**MLAB ANALYSIS USING HIVE AND GOOGLE BIGQUERY**

**Anthony Acosta, Joanne Canizalez,**

**Gustavo Carrillo, Marla Martinez, Sean Lee**

**Department of Information Systems**

**California State University Los Angeles**

|  |  |
| --- | --- |
| **Abstract:** To identify the number of NDT tests performed daily since M-Lab epoch by distinct users, as well as the number of distinct IP addresses that have run tests using both NDT and NPAD tools within the United States and across the globe. This report details the data wrangling and analysis processed via IBM BigInsights to capture visualizations in M-Lab data collections.   1. **Introduction**   Measurement Lab (M-Lab) generates a collection of open-internet performance data sets with statistical value in the context of connectivity priority and performance comparison.  The dataset provided is a collection of on-net and off-net broadband-speed test results using both NDT and NPAD tools, ranging from 2009 and last updated in 2015.  The approach is to summarize the problems related to the term “Last Mile” but more importantly, the types of locations where end-users have performed connectivity tests. In this way, a more precise assessment of bottleneck issues can be determined.  The last mile is typically the bottleneck in a communications network and its bandwidth capacity limits the bandwidth that can actually be delivered to the customer. This is because the topology of most retail communication networks have a tree-like structure, with a few “trunks” representing the main high-capacity lines which branch out to serve the last mile “leaves”; which is subject of constant research to deliver higher speeds to end-users.  **5. Lab Procedure**  The dataset was assorted with IBM BigInsights and analyzed with hive and Google BigQuery.  Data was manipulated through the use of Hive commands for high level of abstraction within the dataset. HiveQL supports custom MapReduce scripts to be plugged into queries. Google BigQuery made the data available at a faster rate, while making it easier to read and understand.  With Hive, tables were created to populate the abstraction results generated by the data manipulation process. Such results estimated the number of NDT tests performed daily since M-Lab epoch by distinct users, as well as the number of distinct IP addresses that have run tests using both NDT and NPAD.  The results yield the greatest concentration of tests performed populated by cities within the United States and globally.  **6. Visualization of Data**  The data extracted of 100 records where the country equals “United States” is downloaded as a csv file and query and analysis.  An Excel spreadsheet was populated with the csv file to make use of the 3D map feature under the Insert tab. With the aid of Excel Power View, the file uploaded into the spreadsheet can be seen to populate across the United States for a geographical layout of the results.  The results from this exercise visualizes the geo-logged cities in the United States reported by M-Lab (NDT tests). Another representation of our visualization can be displayed via Tableau (using the .csv file from Google BigQuery) as it shows the number of NDT tests performed from the years 2009 to 2017. In a broader spectrum, the dataset can be expanded through BigQuery to show the records of the performed tests across the world, rather than a single continent.  With exception of Midwestern United States, the southwestern and northeastern states were found to be of higher concentration of tests performed; possibly indicating bottleneck issues.  Additionally, more data is downloaded from Google BigQuery. In this case, the data extracted was 100 records where the country is not null; allowing for a global scope of the test results as well.  The result displayed a higher concentration of tests performed in Eastern/Southern Europe. | **2. Hive vs Google BigQuery**  Issues were encountered during data-upload and manipulation with Hive. Given the size of the dataset (17.5 GB) and the limited memory available in personal computers; discrepancies were encountered when results were populated.  Google BigQuery however, handled the dataset fairly well during manipulation and assortment of results. This could be due to the dataset coming from google itself and their seemingly unlimited computation power.  The simultaneous usage of Hive and Google BigQuery can be justified as BigQuery was capable of handling the large dataset at faster rates compared to a Hive shell. Google’s services also allowed us to run and generate queries that would provide the detailed information with options to download the dataset as updated files. Overall, the performance of BigQuery allowed us to improve processing times while Hive was utilized for smaller sized queries making data management very flexible.  **3. Network Path and Application Diagnosis Tool**  The NPAD tool is capable of diagnosing a range of common performance issues affecting the last network mile--a popular term used by telecommunications, cable television, and internet industries to refer to the final stretch of telecommunication network delivery services--and end-users’ systems.1  During the course of bulk data transfers, NPAD gathers the performance regulating mechanisms and other detailed statistics for the NPAD server to collect test results along with other network transfer protocols (i.e. IP addresses, download/upload speeds, packet headers, and TCP variables).2  **4. Network Diagnostic Tool**  NDT (Network Diagnostic Tool) provides a sophisticated speed and diagnostic test suitable for both the novice and the network researcher. NDT reports upload and download speeds and attempts to determine what problems limit speeds. It also provides detailed diagnostic reporting on what it finds. While the diagnostic messages are most useful for expert users, they can also help novice users by allowing them to provide detailed trouble reports to their network administrator..3  **Dataset:** M-Lab Dataset: Network Path and Application Diagnosis Tool (NPAD):  Size: 17.5GB  Source: Matt Mathis, with Measurement Lab  Created: December 8, 2009  Last Updated: November 24, 2015  **References:**  **[1]**<https://aws.amazon.com/datasets/m-lab-dataset-network-path-and-application-diagnosis-tool-npad/?_encoding=UTF8&jiveRedirect=1>  [2].Fulldatasource: https://console.cloud.google.com/storage/browser/m-lab/experimental-hive-csv?pli=1  [3].AmbariURL: https://bi-hadoop-prod-4225.bi.services.us-south.bluemix.net:9443/#/main/views/FILES/1.0.0/FilesBrowser  [4].References: https://aws.amazon.com/datasets/m-lab-dataset-network-path-and-application-diagnosis-tool-npad/?\_encoding=UTF8&jiveRedirect=1  https://www.measurementlab.net/data/docs/bq/quickstart/    1Matt Mathis. (2009). Network Path and Application Diagnosis Tool (guide). 1-1  2Matt Mathis. (2009). Network Path and Application Diagnosis Tool (guide). 1-2  3Dodge, T. (2010). Network Diagnostic Tool. 23(2). 2-3 |