McGill | School of Continuing Studies | École d'éducation permanente

Assessment
Data at Scale
YCBS-257

Open Project Assignment

5% due on 23:59 Tuesday Mar 12 2019

Meetup Data Analysis

Introduction

Meetup is an online service used to create groups that host local in-person events. As of 2017, there are about 35 million Meetup users. Each user can be a member of multiple groups or RSVP for any number of events. Users are usually using the website to find friends, share a hobby, or for professional networking. Meetup users do not have "followers" or other direct connections with each other like on other social media sites.

Meetup users self-organize into groups. As of 2017, there are about 225,000 Meetup groups in 180 countries. Each group has a different topic, size, and rules. Groups are associated with one of 30+ categories and any number of more than 18,000 tags that identify the group's theme. The most popular categories are "adventure and outdoor activities, career and business, and parents and family." Most events are on a structured schedule each week or month at a local venue, typically on evenings or weekends. (*source – Wikipedia*)

Meetup API

The Meetup API provides simple RESTful HTTP and streaming interfaces for exploring and interacting Meetup platform from your own apps.

The API is a set of core methods and a common request format. These are combined to form a URL that returns the information you want.

Online API documentation:

https://www.meetup.com

https://www.meetup.com/fr-FR/meetup_api/docs/stream/2/rsvps/?uri=%2Fmeetup_api%2Fdocs%2Fstream%2F2%2Frsvps%2F

Provided files:

- Flume agent configuration file able to connect and collect meetup streaming data
- HiveQL script to create a non-managed table to read ingested data with the Flume agent.

Meetup Data stream ingestion

To ingest meetup data in real time we had configured a Flume agent which connect to the data source and grab all the events in a **json** format. All these events are stored on HDFS.

In order to begin meetup stream data ingestion you need to import the configuration file into your Cloudera VM and run your Flume agent.

Meetup Json Data Reading

To read meetup events stored on HDFS, we created a Hive table, defining simple and complex data type to handle nested json documents. A complex data type represents multiple fields of a single item. Frequently used as the element type of an `STRUCT <>` ,`ARRAY` or the `VALUE` part of a `MAP`.

A `STRUCT` is similar conceptually to a table row: it contains a fixed number of named fields, each with a predefined type.

The `STRUCT` type is straightforward to reference within a query. You do not need to include the `STRUCT` column in a join clause or give it a table alias, as is required for the ARRAY and MAP types. You refer to the individual fields using dot notation, such as `struct_column_name.field_name`, without any pseudo column such as ITEM or VALUE.

In order to read meetup stream json file you need to run the `meetup.hql` script which will create the meetup table (a non-manageable table).

ToDo Work 5%

This exercise is part of the course's participation and it is an open assignment project. Instead of continuing with a project where detailed requirements were predefined by the teaching staff, this project is completely open.

You are the next Data Scientist. Make any kind of data analysis you could decide on the meetup dataset. Follow your inspiration to analyze the data and extract a **Value** from it.

To concretize your ideas and get insights from the data you are invited to use any of: (implementing is a bonus)

- The Hadoop / Spark ecosystem tools.
- Programing language used commonly by data scientists (Java, Python, Scala, R …)
- Visualizations software

What to submit:

One MS Word document <your name> - W2019OP.docx

The file should contain:

- At least **`two analysis ideas`**
- At least **`one idea`** where `machine learning` can be applied

✓ Comments can be submitted in French or in English in the MS Word document.