Providing in-Database Analytic Functionalities to MySQL: A Proposed System

Deeksha M Kumar[1], Harshitha M[1], Muhammed Faheem E[1] , Nesar N[1], Suhaas KP[2]

[1]Dept. of ISE, The National Institute of Engineering, Mysore
[2]Asst.Professor, Dept. of ISE, The National Institute of Engineering, Mysore, Karnataka, India

---***---

Abstract - *Data analytics has taken the tech world by storm. Given the explosion of data in today's world, analytics is needed now more than ever in order for businesses to plan their next move. Conventional Analytic frameworks involve data migration to and from the data store and an analytic tool which proves time consuming with the bulk of data. Our proposed system intends to implement analytic functions within MySQL in order to cut down on the time consumed in the analytics process.*

Key Words: **Database, analytics, Machine Learning, MySQL and UDF.**

1. INTRODUCTION

Quoting Christopher Ré et al in [1], the question "Is there anything fundamentally different about building database system that use machine learning or are designed to support machine learning?" makes you think about how one can go about achieving a database with machine learning like functionalities or analytic functionalities. A traditional method of data analytics involves movement of pre-processed data from the data store, generally a database, to analytic software. The software then performs analytic functionalities following which, data is transferred back to the data store [2]. One can go about their task with this method if he were dealing with a small amount of data. Given the increasing volume of data in today's world, the transportation of data from the data store to the analytic software seems to be one of the biggest challenges as it is time consuming and puts a strain in the network being used to transfer the data.

1.1 MySQL

MySQL is one of the most widely used RDBMS's in the world. It is known as a fully featured RDBMS and it is stable. It is also a multithreaded based server which means it is extremely fast and can support multiple users at once. Each user gets their own thread when they establish a connection with the server. MySQL is also platform independent. It works on almost all platforms. MySQL provides security to data at business level. Data is stored in tables. MySQL also includes Application Programming Interfaces or APIs for Perl, Python, Java. SQL is easy to pick up and use. The SQL commands are pretty concise and easy to learn.

1.2 Machine learning and analytics

Machine learning is a sub-field of Artificial intelligence. It concerns enabling computers to learn without being expressly customized. Throughout the years, Artificial intelligences' ubiquity and request has positively been on the ascent. Machine Learning employs algorithms and statistical models in order to build a mathematical model of data. This data is called training data and is used to make predictions. The types of machine learning are supervised learning, unsupervised learning, and reinforcement learning.

The supervised learning basically has two types: regression and classification. Regression algorithms are employed to predict continuous numeric values. Some common regression algorithms include linear regression, k nearest neighbors.

Machine learning consists of 3 steps preprocessing, where raw data is normalized and all inconsistencies are corrected; training and verification, where cleaned data is trained; classification, where the trained model is used to classify data. Generally pre-processing is done within the database following which the data is migrated to the analytic tool over a network where ML algorithms are applied and analytics is performed. Data is then migrated back to the database [3].

2. EXISTING SYSTEM

Conventional data analytics involves the transfer of pre-processed data from the data store to the analytic tool. The data may be stored in relational databases like MySQL, columnar databases like MongoDB, CSV files or XML files. The Traditional data processing for analytics or prediction requires the user to have the knowledge of programming languages like Python and R. Apart from that, users need to write an external application using programming language and load the data into the application from database.

Analytic functions are then applied to the data by the tool following which data is transferred back to the data store. Transferring data from the data store to the analytic tool is not much of a hassle with small data volumes. But when it comes to large data volumes, we are posed with the problem of increase in network traffic. This load on the network could minimise network performance and lengthen the overall time period of the process.

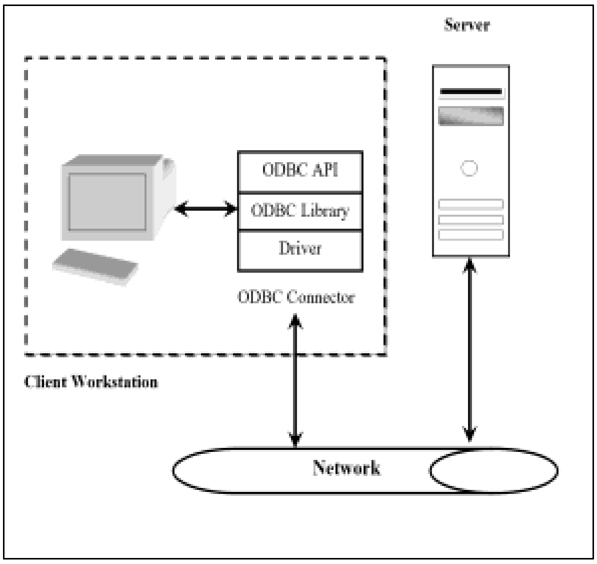

Fig 1: Existing System

Third party applications offer little to no data security. With the vastness of the volume of data being analysed, access to such data could prove disastrous if the data was intended to be confidential. If at all it provides confidentiality, there will be overhead of encrypting and decrypting the data during the data transfer.

