Project Report

Project Title: Tallaght University Hospital

Student Name: Mahnoor Student Id: X00176345

Research Document:

**ETL Tool**:

A diagram of two people

Description automatically generated with medium confidence

ETL Tools stand for Extract Transform and load is used to gather different types of data and locate it in one place. The job of ETL tools is to process collects and refine different types of data and then deliver the data to a data lake or data warehouse such as Redshit, Azure or BigQuery.

They are 3 steps of ETL process such as data extraction, data transformation, and data loading. ETL tools are classified into four main groups as follows: On-premises ETL, Open-source ETL, Cloud-based ETL, and Hybrid ETL.

Examples of ETL tools are as follows MongoDB, Google Cloud, IBM Infosphere DataStage, Oracle Data Integrator, Microsoft SQL Server Integration Services (SSIS), and many more. The best ETL tools that is considered in the industry is MongoDB.

MongoDB is good source as it can handle NoSQL, SQL and various data schema or dataset. MongoDB also has cons and pros are as follows:

Pros:

* Cloud platform, no strain no internal resources
* Large volume of pre-built connectors and compatibility with multiple other data sources
* Built-in MongoDB integration
* Set up data flows with ease.
* Connects to most popular data warehousing solutions.
* Highly secure and stable
* Award-wining customer support.

Cons:

No-code environment might be difficult for engineers to perform their task.

CSViewer:

It is also one of the ETL tools which helps us to read csv files and where we can mange them to make graphs and handle the data in what ever way we wanted.

GeoKettle ETL tools: Based on the Kettle framework, GeoKettle focuses on geospatial data and offers a set of tools for ETL processes related to geographical data, but it can handle CSV files and various other data formats as well.

Downloaded link: <https://sourceforge.net/projects/geokettle/>

CloverETL**:** This open-source ETL tool offers a community edition that supports various data transformations and connectivity options, including handling CSV files. It also handles structured and unstructured data.

Download Link: https://sourceforge.net/projects/cloveretl.berlios/

Data Isolation Tool:

Data Isolation Tool plays important role in transaction the data and it also has four level of data isolation, where each helps users to better access the same data at the same time.

Type of Data

**MariaDB:**   
MariaDB is an open-source, enhanced version of MySQL, designed to be a powerful and compatible database system for various applications.

**MongoDB:**

A NoSQL database storing data as flexible, JSON-like documents (BSON). It's schema-flexible, offers high performance for large datasets, scales horizontally, and suits applications with evolving data structures like real-time analytics or content management systems.

**Comparison of MariaDB and MongoDB:**

* **MongoDB:** NoSQL stores flexible JSON-like documents, great for scalability, dynamic data, and applications handling unstructured data.
* **MariaDB:** SQL relational database, structured tables, focuses on strong data integrity, transactions, and relationships between data, suitable for traditional applications with structured data needs.

Which is best?

MongoDB is good for Unstructured data or semi structure data. Where as Mariadb handles structured and relational data that has well defined schemas. MongoDB is Scalabiliy and cann’t handle large volume of data across the multiple servers where as Mariadb has strong data integrity ability. In MongoDB is flexible schemas whereas Mariadb can handle complex queries, relationships between data and mature SQL-based system.

Which database you have choose and why?

Will decide in the meeting then can write which database we have choose and why.

A diagram of a diagram

Description automatically generated with medium confidence Architecture Diagram:

Describe how we are going to load data?

How we will add csv file in ETL tools

Use case diagram:

A diagram of a person's relationship

Description automatically generated

Wireframe:

ERD Diagram:

Iteration 1:

Database Tables

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name: Patient** | | | |
| **Field** | **Unique** | **Datatype** | **Description** |
| Lab No | ? | Varchar(64) | What is this? Is it lab number specific to a patient? Is it unique? |
| OCS no | ? | Varchar(64) | What is this? Is it unique? |
| MRN | Yes | Varchar(64) | Medical Record Number |
| Forename | No | Varchar(64) | Person first name |
| Surname | No | Varchar(64) | Person sur name |
| DOB | No | Date | Person date of birth |
| Address1 | No | Varchar(256) | Person address first line |
| Address2 | No | Varchar(256) | Person address second line |
| Address3 | No | Varchar(256) | Person address third line |
| Clinician Code | ? | Varchar(64) | What is this? Is this information specific to person? Is it unique? Do we have any other clinician record? |
| Clinician Class | ? | Varchar(64) | What is this? Is this information specific to person? Is it unique? |
| Source Code | ? | Varchar(64) | What is this? Is this information specific to person? Is it unique? |
| Source class | ? | Varchar(64) | What is this? Is this information specific to person? Is it unique? |
| Date of Request | No | Date | Date test is requestd.  Who requests this test? |
| Time of request | No | Date | Time test is requestd.  Who requests this test? |
| Date received | No | Date | Date test request is received  Who requests this test? |
| Time Received | No | Date | Time test request is received  Who requests this test? |
| Tests- @alltests | No | Varchar(64) | How do we present this information? Do they have a separate entry for each test type, such as Na, K, Urea? |
| Fasting? | Yes | Boolean |  |
| Phoning details? | No | Varchar(64) | Is it patient or clinician phone number? |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name: Test** | | | |
| **Field** | **Unique** | **Datatype** | **Description** |
| TLC | ? |  |  |
| TFC Included | ? |  |  |
| test names | ? |  |  |
| Units | No |  |  |
| Age related reference range? | No |  |  |
| Reference Ranges | No |  |  |
| Flag Limit Low | No |  |  |
| Flag Limit High | No |  |  |
| Alert Limit Low | No |  |  |
| Alert Limit High |  |  |  |

Reference:

https://portable.io/learn/best-etl-tools

<https://www.talend.com/resources/what-is-etl/>

<https://www.datacamp.com/blog/a-list-of-the-16-best-etl-tools-and-why-to-choose-them/>

https://mariadb.com/