**ETL and Batch Processing towards real time data**

**Prof. Dhiraj Jadhav**[**[1]**](mailto:dhiraj.jadhav@vit.edu)**, Lajat Manekar**[**[2]**](mailto:lajat.manekar18@vit.edu)**, Deep Patel**[**[3]**](mailto:deep.patel18@vit.edu)**, Abhay Patil**[**[4]**](mailto:abhaysing.patil18@vit.edu)**, Aditya Patil**[**[5]**](mailto:aditya.patil18@vit.edu)**, Vedant Patni**[**[6]**](mailto:vedant.patni18@vit.edu)

***Abstract -* Today, big data is generated from many sources and there is a huge demand for storing, managing, processing, and querying on big data. The Map Reduce model and its counterpart open source implementation Hadoop, has proven itself as the solution to big data processing. Hadoop is inherently designed for batch and high throughput processing jobs. The term ETL which stands for Extraction, Transformation, and Loading is a batch or scheduled data integration process that includes extracting data from their operational or external data sources, transforming the data into an appropriate format, and loading the data into a data warehouse repository and Oracle data warehouse to reduce an execution time and to remove the mismanagement of metadata in an existing ETL process.**

**Key Words *- ETL, Hadoop, HDFS, Oracle, Data Warehouse.***

1. INTRODUCTION

The “Big Data” paradigm is getting an expanding popularity recently. The “Big Data” term is generally used for datasets which are so huge that cannot be processed and managed using classical solutions like Relational Data Base Systems (RDBMS).The project mainly revolves around HDFS and ETL batch processing , so further lets discuss the paradigms.

This ETL Tool is used to simplify the process of migrating data, standardize the method of data migration, store all data transformation logic as Meta data which enable the users, managers and architects to understand, review, and modify the various interfaces and reduce the cost and effort associated with building interfaces. Extraction is the process of reading data from a specified source database and extracting a desired subset of data. Transformation phase applies a chain of rules or functions to the extracted data to derive the data to be loaded. Three forms of transformations are utilized, that is, Subsets of tables, Formatting Data and Primary Keys and Indexes. Subsets are created to remove personally individual information. All tables except the reference table are transferred to the Data warehouses using an ETL process. Primary keys are created to make sure uniqueness within a table and to facilitate the fusion of tables. Indexes are created to expedite queries. Loading is the process of writing the data into the target database.

The ETL process includes designing a target, transforming data for the target, scheduling and monitoring processes. The purpose of using ETL tools is to save time and make the whole process more reliable. The ETL tools are customized to provide the functionality to meet the enterprise requirements. Hence, many of them choose to build their own data warehouse themselves.

Hadoop is an open-source software framework, used in the processing and storage of data for big data applications in clusters of computer servers built from commodity hardware. It provides massive storage for any kind of data, an enormous processing power, and can take concurrent tasks or jobs by using parallel processing. It is the bedrock of big data technologies that support advanced analytics initiatives, including predictive analytics, data mining, and machine learning.

The Hadoop platform has tools that can extract the data from source systems, such as log files, machine data, or online databases, and load them to Hadoop in record time. It is also possible to do transformations on the fly. Complex ETL jobs are deployed and executed in a distributed manner due to the programming and scripting frameworks on Hadoop.

Section 2 of this paper deals with related work done in the Extract, Transformation and Loading into the data warehouses. Section 3 explains an actual process of Extract, Transform and Load. Section 4 explains the steps of proposed work to design an ETL Engine. In section 5

Experimental analysis, results are given, and finally, section 6 presents a conclusion of this paper.

1. RELATED WORKS

Different varieties of approaches for the integration of ETL tool in data warehouses have been proposed. A data warehouse gives a set of numeric values (called facts) that are based on a set of input values in the form of dimensions. Over the years, data warehouse technology has been used for analysis and decision making in enterprises. A concrete ETL service framework was proposed and talked about in metadata management service, metadata definition services, ETL transformation rules service, process definition service etc. Two heuristic algorithms with greedy characteristics were proposed to reduce the execution cost of an ETL workflow. The model defeated the weak points of traditional Extract, Transform and Load ETL tool’s architecture and proposed a three-layer-architecture based on metadata. That built an ETL process more flexible, multipurpose and efficient and finally they designed and implemented a new ETL tool for drilling data warehouse.

Therefore, for a more accurate interpretation of numeric values, business users require an interpretation in meaningful non-numeric terms. However, if the transition between terms is crisp, true values cannot be measured and smooth, transition between classes cannot take place. At last, definition method and related algorithms of ETL rules are designed and analyzed. A data mart contains data from a particular business area and multiple data – marts can form a data warehouse.

1. EXTRACT, TRANSFORM AND LOAD (ETL) and HADOOP DATA FILE SYSTEM (HDFS)

ETL is the process to allow business to combine their data while moving it from source system to data warehouse. Data can be taken from any source. A detailed explanation of the ETL process is Extract, Transform and Load (ETL). The three data base functions are combined into one tool that automates the process to pull data out of one database into another database. Extraction is referred as extracting the data from various heterogeneous systems. Transform means applying the business rules on data which are derived from different sources. The process of pumping the data into the data warehouse for end user access is referred as “Loading”. The Testing of ETL mainly deals with how, from, when, what and where we carry in our data base.

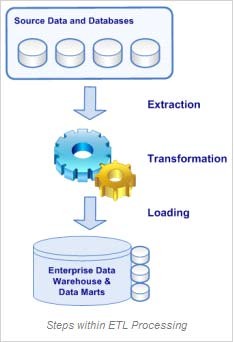


Figure 1. Process of ETL

The ETL tools were created to improve and facilitate data warehousing. ETL eliminates the step of loading the text files into an intermediate storage, saving significant space and time. The purpose of using ETL Tools is to save the time and make the whole process more reliable.

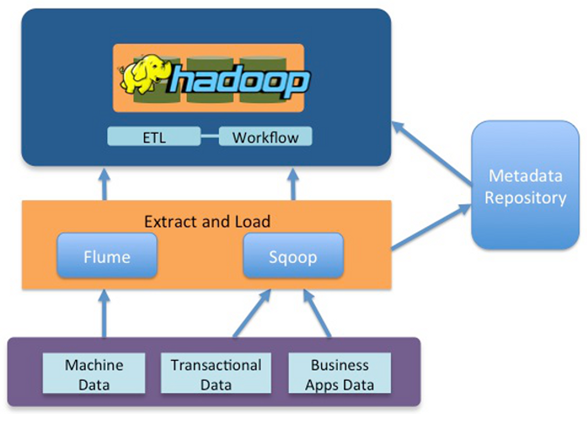
The ETL process consists of the following steps:

1. Initiation
2. Build reference data
3. Extract from sources
4. Validate
5. Transform
6. Load into stages tables
7. Audit Queries
8. Clean Up

Transformation has been applied to achieve migrating from one database to another the goals of the Data warehouse. End-users can access the data via several methods (i.e. DEFAULT; ORC; HIVE-HBASE). The CREATE TABLE schema is the simplest data ware house schema looks like entity – relationship model with points baking from a central table. The center of the star contains number of fact table. This schema, normalizes data dimensions by grouping data into multiple tables rather than one giant table. All tables except the reference table are transferred to the Data warehouse using an ETL process. Many of the tables are split into smaller tables in order to expedite queries. The ETL process includes designing a target, mapping sources to target, extracting data from sources, transforming data for the target, scheduling and monitoring processes, and managing the overall BI environment. Benefits of an ETL tool are given below:

* to simplify the process of migrating data
* to standardize the method of data migration
* to store all data transformation logic/rules as Meta data
* To enable Users, Managers and architects to understand, review, and modify the various interfaces.
* To reduce cost and effort associated with building interfaces.

Hadoop is an open-source software framework, used in the processing and storage of data for big data applications in clusters of computer servers built from commodity hardware. It provides massive storage for any kind of data, an enormous processing power, and can take concurrent tasks or jobs by using parallel processing. It is the bedrock of big data technologies that support advanced analytics initiatives, including predictive analytics, data mining, and machine learning.



1. PROPOSED WORK

It is important in business to understand problems facing an organization, and to explore data in meaningful ways. Data in itself is merely facts and figures. Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data.

We will be able to learn how data is managed in the industry using ETL process while extracting source data from MySQL and then transform and finally load into Hadoop based data warehouse system and further applied to create business intelligence with Hive outputs.

Since this project will emphasize on the migration of data and integration of hive, apache scoop tool n comps

We will also apply compression technique which will increase the performance and throughput. For egg - Text based file may compress up to 40 percent or more.

Let’s learn about the data set used:

India Annual Health Survey (AHS) 2012-13

The dataset comprises a survey conducted in Empowered Action Group (EAG) states Uttara hand, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Odisha, Chhattisgarh & Madhya Pradesh and Assam. These nine states, which account for about 48 percentage of the total population, 59 percentage of Births, 70 percentage of Infant Deaths, 75 percentage of Under 5 Deaths and 62 percentage of Maternal Deaths in the country, are the high focus States in view of their relatively higher fertility and mortality.

A representative sample of about 21 million population and 4.32 million households were covered which is spread across the rural and urban area of these 9 states.

The objective of the AHS is to yield a comprehensive, representative and reliable dataset on core vital indicators including composite ones like Infant Mortality Rate, Maternal Mortality Ratio and Total Fertility Rate along with their covariates (process and outcome indicators) at the district level and map the changes therein on an annual basis. These benchmarks would help in better and holistic understanding and timely monitoring of various determinants on well-being and health of population particularly Reproductive and Child Health. [Source]

Problem statement with respect to the dataset - Ingest the India Annual Health Survey (AHS) 2012-13 data hosted on Amazon RDS into Hadoop correctly and process it to generate the following analyses:

**Analyses**

1. The child mortality rate in Uttar Pradesh
2. The fertility rate in Bihar
3. State-wise child mortality rate and state-wise fertility rate and does high fertility correlate with high child mortality?-
4. Find top 2 districts per state with the highest population per household
5. Find top 2 districts per state with the lowest sex ratio

Such analyses would help in vivid understanding and timely monitoring of different determinants on well-being and health of population particularly Child and Reproductive Health. Based on the analyses, one can also compare India’s position in Global HDI and can suggest ways that can improve it.

Tasks we will perform:

Data Ingestion from the RDS to HDFS using Swoop

* 1. Sqoop import command
  2. Command to see the list of imported data

External table creation in Hive and loading the ingested data into it. Data ingestion verification.

* 1. Command to create the external table
  2. Command to load the ingested data into the external table

Queries to verify that the ingestion is correctly accomplished

Query to count the total number of rows along with the screenshots of the data fetched by the query on MySQL Workbench and Hue

Query to select the top 10 rows and first 8 columns along with the screenshots of the data fetched by the query on MySQL Workbench and Hue

Subset schema creation in Hive to support the analyses

Columns used in the subset schema

Storage format used

[Benchmark the performance before finalizing the storage format to be used. Create one schema using default format and one in any other format such as ORC for the columns to be used. Insert data into both the tables created. Compare the runtimes of the following queries and decide which format to be used.

*select count(\*) from <Table Name>;*

*select State\_Name, count(\*) from <Table Name> group by State\_Name;*

*select \* from <Table Name> where State\_Name = ‘Uttar Pradesh’;]*

You are also expected to benchmark the performance of Hive-Hbase integrated table. Compare the runtimes of the queries in step 2 for the integrated table and the storage format chosen in step 2.

Create and insert command for the default format

Create and insert command for the formats such as ORC

Create and insert command for the Hive-Hbase integrated table

Screenshot of runtimes against each query, given above in step 2, for the default format, for the formats such as ORC as well as for the Hive-Hbase integrated table.

Create and insert command for the partition table for analyses 1 & 2. The partition table should be created using the table created above. This step to be done for the Hive only table.

