Lab Assignment 2

CIS 493/593 Big Data Processing

Sunnie Chung

Transforming Big Data in JSON (JavaScript Object Notation) Format as Semi-Structured Data

One of the most common machine generated Big data format is JSON format. Create a database for the given Yelp Business data by converting the data in a Semi-Structured JSON format to a structured file format (CSV, TSV) so that the converted files can be directly created as database tables in any server in a table format (any SQL server) for the next phase data processing for Big Data Analytics. You can write this lab in your choice of any language.

For this lab, you can assume that you already requested files and received a JSON file in YelpDataSets.zip and business100.json in LabJSON.zip from the Yelp site or the CIS593 class webpage. You can directly download from the Yelp site at https://www.yelp.com/dataset/challenge

1. Usually it is not a good idea to create one big dirty CSV file from the nested complex JSON format. You can create multiple CSV files in Relational database scheme as needed to convert one JSON file to multiple tables in First Normal Form. In general, there are two common ways to process JSON files for any web applications or data mining processes later.

   1. Convert the given Json file into CSV files.
   2. Create a database in your SQL Server from the converted CSV files so that you can build your Web Service Application to display any restaurant information on your webpage per your customer's request **in a real time**.

Download two zip files (LabJSON.zip and YelpDataSet.zip) on the Lab2 JSON Section on the class webpage. LabJSON.zip has 100 business data (business100.json) and a file (OnebusinessDataFormat_yelp.json) that shows the JSON file format for a business data.

To see the Yelp Business Data Format, open the file "OnebusinessDataFormat_yelp.json" using NotePad++.

Automatic database table creation in a SQL Server in your program using JDBC/ODBC is required. You can use MySQL instead of MS SQL Server to avoid the extra JDBC-ODBC Bridge set up if you are using JDBC in JAVA. You can create a Stored Procedure for Bulk Insert to create a SQL table from a CSV/TSV file separately to create database tables from your outputs.

Extra Credit:

Process All ~200,000 Business data from the Original Data file: Business.json in Yelp site.

Submit the following SQL result and show how long the query takes to return the query result (You can find the elapsed time in the message that your SQL server returns for the query)

Select city, COUNT(review_count) From Business Group by city;

Convert JSON data to relational Tables.

You are to design a correct database scheme in **the First Normal Form** to convert the Semi-structured JSON data to table structures.

Two common ways could be:

I. One Way:

1) Create a big dirty table in CSV in your program and

2) Create multiple tables in correct scheme reading from the big table in a Stored Procedure in a SQL Server.

II. Another Way

1) Design a scheme (multiple CSV file formats) and create multiple CSV files in your program and

2) Create database tables directly from each CSV in a Stored Procedure in a SQL Server.

Design your CSV file formats. There is no one strict CSV file format as a solution. Think about what would be a good format to transform those irregular multi valued data or nested data to a table (or connected multiple tables) so that you can retrieve them from a database easily and efficiently without losing information. As long as it is transformed correctly without losing data, any form would be good.

Invalid JSON format handling:

If there is any invalid data format is found in the input file, you can change it to the correct JSON format. For example, $$ in the "Price Range" key value pair in your input json file, the value $$ is not in quotes and this will cause to fail. If , is missing between objects, add it.

Suggested Solution:

Replace $$ with 2 (meaning the price level is 2 in scale 1 - 5) in the file and try to parse the corrected file in your program.

Submit the followings:

- On Blackboard:

1) Lab Report (in doc file) that shows the followings:

> > 1) Your platform set up procedure,
> > 2) The executions to generate the output file in a structured any text file format (in CSV, TSV),
> > 4) Screen captures of your SQL tables created from the converted files, and
> > 5) All your Source codes/scripts.

> 2) Submit Lab2 in one zip file that includes your report file in doc, all your codes/scripts, and your output files.

> - In Class

> 3) Submit a Printout of your report (don’t printout all the output text)