International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Pre-Processing of Big Data: ETL Process for Weather Monitoring Applications

Ms. Kalluri Pratibha Bhaskar Reddy[1], Prof. Trupti K. Dange [2]

[1]TE, Computer department, RMD Sinhgad School of Engineering, Warje, Pune -58, Maharashtra, India
[2]Asst. Professor of Computer Department , RMD Sinhgad School of Engineering, Warje, Pune -58, Maharashtra, India

---

Abstract - *Big Data is new ubiquitous term that is used to describe the massive collection of datasets which are difficult to process by using traditional database and the old software techniques. This data is inaccessible to users. In order to make it consumable for decision-making, we need the technology and tools to find, transform, analyze, and visualize. One aspect of Big Data research is dealing with the Variety of data that includes various formats such as structured, numeric, unstructured text data, email, video, audio, stock ticker, etc. In this paper, our prime focus is on managing, merging, and governing a variety of data. In this paper, data in a weather monitoring and forecasting application is focused that helps in analyzing the global warming parameters, the natural calamities alerts to warn humans and scientists in advance with the help of semantic Extract-Transform-Load (ETL) framework.*

Key Words: **Big Data, ETL Process, Pre-Processing of data, Data warehouses.**

MOTIVATION:

In this virtual era, Data had become the most important factor. In such circumstances, BIG DATA has a boom.

The future scope of BIG DATA had tremendously taking a pace to make the firms much easier to work with.

Big Data market is constantly increasing each year. In March 2012, The White House announced a national "Big Data Initiative" that consisted of six Federal departments and agencies committing more than $200 million to big data research project

1. INTRODUCTION

Big data is a field which is used to extract information, analyze, transform the large amount of data that are complex to deal with the traditional data processing applications. The data with excessive cases tend to offer statistical power with high complexity which could lead to false outputs.

Big data helps in transforming large amount of data to useful data for the different areas of science, business, healthcare, finance, etc.

Big data helps to perform transformation of unstructured data, semi-structured data into structured data. It is defined with the help of 5 V's[6].

They are:

Volume, Variety, Velocity, Veracity, Value.

Volume: The data that is used to perform functions on the very complex due to which the (size) volume of the data is massively large.

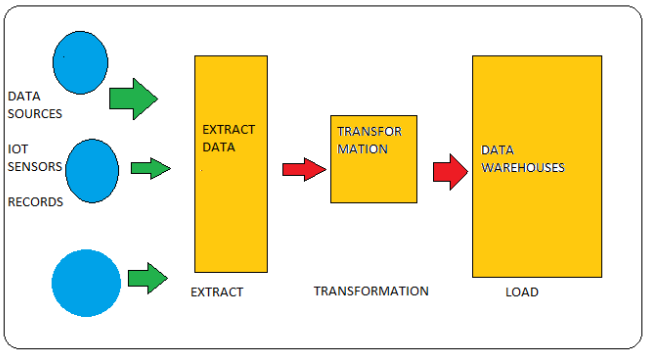
Variety: The data acquired is in different forms. It is unstructured, semi-structured, or even sometimes structured.

Velocity: This aspect deals with the speed and velocity of the data acquired by the applications.

Veracity: This attribute helps in determining the accuracy, effectives of the data.

Value: The data acquired must be of good value. Just abundance of data won't make the data useful.

In this paper we have used the ETL process to perform the process of the extracting the data, transforming the data into useful data and then loading them into the data warehouses which will be used for further weather analysis.



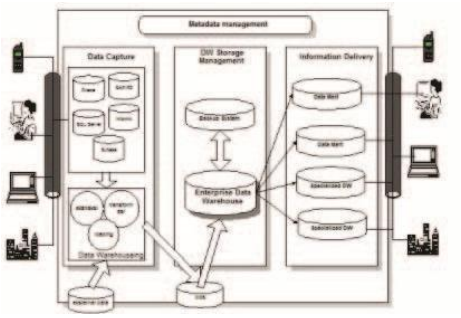
ETL Process for weather applications

The three main phases of ETL process are:

1)Extraction: Large amount of data is extracted for different sources which can be in the form of structed data, unstructured data or semi-structures data. These data can be in any format.

2)Transformation: The data that is being extracted need to be filtered, cleansed, normalized for acquiring the required data.

3)Loading: The data that is acquired after the transformation will be loaded to the allocated data warehouses.



Data warehouse architecture [5]

2. LITERATURE SURVEY.

In the year 2014, S. K. Bansal [2] researched on the semantic ETL process which explained Semantic Extract Transform-Load (ETL) framework that uses semantic technologies to integrate and publish data from multiple sources as open linked data.

