Final Project

Turn in final project deliverables here.

Dataset

- **Description:** The dataset comprises Stackoverflow votes, comments, users, tags, badges, etc related to any online activity occurring on the website.
- **Size:** The total size will be either 42GB or 22 GB. We'll choose a dataset from one of the 2 following sources.
- **Format:** The format of data is in the form of .xml files stored in compressed 7zip files.
- **How you will obtain the data:** We have downloaded the data from the following websites and will choose from either of the following:
    1. https://www.kaggle.com/stackoverflow/stackoverflow
    2. https://www.brentozar.com/archive/2015/10/how-to-download-the-stack-overflow-database-via-bittorrent/
- **What features are available:**
    - **Users**: Reputation, CreationDate, DisplayName, WebsiteUrl, Location, Views, UpVotes, DownVotes, AccountId
    - **Posts**: Id, PostTypeId, AcceptedAnswerId, CreationDate, Score, Body, OwnerUserId, Title, Tags, AnswerCount, CommentCount
    - **Comments** : RowId, PostId, Score, Text, CreationDate, UserId, ContentLicense
    - **Tags**: RowId, TagName, Count, ExcerptPostId, WikiPostId
    - **Badges** : RowId, UserId, Name, Date, Class, TagBased
    - **Votes** : RowId, PostId, VoteTypeId, CreationDate

If you are going to use a new system we didn't cover in class, describe it here. Your project can focus more on using/learning new technologies rather than analysis if you'd prefer. You may also choose to implement a system for processing big data.

We'll be using Google Cloud Dataproc to do pyspark analysis on Google cloud. The files will be stored on Google Cloud Storage which can be replaced with HDFS. We might use Airflow (Google Cloud Composer) to orchestrate jobs. We'll use Google Cloud DataStudio to develop visualizations.

Project Plan

Please provide a high-level description of what you hope to achieve with your analysis. You have flexibility here; doing one cool or interesting thing is better than doing many straightforward things. Imagine that you are going to make a poster to present the project to the rest of the CS department --- what would be the most important insight you'd want others to gain from your work?

Please provide a brief, one or two paragraph description of your project and what you hope to analyze:

There are a lot of things which people discuss on stackoverflow. In this project, we intend to analyze the trending topics, e.g a software, algorithms , new technology etc. There should be a way which stackoverflow can use (or they may be using internally) to know their audience and accordingly promote these topics through forums, discussion boards etc.

The most straightforward way for them to achieve this is to use past statistics (which are logs in our case, in the form of .xml files) and predict trends in the future. We plan to do this, which forms a flow as follows :

Raw Data > Cleaning Data > Data Warehouse > Data Analytics and some ML

We plan to upgrade our work done in Project 2 and 3.

Provide a list of the deliverables you will turn in:

1. **Cleaning the dataset**

   We'll remove the extra data points from files in our dataset that might not be necessary and store it as a new file in GCS.

   Optional:

   We'll generate a simple CSV from "Users.xml" file that will contain the only the user_id and username from a pyspark job. This pyspark job, its output, will be linked to our main/second job with the help of Airflow/Google Cloud Composer.

2. We will do sampling of the files using Spark to assess our dataset.
3. Then we will process the data using Pyspark to derive insights
4. **Finally**,

- We will analyze the comments data and based on postId, we will find the most popular posts and their comments. Out of the text from these comments, we’ll find a trend of things most talked about.
- We ll generate visualizations with Google Data Studio

Optional:

- The above comments data which we analyzed, can also be used to predict some important forums, which stackoverflow can host on their website about topics which people generally like to talk about (using ML).
- We might try to create a pipeline with more than 2 jobs using Airflow
- We might use additional aspects from Project 3 like Machine Learning. Once we complete Project 3, we’ll be clear with what all can be used from Project 3
- Sentiment Analysis of Posts/Comments

This project will be worth about 10 points, so ~4 deliverables is reasonable (2 pts per deliverable). Aim for about 1.5 weeks' worth of consistent effort. In other words, this project will be very focused but you should try to go "deep" in your analysis.