**Vishwakarma Institute of Technology**

**Software Requirements Specification**

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

***1/12/2021***

***Version 0.3***

|  |  |
| --- | --- |
| **Project Responsibility** | **Gr.no & roll no.** |
| Prof.Dhiraj Jadhav | -- |
| Lajat Manekar | 11810486(05) |
| Deep Patel | 11810095(40) |
| Abhaysingh Patil | 11810399(45) |
| Aditya Patil | 11811164(46) |
| Vedant Patni | 11810106(55) |

**Table of Contents**

[**1.**](#_gjdgxs) **INTRODUCTION 3**

[1.1](#_30j0zll) Purpose 3

[1.2](#_1fob9te) Scope 3

[1.3](#_1ci93xb) Scope for Future Work 3

[1.4](#_3znysh7) References 3

[**2.**](#_2et92p0) **OVERALL DESCRIPTION 4**

[Problem Statement 4](#_3whwml4)

[2.1](#_tyjcwt) Product Perspective 4

[Product Position Statement 4](#_2bn6wsx)

[2.1.1](#_3dy6vkm) System Interfaces 4

[2.1.2](#_qsh70q) User Interfaces 5

[2.1.3](#_3as4poj) Hardware Interfaces 5

[2.1.4](#_1pxezwc) Software Interfaces 5

[2.1.5](#_49x2ik5) Communications Interfaces 5

[2.1.6](#_2p2csry) Memory Constraints 5

[2.1.7](#_147n2zr) Operations 5

[**3.**](#_41mghml) **SPECIFIC REQUIREMENTS 6**

3.1SYSTEM FUNCTIONS

3.2 SYSTEM TOOLS

3.3 RESULTS AND CHALLANGES

4.  **END SEGMENT**  **9**

4.1LIMITATIONS

4.2 CONCLUSION 1[2](#_3l18frh)

# 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.

Today, big data is generated from many sources and there is a huge demand for storing, managing, processing, and querying on big data. The MapReduce 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.

## Purpose

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.

* The project mainly revolves around MySQL workbench , Hadoop Data File Systems Apache Sqoop and Apache Hive which are one of the most widely used tools in the industry.
* The process of ETL plays a key role in data integration strategies. ETL allows businesses to gather data from multiple sources and consolidate it into a single, centralized location. ETL also makes it possible for different types of data to work together.
* **Batch processing** helps in handling tasks like payroll, end-of-month reconciliation, or settling trades overnight.
* Batch processing systems can save money and labor over time, but they may be costly to design and implement up-front.

## Scope

Data Set

**India Annual Health Survey (AHS) 2012-13**

The dataset comprises a survey conducted in Empowered Action Group (EAG) states Uttarakhand, 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.Here are some from the many benifits stated below for extract , transform , load process.

* 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.
* Majorly there are some analyses which have done in a very minimal time which would help in vivid understanding and timely monitoring of different dterminants .

## Scope for Future Work

**Streaming incremental loading** moves processed data in real-time. When new data gets processed, the ETL immediately sends it to the target repository. Some organizations prefer streaming incremental loading because they want access to real-time data. Streaming incremental loading, however, can only move tiny amounts of data at a time. In real-world situations, it doesn’t look very different from small-batch incremental loading.

## References

* [Department of Health and Family Welfare](https://nrhm-mis.nic.in/), Govt. of India has published this data in [Open Govt Data Platform India portal](https://data.gov.in/catalog/indicators-annual-health-survey) under [Govt. Open Data License - India](https://data.gov.in/government-open-data-license-india).
* <https://www.upsolver.com/data-lake-platform>
* H Wang, “An ETL Services Framework Based on Metadata,” 2nd International Workshop on Intelligent Systems and Applications, May 2010.
* <https://www.sciencedirect.com/topics/engineering/batch-processing>
* <https://www.talend.com/resources/batch-processing/>

# Overall Description

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.

## System Perspective

The system perspective can be divided into the following sections:

### Economic Feasibility

The project is economically feasible as the only cost involved is having a computer with the minimum requirements mentioned earlier. For the industries to use the particular methodology is quite feasible going with terms of the buisness laws.

### Technical Feasibility

To deploy the application, the only technical aspects needed are as follows:

Operating Environment Windows 7/8/9/10 IOS /Ubuntu

Platform .Net Framework & IIS

Database SQL Server 2016

### Behavioral Feasibility

The system requires very minimal level of technical guidance and all the views available in the system are self-explanatory. The industrialists will be well guided with warning and failure messages for all the actions taken.

