iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38710

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Efficient Transfer of data from RDBMS to HDFS and conversion to JSON format

Dr. C. K. Gomathy [1], K. Vishnu Vardhan [2,] Unnam Pavithra[3]
[1, 2, 3] *SCSVMV University*

Abstract: ***Apache Sqoop is mainly used to efficiently transfer large volumes of data between Apache Hadoop and relational databases. It helps to certain tasks, such as ETL (Extract transform load) processing, from an enterprise data warehouse to Hadoop, for efficient execution at a much less cost. Here first we import the table which presents in MYSQL Database with the help of command-line interface application called Sqoop and there is a chance of addition of new rows and updating new rows then we have to execute the query again. So, with the help of our project there is no need of executing queries again for that we are using Sqoop job, which consists of total commands for import and next after import we retrieve the data from hive using Java JDBC and we convert the data to JSON Format, which consists of data in an organized way and easy to access manner by using GSON Library.***

Keywords: Sqoop, Json, Gson, Maven and JDBC

I. INTRODUCTION

The arrival of Internet has resulted in fast growth of data size endlessly. Distributed, processing and storing of such large data sets has turn out to be a challenge for the database industry. 'Big data' term is normally used for storing large data sets. Relational Database Management Systems can't handle such large data sets. For efficient storage of data and analysis we are shifting to Hadoop. It is a framework for big data management and analysis. For data transferring process we need a connector in between called Sqoop, which loads a table from RDBMS to HDFS and vice-versa. Enterprises are adapting large processing platforms, like Hadoop, to understand actionable insights from their large amount of data. Query optimization is still an open challenge during this environment because of the quantity and different types of data, comprising both structured and semi-structured datasets. Processing big scale data within the many petabytes could also be a very difficult task. Solving the issues related to high volume data requires is usually achieved by dividing the info and work to several computers which may all work together in parallel to finish the task during a reasonable time. Hadoop had gained popularity in Parallel dataflow systems. These systems are mainly used for analyzing.

II. EXISITING SYSTEM

The Query will be executed through HDFS using Sqoop and the data will be transferred but Sometimes for the purpose of updated and newly added rows we need to execute the same query again and again. Executing again and again is very difficult and it results in waste of time and we know Bigdata means a huge amount of data and it needs large and complex queries for analysis.
The format will be in Tables with rows and columns. For the large datasets it is tough to read the exact record column and ID of that record.

III. PROPOSED SYSTEM

In our proposed system, The Sqoop commands (used for data transfer from RDBMS to HIVE) will be stored in a Sqoop job and we just need to execute the job for the data Retrieval from RDBMS. So, we no need to use the commands again and again for updated and newly added rows. After that the data which migrated to hive tables will be taken to Java JDBC program and will be converted to JSON format which is organized and easy to access manner.

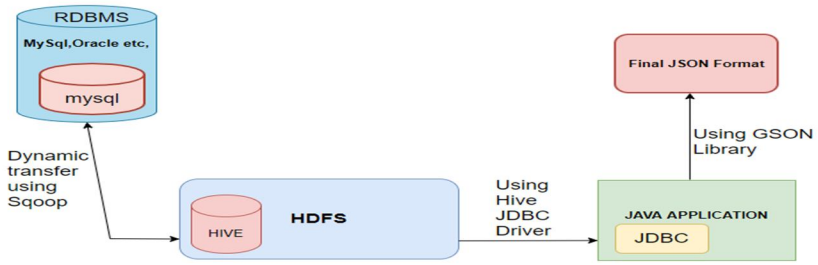


Figure 1: Proposed System Architecture

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

IV. IMPLEMENTATION

The process of implementation of the project:

1) Choosing the Dataset that we want to import from MySQL to HDFS (Hive)
2) Creating a table in Hive with same Format of the Dataset present in MySQL.
3) Writing a Sqoop Job for dynamic transfer of Table data from MySQL to Hive table with same format.

Transferring of the table which present in MySQL database using Sqoop import command to Hive table with same format and later MySQL table gets updated or added news rows to it. For that we are writing a Sqoop job for the dynamic import of data by just executing the Sqoop job the updated and newly added data will be transferred automatically to the Hive table and it will replace the old existing data in the table.

a) Using import command transferring the table data from RDBMS to Hive and check the data is transferred or not.
b) Importing updated Data using Sqoop Job with "Sqoop job –exec " command
c) Adding necessary dependencies in Maven project before accessing the data from Hive
d) Connecting Hive using JDBC and Accessing the Hive data
e) Final step is JSON conversion
f) Process for Conversion of Hive data to JSON Format using Gson library:

Coverting the parameters to java objects. Storing java objects in an Array List. Next converting each object in array list to Json string. Storing all those Json strings in Json Array and converting that Json array to Json Object.

