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.

# Contents

# 1 Introduction

## 1.1 Purpose

The purpose of this requirements document is to address specification for our product. We will clearly explain functionality, external interface, performance, attributes and design constraints of our product while remaining at a high-level to avoid technical details.

The intended audience of this document is the client, our development group, and assessors of our work. Our group and client must share a mutual understanding of the project and all the details it entails, to confirm that our final product meets and matches all expectations. Assessors of our work may reference this document to compare it to our final product, again to ensure our final product matches requirements.

## 1.2 Scope

The name of this product is "Prototype Big Data Archive in a Public Cloud", and will be used to to implement all operations about data from multiple sources and ensure data can be unified and organized onto the consistent cloud platform. There is a functionality of product for retrieving data which makes analysis easier. Besides, product can mine value of data accordingly to integrate them because the data can reflect behaviors of students and staffs. A huge amount of data is generated when students and staffs use various information technologies, such as printers and computers. The product contains several functionality such as ingest, store, manage, some basic reporting and basic analysis. These functionality help OSU Information Services staff directly and expediently understand behaviors of students and staffs. Further, the product will provide a database which can systematically deal with the data. Therefore, one of benefits is our project make the data clearly in the database so OSU Information Services can manage and analyze more easily.

## 1.3 Definitions, acronyms, and abbreviations

| Term | Definition |
|---|---|
| User | The person who interact with our project |
| Big Data | A large and complex data set which is hard to deal with by traditional data processing applications |
| Prototype | A sample of product create for testing the concept |
| Cloud platform | The cloud platform is using the cloud computing technology. And the Platform as a service (PaaS) offered by cloud platform will provide a development environment for the product including OS, compiling and execution environment of programming language. The Infrastructure as a service(IaaS) will abstract the details of infrastructure such as physical computing resources, security.Thus, the user don't need to worry about managing cloud infrastructure |
| AWS | Amazon web service, a Platform as a service(PaaS) offered by Amazon |
| DB | Database |
| Nosql | non-relational database |

## 1.4 Overiew

Following this introduction is a section describing the product as a whole. This includes perspective, function, user characteristics, constraints, and assumptions/dependencies. The purpose of the

section is to provide in-depth details on requirements of the product. Finally, this document is concluded with a section on specific requirements.

# 2 Overall Description

## 2.1 product perspective

The entire process will consist of three steps: ingesting data, managing data and analysing data. The goal of ingesting data is to collect the data from multiple sources. Some of them are created by information technologies on campus. These data will be a limited subset of data which generated by IT systems of OSU Information Services.After ingesting data, the database will be used to manage and store the data. Finally, data will be visualized after they are analyzed by specific technical.

The types of data include log file, clickstream data, and other forms of IT data, thus in the first step, the product will ensure that all of these different types of data can be completely stored in database. In second stage the database is the most important tool to complete tasks because rationality and correctness of database will directly affect performance of whole product. The database will provide enough space to hold vast amount of data, and it is able to build index and process data in batch. Furthermore, the database will implement some basic operations including inserting, deleting, updating and searching for data. Eventually, analysis technical will communicate to database, and then data will be accessed quickly by analysis technical while all of valuable data will be used to represent behaviors of students and staffs.

## 2.2 production function

In our application, the user could search the student information what they want in the database. After that, they could manage it in database. Then, they could load the data from the database into the data analyze tool such as tableau and do some analysis in order to understand the student behaviors.

## 2.3 User Characteristics

There is only one type of user that interact with our product: data analyzer.
Staff who analyze the data can search for the information they want and manage it in the database. They also could extract the data from database into a data analysis tool and do the analysis.

## 2.4 Constraints

In implementation, there are minimal constraints placed on design. There are, however, certain aspects of development that we must keep in mind, including resource limits, implementation language, and development environment. The main aspect being that development on the AWS platform will cost our client money. AWS charges for computing time, temporary data storage, and database usage. Though it is unlikely that we will run into any upper budget limit, we need to have in mind the charges in order to eliminate resource waste. Amazon also offers a budget tool for our client to track our budget usage. It can also provide a forecast for future budget, so we can know in advance if we are approaching any limit. In terms of implementation language, AWS is flexible in the choice of coding language; we are free to choose any programming language that they support. Finally, our development will be done from our personal machines; as a cloud computing service, Amazon's resources can be freely accessed by us wherever on any operating system.

## 2.5 Assumptions and Dependencies

Because the nature of the project is to produce a prototype of a cloud-based big data archive, there is an assumption that our final product will be a competitive solution for OSU's data analytics, compared to a locally-hosted solution. From this base-assumption, there are several dependencies to keep in mind, including scalability, future maintenance, and security.

Our development will be done mostly with small test-sets of data, with student information anonymized before given to us. Therefore, a large concern is scalability - our solution must be able to work with potentially enormous data sets in a reasonable time. Database retrieval time is critical; the solution is almost useless if it takes an inappropriately long time to extract data from our implemented database. Additionally, database manipulation time, such as "joins" is an important factor when considering adopting the final product. Other metrics will be determined as we begin design implementation.

Future maintenance is another concern. Our implementation will be maintained by others of varying levels of expertise. Therefore, our product must have readable code and abundant, clear documentation.

Finally, security is an important attribute. As mentioned before, our development will be using test-sets of user anonymized data, which protects us from liability and knowledge of specific users. Also, databases do have vulnerabilities, like any other software. We will consider prevention of malicious interactions, such as injection attacks. On the same note, if for any reason the database were to go down, we may want to have implemented a backup database. This will depend on the transfer time and quantity of data to store; if it is easily uploaded there is no reason to require a backup database.

Client

Date

Developer 1

Developer 2

Developer 3