Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **903:908***

Automation of Database Replication and Sentiment of The Replicated Data

Gajanan Naik[1], Ashwin kumar Motagi[2]

[1,2]*School of Computing and Information Technology, REVA University*
[2]*ashwinkumarum@reva.edu.in*

**Corresponding author's E-mail: gajananpna@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

This paper deals with 2 topics 1) The automation of database replication 2) performing sentiment Analysis on the replicated data. The process of automation is continuously evolving across all the industries and more so in the IT industry, it increases the value of time and indirectly saves a lot on costs, and also on tasks that increase the value of your IT ecosystem exponentially. It also makes your day-to-day workload more interesting: learn, code, test, automate and learn something new. The databases need to have data from other databases for various purpose either partly or wholly, may be table, schema or database level. The replicated data may be useful for review and hence we may use sentiment analysis on the dataset using python and related tolls of ML. The data will be going through tokenizing, removing stop words, normalizing, vectorizing.

Keywords: *Automation, Database replication, Heterogenous Databases, Machine learning, Python, Sentiment Analysis, Graphical output*

1. Introduction

Database Replication is used to transfer data to/from between databases either unidirectional, bidirectional or multidirectional. Oracle Golden Gate is a software package that allows the real-time replication of data between homogeneous databases (Oracle databases) or heterogeneous databases (MySQL, PostgreSQL, SQL Server, Sybase and others) over a network connection. You can use it to consistently replicate live data between a source and target database.

When migrating data from your own data centre to a public or private cloud, a key success factor is 100% data integrity between the source and the target database at all times. Also, some applications may not allow any /limited downtime for migration, however Goldengate allows to replicate in real time and automating this from premise to premise or cloud allows us to leverage the benefit.

On the path, you will most likely set up different environments to replicate data (dev, test, etc.), test your application, and perform a few mock migrations and switchovers before the final. This means setting up a new replication each time, testing your application, and resetting the environments repeatedly. Depending on your database size, this process will be time-consuming for your DBAs, potentially error-prone, or highly inefficient at best. A flexible and automated process that allows you to set up a new replication stream effortlessly will allow you to focus on other important tasks that contribute to the final success of the migration.

a) Automation of database replication.

The replication is used between databases for various reasons like maintain standby database, transfer particular set of data in a distributed database, for taking backup etc. However, setting up replication can be tedious and would need specific expertise, which will be time consuming as well expensive, hence we need a method to set up replication based on the project need with minimum manual intervention. There are some tools available in market for automation tools for infrastructure like Puppet, Ansible etc, however these kinds of tools haven't been much used for Database Replication, in addition there is cost for Licensing also. Hence, we will try to do this automation with scripting languages. In this paper we will take two Oracle databases and use golden gate replication software for the same. We will use bash scripting to automate the replication, in this paper we will see DML and DDL replication

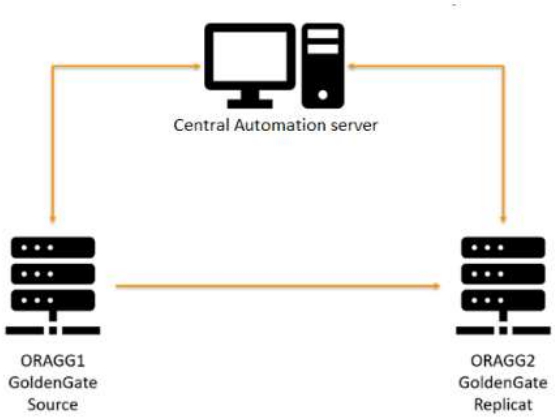

Fig 1. Overview of Replication Automation

B) Sentiment Analysis of the replicated data.

A sentiment is an opinion of expression and feelings, it can be positive, negative or neutral, such happy, sad, good, bad etc. Sentiment analysis is use of NLP and ML methods to find out the emotions from literature.

It involves use of various techniques and algorithms to classify text as positive, negative or neutral.

In the age of Internet, there is huge data that is collected from different sources from different parts of the world. We will visit the sites that provide opinions and reviews, online sites which provide feedback (TripAdvisor, Rotten Tomatoes), , e-commerce sites (flipkart, Amazon, Myntra), and social platforms (snapchat, Facebook, Twitter) to get response from as many different people about their view about a product or a facility. .

