**TY. B. Tech.**

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

**Project Statement of Work**

***01-12-2021***

***Version 1.0***

|  |  |  |  |
| --- | --- | --- | --- |
| Project Group Information | | | |
| Roll. No. | **Gr. No.** | **Name** | **Roles** |
|  | -- | Prof.Dhiraj Jadhav | Mentor / Guardian |
| 05 | 11810486 | Lajat Manekar | Front End , Data Ingestion , data enriching |
| 40 | 11810095 | Deep Patel | Frontend, Backend , compression query analysis , ASP |
| 45 | 11810399 | Abhaysingh Patil | Data Pre-processing , testing |
| 46 | 11811164 | Aditya Patil | Back End , Loading the dataset |
| 55 | 11810106 | Vedant Patni | Back End, Importing and verifying |

**Academic Year: 2020-21 Semester: I**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page** |
| 1 | Title | **2** |
| 2 | Background | **2** |
| 3 | Objective | **3** |
| 4 | Definitions and Applicable Documents | **4** |
| 5 | Business and/or Technical Environment | **5** |
| 6 | Description and Scope Of Work | **6** |
| 7 | Approach and Methodology | **7** |

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.

# BACKGROUND

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.

* 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 labour over time, but they may be costly to design and implement up-front.

# OBJECTIVE

Some organizations use more than one data warehouse or partner with organizations that use different data warehouse services. Even when using a single data warehouse, organizations often need the ETL process because they want to:

* Locate specific types of data.
* Separate data into a variety of categories.
* Transform multiple data formats into a common format.
* Add data to data analysis tools that use algorithms and artificial intelligence to identify trends.
* Load data to visualization tools that turn enormous amounts of information into graphs that most people can understand.

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 benefits stated below for extract, transform, and load process.

# DEFINITIONS AND APPLICABLE DOCUMENTS

### Step 1 – Extract

The extract step lets an ETL platform pull data from a source or multiple sources. Sophisticated ETL platforms can target specific types of data to extract from sources. Xplenty, for example, can identify data in non-relational NoSQL databases like MongoDB as well as relational SQL databases like Amazon RDS.

### Step 2 – Transform

Xplenty gives you [no-code and low-code options](https://www.xplenty.com/blog/low-code-vs-no-code/) for transforming data before loading it to a destination. When building a data pipeline that connects the extraction source and the load destination, you can use transformations like:

* **Duplication** that identifies and deletes duplicate data.
* **Format revision** that reformats data into a consistent format.
* **Cleansing** that deletes old and incomplete data that doesn’t add value to the data set.
* **Joining** that combines data from more than one source.
* **Splitting** that divides a column into multiple columns.
* **Integration** that standardizes data elements throughout the data warehouse.
* **Validation** that lets users create unique rules for the ETL to follow when it encounters specific instances, such as reporting an alert when it encounters a blank row.

Transformation is arguably the most important part of the ETL process because it makes changes to data before the final step, loading the processed data to a target destination.

### Step 3 – Load

Data loading is the process of moving data from the ETL platform to a destination or multiple destinations. Data loading usually happens in [batch increment or streaming incremental loads](https://www.xplenty.com/blog/ways-to-build-an-etl-process/).

* **Batch incremental loading** moves data to the target repository in a batch or batches. Many organizations use batch incremental loading outside of peak hours. Batch incremental loading can take several minutes or hours, depending on how much data gets moved. Letting the ETL platform work during off-peak hours helps prevent system overloads that slow down other processes.
* **Small batch incremental loading**works slightly differently. Instead of trying to load all of the data in a single batch, it breaks the transformed data into sections and loads then minute by minute. This approach creates a smaller burden on the system and behaves much like real-time updates.

1. **Design -** The process of defining the architecture, interfaces, and other characteristics of a system or component. The phase in the [software life-cycle](http://soft.vub.ac.be/FFSE/SE-contents.html#SoftwareLifeCycle) that emphasises a logical solution, i.e. how the system fulfils the requirements.
2. **Domain -** A problem area. Typically, many application programs exist to solve the problems in a single domain. An area of knowledge or activity characterized by a set of concepts and terminology understood by practitioners in that area.
3. **Framework** - A framework, or software framework, is a platform for developing software [applications](https://techterms.com/definition/application). It provides a foundation on which software developers can build programs for a specific [platform](https://techterms.com/definition/platform).
4. **Model -** A model is a simplified representation of a system or phenomenon with any hypotheses required to describe the system or explain the phenomenon, often mathematically. It is an abstraction of reality emphasizing those aspects that are of interest to someone.
5. **Platform -** A set of subsystems and technologies that provide a coherent set of functionalities through interfaces and specified usage patterns, which any application supported by that platform can use without concern for the details of how the functionality provided by the platform is implemented.
6. **Software -** Software is part of a system solution that can be encoded to execute on a computer as a set of instructions; it includes all the associated documentation necessary to understand, transform and use that solution. Software is the collection of computer programs, procedures, rules, and associated documentation and data.
7. **Statistics -** The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.
8. **User** – An authority that can use the application, resources or system.
9. **User Interface** - The user interface (UI) is the point of human-computer interaction and communication in a device. The user interface in the [industrial design](https://en.wikipedia.org/wiki/Industrial_design) field of [human–computer interaction](https://en.wikipedia.org/wiki/Human%E2%80%93computer_interaction), is the space where interactions between humans and machines occur.

**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 SQOOP**: 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.

# DESCRIPTION AND SCOPE OF WORK

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 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 benefits stated below for extract, transform, and 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 determinants.

# APPROACH AND METHODOLOGY

1. ***Upload the dataset in MySQL RDBMS***
2. ***Data Ingestion from the RDS to HDFS using Sqoop:***
3. ***External table creation in Hive and loading the ingested data into it. Data ingestion verification:***
4. ***Queries to verify that the ingestion is correctly accomplished***
5. ***Subset schema creation in Hive to support the analyses:***

***> HiveHbase****, ORC, Default format*