

CA675: Cloud Technologies

Assignment 1: Data Analysis on google cloud and local hadoop deployment

Assignment implemented both on **Google Cloud Platform** AND local machine

Contents

1) Task 1: Data Extraction, Transform, Load

2) Task 2 & 3: Cleaning, Loading and Querying Data

3) Task 4: TF/IDF

4) Screenshots

GITHUB Repo:

Task 1: Data Extraction

1. Data Queried from following link:

<https://data.stackexchange.com/stackoverflow/query/new>

1. Query Used – Double click object file



Task 1: Transform

1. File combined by Linux cat command.
2. Data cleaned in Python for final export.

**Note: Double click object**



**Reasons to use Python**

* Access to regex.
* Pandas DataFrame to load data.

Task 1: Loading

**Reasons to use Hive:**

1. Hive has similar syntax to SQL
2. Already have experience in querying data on Teradata.

**Steps Taken**

1) Create Table in Hive: Attached SQL



2) Create view on top of table with casting to int for score:

**Reason**: Use Serde format while loading data instead of delimited for loading data correctly.

Attached SQL for view



Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Above queries output are executed directly in Jupyter Notebook along with output and can be viewed on GitHub with following link

Task 4: TF/IDF

Reasons to use Python for calculating TF/IDF and Hive for storage

1. Python has inbuilt libraries to tokenize and calculate TF/IDF.
2. Pandas can be used for data manipulations and DataFrame can be used as temporary data storage while code executions.
3. Hive querying has similar syntax to SQL.
4. Already have experience in querying data on Teradata hence familiar with SQL syntax.

Code along with output can be found at following link: