Oregon State University

CS 461

Fall 2016

Prototype Big Data Archive in a Public Cloud

Developer:
Zhi Jiang
Isaac T Chan
Zhaoheng Wang

Instructor:
D. Kevin McGrath
Kirsten Winters

Client:
David Barber

Abstract

OSU campuses generate data constantly from multiples sources, including computer labs, wireless usage, student devices, and many others. This quantity of data, also known as big data, can effectively represent all kinds of behaviors of students for information technology. For example, analysis can be run to determine common student behaviors in order to allocate OSU resources more effectively. Currently, the data is very difficult to manage because it is collected from multiple sources and is impossible to analyze. The data is neither stored in the same formats nor in the same locations, meaning it is inaccessible and useful information is unable to be extracted.Our goal for this project is to unify and organize the data onto 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 allow OSU big data to be analyzed and eventually it can be scaled to analyze all the data that OSU collects.

1 Problem definition

In age of the Internet, data has become a valuable resource for everyone, because the data provides insight into user behaviors and performances. There are three main problems the OSU Information Services need to solve when they process these data. Firstly, they need to ensure consistency of all kinds of data and ingest information from them. For instance, there are two kinds of data which are log files and clickstream data, and the difference of format among data will cause they need to ingested by different tools. Secondly, these data need to be stored in database, because the purpose of building database is effectively management these formatted data.For example, staffs of OSU Information Services can get average of GPA quickly by using corresponding database query. Last problem is these data must be analyzed and visualized. The results of data analysis can help OSU Information Services understand behaviors of user. For example, if a result of log file analysis reflects some websites always appear error when students click them, OSU Information Services can directly and quickly find these websites and then fix errors. Thus, benefit of the product is to provide a complete workflow built on cloud platform for solving these three problems.

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 three specific steps in this entire process: to collect formatted and cleaned data through format conversion and ingestion; to manage and organize the data in database; and to analyze data and visualize results of analysis results.

First step is we will focus on collection of formatted and cleaned data. The format conversion is primary we need to do, because it is useful for uniformly processing data later. And then we will provide some efficient methods of data ingestion to parse important information from these formatted data since accuracy of analysis results largely depend on quality of these information.

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 a non-relational model for the data because different types of data will be stored in the database and their relationship are likely indistinct. On the other hand, flexibility of NoSQL databases 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. As for implementation of NoSQL database, we believe that DynamoDB is the best choice. DynamoDB can complete storing and accessing data at a low-latency response time, because the its 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. We still have other alternatives to implement NoSQL database such as SimpleDB and MongoDB. For example, the SimpleDB 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. MonogoDB will apply a larger disk space when it needs more space. We will compare and discuss these different ways to implement NoSQL in the future.

In the last step, we will do implementation to analysis and visualize data. The main task of analysis techniques is to find valuable data which can really represent users behaviors and performance. On the other hand, the goal of visualization is to clearly present analysis results to users for understandings of their behaviors.

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 the first objective 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 a cost-value comparison between the Amazon cloud solution and locally-hosted hardware. The client will ask 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 Engineering Expo in June 2017, 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