Project Title:	**Dynamic Data Ingestion and Storage in HDFS with Automated Hive Integration**
Technologies	**Python, API, HDFS, Hive, Shell scripting**

Problem Statement:

The task aims to fetch data from a specified link, store it in HDFS, and create a Hive table to visualize the data. Initially, ensuring access to the provided link and successful data retrieval is essential. Subsequently, determining the data format and schema, if structured, is pivotal. Utilizing tools like wget or curl for data retrieval and HDFS CLI for storage follows. Finally, creating a Hive table, loading data, and verifying correctness conclude the task, with an optional script for automation.

Virtual Hadoop Machine - platform

Approach:

To accomplish the task of fetching data from the provided link (https://www2.census.gov/programs-surveys/popest/datasets/), storing it in HDFS, and creating a Hive table, follow these steps:

1. Verify accessibility and download capability from the provided link.
2. Determine the data format and structure (single or multiple files, structured or unstructured).
3. If structured, ascertain the data schema for Hive table creation.
4. Employ wget or curl to download data and pipe output to HDFS CLI (hadoop fs -put) for storage.
5. Use Hive CLI to create a new database and table.
6. Utilize LOAD DATA INPATH HiveQL command to load data from HDFS into the created Hive table.
7. Confirm data integrity using SELECT HiveQL command to view and validate loaded data.

8. Optionally, automate the process via scripting for future data refreshes.

Results:

Following the outlined approach should lead to successful data retrieval from the specified link, storage in HDFS, and the creation of a Hive table for data visualization. Validation of data integrity within the Hive table ensures correctness. Optionally, a script can facilitate automated data refreshes for enhanced efficiency .

Submission:

- Provide a well-commented Colab file containing the complete code for the project, organized into sections for data Pipeline and Analysis.
- Upload the same into github with a proper Readme file.
- Presentation on the entire project, including Problem Statement, Tools Used, Approaches and Insights Found.

Evaluation Metrics:

- Project evaluation will be done in the live session and have to showcase the approaches done to complete the project
- You are supposed to write a code in a modular fashion (in functional blocks)
- Maintainable: It can be maintained, even as your codebase grows.
- Portable: It works the same in every environment (operating system)
- You have to maintain your code on GitHub.(Mandatory)
- You have to keep your GitHub repo public so that anyone can check yourcode.(Mandatory)
- Proper readme file you have to maintain for any project development(Mandatory)
- Follow the coding standards:
    - https://www.python.org/dev/peps/pep-0008/
- You should include basic workflow and execution of the entire project in the readme file on GitHub

GitHub Repo:

The attached reference document will help you use GitHub effectively. - Link

Reference Material:

Official Documentation:

- Apache Hadoop Documentation
- Apache Hive Documentation

Official Websites:

- Apache Hadoop
- Apache Hive