Google BigQuery/Dataflow

Group 1

Rubric Deliverable 1

Objective: Learn about a Google Service or AWS service we didn’t cover in our class and build your project around this service.

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. As large-scale data is gathered and processed, more tools are required to manage and interpret this data. Google Cloud Services, housing many useful tools regarding infrastructure of the cloud also provides powerful tools for large scale data management. A useful tool that our group wanted to explore was Google Big Query as well as Google Dataflow. Big Query is a fully managed, serverless data warehouse that enables scalable analysis over petabytes of data. It is a Platform as a Service that supports querying using ANSI SQL, making it easy to pick up and use for new SQL users and veterans alike. Being as dataflow management was not necessarily something we covered too heavily in class, we felt as though it was the perfect challenge for us as students and provided us with useful tools and knowledge we will carry beyond this course. Regarding the primary use of Big Query: as described, it is a tool to manage data on both a small scale and large scale alike, for understanding how Big Query works we developed a small SQL command that filtered a database of names in the US by NAME and GENDER, as well as total populum with selected name. (See demo presentation of Google Big Query on attached GitHub). With this tool, we are now able to sift through any dataset we desire in SQL fashion all through Google Cloud Platform itself. This now opens the opportunity to use Google Dataflow in conjunction with our Google Big Query tool. Regarding Google Dataflow, Google Cloud Dataflow is a cloud-based data processing service for both batch and real-time data streaming applications. It enables developers to set up processing pipelines for integrating, preparing and analyzing large data sets, such as those found in Web analytics or big data analytics applications. Utilizing Google Dataflow in unity with Google Big Query we can now analyze large scale databases, from either batch or real time and filter it to meet whatever standards we desire as users. A unique use and demonstration of this would be to filter tweets on Twitter by certain phrases or terms, and this is very much possible with Google Dataflow. Throughout our semester we covered many tools in AWS and GCP alike. A difficult decision in deciding what platform or structure to use boiled down to what platform provided the tools necessary to accomplish what we wanted. As a group we wanted to understand dataflow and explore the different ways to manage it, Google Cloud Platform provided all of these tools with added help pages and manuals for new users which really appealed to us as students.