CSE 511 Term Project Final Document: Analysis of Arizona Businesses using Yelp Dataset

Introduction

We would be doing the analysis of the user reviews of the local businesses on the Yelp website for the state of Arizona. Since the datasets are large, we would be using distributed computing frameworks such as Hadoop and Spark. For the sake of the term project we will use the local installation.

Setup Instructions

We would be using Ubuntu 22.04 LTS in a virtual machine (VM) (details provided in the later section). The VM contains all the necessary files already set up to perform the assignment. Downloading the VM may take a while but once it's setup, you can focus on the project itself without going through the hassle of installing everything from scratch.

Summary of What You Need to Install:

1. **Java Development Kit (JDK)**
   - Required for running Hadoop and Spark.
2. **Hadoop**
   - For distributed storage and processing.
3. **Apache Spark**
   - For distributed data processing, can be integrated with Hadoop.
4. **PySpark**
   - Python API for Spark, allowing you to write Spark applications in Python.
5. **Jupyter Notebook (Optional)**
   - For an interactive development environment, particularly useful when working with PySpark.

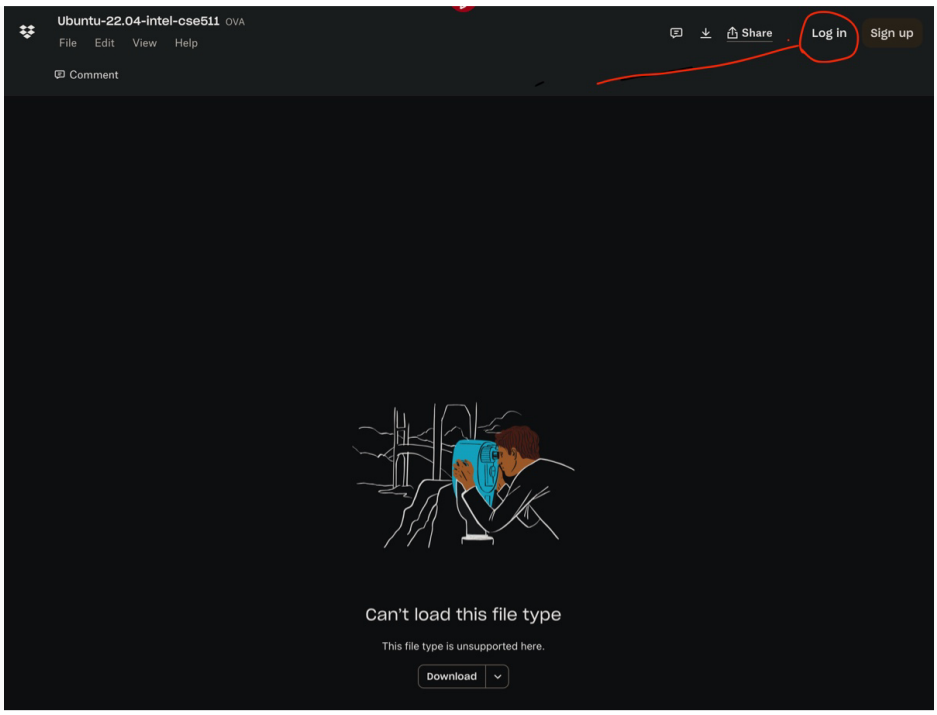
> **For students to focus on the project and not the installation, virtual machines (both for Apple Silicon and Intel processor based systems) have been provided.**

Instructions to install the virtual machine

Steps for Intel-based processors

1. Install VirtualBox (Intel processors), which is a cross-platform compatible Virtual Machine.
2. Download VirtualBox suitable for your system from here.
3. Download the VM file from the Dropbox link here

- You will reach this page



- Click on "Log In" and use your ASU credentials to log into Dropbox.
- You will now be logged in, then click on "Download" button to start the download.

