Oregon State University

CS 461

Fall 2016

Prototype Big Data Archive in a Public Cloud

Author:
Zhi Jiang
Isaac T Chan
Zhaoheng Wang

Instructor:
D. Kevin McGrath

Client:
David Barber

Abstract

OSU campuses generate data constantly from multiples sources, including computer labs, wireless usage, student devices, and many others. This data can effectively represent all kinds of behaviors of students for information technology. Analysis can be run to determine how to prioritize funding in order to use OSU resources more effectively. But the most important problem currently is that the data is very difficult to manage from multiple sources and impossible to analyze. The data is not stored in a unified format and in the same locations, meaning it is inaccessible and useful information is not able to be extracted. Our goal for this project is to unify and organize the data into the consistent cloud platform of Amazon Web Services, which additionally provides utilities to manage and analyze. To achieve this, we plan to have a working prototype at the Engineering Expo that demonstrates the value of analyzing OSU big data and how the cost-to-value of our Amazon cloud solution compares to locally-hosted hardware. Our prototype will be used to maximize the value of OSU big data and eventually it can be scaled to analyze all the data that OSU collects.

1 Problem definition

In age of the Internet, the data has become a valuable resource for everyone, because the significance of the data is to express behaviors and performances of users. On campus, students are creating many types of data through multiples source including laptops, phones, web applications and others. Thus the problem OSU Information Services has to face is diversity of data, which impacts the methods that which they effectively manage the data, because the data formats are more diverse, including text, audio, log files, clickstream and others. It is difficult to manage and utilize the huge amount of data resources, because we do not currently have any appropriate tools which are able to complete tasks for data such as integration and analysis. Further, the cause of solving this problem is that OSU Information Services would like to achieve specific improvement and optimization for users demands according to results of data analysis.

2 Problem solution

Our proposed solution is an implementation to complete multiple operations for the data on the cloud platform of Amazon Web Services. There are specific three steps in this entire process. First of all, we will focus on discovery and collection of data. Ensuring accuracy of results largely depends on quality of the data, therefore requirements for data we would like to discover and collect should be that they are multifarious and the capacity of the data is large. We will provide efficient method of data ingestion to institute sweeping data collection for all sources. Further, we need to ensure the consistency of data format when we collect them, so format conversion will play an important role in this process.

Secondly, we would like to utilize NoSQL database to manage and organize the data as main tool. The reason we choose this kind of database is that it can build non-relational model for the data because different types of data will be store in the database and their relationship are probably indistinct. On the other hand, flexibility of NoSQL database means it can easily adapt to the new data types, and is not be affected by changing of content structure from the third party data provider. These advantages can greatly deal with current condition of data resources. As for implementation of NoSQL database, actually there are many options for us to implement database such as DynamoDB, SimpleDB and MongoDB. The disadvantage of SimpleDB is the volume of each domain is limited, and which is only 10GB, thus it is not an appropriate choice for managing big data. One of characteristics of SimpleDB is it does not require the user to specify the primary key and to create an index, because it will be the default to create an index for all attributes. But in fact this process of creating index for all attributes will spend a lot of time. On the other side, MongoDB could be another option for implement NoSQL database. However, The negative influence of MongoDB is it will occupy much space, because once the space is not enough, it will apply a larger desk space increasing from 64MB to 2GB(the largest size supported for single file). On the Amazon platform, DynamoDB is the best choice to implement NoSQL database. It is able to complete storing and accessing data at a low-latency response time, because the DynamoDB service is built on Solid State Drives (SSDs), and its construction method is designed to ensure that the performance is stable and the delay is reduced. Different from many non-relational databases, DynamoDB allows developers to use strongly consistent to ensure the latest values will be always read, thus this characteristic makes the development more convenient.

In the last step, we will show our outcomes according to utilize and analysis the data we have stored in database, thus this implementation is able to have analysis techniques such as machine learning. The main task of analysis techniques is to mine valuable data which can really represent users behavior and performance. At the Expo, we hope to have a working demo of the project, where we can demonstrate the processes of ingestion, conversion, and storage. We also hope to be able to show the worth of the solution with a drawn conclusion from the newly organized data.

3 Performance metrics

There are three objectives to fulfill for this project. First being a method of ingestion and management of sample data into Amazons cloud platform. It is clear when this is complete, when the data is loaded from multiple sources into a cloud database. However the method of determining the quality of our ingestion solution is tied closely to the second objective: rudimentary analysis, reporting, and visualization. Once we are able to visualize, extract, and analyze data, we will know if we ingested and stored the data in a functional way. At the completion of our project, and the third objective, we will know if our implementation solves the clients needs by providing an unbiased and true comparison between the Amazon cloud solution and locally-hosted hardware. On the other hand, the cost is also one of performances metrics we need to consider. The client will ask an IT staff to estimate the cost of implementing local solution, and then we will provide our information about how we implement in AWS. After we analyze the cost and value ratios for Amazon platform, we will compare it with the estimate of the cost of a local implementation to determine which is superior. With the added utilities of analysis and data management, our cloud solution should provide additional value, although it remains to be seen which is a more scalable and efficient solution. At the Expo, we plan to demonstrate the entire process from start to finish, and provide an example of a conclusion drawn from analysis of the data, showing how the prototype big data archive will benefit OSU if the project is adopted by OSU IT and scaled to analyze more data.

______________________________ ______________________________

Client Date

____

Developer 1

____

Developer 2

____

Developer 3