Prototype Big Data Archive in a Public Cloud

Group 56: Pathfinder of Big Data

Zhi Jiang, Isaac T Chan, Zhaoheng Wang

CS 463: Senior Capstone Spring 2017

Oregon State University

Abstract

This report presents a thorough review of our 2017 capstone project. Included are previous documents: requirements, technology review, and design. Weekly blogs are also included to track our group progress by weekly basis. Finally, there is a copy of the poster that we presented at the Engineering Expo, project documentation, and personal statements from each member of our group. The purpose of this report is to coalesce all work over the year into a singular document.

CONTENTS

1 INTRODUCTION

OSU campuses generate data constantly from multiples sources. This quantity of data, also known as big data, can effectively represent student behaviors for information technology. David Barber, from Oregon State University Information Services, requested this project to unify the data onto the consistent cloud platform of Amazon Web Services, in order for queries and analysis to be performed.

Zhi Jiang, Isaac Chan, and Zhaoheng Wang took on this project and collaborated together to architect and develop a workflow, starting with data ingestion, processing, loading into a database, and finally to analysis and visualization. Our group did all of the development work, but constantly were in communication with our client for permissions regarding AWS and progress updates.

2 REQUIREMENTS

2.1 Original Document

Oregon State University

Software Requirements Specifications

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, 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 a consistent cloud platform. The product can mine valuable information from data 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 will provide a database which can systematically deal with the data, so it contains several functions about operation for data such as ingest, store, manage and retrieve. On the other hand, the product is also able complete basic reporting and analysis. Therefore, these functions of product can help OSU Information Services staffs conveniently manage data, and they can directly and expediently understand behaviors of students and staffs according to analysis.

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 References

References

[1] IEEE Software Engineering Standards Committee, "IEEE Std 830-1998, IEEE Recommended Practice for Software Requirements Specifications", October 20, 1998.

1.5 Overview

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

In terms of perspective, our product is the most important part of a larger system. The larger system includes three parts, and which are collecting data, managing data and analysing data, so main functionality of the product are used to manage data. The produce is independent of other parts of whole system because client will provide sample of data to test this product, thus there is no interconnection between our product and collecting data part. On the other hand, we need to complete basic reporting and analysing function. Hence there should be data transmission between database and data analytics.

Figure 1: work flow of entire system

2.2 Hardware Interface

The database is built on public cloud so AWS platform will provide all aspects of the hardware supports. For example, high capacity solid state disks will used to store data items as a storage of product. And low-latency network connection can ensure maximum speed of accessing database.

2.3 Software interface

The AWS platform contains a variety of software provisioning for our product. On the one hand, developers can use proper programming language to implement functions of product base on via SDKs provided by AWS. On the other hand, the management console can help developer monitor all kinds of condition of database such as calculating throughput and cost.

2.4 Production function

The product will store different types of data such as log file, clickstream data in database. Besides, the product 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.

2.5 User characteristics

There is only one type of user that interact with our product: data analyzer. Staff who analyze the data can insert different types of data into database. Then, they could search the information base on the specific condition. After searching, data analyzer could do some management for the data

such as sorting data. Finally, data analyzer could also extract the data from database and load it into a data analysis tool to do the analysis.

2.6 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.

2.7 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.

3 Specific requirements

3.1 Functional requirements

- The user can insert any types of data they want into database.
  Description: In this function, the user should be able to insert multiple data types in database such as log file and click stream. These data could be non-relational.
  Sequence of operations: 1) access the management console of database. 2) upload data to management console 3) use inserting function to store data in database 4) output result of inserting data.
  Test: When inserting different types of data in database, there should not exist any error about inserting. On the other hand, we can do random testing to randomly choose a set of data and then check diversity of them.

- The user can find information in database.
  Description: In this function, the user should be able to find information in database.
  Sequence of operations: 1) access the management console of database. 2) enter information which needs to be searched. 3) use searching function to search information in database 4) output related data items.
  Test: We can use unit testing to check correctness of results after we find a item.