1. Next, open VirtualBox and import the VM file (ending in `.ova` extension) using instructions found here

The instructions are:

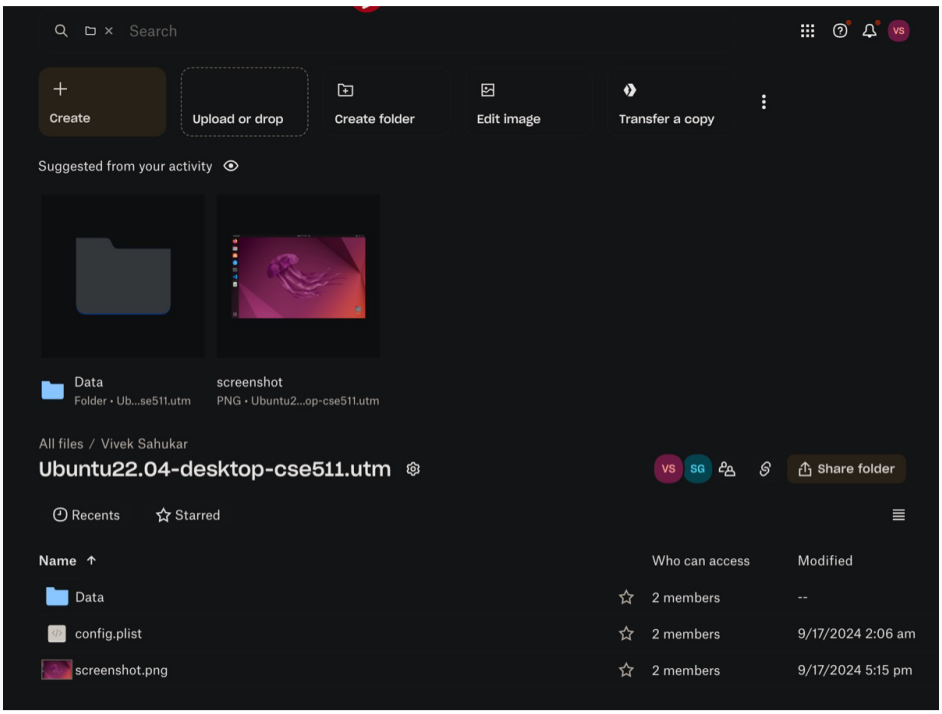
- Click Import Appliance from `File`.
- Choose downloaded VM file
- Click on `next/continue` and stop at the appliance setting page.

5. Change the RAM/processor-cores to meet your PC specifications
   - Adjust the RAM or/and the number of processors to make the VM compatible with your system. VirtualBox may not allow you to start the VM otherwise.
     - Make sure to allocate just enough RAM for the VM to start and spare enough RAM for your computer.
     - You can also change these settings later (after importing the VM) in the settings of theVirtualBox (system settings).
   - Finally, click on `import`.
6. Start the VM (password is **dps**)

