Title: Crime Analysis By Implementing Document Oriented Data Warehouse

Abstract:

1: Introduction:

Analysis of data is the process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making [1]. Data analysis have diverse techniques and approaches and is used in different domains such as social science and business [2].

A data warehouse is a system used for reporting and data analysis. Traditional data warehouses are unable to meet the growing needs of the modern enterprise to integrate and analyze a wide variety of data .These data can be structured, unstructured or semi-structured. These data are defined as collections of data that are as large and complex to manage as it has become almost impossible to store and process using traditional database management tools [6].

Unstructured data is a significant challenge for scientists because it often takes a lot of time to structure and prepare for analysis [4]. The analyzing and processing the unstructured data, formatting and merging it with traditional structured data provides decision makers with a better insights[5]. Therefore, the relational databases, which have been the perfect support for storing data for many decades, are no longer suited to the Big Data phenomenon. In this context that we are witnessing the appearance of the Not-only-SQL systems, which present an alternative to relational databases. This new generation of databases has the power not to handle only a large amounts of data, but also the variety and the velocity of data. These databases are used by the largest organizations in the world such as Google, Amazon, etc. Most NoSQL databases are schema-free or at least have very relaxed schemas. There is often no need to define any sort of schema for the data.

This paper deals with the design of a data warehouse from NOSQL document oriented database for the analysis of the crime dataset.

This article is structured as follows , in section-2 , we present some literature review about the data warehouse. Section-3 summarizes about the platform and technology used.

Keywords: NOSQL , Data Warehouse , MongoDB, MERN

2: Literature Review :

In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis and is considered a core component of business intelligence[3].A traditional relational database is well known and useful for a long time. However, its rigid structure is inflexible and inefficient for semi-structured and structured data. Nowadays, there are popular alternative databases such as NoSQL. A NOSQL data warehouse is designed generally for non tabular data and they are able to combine data of any structure without the need to modify the schema[7][8].

The four types of NOSQL databases are key- value,columnar,document and graph databases[9].

Key-value databases are simplest NOSQL models in which data is represented as key value pairs . A key may be simple (e.g. hash or filename) or structured (e.g. composite) . A data represents arbitrary data type , structure and size (e.g image , document, string) that is uniquely identified by indexed key[10].

Columnar are extended key value databases in which key is represented as row and value is represented as sequence of nested column families (key-value pairs) .Column families is constructed by arbitrary number of columns that are logically related to each other and usually accessed together. Data in this type of database an be effectively partitioned horizontally and vertically, making them ideal for storing huge datasets and performing efficient aggregate queries [11].

Document databases are extended key value databases in which value is represented as data encoded in formats such as XML , JSON or BSON (binary JSON) . Contents of document is represented as attribute-value pair which can be modified at run time. This type of database provided flexible , self-describing schema ideal for modern web applications with the focus on high development productivity and low maintainence cost [12].

Graph databases maintain graph oriented datasets . The content about the graph theoritical foundation is stored in these databases , consisting of vertices representing entities and edges representing relationships between then . Thus they store datasets in efficient manner and perform querying and analyzing by traversing through entity relationships [13].

There are some characteristics of traditional and modern data warehouses such as purpose, scope, data source , technology , end user and architecture [14].

|  |  |  |
| --- | --- | --- |
| Characteristic | Traditional Data Warehouse | Modern Data Warehouse |
| Purpose | Used for decision making process by treatment of collected data for a specifi business area which is integrated , non volatile and time varying. | Processing any structured data from numerous sources . The volume of data exceedes the traditional tools ability to operate on. |
| Scope | Business Initelligence (BI) , Online Analytical Processing (OLAP) | Analyzing and discovering insight from large volumes of data characterized by 4V’s (volume , velocity , variety , veracity ). |
| Technology | Matured technology with free or licence software | Technology is still progressing .E.g. Hadoop |
| End user | Business analysts , top managements that do not require the specific knowledge of technology. | Data scientists who have the knowledge of mathematical statistics , algorithm s. |
| Architecture | Extract, Transform , Load . Star-schema is the commonly used architecture model | Depends on problem . There is still no reference architecture. |

3: Platform and Technology used :

The structure of NoSQL is different from that of a RDMBS . However , most of the work to create and design a data warehouse deals with RDBS based approaches. In the field of emerging NoSQL databases , few works have mentioned the technologies of building a data warehouse.

Recent studies that seek to improve data warehouse efficiency explores the use of various technologies including one we used in building data warehouse , MongoDB.MongoDB is recognized as a promising new technology . It is NoSQL document oriented database with extreme query aggregation properties. It supports high availability , horizontal scaling , ad hoc queries , indexing and real time aggregation.

A nodejs express server is deployed for listening to the requests of multiple clients. Express is a minimal and flexible Nodejs web application framework that provides a robust set of features for web and mobile applications. Express provides a thin layer of fundamental web application features, without obscuring Nodejs features. It is a lightweight server which handles communication with the data layer and processes integrated apis.