- The user can do conditional find for information.
  Description: In this function, the user should be able to search information according some specific condition. For example, user can set some restrictions like they can only find data about senior students or engineering students.
  Sequence of operations: 1) access the management console of database. 2) enter information which needs to be search and set specific condition. 3) use search function to search information bases on condition. 4) output related data items.
  Test: We can create specific unit testing according to finding condition. Then to use these unit testings to check correctness of results when we conditionally search an item.

- The user can aggregate data and then find specific result.
  Description: In this functions, the user should be able to use aggregate methods to find some specific results such as average of GAP or the highest utilization ratio of certain printer on the campus.
  Sequence of operations: 1) access the management console of database. 2) enter result which needs to be aggregated. 3) use aggregation function to specific result. 4) output specific results.
  Test: We can use unit testing to check correctness of results after aggregate data. But we need to do some extra computing by ourselves for those specific results like average of GAP.

- The user can sort some specific data.
  Description: In this function, the user should be able to sort specified data. For example, user can sort amount of credits for all engineering students.
  Sequence of operations: 1) access the management console of database. 2) target a set of data which need to be sorted. 3) use sort function to sort a set of data. 4) output sorted list of data.
  Testing: When database return sorted list, we can use unit testing to traversal the entire list and check correctness of relationship between adjacent items.

3.2 Performance requirements

Performance metrics will be determined and assessed as we begin implementation. Currently there is no reliable basis of comparison for any performance metrics regarding database operations, such

as data insertion, manipulation, and searching. A basis of performance comparison is subjective and may not fit our implementation exactly, thus are not defined at this point. Performance times for operations can be assessed by the client for acceptable runtimes and after client response, operating methods may be altered.

3.3 Software system attributes

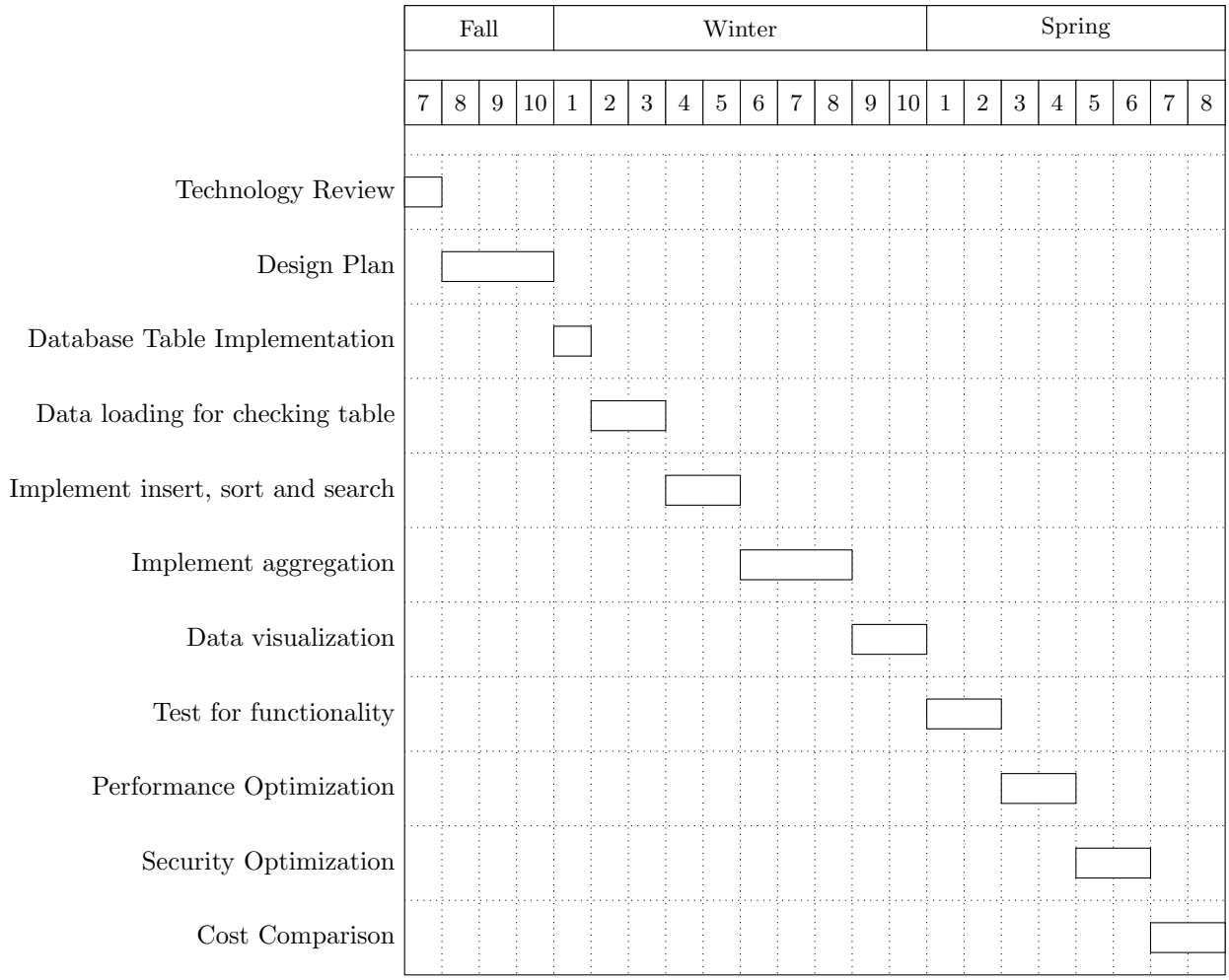
Reliability:
Compared to a locally hosted solution, a cloud platform offers additional benefits regarding reliability. When it comes to server maintenance, updates, and upgrades, downtime is a concern. Server downtime can interrupt analysis database access. Cloud platforms provide a reliable software system with minimal downtime. There are agreements and certifications required by cloud providers for acceptable amounts of downtime. As we implement our big data prototype with a cloud platform, reliability requirements are fulfilled by the cloud platform.