3. PROPOSED SYSTEM

The main objective of this system is to eliminate the need of the external applications which import data from the database through socket connections and using MySQL connectors. This system integrates Machine Learning algorithms like linear regression within the MySQL database server itself so that there will be no need for data migration to and from application and database server. Thus it eliminates the risk of data leakage or data loss [4].

As MySQL is relational database management system, the RDBMS can prevent data corruption through ACID or Atomicity, properties; it can automatically manage data storage for the user and make data easier to reason about by enforcing a rigid schema. In addition, the RDBMS can perform efficient execution on larger-than-memory data by only loading required columns. This system uses in-database processing methods, which does the computation where the data resides.

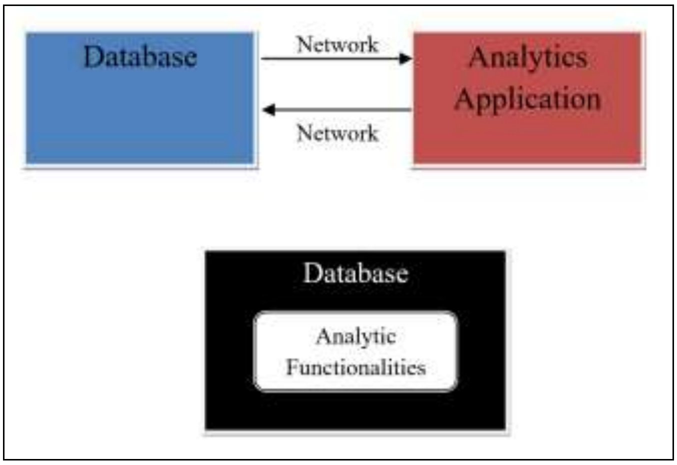

Fig 2: Comparison of Existing and Proposed System

In our proposed system, the best way to plug the functionalities or the algorithms of ML is by using User Defined Functions (UDF) because of its own advantages. UDFs are much easier to develop than is hacking raw code into the MySQL server. If our function were hacked into the server, we would need to change the MySQL source every time we upgraded, which is never easy. MySQL code base evolves quite rapidly and to implement the same function we would need different code changes in every new version. They will not even need to be recompiled when the server is upgraded. And other reason to use UDF is that they are designed for development speed, the API or Application Program Interface is easy to access, and compilation is much quicker than rebuilding the entire server just to add a tiny function.

Since it is in-database processing, it reduces the network load and traffic, which can be effective in cloud data centres where thousands of rows of data should be moved from one geographical location to another and eliminating the need to encrypt and decrypt the data during transmission, which is the overhead in traditional data processing or existing systems.

It also eliminates the need to know python and R for doing predictive analysis in the application and provides these features in Structured Query Language itself, which is widely known by the people. And UDF can be installed whenever it is needed or can be dropped when it is not needed thus providing the opportunity to upgrade or modify the functions as needed by the user, by linking the binaries to MySQL server even during run time.

This system is expected to give considerable advantages over existing systems and provide users with easy and simple way to do predictive analytics.

3.1 Linear Regression

To achieve simple predictive analysis, the linear regression with one variable algorithm is used. Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables, with one variable, x, regarded as the predictor, explanatory, or independent variable. The other variable y, is regarded as the response, outcome, or dependent variable.

The formula used for predicting the value of outcome variable is Y=mX+C ,where m is the slope of the line and C is constant value, Y is dependent variable and X is independent variable.