Note: While doing the benchmarking, you might get the run times of formats such as ORC different than expected. These anomalies exist if the dataset is small. The main idea is to let you know that benchmarking is an important step in the end to end engineering process. Generally, you would benchmark the performance against different formats and choose the best. However, here go ahead with the formats such as ORC only.

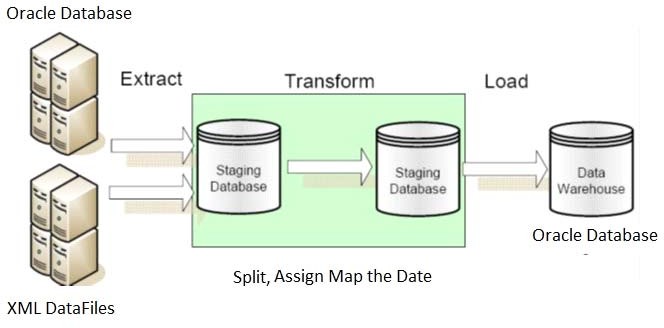


Figure 2.The process of Proposed ETL

The steps for designing the ETL are given below.

1. Extract the data from operational data source. Data extraction is one of the three main functionalities of the ETL tools. A main consideration to assess is the product's ability to extract from a variety of various data sources.
2. Creating Table with relevant attributes based on user requirement.
3. Transform it to fit operational needs. Generate the xml document file for the collected data.
4. Meta-Data for XML document file

This Research work implements three protocols namely Oracle Database, XML Data File and JDBC. The Protocol will be part of the url attribute of the target or source node. Every transformation will have a source and target.

<source url="xml://localhost/etl/test.xml“>

…

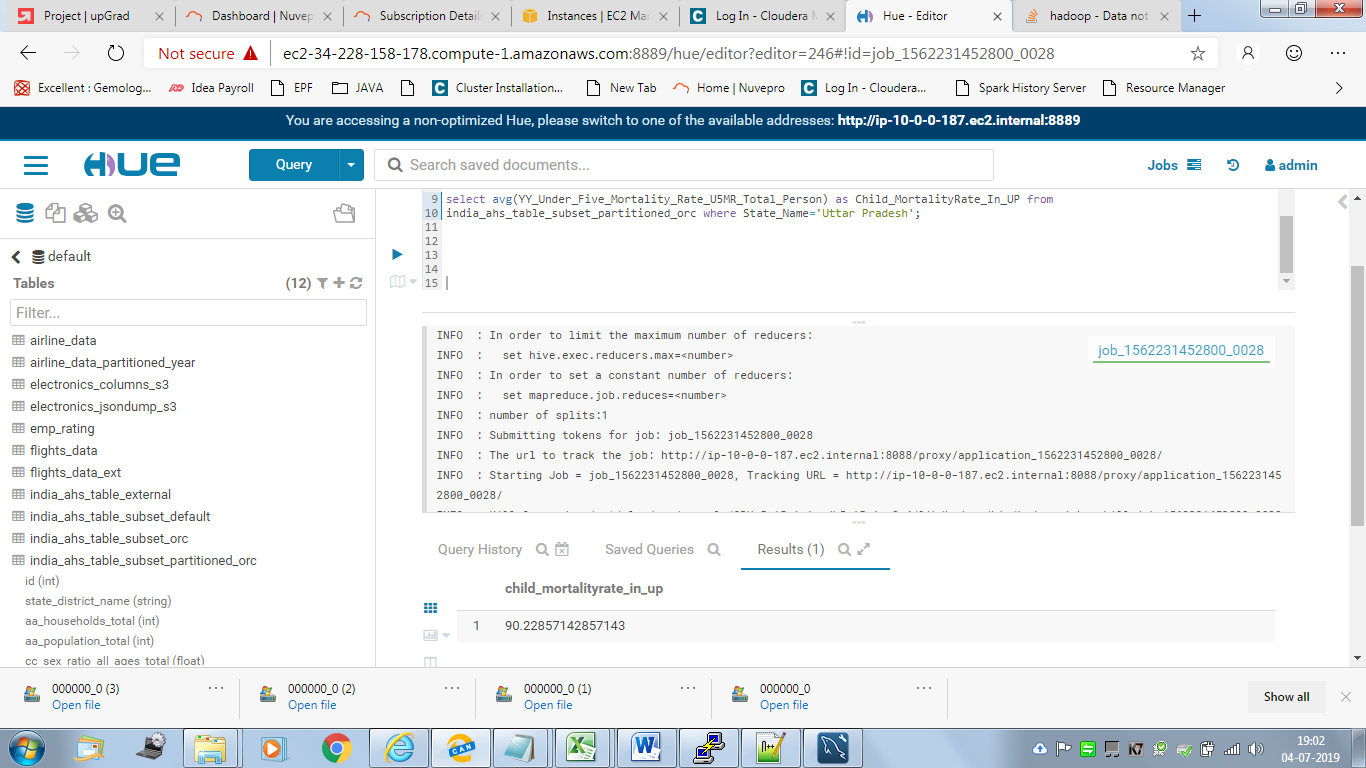
<target url="jdbc:oracle:thin:@localhost:1521:XE"

1. Eliminate the inconsistent data.
2. Split the table.
3. Assign the data.
4. Loading it into the end target. Pump the data into Oracle data warehouse. The loading phase is the last step of the ETL process. The information from data sources are loaded and stored in a form of tables. There are two types of tables in the database structure: Fact tables and Dimensions tables. Once the fact and dimension tables are loaded, it is time to improve the performance of the Business Intelligence data by creating Aggregates.
5. [Sqoop Import](https://sqoop.apache.org/docs/1.4.2/SqoopUserGuide.html#_controlling_the_import_process)
6. [Create Table in Hive](https://www.cloudera.com/documentation/enterprise/5-8-x/topics/impala_create_table.html)
7. [Load Data in Hive](https://www.cloudera.com/documentation/enterprise/5-2-x/topics/impala_load_data.html)
8. [Hue](https://www.cloudera.com/documentation/enterprise/5-9-x/topics/hue.html)
9. [Hive-HBase Integration](https://cwiki.apache.org/confluence/display/Hive/HBaseIntegration)
10. Orc Compression
11. Clean up.
12. EXPERIMENTAL ANALYSIS AND RESULT