Availability:
Additionally, cloud platforms provide a method of access from anywhere, only requiring user credentials and an Internet connection. This is another benefit to the cloud solution.

Security:
Security is an important attribute. 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.

Maintainability:
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.

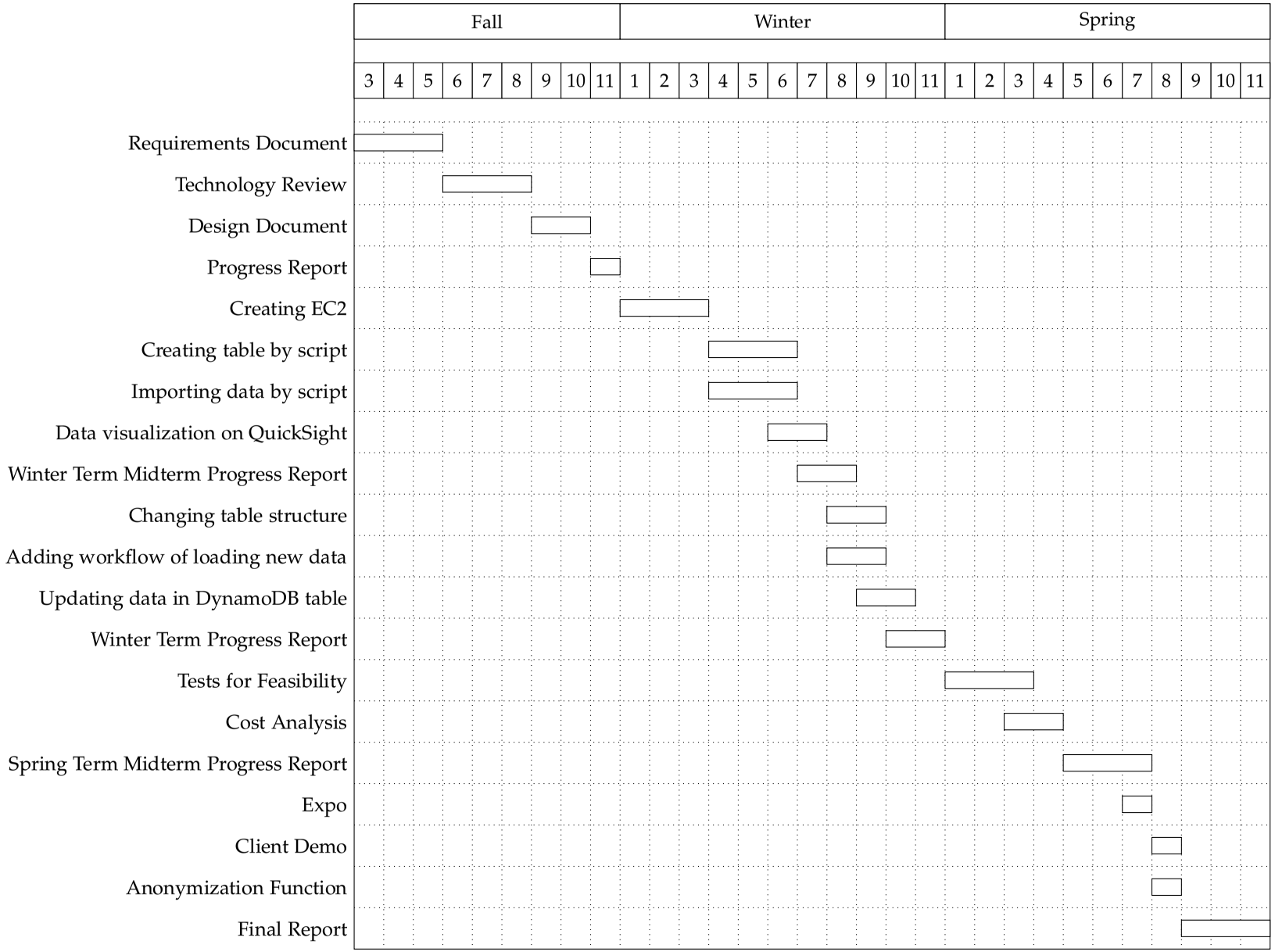
4 Schedule


Fall
Winter
Spring
7 8 9 10 1 2 3 4 5 6 7 8 9 10 1 2 3 4 5 6 7 8
Technology Review
Design Plan
Database Table Implementation
Data loading for checking table
Implement insert, sort and search
Implement aggregation
Data visualization
Test for functionality
Performance Optimization
Security Optimization
Cost Comparison

2.2 Requirements Revision

1	The user can insert any types of data they want into database	Changed	The user cannot insert data into database directly, all data are imported by script from data storage to database.
2	Create table in database	Added	In order to achieve better automation, we use script to create table when we import data.
3	Import data from data storage to database	Added	All data are stored in data storage at the beginning, and then we use script to import them into database.
4	Data anonymization	Added	This is a new requirement of client. We separate Mac address, ONID and the rest of data into three different table in database.
5	Data Visualization	Added	According to the requirements from our client, we need to do the rudimentary analysis on visualization tool
6	Database Performance	Changed	Due to the small volume of data ingestion, this requirement had reduced priority.
7	Tests for Functionality	Changed	Instead of testing for viability of analysis, our client requested a data standardization step, which was run on data that would be loaded into the database.

2.3 The Entire Process



3 TECHNOLOGY REVIEW

3.1 Original Document

Prototype Big Data Archive in a Public Cloud

Group 56: Pathfinder of Big Data

Zhi Jiang, Isaac T Chan, Zhaoheng Wang

CS 461: Senior Capstone Fall 2016

Oregon State University

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. Currently, the data is very difficult to manage because it is collected from multiple sources and is impossible to analyze. Therefore, this project requires the use of various technologies for support. There are nine pieces of technologies required:1) Methods to measure performance metrics of database functionality. 2) Methods of database security. 3) Methods of user interaction with the system. 4) The framework and storage of processing unprocessed data 5) The ingestion and parsing for unprocessed data. 6) The operation for formatted and cleaned data in data storage 7) The storage way for dealing with processed data. 8) The programming language for achieving database functionality. 9) The visualization tool use to display the data. For each piece, we will provide the best three technology options. Our goal for this paper is to analyze each technology option and determine the optimal technology that we will implement.

