
A. Information Gathering

Data collection stage mainly focus of gathering data from various sources. Analytics requires lot of data. We have made a web crawler using python to acquire data from

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

websites. The data obtained from web crawler is in XML format. We have also collected inventory data from various websites which is in form of flat files, excel sheet.



B. Data Cleaning

The raw data that is obtained may contains lot of inconsistencies. Such data cannot be directly used for analysis. Even when the data obtained may contains some noise (age of car is negative). We need to ensure that all such data is cleaned before it is used for analysis. For cleaning of data we are using OpenRefine a data cleaning tool.

C. Data Integration

We have integrated the cleaned data into a structured database using Oracle. For performing the data management NoSQL databases like MongoDB, Hadoop are preferred as they have ability to retrieve data a lot faster than structured databases like Oracle. But in our case study the data is not very big. Considering the size of data, structured databases like oracle will also give a very good query performance as any NoSQL database. Hence we are integrating data into a centralized Oracle database using Oracle 10g. We will be writing scripts to load cleaned data from various disparate sources into Oracle database.

D. Data Visualization

The data is analyzed using D3.js framework. D3 has a wide gallery of visualizations which can be used to analyze data of any format as per the requirement. D3 requires input in the form of json file. So we will be constructing REST API using Spring framework and publish web services using Apache tomcat as webserver. To build REST API we will be using MVC architecture. There will be a controller layer which will act as an endpoint to interact with the front end. The Service layer which will do all the processing of the data and return controller a JSON in the require format for a given query. The data access layer will perform actual data extraction from database.

V. ALGORITHM IMPLEMENTATION

A. *Proposed algorithm for web crawler*

```
Urls[]
pageVisited[]
xmlText=open("output.xml","w",buffering=20*(1024**2)
)
Add url of the first page to Urls[]
For url in  urls
    Open the URL
    Convert html to text
    Links = Find all <a> "anchor tags from the text
    Remove the url from  Url[]
    For link in Links
        If link contains 'limit=' and not 'lang='
        Get "href" for this <a>
        If page (link) not in pageVisited
        Add page to pageVisited
       Extract all content on page
       Generate xml element of the content     extracted.
      Append the xml generated to the root xml
Convert xml generated to String
Write string to the file
 Close file.
```

VI. CONCLUSION

This paper has discussed the technique of Big Data Analysis. The proposed work is an effort to suggest an approach for handling the Big Data. Approach suggested from the beginning of making a web crawler then retrieving information then cleaning and integration of the data and the visualization of the data has been stated. This work will surely be useful for organizations to manage the data.

ACKNOWLEDGMENT

We are thankful to all the authors for blooming our knowledge about big data analytics and providing us with information about web crawlers and data preprocessing techniques. We thank **Prof. Sachin Gavhane** for providing us with all the resources and for his continuous support and motivation. We thank principal and HOD Information technology engineering department, ATHARVA. We are extremely thankful to all staff and the management of the college for providing us all the facilities and required resources.

REFERENCES

1. Mini Singh Ahuja, Dr Jatinder Singh Bal and Varnica (2014), "Web Crawler: Extracting the Web Data", International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014, ISSN: 2231-2803.
2. Kamran Ali and Mubeen Ahmed Warraich, "A framework to implement Data Cleaning in Enterprise Data Warehouse for Robust Data Quality", 978-1-4244-8003-6/10 ©2010 IEEE.
3. Maurizio Lenzerini ," Data Integration: A Theoretical Perspective", Dipartimento di Informatica e Sistemistica Universit `a di Roma "La Sapienza", Via Salaria 113, I00198, Roma, Italy, lenzerini@dis.uniroma1.it.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

4. Michael Bostock, Vadim Ogievetsky and Jeffrey Heer, " D3: Data-Driven Documents", (2011).
5. Nemeslaki, András; Pocsarovszky, Károly (2011), "Web crawler research methodology", 22nd European Regional Conference of the International Telecommunications Society.
6. A. Y. Halevy. Answering queries using views: A survey Very Large Database J., 10(4):270–294, 2001.
7. R. Hull. Managing semantic heterogeneity in databases: A theoretical perspective. In Proc. of 16th ACM SIGACT SIGMOD SIGART Symp. On principles of Database systems (PODS'97),1997.
8. F. Naumann, U. Leser, and J. C. Freytag. Quality- driven integration of heterogenous information systems. In Proc. of the 25th Int. Conf. on Very Large DataBases (VLDB'99), pages 447–458, 1999.
9. Vladislav Shkapenyuk, Torsten Suel, "Design and Implementation of a High-Performance Distributed Web Crawler", NSF CAREER Award, CCR-0093400.
10. S. Chaudhuri, K. Ganjam, V. Ganti, "Data Cleaning in Microsoft SQL Server 2005", In Proceedings of the ACM SIGMOD Conference, Baltimore, MD, 2005.