We have proposed a new ETL and batch processing methodology for increasing the performance of the ETL which is different from the traditional ETL tool. This proposed Hyper ETL transforms the XML document file into the Oracle data base and we found that, 1 hour 57 minutes for nearly 10000 records for which execution time is less than the previous one, were transformed and loaded into relevant tables. Experiments were conducted in MySQL workbench, Oracle and XML language. The significance of this ETL was demonstrated through some sample Sales records and we used high configured system Intel® Xeon® E7-4600 Series Processor, 64GB DDR-III RAM for testing the Hyper ETL . In this paper, a systematic, an uncomplicated, and understandable Extract Transform Load Methodology is proposed. Experimental and analysis results are given below.

## Result:

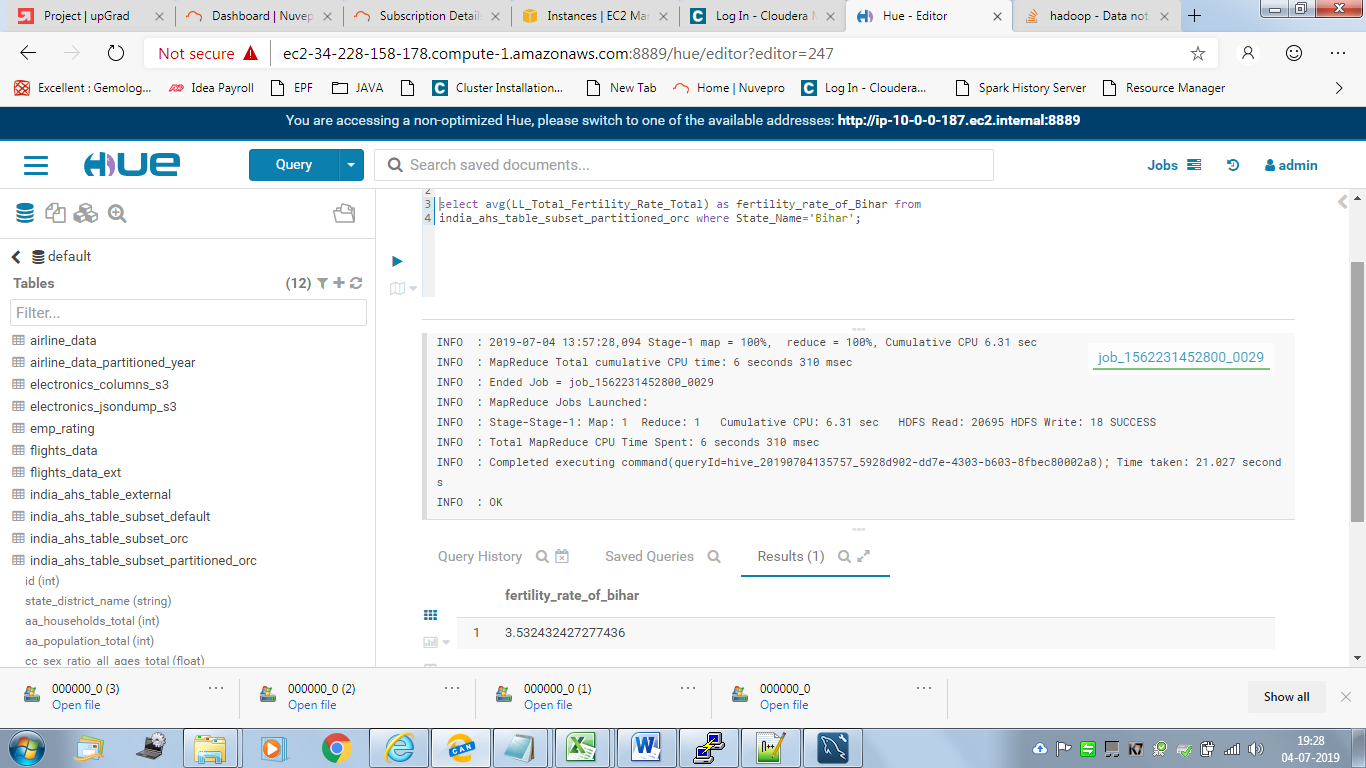
**1. The child mortality rate of Uttar Pradesh**