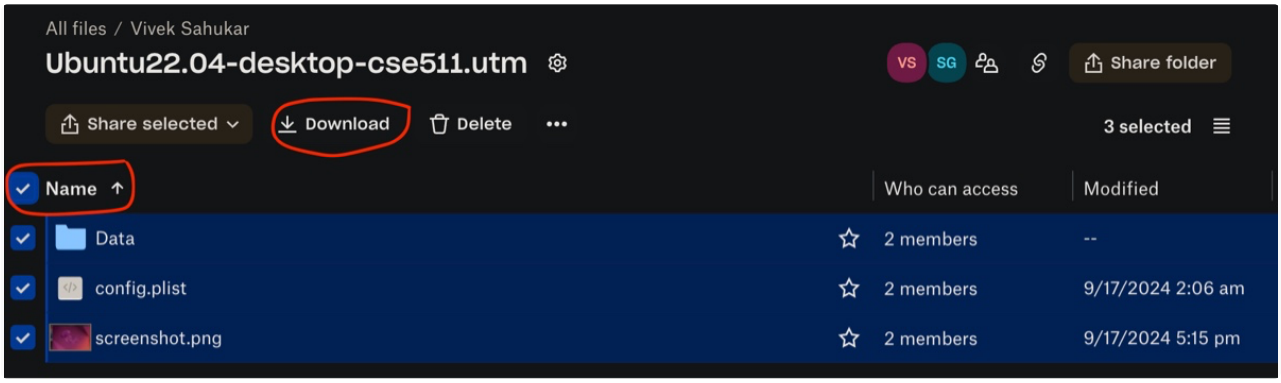
Steps for Apple Silicon processors

1. Install UTM (Apple Silicon processors), which is a Virtual Machine simulator for Apple Silicon processors (M1/M1 Pro/M1 Max/M2/M3).
2. Install UTM from here.
3. Download the VM file from the Dropbox link here.

- You will reach this page

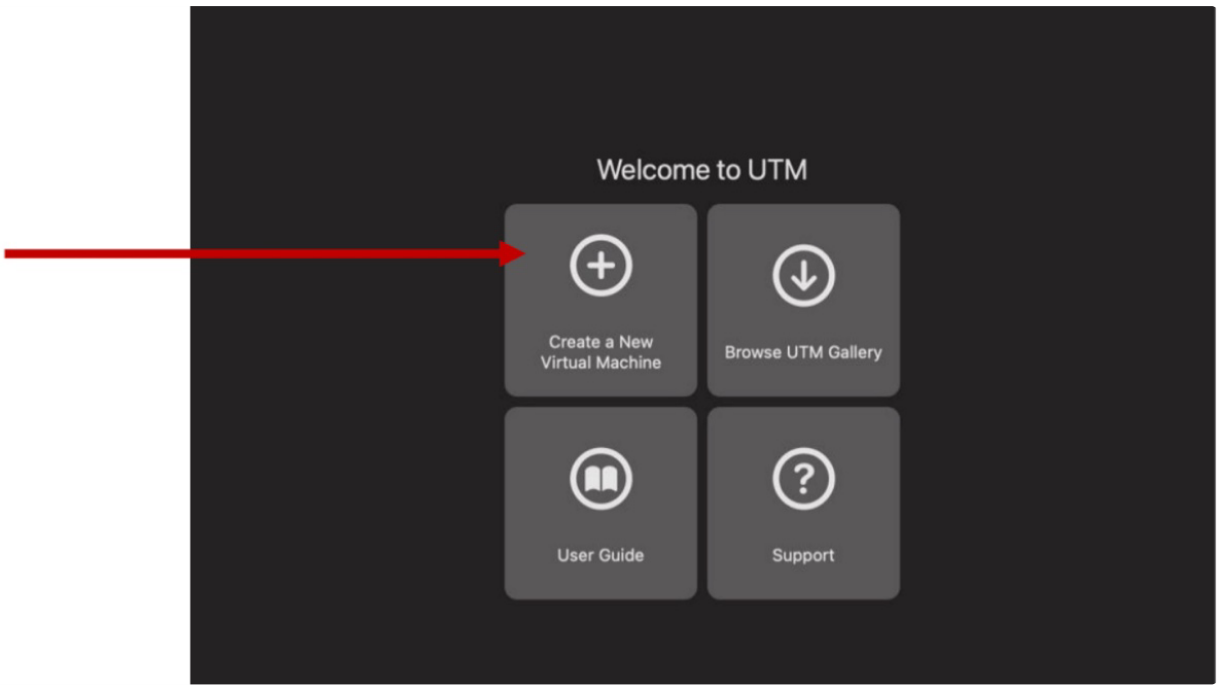


- You will need to login into Dropbox with your ASU credentials.
- Then once you are logged in, click on the checkbox on the left side of the "Name" as shown below to select all the 3 files: `Data`, `config.plist`, `screenshot.png`