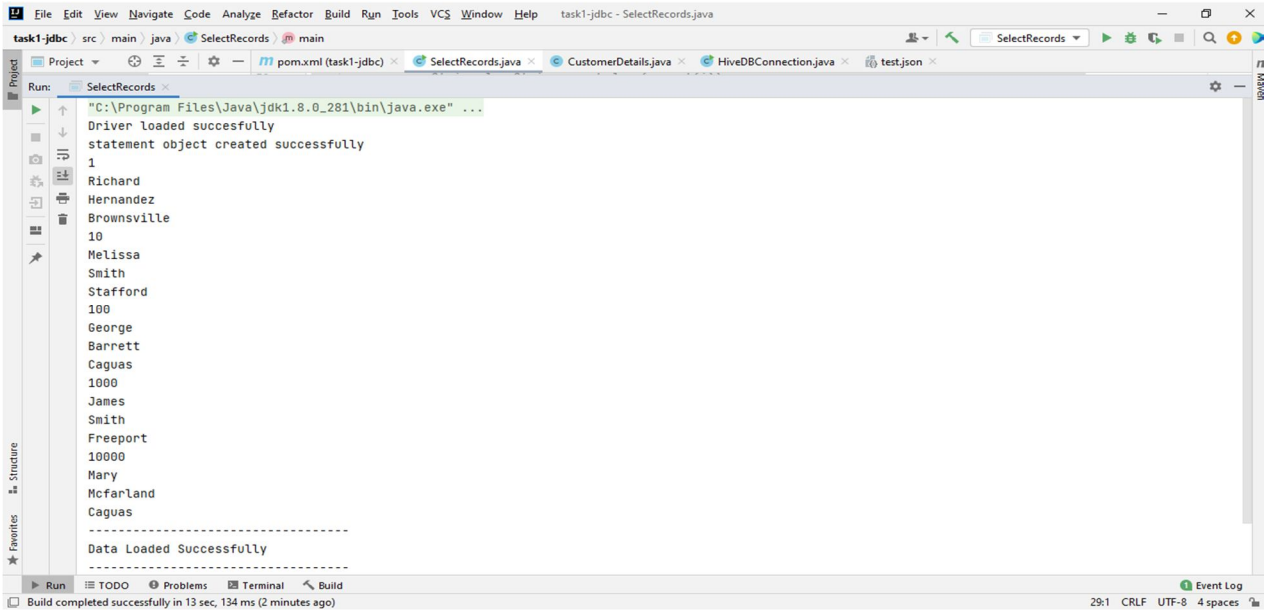


Figure 2: Data loaded from Hive Table using JDBC

A. *Final Output for 3 records*

{
"JsonFormatData": [
{
"id": 1,
"fname": "Richard",
"lname": "Hernandez",
"state": "Brownsville"
},
{
"id": 10,
"fname": "Melissa",

©IJRASET: All Rights are Reserved

"lname": "Smith",
"state": "Stafford"
},
{
"id": 100,
"fname": "George",
"lname": "Barrett",
"state": "Caguas"
}
]
}

V. RESULT

Finally, creation of table in Hive, transforming data from RDBMS to HDFS using Sqoop Import, Creation of Sqoop job using incremental last-modified for automatic import of data for updated and newly added data, accessing data from Hive table using JDBC and conversion of accessed data to Json format are Completed.

VI. CONCLUSION

In this work, we proposed a Dynamic transformation of data and JSON conversion. The main benefit of our proposed system is Time efficiency and to improve the overall performance of the Execution, command need to be processed efficiently. As data is being increasing day by day due to wide range use of social networking sites, the problem will raise like how to store, process, manage, use all these large amounts of data by the use of big data a user can access the past, present data which has been stored and analyzed from past years. Transforming of data from RDBMS to HDFS is easy now through Apache Sqoop efficiently. Sqoop exchanges the data in little time and also the performance will be high.