**2. The fertility rate of Bihar**

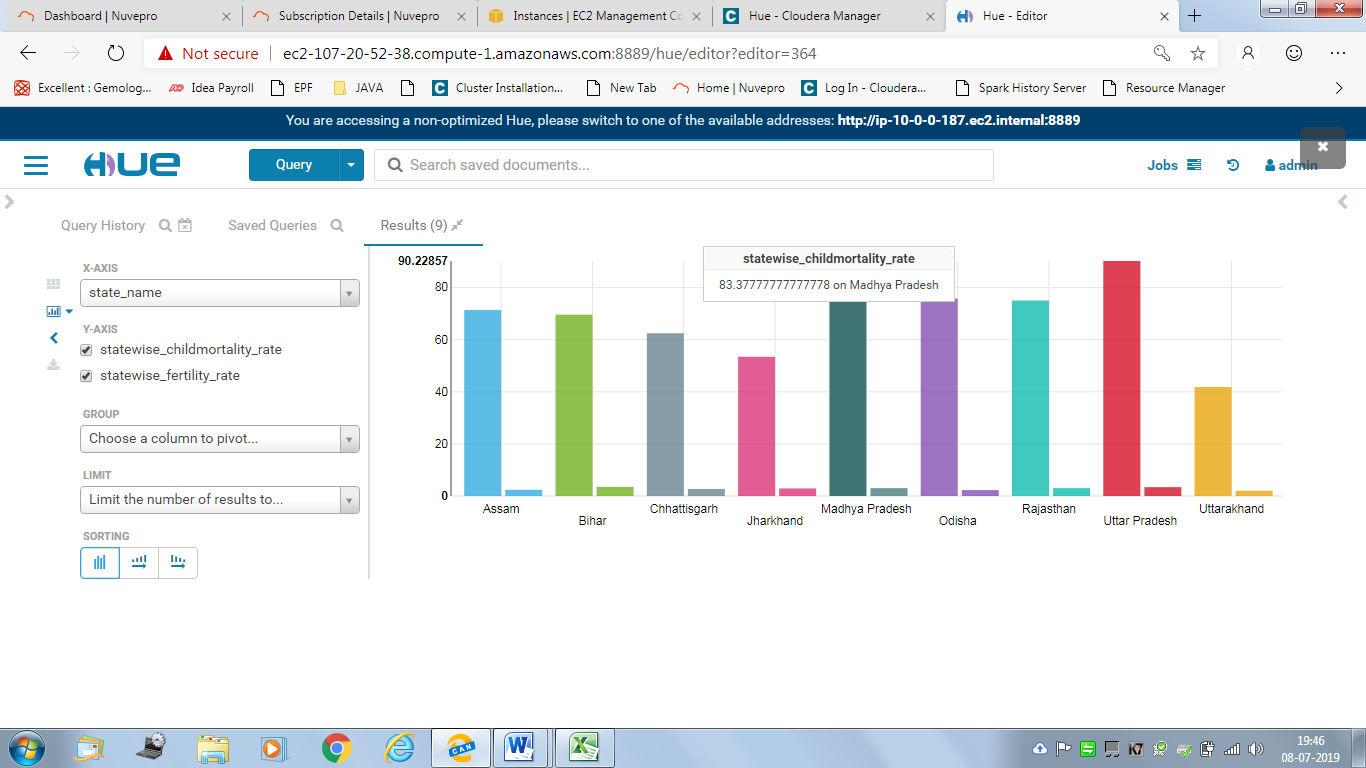
Output: 3.5534789

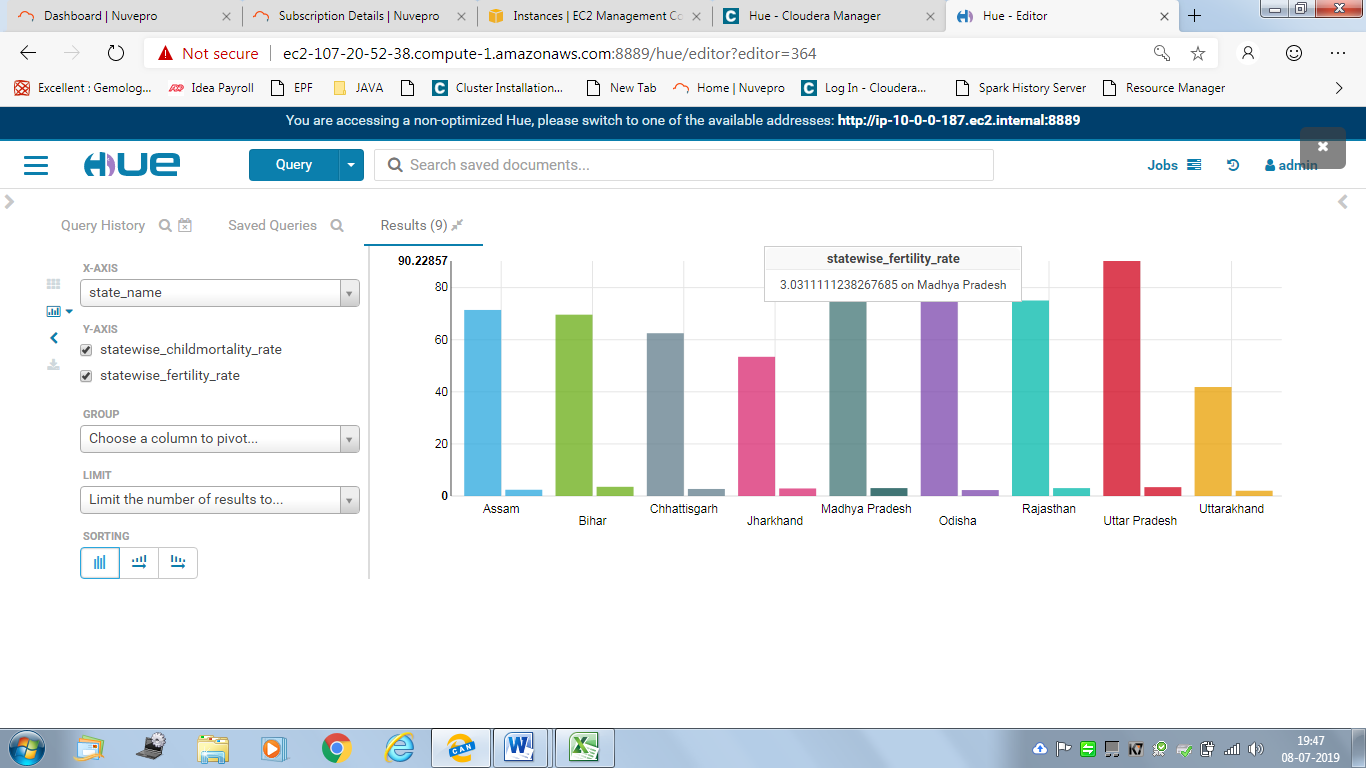
Time taken to analyse: 25 sec



**3. State wise child mortality rate and state wise fertility rate and does high fertility correlate with high child mortality?**

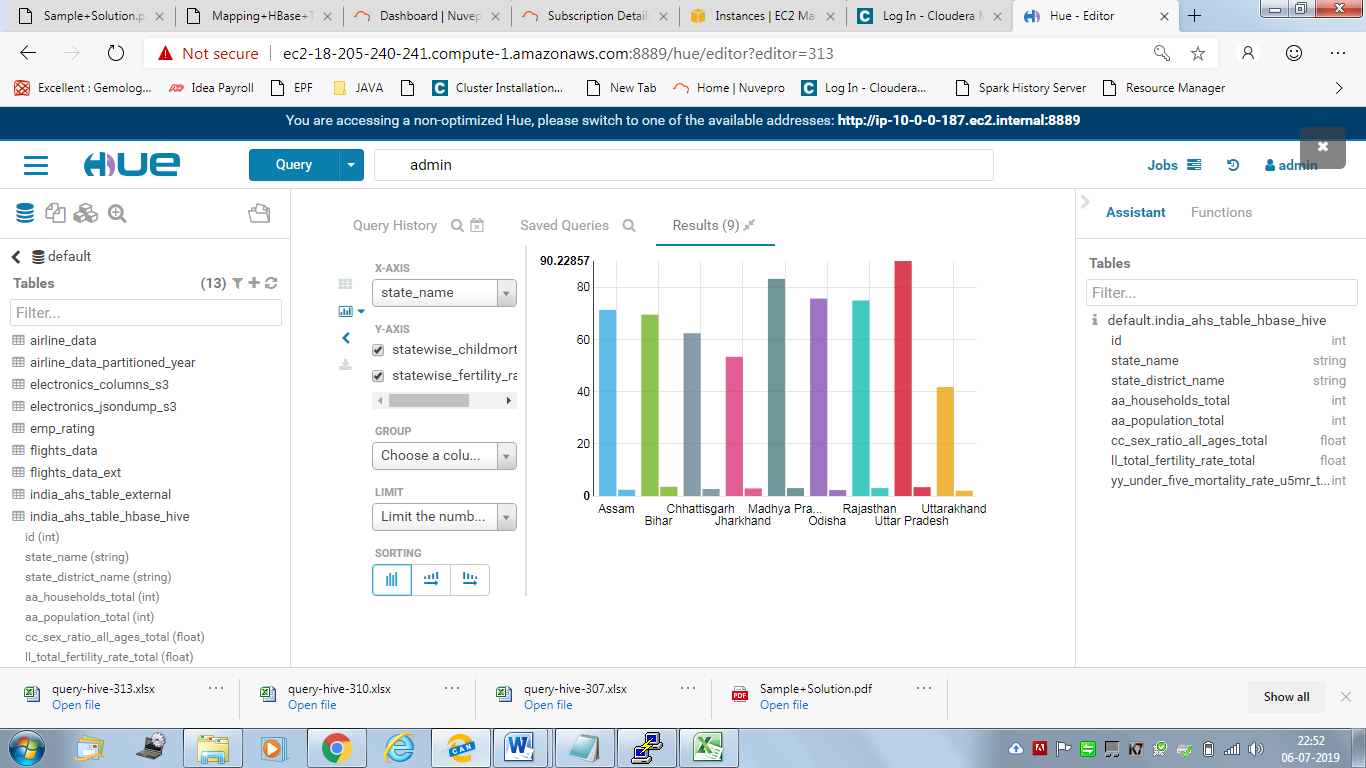
|  |  |  |
| --- | --- | --- |
| state\_name | statewise\_childmortality\_rate | statewise\_fertility\_rate |
| Assam | 71.43478261 | 2.399999997 |
| Bihar | 69.62162162 | 3.532432427 |
| Chhattisgarh | 62.5 | 2.701249987 |
| Jharkhand | 53.44444444 | 2.894444452 |
| Madhya Pradesh | 83.37777778 | 3.031111124 |
| Odisha | 75.8 | 2.279999991 |
| Rajasthan | 75.0625 | 3.028125003 |
| Uttar Pradesh | 90.22857143 | 3.397857138 |
| Uttarakhand | 41.84615385 | 2.022307708 |