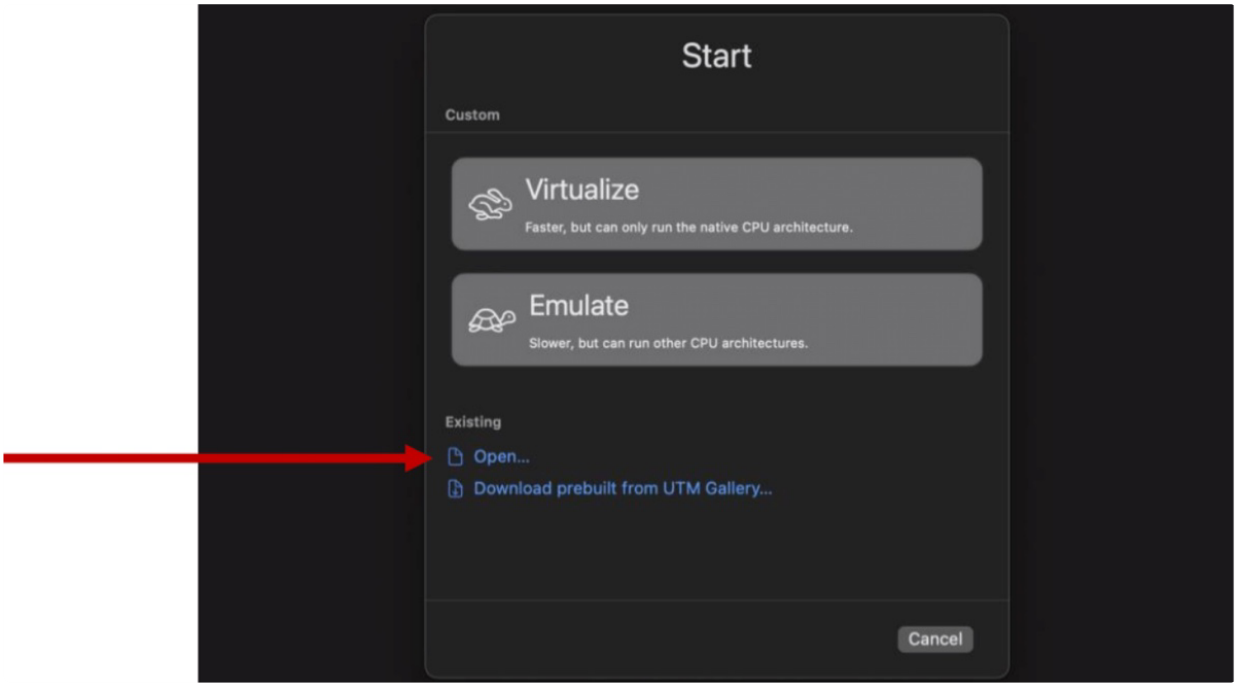


- Then click on "Download" button as shown in the above picture; File ending in .utm.zip will start downloading.
- After download unzip the file and you should get the file with `.utm` extension. This is the virtual machine file that you would use in the Virtual Machine Software.