CONTENTS

1 Introduction

The purpose of this document is identify the best way to solve problem we stated before. So in this document, we separate the whole product into nine distinct pieces and each member will discuss three pieces individually. Isaac Chan is in charge of these pieces: methods to measure performance metrics of database functionality, methods of database security, and methods of user interaction with the system. Zhi Jiang is in charge of these three pieces: the framework and storage of processing unprocessed data, the ingestion and parsing for unprocessed data and the operation for formatted, and cleaned data in data storage. Zhaoheng Wang is in charge of the storage way for dealing with processed data, programming language for achieving database functionality, and the visualization tool to display the data.

2 Methods to measure performance metrics of database functionality

The following are the three best options for benchmarking and measuring performance of our implemented NoSQL database: YCSB, AWS Cloudwatch, and TPC-H. YCSB, or Yahoo! Cloud Serving Benchmark, is an open source framework for evaluating and comparing the performance of multiple types of data-serving systems[1]. AWS Cloudwatch is a built-in utility of AWS and can be used to collect and track metrics. TPC-H benchmark consists of a suite of business oriented ad-hoc queries and concurrent data modifications[2]. With the benchmarking and performance measurement utility, we hope to obtain a baseline for our database performance and examine how various data and query loads compare to the baseline. We will be evaluating operation speed of operations such as database inserts, updates, and reads.

	Inserts	**Updates**	**Reads**	**Visualization**	**Extra Notes**
YCSB	Yes	Yes	Yes	Raw data, can be plotted	Open-source utility means we can customize tests to fit our use-case
AWS Cloudwatch	Yes	Yes	Yes	On AWS UI and also provides raw data	Built-in utility eliminates complexity of implementation
TPC-H	Yes	Yes	Yes	Raw data	Lack of customizable tests

YCSB is a very customizable, open-source utility that can produce relevant and informational metrics for our database. It has a wide user-base and should be easy to implement.

AWS Cloudwatch is a built-in tool that can deliver relevant metrics and should work well with our database on AWS. It also has customizable metrics, which we would implement using AWS CLI (command line interface).

Finally, TPC-H is an enterprise-grade option, mostly used for server-production companies to measure how their products compare to alternatives. There is a lack of customization and a lack of a community for troubleshooting. Documentation is minimal. Implementing TPC-H is likely to be a challenge.

We will select AWS Cloudwatch as the best option. Cloudwatch can also use customized metrics tests which is important in order to know metrics to fit our use case. Although YCSB may be more easily customizable, the lack of installation makes Cloudwatch the best option.

3 METHODS OF DATABASE SECURITY

Traditionally, NoSQL databases have minimal security. In most implementations, the only security is to allow access to the database via trusted machines. However, relying solely on the network is almost certainly an invitation for a breach to sensitive information. NoSQL databases also cannot use external encryption tools, such as LDAP or Kerberos. Our best options for database security is AWSs user authentication policies, encrypting sensitive data fields, and using sufficient input validation to avoid injection attacks.With methods of database security, we hope to improve security of our database by restricting access and preventing malicious utilization of the data. We will evaluate database security by ensuring only approved user access, whether or not sensitive data is encrypted, and resistance to injection attacks.

	User restricted access	**Sensitive data encryption**	**Resistance to injection attacks**
AWS user authentication	Yes	No	Yes
Sensitive data encryption	No	Yes	No
Input validation	No	No	Yes

AWS offers a useful utility for user authentication, where users of the system can be granted different levels of access. Naturally, with authentication there is a built-in resistance to injection attacks because hopefully authenticated users are less prone to malicious intent.

There will be sensitive data in our database, most notably user-identification information, such as student IDs. This will be encrypted prior to given to us, and will most likely remain encrypted as the database is implemented and real data is inserted.

Finally, input validation is a minimal concern if we decide to implement NoSQL using AWS DynamoDB. DynamoDB does not support multiple actions with a single command, removing the risk of injection attacks. If we choose to implement the database using another tool, then there will need to be test cases to identify and ignore injection attacks.