In 2015, Taleb [1] clearly stated the step to be undertaken to get a required flawless structured data to perform the weather analysis for further assistance. The data must undergo process like data normalization, removal of duplicates, etc. to acquire the required data.

The following functions are performed in order to

a) Data integration

b) Data Enhancements and Enrichment

c) Data transformation

d) Data reduction

Extract-Transform-Load (ETL) as shown [3][4] is one of the popular approaches to data integration. The authors described a taxonomy of activities in ETL along with a framework that uses a workflow approach to design ETL activities. A declarative database programming language called LDL was used in order to define the semantics of ETL activities.

Wijaya in 2015 [6] along with his colleagues explained the different strategies that are required when the load of data into the warehouse. There are specific requirements to be met like metadata requirement, etc.

Ashish Juneja as with the co-authors [7] did as amazing work by proposing addressing various aspects of the raw data to improve its quality in the pre-processing stage, as the raw data will be in an unstructured or semi structured format.

Last year in 2019, Petre Lameski with his co-authors [8] explained the cloud warehouses which are another source of storing the large structured data which mainly focuses on cluster-optimization

3. LIVE SURVEY

In this research, the ETL process in used to perform data cleansing to acquire the required data that will be used to predict the weather status of the region which will help in global warming analysis, flood forecasting, etc. The data of different attributes are collected like temperature, humidity, moisture to perform the analysis.

A Weather Monitoring and Forecasting Application are conceptualized to understand the pre-processing of data and further ETL process. The data is being received and captured from many sources like IOT Sensors for various weather parameters.

The application will extract the data of the last 10 decades along with the present weather data of different location to monitor the changes in the temperature.

According to UN If the difference in the temperature exceeds by even by 2 degrees, the chances of all the corals to vanish are about 70-90% by 2100.

The data of small regions could be acquired to make effective models and then add them up to work it over a large geographical area.

The data loaded after the transformation for the rainfall analysis helps the forecasting applications to predict the harmful weather conditions and help the users to stay safe from the circumstances.

4. CONCLUSIONS

Thus, we have successfully acquired the structured data by performing the process of (ETL) Extract- Transform–Load for monitoring the weather. Big Data is a powerful area of work where Integration of data from various heterogeneous sources into a meaningful data model that allows intelligent querying Traditionally, Extract-Transform-Load (ETL) process has been used for data integration in the industry. In this paper, semantic ETL framework that uses semantic technologies are proposed to produce rich and meaningful important knowledge about data integration and also

produce semantic data that can possibly be published on the internet and contribute to the community of data. Successful creation of such a framework will be of tremendous use to various innovative Big Data applications as well as analytics. In order to test the proposed technique, we used the semantic ETL process to integrate a few public data sets with information on temperature, area, moisture, etc.

REFERENCES

1. Ikbal Taleb, Rachida Dssouli and Mohamed Adel Serhani, "Big Data Preprocessing: A Quality Framework" in 2015 IEEE International Congress on Big Data (Big Data Congress),2015

2. S. K. Bansal, "Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration," in 2014 IEEE International Congress on Big Data (Big Data Congress), 2014, pp. 522–529.

3. P. Vassiliadis, A. Simitsis, and E. Baikousi, "A Taxonomy of ETL Activities," in Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP, New York, NY, USA, 2009, pp. 25–32.

4. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, and S. Skiadopoulos, "A generic and customizable framework for the design of ETL scenarios," Information Systems, vol. 30, no. 7, pp.

5. A knowledge-based approach for quality-aware ETL process Imen Hamed; Faiza Ghozzi 2015 6th International Conference on Information Systems and Economic Intelligence (SIIE)Year: 2015 | Conference Paper | Publisher: IEEE

6. An overview and implementation of extraction-transformation-loading (ETL) process in data warehouse (Case study: Department of agriculture) Rahmadi Wijaya; Bambang Pudjoatmodjo 2015 3rd International Conference on Information and Communication Technology (ICoICT)Year: 2015 | Conference Paper | Publisher: IEEE

7. Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application Ashish Juneja - Nripendra Narayan Das Department of Computer Science & Engineering Faculty of Engineering & Technology, Manav Rachna International Institute of Research and Studies, Faridabad, India 2019

8. Cluster-size optimization within a cloud-based ETL framework for Big Data Eftim Zdravevski; Petre Lameski; Ace Dimitrievski; Marek Grzegorowsk0069; Cas Apanowicz, 2019