4. To load the pre-configured VM:
   - Click + "`Create a New Virtual Machine`".



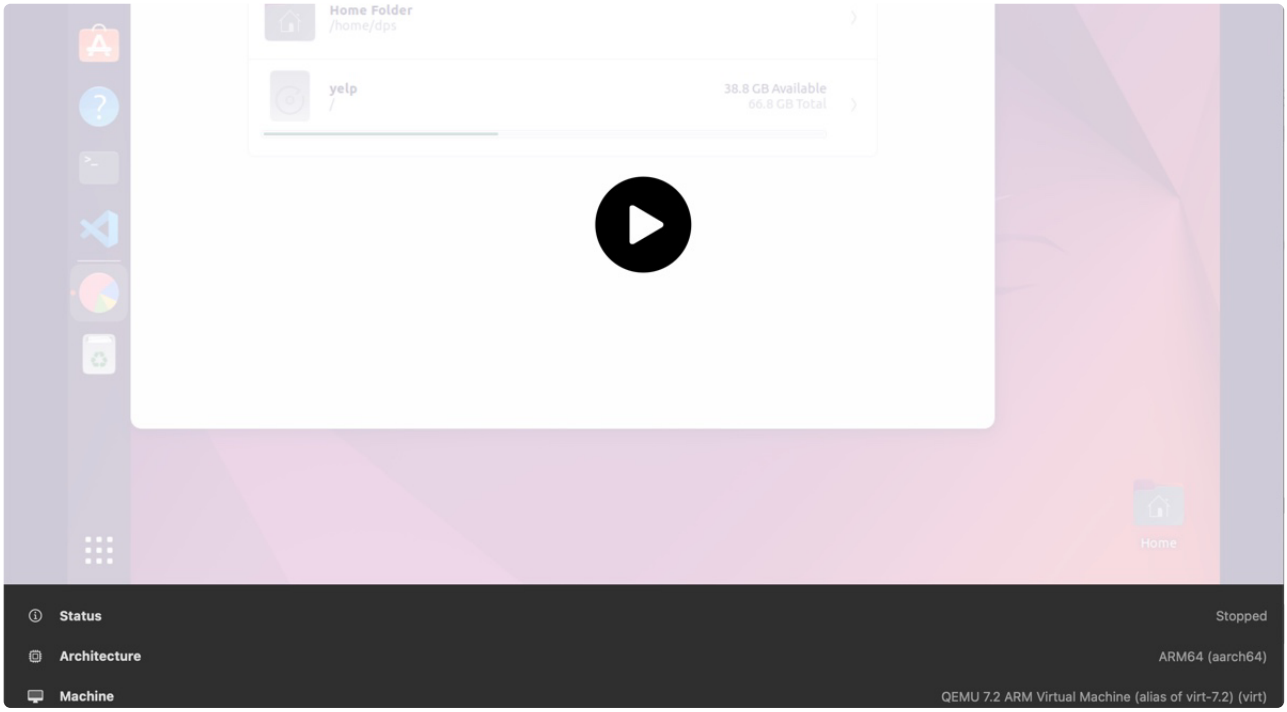
- Click " `Open` " under the existing tab.



- Select the downloaded `.utm` file.

5. Change the RAM/Processor-cores to meet your MAC specifications
   - Adjust the RAM or/and the number of processors to make the VM compatible with your system. UTM may not allow you to start the VM otherwise.
   - Make sure to allocate just enough RAM for the VM to start and spare enough RAM for your computer.
   - To adjust the configurations, right click on the VM and click on " `Edit` ".
6. Start the VM by clicking on the play button

> Ignore "Display Output is not Active" message and wait for some time for VM to start
>
> Password is `dps` whenever required for login or installing software.



Some common commands

Format Hadoop Filesystem:

Format the Hadoop filesystem (only required when running Hadoop for the first time):

```
hdfs namenode -format
```

Start Hadoop:

Start the Hadoop daemons:

```
start-dfs.sh
start-yarn.sh
```

Verify Hadoop Installation

To check that Hadoop is running correctly, open the following URLs in your web browser:

- **HDFS NameNode Web UI:** http://localhost:9870
- **YARN ResourceManager Web UI:** http://localhost:8088

You should see the Hadoop interfaces, indicating that Hadoop is running successfully.

Apache Spark

Spark is a fast and general-purpose cluster computing system.

Verify Spark Installation:

Start the Spark shell to verify the installation:

```
spark-shell
```

PySpark

PySpark is the Python API for Spark, allowing you to write Spark applications in Python.

Verify PySpark Installation:

You can start a PySpark shell by running:

```
pyspark
```

5. Coding Environment: Jupyter Notebook + VS Code (Optional but Recommended for Interactive Development)

Jupyter Notebook provides an interactive environment where you can run PySpark code. VS Code with Python, Jupyter and other required libraries has been provided. Students are free to install other libraries of their choice.

Start a Jupyter Notebook with PySpark integration by running:

```
pyspark
```

> By following these steps, you'll set up a powerful local environment for big data processing with Hadoop, Spark, and PySpark, all within a Jupyter Notebook if desired. Please reach out during the office hours to help with the installation. I have also provided a video which goes over the installation and download of the virtual machines.