For the user interface , the react library is ideal for creating interactive html pages to provide satisfying experience to users. React is also an architecture , a framework that helps in implementing asynchronous flow of program.

4: Designing methodology:

4.1: Database collections:

In this paper , the dataset is provided by an open source database [15]. It contains government recorded files about different kinds of criminal cases and activities around a country in tables.This data needs to be cleansed, transformed , catalogued and made available for use for interested professionals. In NoSQL database , three collections are generated , a collection to store raw data from main source , a collection for searching purposes and a collection for storing operable data .

Operational Collection

(Data Layer)

Raw Collection

(Staging Layer)

raw\_id wc\_id

Words Collection

(Search Layer)

The raw collection stores the files from the main source and generates a unique id for each file this unique id is used as a reference in words collection. The data of the file is cleaned and structured each row is transformed into a document with a reference id from words collection and stored in data layer.

4.2 Algorithms:

4.2.1 Search Component:

The dataset contains files with specific names that is used as the data source for search layer . The filename is a string which is cleaned , transformed and stored in the words collection. Each filename has words which are seperated and saved as a list . A document is created which have key values pairs . For the search algorithm this document have a special key value pair named “words array” which references to the list of words.

The user interface has a search bar which queries the related searches about the typed content . The typed content is sent to the server where server unpacks this content into a list of words and fetches the data from the search layer using aggregated queries and selects those documents whose “words array” key’s value , which is a list of words itself matches with the typed content partially. The pseudo code of this algorithm is

Function match(from\_user , from\_collection)

{

Match = false

Travel in from\_user

{

Test = Pick current word from from\_user

Travel in from\_collection

{

Target = Pick current word from\_collection

If test matches with target partially (substring)

Or vice-versa then set match to true and break this traversing

}

If match is true then break this traversing

}

Return match

}

4.2.2 Filters component:

Dataset contains files in which data is present in table. Each row of table gets converted into a document and stored in operational collection. The various key-value pairs of these documents have the value attribute as either a string or a number. An example of such a pair is State/UT = Gujrat , Year = 2012. So an algorithm is required which first selects those keys which have values only as string while making exception for the Year. Once those keys have been identified then a list of unique values is created. The pseudocode of filters list generation is :

Function Create Filters (operational documents list)

{

Filters = [] // list of filters selected

Keys = [] // list of keys in first document

Travel in keys

{

Initialize current key

document = Pick first document from the list

If (document[current key] is not a number)

{

// Initialize a new entry in filters as

Filters[current key] = [];

}

Else

{

If ( current key is year)

{

Filters[current key] = [];

}

}

}

Travel in list of documents

{

Initialize current document

// travelling in a document will provide a list of all //the key value pairs that document have

Travel in current document

{

Initialize current key and current value

If (current key is present in filters)

{

If (current value is not present in filters[current key])

{

Push current value in the filters[current key]

}

}

}//close current document traverse

}//close all documents traverse

Return filters

}

4.3: User Interface :

To request the server for accessing the datasets for analytical processing we proposed a simple yet effective user interface for searching the content and visualizing the data. A search bar is created which uses the search algorithm on every keystroke to list down suggestions exactly as how the commercial search engines generates suggestions based on user typed content in the search bar input field. The search bar and suggestions list is shown in figure 1



Figure 1

After searching , a list of related content is generated as shown in figure 2. User can click and the data from the data layer for that specific topic is fetched for further categorization.



Figure 2

A List of filters is placed on the left side of the screen.Each filter have select all , clear and already active values will be displayed as shown in figure 3.

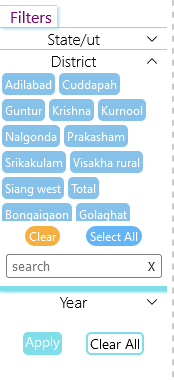
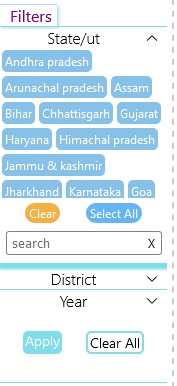
 

Figure 3

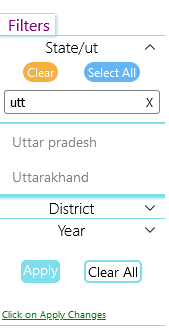


Figure 4

Each filter have their own search bar which user can type into to look for a specific value as shown in figure 4. A list showing the related values is generated in which user can select multiple values.

Once the filters have been selected then a reminder will pop to remind the user to apply the selected settings to generate graphs as shown in figure 4.

After the filters have taken effect the right side of the screen is updated to show the list of contents on the current screen page (figure 5)

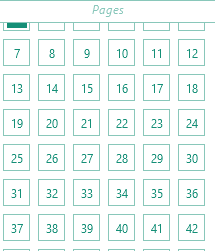


Figure 5

The contents list shows the title of the graph which user can click to auto-scroll the central part of the screen for viewing the results. The datasets is of huge volume and rendering such information will slow down the system so pagination is required to limit down the memory usages of the browser. A grid of page numbers is created which updates based on the filter settings , user can click on any page to jump to that page.