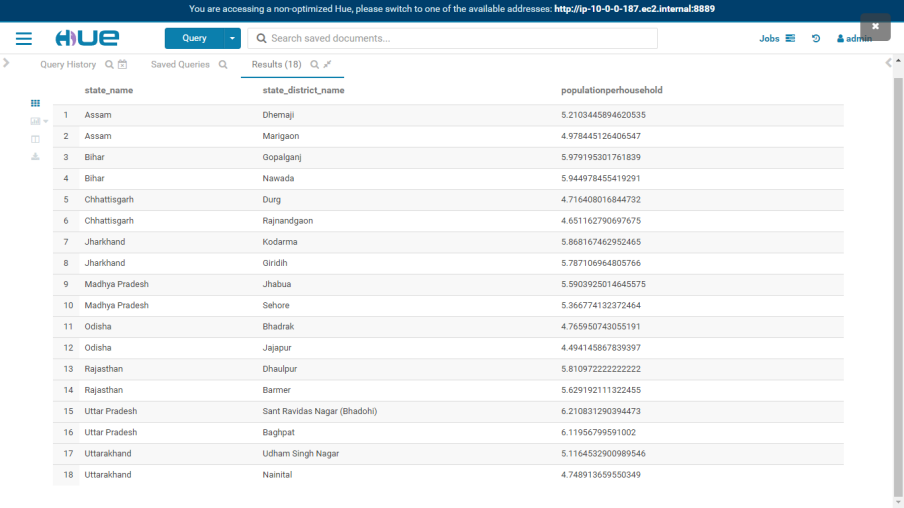


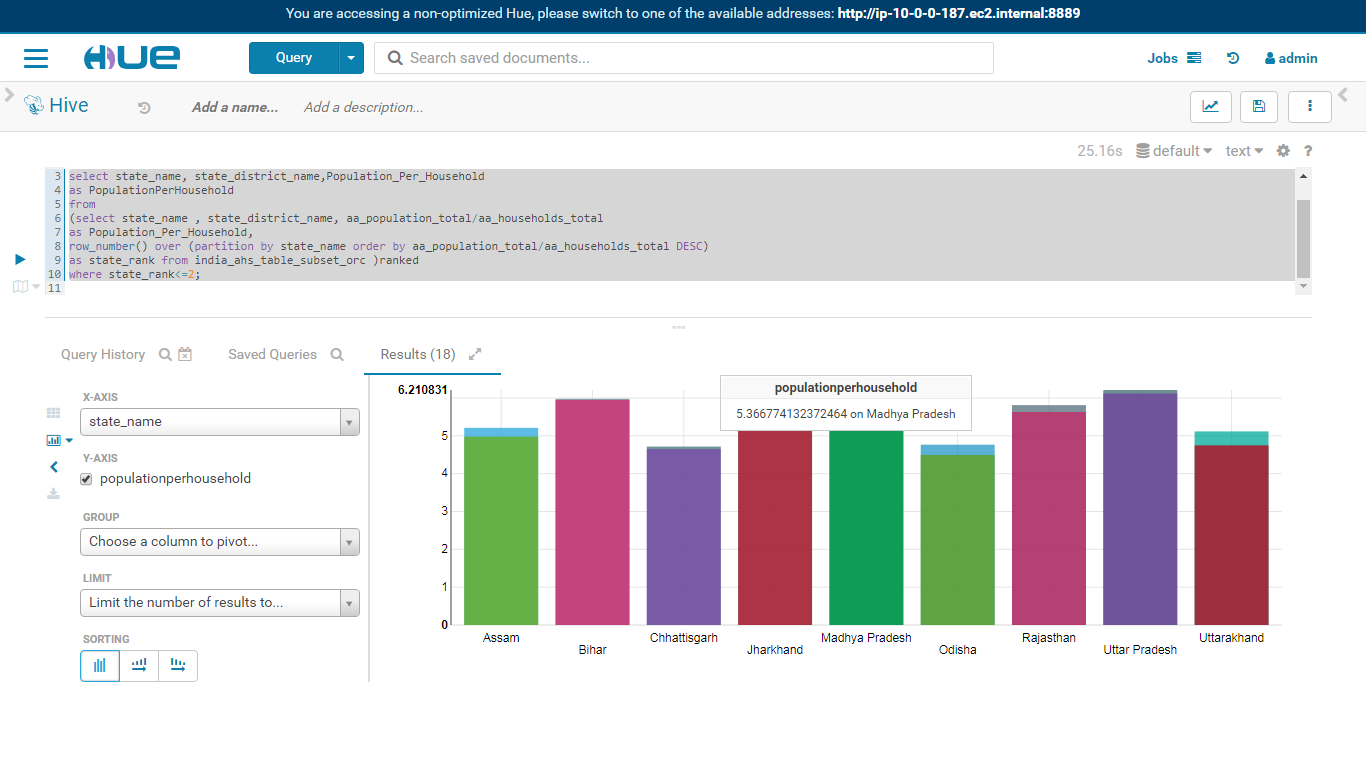
**<Query on the table with the chosen format such as orc>**

|  |  |  |
| --- | --- | --- |
| state\_name | statewise\_childmortality\_rate | statewise\_fertility\_rate |
| Assam | 71.43478261 | 2.399999997 |
| Bihar | 69.62162162 | 3.532432427 |
| Chhattisgarh | 62.5 | 2.701249987 |
| Jharkhand | 53.44444444 | 2.894444452 |
| Madhya Pradesh | 83.37777778 | 3.031111124 |
| Odisha | 75.8 | 2.279999991 |
| Rajasthan | 75.0625 | 3.028125003 |
| Uttar Pradesh | 90.22857143 | 3.397857138 |
| Uttarakhand | 41.84615385 | 2.022307708 |