Dataset

We would be using the Yelp dataset, which is a dataset of local businesses, reviews and user data released earlier as part of academic research challenge. There are 2 sub-datasets: `yelp_photos` which is a collection of photos of the businesses and `yelp_dataset` which is a collection of 5 json files. `business_id` and `user_id` are the common identifiers across the datasets. We would be using only `yelp_dataset` and not the `yelp_photos` dataset. Details are as follows:

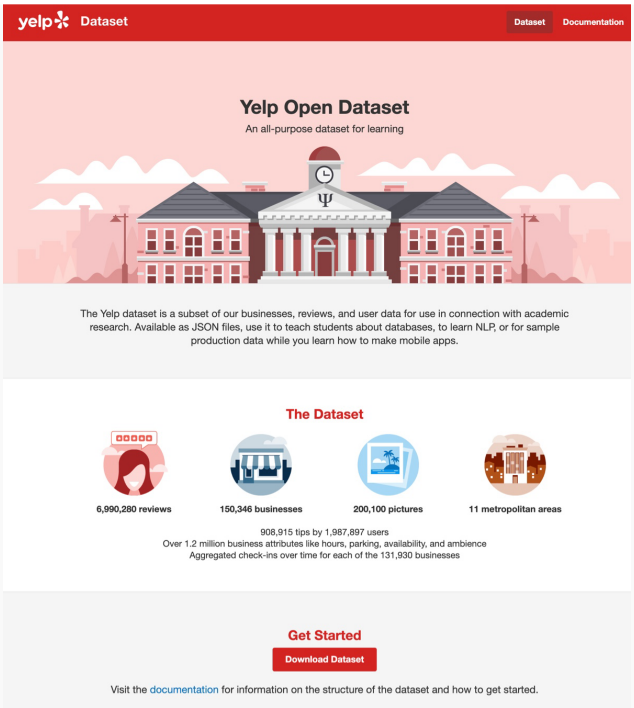
Dataset: `yelp_dataset`

Files	Description	Fields
`yelp_academic_dataset_user.json`	Reviewer information	'user_id', 'compliment_writer', 'cool', 'useful', 'compliment_list', 'fans', 'compliment_more', 'average_stars', 'name', 'friends', 'elite', 'compliment_funny', 'compliment_cute', 'compliment_cool', 'compliment_hot', 'review_count', 'compliment_plain', 'compliment_profile', 'compliment_photos', 'compliment_note', 'funny', 'yelping_since'
`yelp_academic_dataset_tip.json`	Information about tip	'business_id', 'user_id', 'compliment_count', 'date', 'text'
`yelp_academic_dataset_review.json`	Reviews given by the user	'business_id', 'user_id', 'review_id', 'date', 'stars', 'useful', 'funny', 'cool', 'text'
`yelp_academic_dataset_checkin.json`	Check-in information	'business_id', 'date'
`yelp_academic_dataset_business.json`	Restaurant business information with location and city	'business_id', 'name', 'stars', 'is_open', 'address', 'latitude', 'categories', 'state', 'hours', , 'city', 'review_count', 'attributes', 'longitude', 'postal_code'