As shown in the comparison table above, no one method can cover the security of the table. We will need to implement all three methods. Due to the lack of external tool support by NoSQL databases, we must resort to using modular security methods.

4 METHODS OF USER INTERACTION WITH THE SYSTEM

There will only be one type of user that interacts with the system - OSU staff that are able to manage and analyze the data. There are many analysis tools that can be used in conjunction with AWS. Most notably, Amazon Machine Learning (AML), Amazon EMR, and Amazon QuickSight. AML is a service that provides easy-to-use machine learning technology. Amazon EMR is a comprehensive utility that allows users to interact with databases, data warehouses, and customize their analysis. Amazon QuickSight is a fast business intelligence service that allows users to visualize data and provide responsive analysis. Our goal is to shape our implementation database to work with a utility that allows users to interact with our system. The ideal utility would provide tools to manage the data, provide analysis, machine learning support, and visualization.

	Data Management	Analysis	Machine Learning	Visualization
AML	No	No	Yes	No
EMR	Yes	Yes	Yes	No
QuickSight	No	Yes	No	Yes

AML is a very specific utility that only really offers machine learning analysis in a simple interface. It does contain much else within the tool, but the user can visualize the results from machine learning models with Amazon Cloudwatch.

Amazon EMR is a comprehensive tool that can manage data and provide analysis, through conventional queries or machine learning technology. It is more difficult to use, and does not provide native visualization.

Lastly, Amazon QuickSight is a business intelligence tool that can deliver fast analysis and boasts a very attractive visualization tool. However, more in-depth analysis and methods of data management are unavailable.

Amazon EMR is the technology that we will primarily be considering. After the conclusion of our project, maintaining the database, as well as analysis, is critical to the survival of our prototype. Our chosen technology must support all of these, and Amazon EMR is the most comprehensive tool. However, in the end, all of these technologies can be used in conjunction with our final big data prototype, but it is important to have a primary tool in order to ensure database maintenance and analysis is continuable and extensible after the Expo.

5 The framework and storage of processing unprocessed data

In the entire large system, the client will collect data for us firstly, afterward we should do some analyzing and parsing for these unprocessed data. Meanwhile, we need to provide enough space to store these unprocessed data. In this step, the primary we must consider is to find a proper framework to build storage and then perform more operation like parsing data, so we will discuss framework and storage in this section.

According to clients requirement, we should use Amazon Web Service to complete these tasks. Although we have talked about advantages of EMR above, we also would like to choose Amazon EMR (Elastic MapReduce) in this part and we focus on Hadoop framework of Amazon ERM. Specifically, Hadoop is software framework that perform distributed processing a large amount of data across hundreds of inexpensive servers[3]. The framework contains two kinds of tool. One is storage, and which is used store unprocessed data due to data cannot be stored in database directly. Another tool is used parse data. The advantage of Hadoop is obvious, because Hadoop can provide a high level of durability and availability while still being able to process computational analytical workloads in parallel. The combination of availability, durability, and scalability of processing makes Hadoop a natural fit for big data workloads[4]. In Hadoop, there are many kinds of tools, so according to our researches, we find two effective tools which are used to store unprocessed data as a storage, and they can also interact with Hadoop framework.

First of all, the Amazon Simple Storage Service (Amazon S3) is one of best choices for us. Our main part of product database is built on AWS cloud platform, thus one advantage of S3 is that it has high interactivity with our database.

In the other words, it is effortless to build connection and transform data among them. In addition, according to clients description, our product should be able deal with many kinds of data such as log file and stream, hence S3 is scalable and it can satisfy as much needs as our data. The cost of entire product is also an important criterion we need to consider, because client wants to compare cost between cloud product and local hardware, so one crucial benefit of choosing this service is its cost is low.

