CIS4560 Term Project Tutorial

Authors: Ruben Chagollan

Instructor: Jongwook Woo

Date: 12/3/2025

Lab Tutorial

Ruben (rchagol6@calstatela.edu)

Yelp Data Analysis using Spark (your Title)

Objectives

In this hands-on lab, you will learn how to:

- Download a large real-world dataset
- Convert JSON data into CSV using Python
- Upload large data files into HDFS
- Create Hive tables and run distributed SQL queries
- Perform tempo-spatial analysis on Nevada businesses
- Visualize results using charts

Platform Spec

- Hadoop Cluster with HDFS + Hive
- Nodes: 5 total (2 Master, 3 Worker)
- CPU / Memory: Distributed resources across nodes
- OS: Linux (SSH access)

Step 1 Download Yelp Dataset

This step is to get data manually….

Download from official source: https://www.yelp.com/dataset

Extract files locally using Git Bash:

cd ~/Downloads

unzip Yelp-JSON.zip

cd "Yelp JSON"

tar -xvf yelp_dataset.tar

Files extracted include:

- yelp_academic_dataset_business.json
- Review, Checkin, and other datasets

Step 2: Convert JSON to CSV using Python

Create script: convert_business_to_csv.py

```
import json
import csv

INPUT_FILE = "yelp_academic_dataset_business.json"
OUTPUT_FILE = "yelp_business.csv"

FIELDNAMES = [
   "business_id", "name", "address", "city", "state",
   "postal_code", "latitude", "longitude",
   "stars", "review_count", "is_open", "categories"
]

def main():
   total = 0
   with open(INPUT_FILE, "r", encoding="utf-8") as fin, \
      open(OUTPUT_FILE, "w", encoding="utf-8", newline="") as fout:
      writer = csv.DictWriter(fout, fieldnames=FIELDNAMES)
      writer.writeheader()
```

```
        for line in fin:
            data = json.loads(line)
            writer.writerow({col: data.get(col) for col in FIELDNAMES})
            total += 1

    print(f"Done. Wrote {total} rows to {OUTPUT_FILE}")

main()
```

Run Script: python convert_business_to_csv.py

Step 3: Upload Data to Linux and HDFS

This step is to… upload data to Linux and HDFS

scp yelp_business.csv rchagol6@129.153.113.98:~/

mv yelp_business.csv /tmp/

hdfs dfs -mkdir -p /user/rchagol6/yelp

hdfs dfs -put /tmp/yelp_business.csv /user/rchagol6/yelp/

hdfs dfs -ls /user/rchagol6/yelp/

Step 4: Create Hive Table

This step is to…

Launch Hive and select database:

Hive

Use rchagol6:

Create external table: CREATE EXTERNAL TABLE yelp_business (

business_id STRING,

name STRING,

address STRING,

city STRING,

state STRING,

postal_code STRING,

latitude DOUBLE,

longitude DOUBLE,

stars DOUBLE,

review_count INT,

is_open INT,

categories STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/rchagol6/yelp'

TBLPROPERTIES ("skip.header.line.count"="1");

Step 5: Queries for Tempo-Spatial Insights

Nevada Businesses Only:

CREATE TABLE yelp_business_nv AS

SELECT * FROM yelp_business WHERE state = 'NV';

Query 1 – How many businesses in Nevada: SELECT COUNT(*) FROM yelp_business_nv;

Query 2 – Top 10 cities by business count:

SELECT city, COUNT(*)

FROM yelp_business_nv

GROUP BY city

ORDER BY 2 DESC

LIMIT 10;

Query 3 – Top Categories in Nevada:

SELECT categories, COUNT(*)

FROM yelp_business_nv

```
GROUP BY categories

ORDER BY 2 DESC

LIMIT 10;
```

Query – 4 Average Rating by City:

```
SELECT city, ROUND(AVG(stars),2), COUNT(*)

FROM yelp_business_nv

GROUP BY city

HAVING COUNT(*) >= 20

ORDER BY 2 DESC

LIMIT 10;
```

Query – 5 Business Status (Open/Closed):

```
SELECT is_open, COUNT(*)

FROM yelp_business_nv

GROUP BY is_open;
```

Step 6: Visualization

After exporting query results from Hive:

Visualization A – Top Cities in Nevada

1. Copy the Hive query results (city + count)
2. Paste into a small table in PowerPoint
3. Insert → Chart → Bar Chart
4. Use bold labels and increase font size (≥ 28)
5. Add chart title:

   Top Nevada Cities by Number of Businesses

Visualization B – Top Business Categories

1. Copy the query results into a table
2. Insert → Bar Chart
3. Sort from largest to smallest
4. Title the slide:

   Most Common Business Categories in Nevada

Visualization C – Average Rating by City

1. Copy average ratings & city list
2. Insert → Bar Chart
3. Label chart clearly
4. Title:

   Average Yelp Rating by City

Visualization D – Open vs Closed Businesses

1. Use results from Hive is open query
2. Insert → Pie Chart
3. Set green = Open, red/gray = Closed
4. Add % labels
5. Title:

   Business Status — Open vs Closed in Nevada

All charts were created using PowerPoint's built-in chart editor by importing the query results into a table and selecting bar or pie chart types for clear visual communication.

References

1. Yelp Open Dataset: https://www.yelp.com/dataset
2. Apache Hive Documentation: https://hive.apache.org/
3. Hadoop Distributed File System (HDFS Design): https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
4. Python JSON and CSV Libraries – Official Documentation: https://docs.python.org/3/library/json.html https://docs.python.org/3/library/csv.html
5. GitHub Repository – Project Code and Documentation: https://github.com/rchago7/yelp-nevada-analysis