Some of applications of sentiment analysis include, analysis of viewpoint in social media, scrutiny for digital media, or assist for researching of sentiments in Literature and/or Natural language processing.

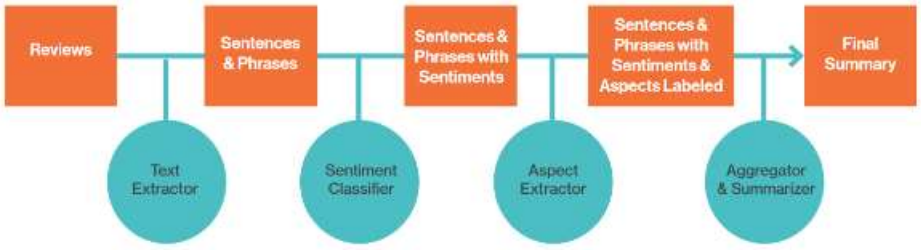

Fig 2. System overview Finding sentiments for Aspects

2. ML based sentiment analysis on the replicated data.

We will have to make some changes to the database before we can start replication.

- We need to create new user for Goldengate on both source and target, which is responsible to read data from multiple schemas.
- DB must be in archivelog mode.
- We need to add supplemental logging.
- Set couple of database parameters.

For the rest of this article, the following assumptions are made: The two databases required for replication are created, are created and running; the GoldenGate software is installed on both instances, with the manager process already configured and running; any firewall between the two hosts has the proper ports open (usually 7809), and the GoldenGate user has been created in both databases.

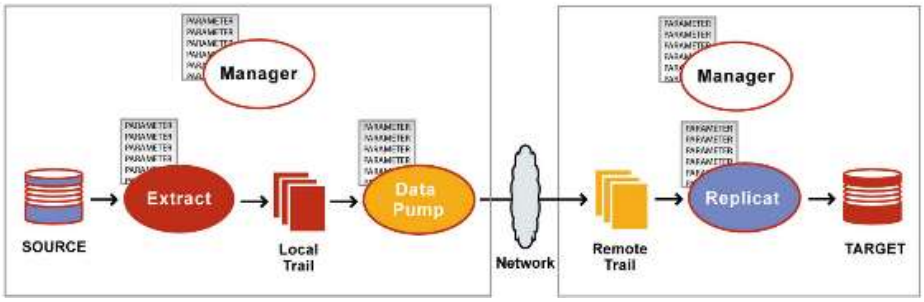

Fig 3. Components of current Database Replication

The Golden Gate Software Command Interface (GGSCI) is used to communicate between different components for the process of replication. The basic components are shown in the figure above.

- The extract process is present on the primary side of the system, its job is to pick out the committed transactions from both committed and uncommitted ones and write them to corresponding files used to transfer the data.
- A trail is type of files which gets its data from the extract process and then forwards it to its destination.
- Replicate process picks the extracted data in the from the trail files and converts them into DML, DDL and then applies it to the destination database.
- The collector process is present on the destination side, there is one to one mapping between the extract and the collector process, it receives the data, which it writes to trail files.

Here are high-level steps to configure DML Replication Define Capture Process to Write to Local Trail

- Define Pump Process, Remote Trail and end data to Remote Trail from Pump Process over TCP/IP
- Define Delivery Process to Apply changes to Target Database
- Start GGSCI (GoldenGate Software Command Interface) and connect to Source Database
- Add Transaction Data Capture on Tables that needs to be replicated
- Once Capture/Extract Process is added in Golden Gate Register this Capture/Extract Process with Database using GGSCI> register extract <Extract/Capture_ProcessName> DBname
- Depending on the number of databases, you will have to use the number of extract and capture processes.
- Monitor the alert log for mining process start

High level steps for DDL replication.

- Create DDL Statements Replication Tables Support
- Create GoldenGate Database Role and Grant Privileges
- Enable DDL Replication Support
- Set GoldenGate DDL Parameter Syntax and Options
- Edit Extract and the Replicate Parameters Files

Automation in the paper: We will be using the **ggsci** command line interface for to be called from bash script and use the available processes to be invoked through for all the DML and DDL replication, we will be scanning the log and error files through our scripts to make note of any error and warning and then take a corrective action and email to the IT consultant handling the project**.** We will be allowing interactive tool to ask questions like 1) Choose the Source and the Target 2) Choose What type of replication do you need a) DML, b) DDL, the user will have to provide input 3) Choose whether it is a) new object b) existing object 4) Whether monitor log or any other kind of book keeping. All this will be done our scripts and the background work where IT team will be needed

To configure all the details. Below is the link where a demo of configuring is given: https://www.oracle-scn.com/installation-and-dml-replication-configuration-of-oracle-goldengate-11g/

However in our case all this is hidden from the user and he need to have expertise on the replication technology.