Second technology about storage we find is Hadoop Distributed File System (HDFS). Compare with S3, the scalability of HDFS is not better than the former. The main difference is that HDFS depends on local storage, so it has to add larger hard drives or more machines to the cluster when it needs to expand storage space for more data[5]. Meanwhile this weakness will cause that cost of it will be increased obviously. As for size limitation, any size of files can be allowed to store in HDFS, but the maximum size of single data element is only up to 5GB. We have not yet known all information about sample data client will provide, thus it is unclear whether this limitation of size on S3 will affect our product. Overall, S3 is better than HDFS because it can maximally decrease cost and ensure effective connection with database.

6 The ingestion and parsing for unprocessed data

Our product must be able to ingest and parse these unprocessed data such as format conversion, thus we need to find proper tool base on Hadoop framework for corresponding type of data. In section, we will compare two tools for parsing log file firstly, and then we will discuss a tool deal with stream data.

Dealing with log files is indispensable to our product, so Apache Spark is a remarkable tool for parsing log file. Apache Spark, as source processing engine for a large-scale data, can be easily used to parse log files. Specifically, it supports multiple programming language to write application of data analyze such as Java, Scala and R, so which means we have many choices to develop application quickly. On the other hand, the processing speed is very important to Big Data, so Apache Spark still has high speed of processing data for this aspect. According to 6 Sparkling Features of Apache Spark written by Lijin Joseji, Spark enables applications in Hadoop clusters to run up to 100x faster in memory, and 10x faster even when running on disk. Spark makes it possible by reducing number of read/write to disc[6].

MapReduce is one component of Hadoop and it is also used to process and generate large data sets. The obvious restriction of MapReduce is that it only provides two kinds of operations: Map and Reduce. Because the core concept of MapReduce processing a data is that it will separate the data into a series key/value pairs by Map function firstly, and then using Reduce function to sort each key/value. But in fact, many calculating for data cannot fit this kind of operation model. On contrary, Apache Spark can provide more operations to deal with data. As for programming language, MapReduce only supports Java, so these attributes make writing program more complicated for developers. On the other hand, all of data need to be store disk while MapReduce is processing them and Apache Spark process data in memory. Although Apache Spark is faster than MapReduce, it also means Apache Spark needs a lot of memory[7]. If size of data we want to process is not large, probably we do not need to provide more memory for Apache Spark. Overall, Apache Spark has more advantages than MapReduce on many aspects like speed of processing data and methods of operating data, so we would like to choose Apache Spark to parse log files.

The stream is another important type of data we will process, so we would like to choose an effective tool to analysis streaming data specially. Amazon Kinesis is also provided by AWS, therefore it has can commendably interact with other AWS products we choose such as data storage S3. Amazon Kinesis Streams, as one function of Amazon Kinesis, can support developers to build custom applications that process or analyze streaming data for specialized needs[8]. Another advantage is Amazon Kinesis Streams supports real-time data processing. Actually we are not sure this benefit is necessary for us because client will provide sample data rather than real-time data, but we will meet with client and determine whether the client needs function later. Amazon Kinesis API can be used in Amazon Web Services SDKs, and Amazon Web Services SDKs contains multiple programming language such as Java and PHP, so developers could use familiar programming language to complete tasks in this part.

7 THE OPERATION FOR FORMATTED AND CLEANED DATA IN DATA STORAGE

After data are formatted and cleaned by corresponding tools, we should be able to do some operations for these data. For example, we need to transfer these data from data storage to database. In section, we will discuss differences and functions of three tools operating data.

The main purpose of Hive is control data in storage on Hadoop framework such as HDFS or S3. After data is formatted and cleaned by corresponding tool, they will be stored in data storage again, and then next step is move these data from data storage to database. In general, the purpose of Amazon ERM Hive is to make connection between storage and database. The developer can write appropriate Hive command or Hive script to operate data in storage. For instance, developer can use Hive to make table for data on storage after a log is pursed, and then this table can be imported to external database. On the contrary, Hive can also perform the same operation for data from external database to data storage, so the interoperability of data is improved between these tools. In addition, it is worth nothing that Hive scripts use an SQL-like language called Hive QL[9], so it can help developers who are familiar with SQL to complete corresponding tasks quickly.