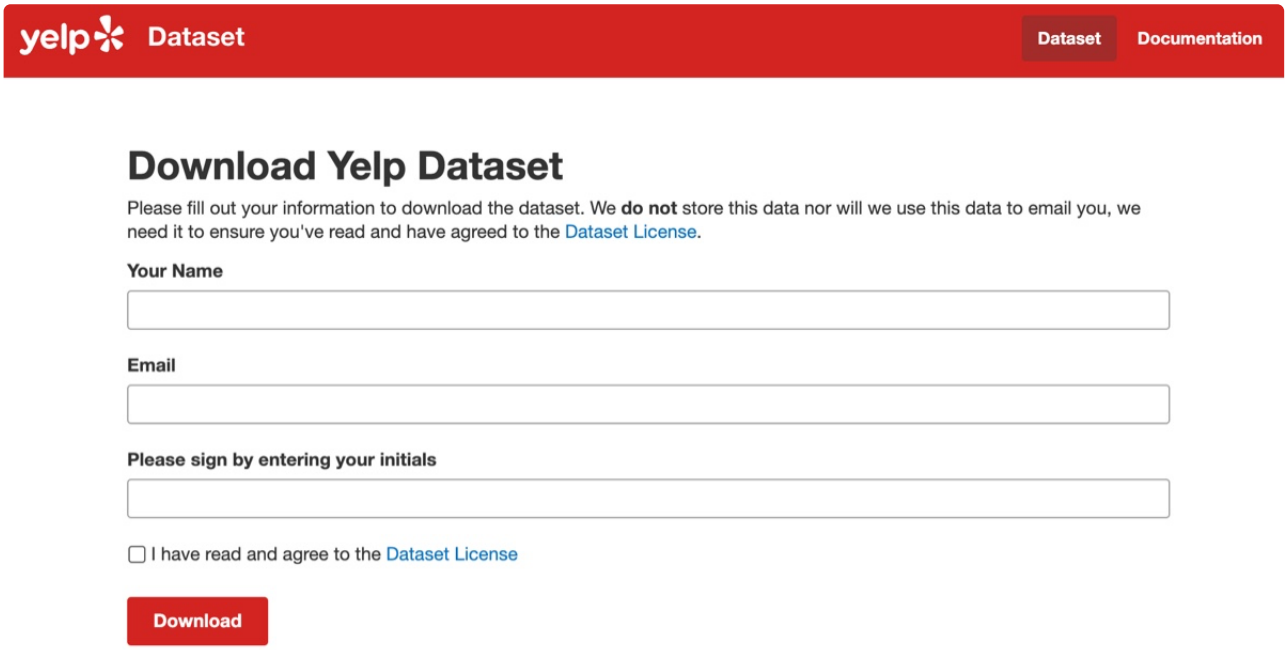
To gain deeper understanding of the dataset structure, please visit here.

Dataset Download (not required, already provided in the Virtual Machines)

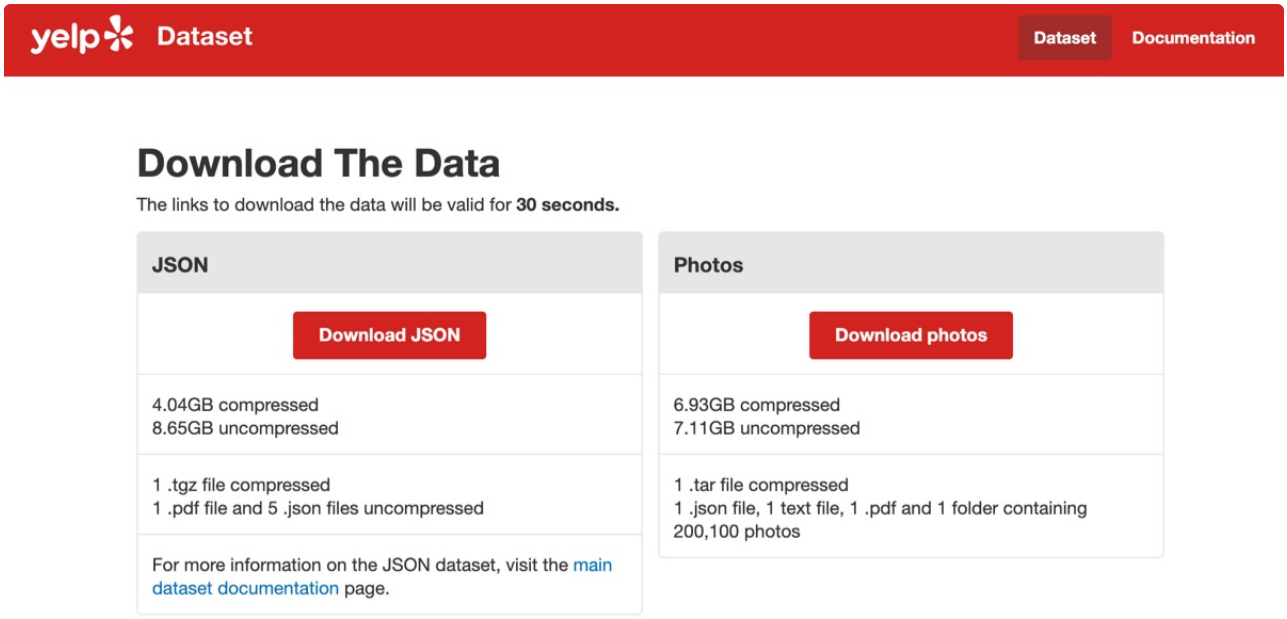
1. To download the Yelp Open Dataset, go here.