The central section of the screen shows the graphical representation in order to summarize the data.A list to the references is also generated to simplify the layout settings. The list and graphs are shown in figure 6.

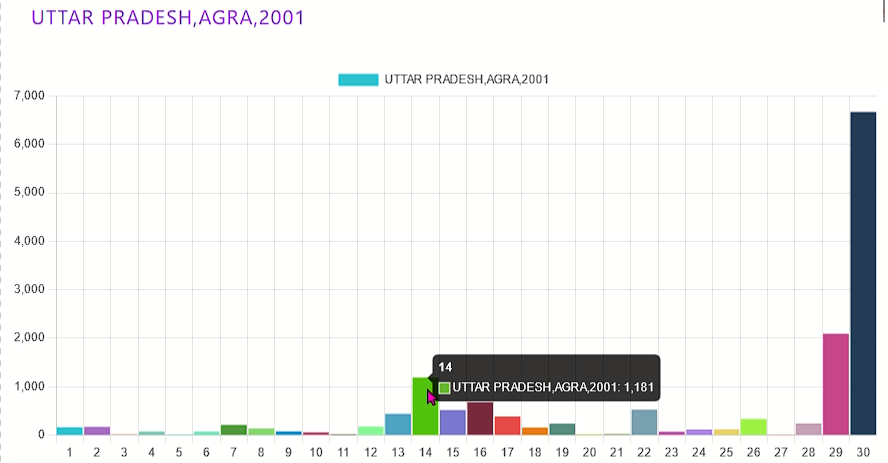




Figure 6

References:

1. "Transforming Unstructured Data into Useful Information", Big Data, Mining, and Analytics, Auerbach Publications, pp. 227–246, 2014-03-12, doi:10.1201/b16666-14, ISBN 978-0-429-09529-0, retrieved 2021-05-29
2. "The Multiple Facets of Correlation Functions", Data Analysis Techniques for Physical Scientists, CambridgeUniversity Press, pp.526576,2017,doi:10.1017/9781108241922.013, ISBN 978-1-108-41678-8, retrieved 2021-05-29.
3. Dedić, Nedim; Stanier, Clare (2016). Hammoudi, Slimane; Maciaszek, Leszek; Missikoff, Michele M. Missikoff; Camp, Olivier; Cordeiro, José (eds.). An Evaluation of the Challenges of Multilingualism in Data Warehouse Development. International Conference on Enterprise Information Systems, 25–28 April 2016, Rome, Italy (PDF). Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016). Vol. 1. SciTePress. pp. 196–206. doi:10.5220/0005858401960206. ISBN 978-989-758-187-8. Archived (PDF) from the original on 2018-05-22.

[4] Orobor, I.A., 2016. Integration and analysis of unstructured data for decision making: Text analytics approach. International Journal of Open Information Technologies 4, 82–88

[5] Gupta, V., Rathore, N.S., 2013. Deriving business intelligence from unstructured data, in: International Journal of Information and Computation Technology, pp. 971–976

1. “Design a Data Warehouse Schema from Document-Oriented Database” . Senda Bouaziz, Ahlem Nabli, Faiez Gargouri , in : 23rd International Conference on knowledge based and intelligent information & engineering systems.

[7] Mior, M.J., Salem, K., Aboulnaga, A., Liu, R., 2017. Nose: Schema design for nosql applications. IEEE Trans. Knowl. Data Eng. 29 (10), 2275–2289

[8]Strohbach, M., Daubert, J., Ravkin, H., Lischka, M., 2016. Big Data Storage. Springer International Publishing, Cham, pp. 119–141.

[9] Cattell, R., 2011. Scalable sql and nosql data stores. Acm Sigmod Rec. 39 (4), 12–27.

1. Davoudian, A., Chen, L., Liu, M., 2018. A survey on nosql stores. ACM Comput. Surv. 51 (2)

[11] Samaraweera, G.D., Chang, J.M., 2021. Security and privacy implications on database

systems in big data era: A survey. IEEE Trans. Knowl. Data Eng. 33 (1), 239–258.

[12] Eyada, M.M., Saber, W., El Genidy, M.M., Amer, F., 2020. Performance evaluation of iot data management using mongodb versus mysql databases in different cloud environments. IEEE Access 8, 110656–110668.

1. Pusala, M.K., Amini Salehi, M., Katukuri, J.R., Xie, Y., Raghavan, V., 2016. Massive Data Analysis: Tasks, Tools, Applications, and Challenges. Springer India, New Delhi, pp. 11–40.
2. Leo Williyanto Santoso , Yulia , Data Warehouse with Big Data Technology for Higher Education , 4th Information systems International Conference 2017 ISICO 2017 , 6-8 November 2017 , Bali , Indonesia.

[15] Crime in india, https://www.kaggle.com/datasets/rajanand/crime-in-india