**HU Extension School  E-90 Cloud Computing**

# Final Project Topic: Google BigQuery

**Hanjiao Qiu**

**Problem:**

1) Query massive dataset using Google BigQuery. Search and sort the popular CDN (Content Delivery Network) providers, top sites to have content security policy to prevent XSS and explore other web behaviors. Understand the web performance from the HTTP Archive data since the last two months. 2) Do real-time system log analysis, execute queries and visualize the log data through interacting with the BigQuery in Google sheets. This problem will examine the super fast performance, high reliability and security and low-cost in big data analysis of Google BigQuery.

**Description:**

Google BigQuery provides three ways to access it: Web UI, bq command-line tool and BigQuery REST API using a variety of client libraries such as Java, Python. It also enables a variety of third-party tools to interact with it, such as visualizing or loading the data. This project details the three ways to analyze the large scale HTTP Archive data using BigQuery. It also demonstrates the data loading and visualization by interacting with Fluentd and Google sheets.

**Benefits:**

1) Analytics as a service, use SQL-style query and can be accessed by RESTful API.
2) Reasonably fast, process multi-terabyte datasets in seconds.
3) Scalable, reliable and secure.

**Dataset:**

HTTP Archive data (http://httparchive.org/), also available through BigQury. (200+ GB, csv)
Google BigQuery public data sample. (50+ GB, csv)
Real-time log generated from Google Compute Engine Instance (78 MB, csv)

**Operating System:**

Mac OS, Ubuntu 14.01

**Software:**

Google Cloud SDK, Google BigQuery API Client Library for Java, Java version 1.8.0.

**Overview of steps:**

1) Use the BigQuery Web UI to analyze the data.
2) Use the "bq Command-Line Tool" to analyze the data.
   2.1. Google Cloud SDK installation.
   2.2. Create table, schema, load data and execute queries through command-line tool.
3) Use the "BigQuery API" to analyze the data.
   3.1. Download the Google BigQuery API Client Library for Java.
   3.2. Import data from local file and Google cloud storage, query data and export results in Java.
4) Real-time logs analysis using Fluentd and BigQuery.
   4.1. Create Google Compute Engine instance and load data using Fluentd.
   4.2. Use BigQuery and visualize data in Google sheet.