2. Click the `Download Dataset` at the end of the page. Fill the required information (name, email, sign and agreeing to terms and conditions). Then click the `Download` button at the end of the page.



3. You will reach this page.



Click on the left `Download JSON` to download the dataset. Since the download is large, please be patient and use a stable internet connection while downloading the dataset. The `Photos` dataset is required only for the Bonus section of the project.

Project

The project consists of two milestones

Milestone 1: Business Level Analysis

Deadline: Nov 17, 2024

Objective: Analyze businesses based on their attributes, reviews, ratings, locations, and other relevant data.

Deliverables

1. Install Virtual Machine on your local machine using the above mentioned guide.
2. Convert `.json` to format (such as `Parquet`) for easy analysis; Filter the Yelp Dataset to include only the data relevant to the businesses in the state of Arizona (AZ).
3. Choose one category of the business.
4. Use Spark SQL Queries to combine the `.json` files and retrieve the required information.
5. Present the analysis in the form of a maximum of 2 pages report (excluding the introduction, code, figures etc.). The example analysis could be focus on a particular category of the business and see how they are doing in the AZ based on the user reviews. See which attributes of the business are more attractive for the customers. Are there any particular locations (zipcode) where the business are doing well.
6. See Jupyter notebook `Project1Milestone1.ipynb` for getting started on the project. The code mentioned is to help the students to do the analysis. The queries from the notebook cannot be copied and presented in the assignment.

Milestone 2: User Level Analysis

Deadline: Nov 24, 2024

Objective: Analyze user behavior, contributions (reviews, tips), and influence within the Yelp community.

Deliverables

1. For the category of the business chosen in AZ, now do a user level analysis. Example study could be analyzing user activity, sentiment analysis of the user reviews, how sentiment varies by user characteristics etc.
2. Present the analysis in the form of a maximum of 2 pages report (excluding the introduction, code, figures etc.)
3. See Jupyter notebook `Project1Milestone2.ipynb` for getting started on the project. The code mentioned is to help the students to do the analysis. The queries from the notebook cannot be copied and presented in the assignment.

Grading Rubric

Final submission would a project report (maximum 4 pages) and the accompanying code showing which queries you ran and how you did the analysis. Recommended way for submitting the code is to do the analysis in a Jupyter notebook and submit it after you have run all the cells.

Grading Rubric

Milestone 1: Business based analysis:

Total 5 queries, out of which 2 queries could be simple using 1 business dataset only, however, remaining 3 queries should be complex combining multiple datasets and query filters.

10 points for each query and 50 points for the report (graphs, figures, analysis) = Total 100 points

Milestone 2: User based analysis:

Total 10 queries, out of which 4 queries could be simple using 1 user dataset only, however, remaining 6 queries should be complex combining multiple datasets and query filters.

5 points for each query and 50 points for the report (graphs, figures, analysis) = Total 100 points

> The queries mentioned should not be random and should show the analysis in a structured format, so that the report portion can be evaluated fairly. Use plot, and figures to show your results. Please do not use the queries as it is from the code in the given Jupyter notebooks. The code and the report will be checked for plagiarism.

Video

Please find the link to the project video here, where I explain how to start the VM, access the data files and run the basic analysis using the template code provided for both the milestones.