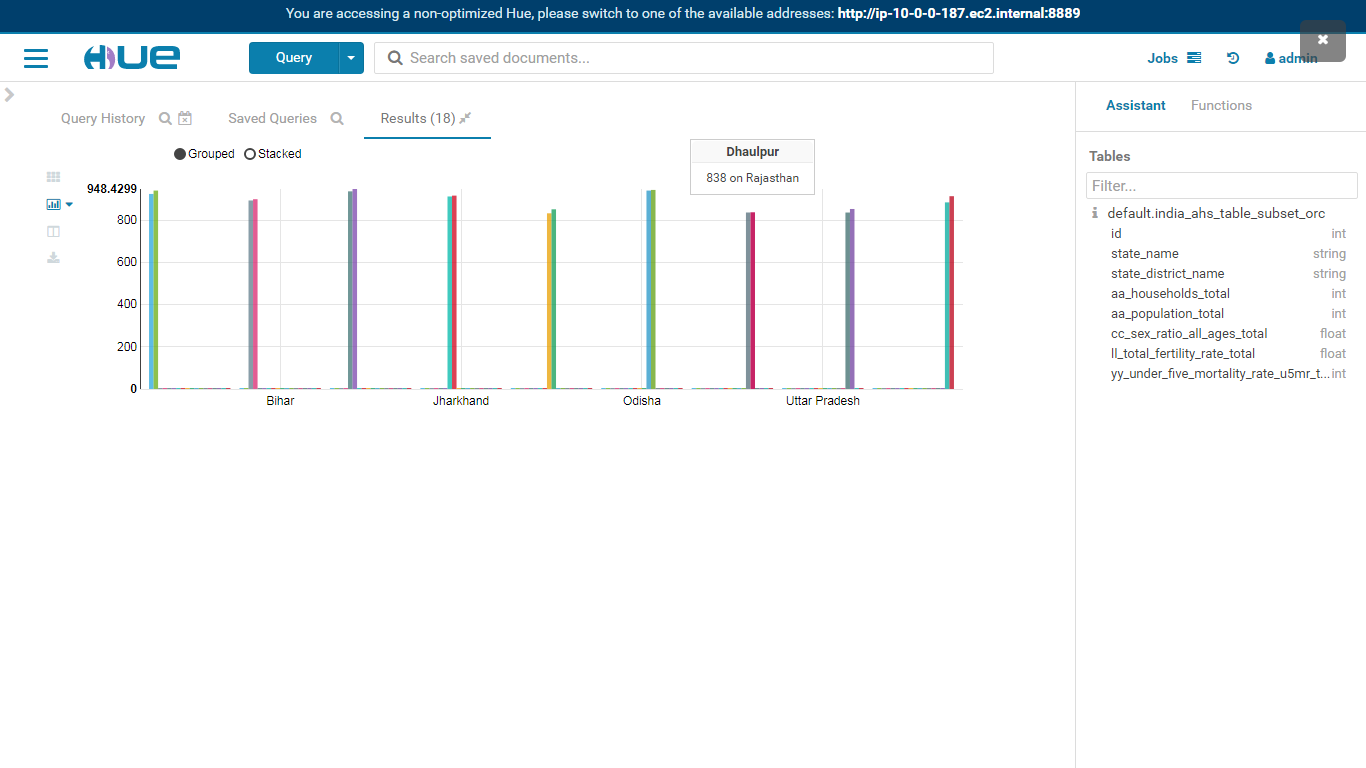


**4. Find top 2 districts per state with the highest population per household**

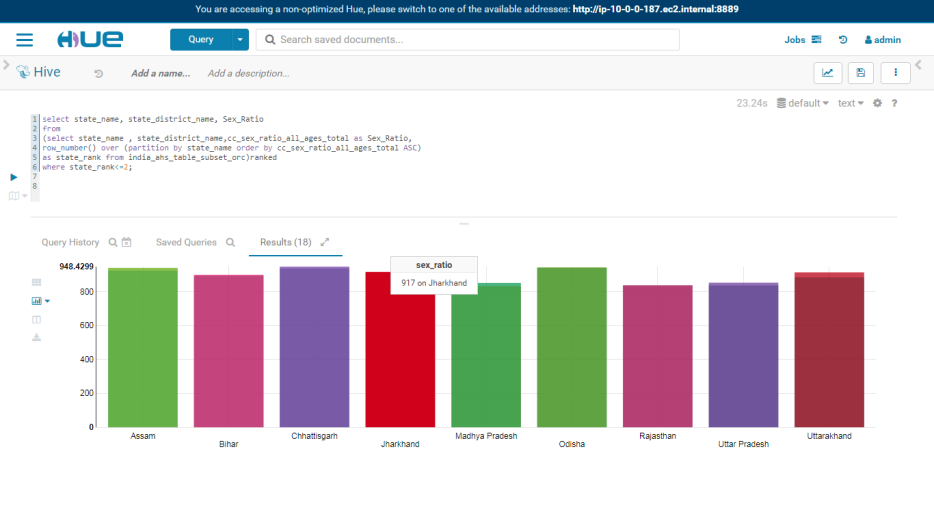




1. **Find top 2 districts per state with the lowest sex ratios**



|  |  |  |
| --- | --- | --- |
| state\_name | state\_district\_name | sex\_ratio |
| Assam | Kamrup | 925 |
| Assam | North Cachar Hills | 941 |
| Bihar | Pashchim Champaran | 894 |
| Bihar | Khagaria | 900 |
| Chhattisgarh | Koriya | 937.2999878 |
| Chhattisgarh | Bilaspur | 948.4299927 |
| Jharkhand | Dhanbad | 913 |
| Jharkhand | Bokaro | 917 |
| Madhya Pradesh | Morena | 833.1300049 |
| Madhya Pradesh | Datia | 852.1199951 |
| Odisha | Sonapur | 941 |
| Odisha | Jharsuguda | 944 |
| Rajasthan | Karauli | 837 |
| Rajasthan | Dhaulpur | 838 |
| Uttar Pradesh | Gautam Buddha Nagar | 836.8200073 |
| Uttar Pradesh | Shahjahanpur | 853.6699829 |
| Uttarakhand | Haridwar | 884.9299927 |
| Uttarakhand | Udham Singh Nagar | 914.3099976 |



1. CONCLUSION

We have proposed the sophisticated structural design of ETL which accomplishes, the mapping multiple sources into multiple targets, merging the relevant field from three different tables and purging of all data and eliminating the duplicate fields from the table.This ETL was demonstrated through sample Sales records and it is suggested that this ETL reduces the time, and improves the decision-making process and also automates the key activities of the process.We have learnt about the Hadoop Ecosystem such as HDFS , Apache Hive , Sqoop ,  HUE (which are one of the widely used tools in the industry) . How data is managed in the industry using ETL process while extracting source data from MySQL RDBMS and then transform and finally load into Hadoop based ware house system and further applied to create business intelligence with HiveQL outputs. Also have applied ORC (optimized row columnar) to increase the efficiency of the performance and reduce disk storage. The same approach and methodology can be used in other industries as well. It will help in vivid understanding and timely monitoring of different determinants. Hence, we conclude with the expectation of the above mentioned delivers will be tackled in an enterprise and industry.

1. REFERENCES

1. Fasel and D. Zumstein, “Fuzzy data warehouse approach for web analytics,” In MD.
2. Lytras, E. Damiani, J. M. Carroll, R. D. Tennyson, D. Avison, A. Naeve, A. Dale, P. Lefrere,F. Tan, J. Sipior, and G. Vossen, “Visioning and Engineering the Knowledge Society - A Web Science Perspective,” volume 5736 of Lecture Notes in Computer Science, pages 276–285. Springer, 2009.
3. Hariprasad T, “ETL testing Fundamentals,” on March 29, 2012.
4. Huamin Wang, “An ETL Services Framework Based on Metadata,” 2nd International Workshop on Intelligent Systems and Applications, May 2010.
5. H. Inmon, “Building the Data Warehouse,” Wiley Publishing, Inc., 4 edition, 2005.
6. Dr.Mitesh Joshi ‘A framework of study of ETL process optimiztion based on metadata repository ‘ , IEEE Conference on computer and technology APRIL 2013

**THIS PAGE HAS BEEN LEFT BLANK INTENTIONALLY.**