Sentiment Analysis:

There are three types of classification possible as below:

- Document-level classification:
- Sentence-level classification:
- Aspect level sentiment classification

The first step for sentiment analysis is the cleaning of data and preprocessing like removing the stop words, the noise in the data can affect the accuracy of the analysis. We will be using python for coding the project. Python has NLP libraries present such as Spacy, Pandas, matplotlib. pyplot, numpy, NLTK, cx_Oracle, TextBlob, etc which we will be using for various tasks. Then we will be using the analysed data for plotting graphs so that it can be easy to understand using visuals.

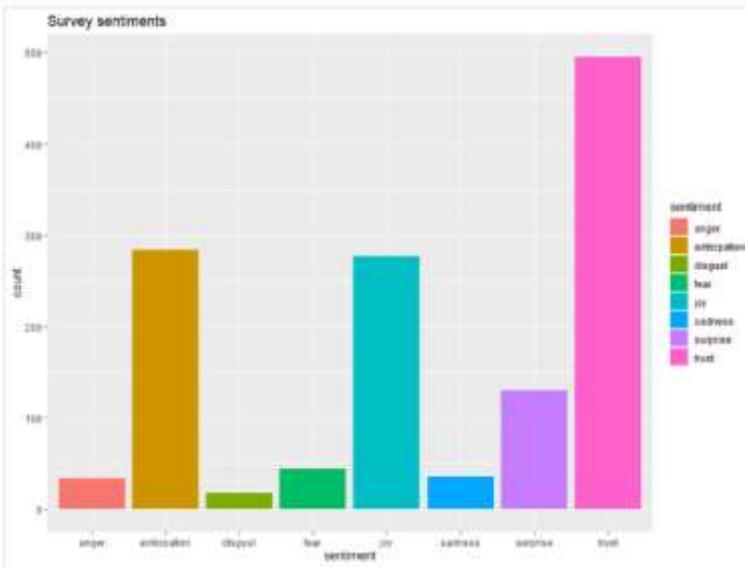

Fig 4. Representation of output in graphical form

Some of steps used before sentiment analysis are which we will conducting in our paper:

Cleaning: The input mostly includes contains content like tags, which will not help in the analysis of the sentiment. So, we must make sure that such contains are removed before it is sent for analysis.

Removing characters from different accent: In this paper we are looking at English language, so there can be accent from other language or English spoken in different country, if such words are allowed then, it may not be able to identify such characters, hence we have to remove them.

Removal of special characters: The special characters and symbols can make it difficult for analysis and hence it needs to be removed, there are some tools available in Python for the same.

Stemming: is the way to reduce the particular word to their stem or base, it is a technique to assemble related words together.

Lemmatization: is a method that converts any kind of a word to its base root mode.

Removing stopwords: There are words in language that do not add much value to the meaning in the analysis, words such a, an, the, in, they and hence such need to be removed. Once the above processing is done, then we will have to calculate sentiment score, which involves sentiment polarity and subjectivity score.

The output will be in two parts one is the form of text and other in the form of graphs, also it can be segregated at each comment level, sentence level or bigger text. The below is small snippet of the output.

The Analysed Text from the DB is

I am Sad', 'it is very bad', 'I had horrible experience' << **This is exact input**

The polarity is *The Subjectivity is*

-0.8033333333333332 0.9555555555555556 << **These are values for polarity and subjectivity**