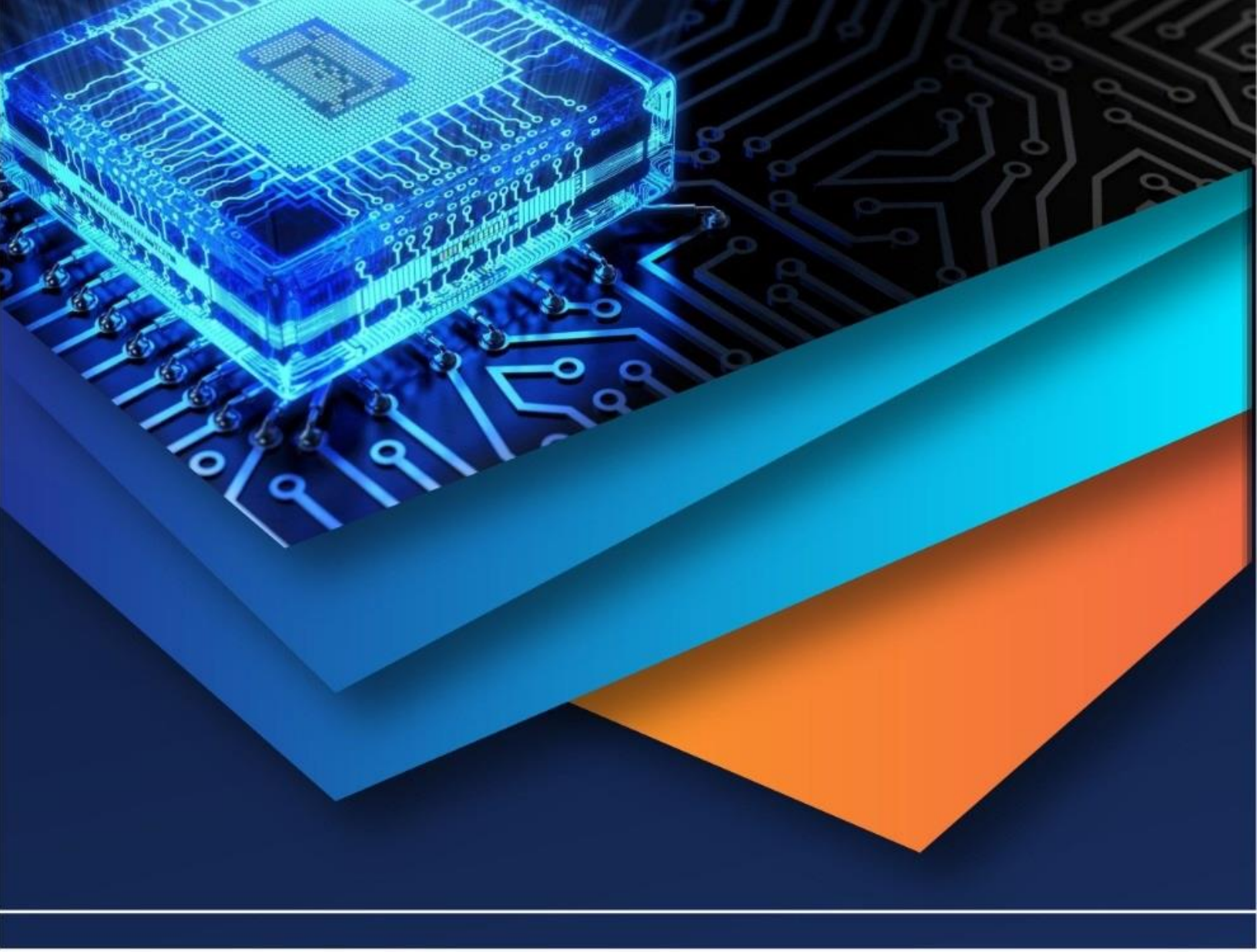
REFERENCES

[1] Big Data Analytics Using Hadoop Map Reduce Framework and Data Migration Process Published in: 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) Date of Conference: 17-18 Aug. 2017 Date Added to IEEE Xplore: 13 September 2018
[2] Hadoop Distributed File System (HDFS), http://hortonworks.com/hadoop/hdfs/
[3] Generalized Big Data Test Framework for ETL migration Published in: 2016 International Conference on Computing, Analytics and Security Trends (CAST) Date of Conference: 19-21 Dec. 2016 Date Added to IEEE Xplore: 01 May 2017
[4] Apache Hive-Based Big Data Analysis of HealthCare Data Zhenlin Kan1*, Xinru Cheng2 , Seung Hyun Kim3 ,Yuting Jin4
[5] Towards Efficient and Unified XML/JSON Conversion - a New Conversion Method. Transactions on Internet Research (TIR), 13(1):58–64, January 2017.
[6] Transformation of Data from RDBMS to HDFS by using Load Atomizer IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 10, October 2016
[7] Optimization of Multiple Queries for Big Data with Apache Hadoop/Hive 2015 International Conference on Computational Intelligence and Communication Networks
[8] Review of Hadoop performance optimization 2016 IEEE International Conference on Computer and Communications
[9] Dr.C K Gomathy, Article: An Effective Innovation Technology In Enhancing Teaching And Learning Of Knowledge Using Ict Methods, International Journal Of Contemporary Research In Computer Science And Technology (Ijcrcst) E-Issn: 2395-5325 Volume3, Issue 4,P.No-10-13, April ’2017
[10] Dr.C K Gomathy, Article: A Semantic Quality of Web Service Information Retrieval Techniques Using Bin Rank, International Journal of Scientific Research in Computer Science Engineering and Information Technology (IJSRCSEIT) Volume 3 | Issue 1 | ISSN : 2456-3307, P.No:1563-1578, February-2018
[11] Dr.C K Gomathy, Article: A Web Based Platform Comparison by an Exploratory Experiment Searching For Emergent Platform Properties, IAETSD Journal For Advanced Research In Applied Sciences, Volume 5, Issue 3, P.No-213-220, ISSN NO: 2394-8442, Mar/2018
[12] Dr.C K Gomathy, Article: A Study on the Effect of Digital Literacy and information Management, IAETSD Journal For Advanced Research In Applied Sciences, Volume 7 Issue 3, P.No-51-57, ISSN NO: 2279-543X,Mar/2018

AUTHOR’S PROFILE

1) **Dr.C.K.Gomathy** is Assistant Professor in Computer Science and Engineering at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya deemed to be university,Enathur,Kanchipuram,India. Her area of interest is Software Engineering,Web Services, Knowledge Management and IOT.
2) **K.Vishnu Vardhan**, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university,Enathur,Kanchipuram,India.
3) **Unnam Pavithra**, student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university,Enathur,Kanchipuram,India.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)