# Practical 1

* Aim :- Creating Data Models Using. Cassandra.

Definition :-

* Cassandra.

Cassandra is a distributed database from Apache that is highly scalable and designed to manage very large amounts of structured data. It provides high availability with NO SQL database.

Apache Cassandra is highly scalable, high performance distributed database designed to handle large amounts of data across many commodity many commodity servers

* NO SQL Database.

A NO SQL database is a database that provides mechanism to store and retrive data other than tabular relations used in relational Databases.

The primary Objective of NOSQL database. is to have :

- Simplicity of design.
- Horizontal Scaling
- More Control over Availability.

* Software requirement :- Cassandra, Python GUI 2.7.15

# Practical 2

* Aim :- Homogeneous Ontology for Recursive Uniform Schema. (HORUS)

* Definition :- HORUS

The Homogeneous Ontology for Recursive Uniform Schema (HORUS) is used as internal Data format structure that enables framework is The Use of HORUS methodology results in a hub-and-spoke data transformation approach. The basic concept is to alternative raw data and then transform it first to single format. That means that there is only one format for text files, one format for JSON or XML one format for images and video. Therefore to achieve any to any transformation leverage the framework is only requirements are a data format-to-HORUS and HORUS-data-format converter.

* Software requirement :- python GUI 2.7.15

* Libraries and Package :-

Pandas, matplotlib, numpy, Scipy

# Practical 3.

* Aim :- Utilities and Auditing.

* Definition :-

- Data Processing Utilities :-

* Retrieve Utilities

Utilities for this superstep contain the processing chains for retrieving data out of raw data lake into a new structured format. It is used by Data Science framework to enable the reduction of development work required to achieve a complete solution that handles all data format.

* Organize Utilities.

For this superstep it contains all processing chains for building data marts. The organize utilities are mostly performed to create data marts against data scientist result stored in data ware house and facts.

* Report Utilities

The report utilities are mostly used to create data visualization against the data science results stored in data marts.

Maintenance Utilities

* ~~Back up and Restore Utilities~~.

This data science solutions you are building are a standard data system consequently

* Backup and Restore Utilities
These perform different types of Database backups and restores for solution.

* Checks Data Integrity Utilities
These utilities check allocation and structural integrity of database objects and indexes across ecosystem to ensure accurate processing of data into knowledge.

* History clean up Utilities
These utilities archive and remove entries in history tables in Database.

* Software requirement :- Python GUI 2.7.15.

* Libraries :-
Pandas, matplotlib, Scipy.

# Practical 4.

* Aim :- Retriving Data.

* Definition :- Retrieve Superstep
The Retrive superstep is a practical method for importing completely into processing ecosystem a data lake consisting of various external data sources. The successful retrival of data is a major stepping stone to ensuring that you are performing well. Data lineage delivers audit trails of data elements at lowest granular level to ensure full Data governance. Data governance supports metadata management for system geidelines. The Retrive superstep supports edge of ecosystem where it makes direct contact with Outside data world.

* Software requirement :- R Studio.

* Libraries :- R studio.
read (r), tibble, data table.

# Practical 5

* Aim :- Assessing Superstep.

* Definition :- Assess Superstep.
Assess Superstep Data quality refers to the condition of a set of qualitative or quantitative variables. Data quality is multi-dimensional measurement of acceptability of specific data sets. Data ~~quality~~ profiling involves observing your data sources all the viewpoints that information offers. The main goal is to determines what additional processing to apply to entries.

* Errors.

- Accept the error.
If it falls within an acceptable standard can decide to accept it and move on to next data entry

- Reject the error.
Occassionally predominantly with first time data imports the information is so severely damaged that it is better to simply delete data entry and not try it to correct it.
Note :- Removing is a last sort.

- Correct the error.
This is the option that major part of assess step is dedicated to spelling mistakes in

in customer names, addresses and location are common source of errors which are methodically corrected.

- Create a Default value.
This is an option that I commonly used in my consulting work with companies. Most system developers assume that if business doesn't enter value they should enter a default.

* Software requirement :- Python 3.12.

* Libraries :-
Pandas.

# Practical 6.

* Aim :- Processing Data.

* Definition :- Process Superstep
The Process Superstep adapts assess results of retrieve versions of data sources into a highly structured data vault that will form the basic data structure for rest of data science steps this data vault involves the formulation of a standard data amalgamation formed across a range of project.

* Software requirement :- Python 3.12.

* Libraries :-
pandas, sqlite3, uuid.

# Practical 7.

* Aim :- Transforming Data

* Definition :- Transform superstep
The transform superstep allows you as a data scientist to take data from data vault and formulate answers to questions raised. The transformation steps the data science process that converts results into insights. It takes standard data science techniques and methods to attain insights and knowledge.
The transform superstep uses the data vault from process steps as its sources data. The transformations are aimed to work with five dimension of data vault.

* Software requirement :- Python.

* Libraries :-
pandas, sqlite3, uuid, numpy, mat plot lib.

# Practical 8.

* Aim :- Organizing the Data

* Definitions :- Organize Superstep.
The Organize Superstep takes complete data warehouse and it is then passed to Report Superstep. It collects the relevant information from your prepared data warehouse to match requirement of a specific segment at customers decision makers. The Organize Superstep keeps your data into smaller data structure called data mart.

* Software requirement :- Python 3.12.

* Lab names :-
Pandas, sqlite 3, networkx, matplot, lib, numpy.

# Practical 9.

* Aim :- Generating Data.

* Definition :- Report Superstep.
Report Superstep is step in the ecosystem that enhances the data science findings with art of Storytelling and Data visualization.
You can perform the best data science but if you cannot execute report step by turning your data science into actionable business insights you have achieved no advantage for your business.

* Software requirement :- python 3.12.

* Libraries :-
Pandas, networkx, matplotlib, sqlite3

# Practical 10.

* Aim :- Data visualizations with Power BI

* Definition :- Power BI
Power BI is a unified, scalable platform for self service and enterprise business intelligence (BI). Connect to and visualize any data and seamlessly infuse the visuals into apps you use every day.
Easily connect to model and visualize your data creating memorable reports. personalized with your KPIs and brand. Get fast AI powered answers to your business questions even when asking with conversational language.

* Software requirement :- Power BI