3.2 User-Defined Functions –MySQL

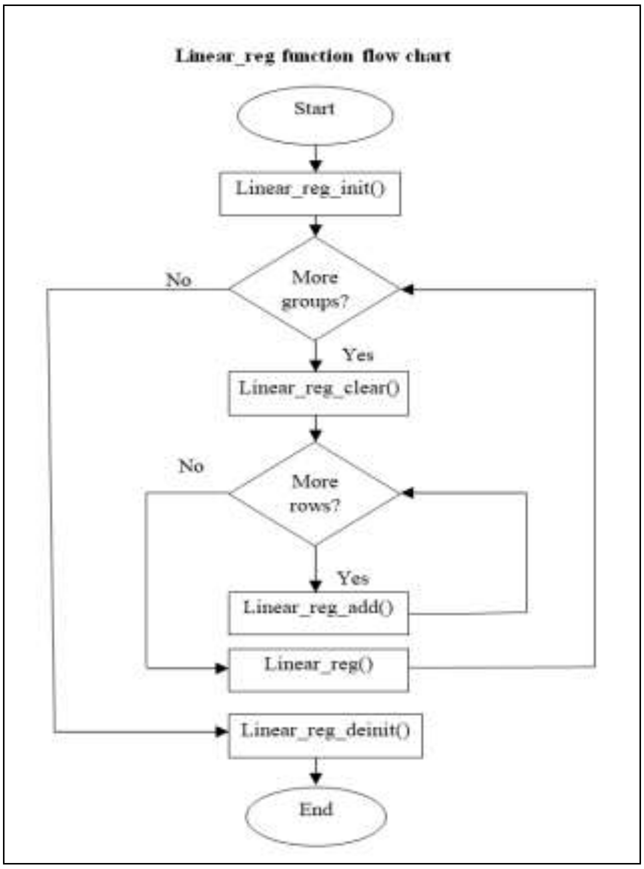

Fig 3: Flowchart of Working of Proposed System

User-Defined functions are the best way and the easiest way to plug the new functions to existing MySQL server. It works faster than the stored procedures [5] ,which is other way of plugging the new functions to MySQL server. User-Defined functions can be dynamically installed by the CREATE FUNCTION command and uninstalled by DROP FUNCTION command. To implement the function, we need to design the aggregate type of UDF and the linear regression with one variable algorithm into the UDF methods [3].

The Linear_reg_init() method is used to check for metadata, verify the required number of arguments and allocate the memory if needed by the function. Linear_reg_deinit() method is used to free the memory used by function after the execution of the query. Linear_reg_add() and Linear_reg_clear methods are used to add up each row in the database and clear the values after each group of values mentioned in the query. Linear_reg() method does the actual work of the function and linear regression with one variable algorithm is implemented in this method.

4. CONCLUSIONS

Data migration and network traffic are two of the biggest concerns and reasons for inefficiency in conventional analytic workflows. This system proposes to cut down on data migration and network traffic and reduce them by implementing analytic functionalities within MySQL itself. Knowledge of SQL commands is only necessary in order to operate the system. As MySQL is a widely used RDBMS, our system could be utilised by many in the IT sphere.

The system will consume a smaller amount of time overall in comparison with existing systems. The ease of utility guaranteed by UDFs is also an added advantage.

ACKNOWLEDGEMENT

We are proud and grateful to have been given an opportunity to present an idea such as ours. To our principal, Dr. G S Ravi, we are extremely grateful for his kind permission to carry out our work. Our sincere thanks to the Department of Information Science and Engineering, NIE, Mysuru.

We are indebted to Mr. Suhaas K P, our project guide for his constant support and valuable insights to the project. When we were lost, he helped us find a path and we will be ever grateful.

REFERENCES

[1] C. Re, D. Agrawal, M. Balazinska, M. Cafarella, M. Jordan, T. Kraska, R. Ramakrishnan, “Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype?” SIGMOD ’15, May 31-June 04, 2015, Victoria, Australia. ACM 978-1-4503-2758-9/2015

[2] U. Syed, S. Vassilvitskii, “SQML: Large-scale in-database machine learning with pure SQL” SoCC_'17 Proceedings of the 2017 Symposium on Cloud Computing, Santa Clara, CA, USA

[3] M. Raasveldt, P. Holanda, H. Mühleisen, S. Manegold, ”Deep Integration of Machine Learning Into Column Stores” 21st International Conference on Extending Database Technology (EDBT), March 26-29, 2018, ISBN 978-3-89318-078-3

[4] J. Vinish D’silva, F. De Moor, B. Kemme,” AIDA - Abstraction for Advanced In-Database Analytics” PVLDB 2018 DOI:10.14778/3236187.3236194

[5] C. Ordonez, C. Garcia-Alvarado, “A Data Mining System Based on SQL Queries and UDFs for Relational Databases” CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management, DOI 10.1145/2063576.2064008

[6] https://www.analyticsvidhya.com