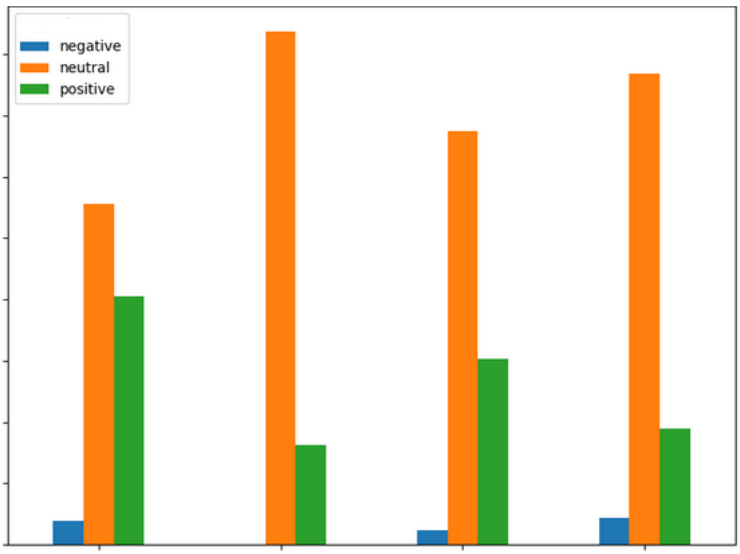

Fig 5. Typical output

3. Advantages and application of SA and Replication Automation

The Automation will have following benefits.

The team of IT professionals required will be less.

The cost of the project will be reduced.

The human errors will be less.

The monitoring will be easy and 24/7.

The learning curve and time spent on the project will be less.

Few of the applications of sentiment analysis are as given below:

- Keep a check on market.
- Get the feedback about product from the customers.
- It helps in boosting the quality of support.
- Keeps check on the challenger.
- It is used for recommending product in shopping.
- Helps in advertisement based on interest.
- Get rid of spam mails.
- Evaluation of mental aspects.
- Emotional analysis of social platforms like Twitter.
- Understanding of TV shows ratings.

I. Future scope

Constraints and issues in Sentiment Analysis and Replication Automation

Sentiment analysis is very promising for digital platform problems. Although there are hurdles in the same.

Below are the challenges in the sentiment analysis:

- It is quite difficult for computers to identify things like jokes, sarcasm, irony.
- The same words can have different meaning in different scenario.

Other challenges of sentiment analysis:

- The language can portray different meaning based on body language, facial expressions and also cultural and regional differences.

- The negation words can completely change the meaning of the sentence.

For example, humans are still better than computers at gauging the emotions. Software can get confused at differentiating from sarcasm from regular text, also they are not great at the meaning of a word with respect to the context also mixed opinions like the Movie was great, but the music could be better. These are some of the problems in sentiment analysis:

- It can difficult to know what technique to use because text may be few words, few sentences or complete document.

4. Conclusion

It Sentiment analysis tries to understand the emotions from a piece of document, from the huge amount of data that is available in the digital domain. In other words, we can generally use a sentiment analysis approach to understand opinion in a set of documents.

It is sometimes referred to as opinion mining, where we can use various tools of ML, NLP to extract the data in a set and identify the emotions and then classify them. Although the sentiment analysis is complex, much progress has been made in last few years due to Advances made in NLP, companies want to take feedback from customers for their product and services so that they can make improvement. It also helps customers on having an idea about the things on which they will be using their money.

For example, someone planning to go out for a movie can decide, how the movie will be based on review, perceiving a sentiment is natural for humans. Hence, sentiment analysis is a great tool to understand the documents underlying subjective nature, in which ML and its tool plays an important role.

Acknowledgment

It is a matter of pride for me to acknowledge my heartfelt gratitude to the School of CSE, Reva University for allowing me to explore my capabilities through this paper. I would like to express my sincere gratitude to our Guide, Dr. Ashwin kumar Motagi for his valuable guidance and advice towards completing this paperwork.

References:

[1] https://thesai.org/Downloads/Volume10No2/Paper_48A_Study_on_Sentiment_Analysis_Techniques.pdf
[2] https://www.sciencedirect.com/science/article/pii/S2405959520300394
[3] https://www.aclweb.org/anthology/S15-2093.pdf
[4] http://cse.iitkgp.ac.in/~saptarshi/courses/socomp2020a/sentiment-analysis-survey-yue2019.pdf
[5] https://www.oracle.com/a/ocom/docs/con6569-auto-gg-microservices-3960430.pdf
[6] https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13217
[7] https://www.globallogic.com/wp-content/uploads/2019/12/Introduction-to-Sentiment-Analysis.pdf
[8] http://www.ijstr.org/final-print/dec2019/Sentiment-Analysis-Of-Product-Reviews-A-Survey.pdf
[9] https://www.ijcaonline.org/research/volume125/number3/dandrea-2015-ijca-905866.pdf