### Hardware Interfaces

Processor: Intel Core i5 7400 – LGA1151 – 7th Generation Core Desktop Processor

RAM: 6 GB

Minimum Space Required: 3 GB

Display: 64 bit color

GPU : AMD RYZEN 5 1070 / Nvidia 1080ti (minmum)

### Software Interfaces

Operating Environment: Win 7/ 8/ 8.1/ 10 / IOS

Platform .Net Framework & IIS Visual Studio 2019

Database SQL Server 2019

MySQL workbench 2.1.0

Oracle Virtual Machine / VMware machine work station

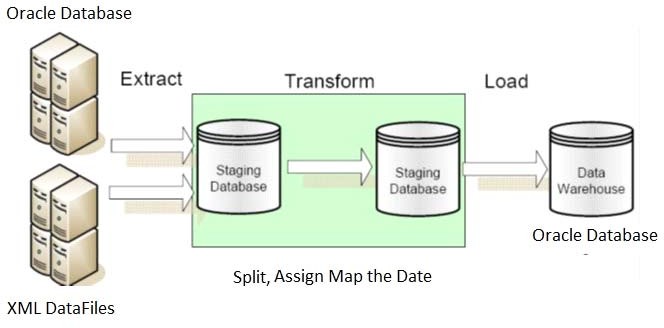
### Communications Interface

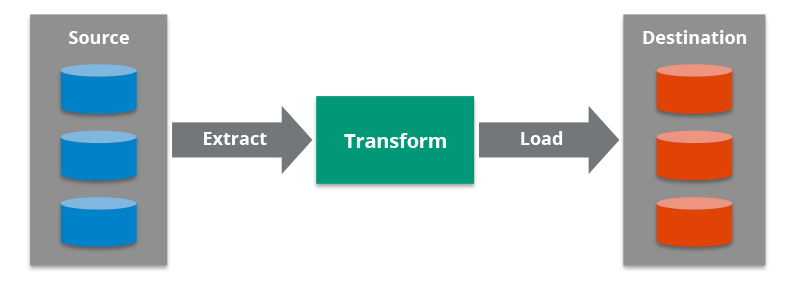
Stable Internet Connection (preferably at least 5mbps)

### Operations

After carefully analyzing the requirements and functionality of the web application, I had two important diagrams by the end of the analysis phase. They are the ER diagram and data flow diagram which were the basis for finding out entities and relationships between them, the flow of information.

## 3.1 System Functions





## 3.2 System Tools

Some main softwares and applications required are as follows :

**Apache Hive 3.1.2** : Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

**Apache SCOOP** : Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

### HUE :Hue is an open source SQL Assistant for Databases & Data Warehouses.

Hue brings the best [Querying Experience](https://docs.gethue.com/user/querying/) with the most intelligent autocompletes, query sharing, result charting and download for any database. Enable more of your employees to level-up and perform self service analytics and also visually discover insights .

**MySQL RDBMS** :MySQL is an open source, relational database management system (RDBMS) based on structured query language (SQL). MySQL is available on all major operating systems, including Windows, Linux and Solaris. ... MySQL, like other relational databases, stores data in tables, columns and rows.

**PuTTY (64Bit)** : PuTTY is a free and open-source terminal emulator, serial console and network file transfer application. It supports several network protocols, including SCP, SSH, Telnet, rlogin, and raw socket connection. It can also connect to a serial port.

### 3.3 Result and Challenges

Analyses to be done :

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 ratios

All the analyses gave the following output under 60 to 80 seconds with the visualizing of bar graphs and charts depending on the size of the data base we are filtering and finding . Henceforth with these analysis we can now suggest ways that can improve it .

The overall idea of doing this project was to get a real time experience. Learn new technologies and analyse the results in a minimal time setting the benchmark for batch etl.

## 3.4 Limitations

This system does not have a built in streaming etl process. Streaming ETL is the processing and movement of data which is continously streaming from one place to another. a streaming ETL architecture (or real-time ETL architecture) is fundamentally the same as a traditional ETL architecture.

Solution : (a general idea)

The dataset has to been hosted on Amazon RDS , and then from the cloud we can apply the same methodology it can process the streaming data in batch and perform the certain analysis for the betterment of society.

### 3.5 Conclusion

Our main object of the project was to ingest the indian annual health survey data hosted on MySQL database into Hadoop correctly an process it to generate analysis and The overall idea of doing this project was to get a real time experience. Learn new technologies and analyse the results in a minimal time setting the benchmark for batch ETL.We applied an ORC (optimize row columnar ) which stores data in a compressed form and perpaps because of that we were able to save a significant disk storage and huge cost in terms of storage and data preprocessing . Henceforth were able to generate better analytical methodology for Batch processing of the given data set and further this methodlogy can be used to create buisness intelligence with HiveQL ouputs.