Impala, as real-time interactive SQL query tool, has the similar functions with Hive in Amazon ERM. There is difference between methods to execute SQL queries for them. The way of Impala executing SQL queries is using a massively parallel processing (MPP) engine. On the other hand, Hive executes SQL queries using MapReduce. Hence Impala does need to create MapReduce jobs and then it will spend faster query times than Hive[10]. The advantage of Impala can help developer to implement quickly some ideas about operation of data, so we will consider use these two tools together in this part.

Amazon EMR also supports Apache Pig, and Apache Pig is used to operate data on top of Hadoop as well. Firstly, Pig is different from SQL, so the developer need to speed more time learning it. Secondly, Apache Pig, as a dataflow language, can control and optimize each step while it processing data. In fact, Apache Pig is unfitting for this product, because the purpose of this part is to operate data between data storage and database, nevertheless the Apache Pig is unable to interact with external database. We will consider use advantages of Apache Pig and Hive together, for instance, Apache Pig processes data and then Hive transfers data between data storage and database. Overall, these three tools have the respective advantages, so the best way is to combine advantage of each to operate data in this part.

8 THE STORAGE WAY FOR DEALING WITH PROCESSED DATA

The goal for this part is to figure which is the optimal option of storage way for dealing with processed data.our product requires to use the NoSQL database however it may not be the optimal option for storage. As a result, we will first compare other storage way with database to check whether database is the optimal choice. If the database is the optimal choice, we will compare the Sql with NoSQL databases. After that, it is necessary to figure out which type of SQL or NoSQL is the best option.

There are many ways for storing data such as Database, Files, Cookies and other ways. we would like to start with cookie storage. Usually, cookies are used to store tiny bits of data. These data are very small only below 4 kilobytes per domain[11]. Besides. the cookies are pass the data by request which is not suitable for our product. Therefore, the cookie storing is not a good choice for our product. Another option is using Files such as XML to store the data. The data will be very large quality in our product thus there will be a lots of XML files for storing. Because of this, the management will be complex by using XML files. So the file storing is also not the suitable option for storage. The next option is using database. The database is used to store the data. The data could be managing, retrieving and organizing in the database[12].The database gives promote accessibility for the data and the data could be easily accessed by using query. Besides, it makes the data more security and reducing the cost for data insert, storing[13]. As a result, the database is the optimal choice for our product.

The database could divide into two types in general: the Relationship database and Non-relational database.The Relational database is usually represent the data base on the table however the Non-relational database use dynamic schema for data. When dealing with multiple type of data, the Relationship database is not the optimal choice because it is not the optimal choice if the data is storing in hierarchical. However, the Non-relational database is the optimal choice for dealing large quantity of data which stores in hierarchical. For scalability, the Relationship database could increase its scalability by promoting the power of hardware however the Non- relational database will increase its scalability by reducing the load. After comparing with the Relationship database and Non-relational database, we find that the Non-relational database Is the optimal choice for our product because the Non-relational database is suitable for dealing with large quantity of data which stores in hierarchical[14].

The Non-relational database we want to evaluating are: SimpleDB, MongoDB, DynamoDB, Cloud Datastore. we generate several criteria for evaluating these NoSQL databases. These criteria will evaluating the NoSQL database in various features such as cost,platform support, security,loading speed and many other features.

Tool	**Platform**	**Security for the data**	**Speed for load-ing data**	**Features**	**cost for the tool**
SimpleDB	Amazon web services	Yes	Fast	High availability and simple to use	Low

MongoDB	Cross-platform	Yes	Fast	Document database, high performance and high availability	Low
DynamoDB	Amazon web services	Yes	Fast	Support key-value model,High availability, free-text search,flexible database schema	25GB for free and the cost is low
Cloud Datastore	Google cloud platform	Does not mention on the product page	Fast	High availability and high scalability	Low

SimpleDB is offered by Amazon web services. It is security for the data and the speed for loading is fast. Besides, the cost for it is low.

MongoDB is supported by cross-platform. It is also security for the data and the speed for loading is fast. It has the high availability and high performance. The cost for it is also low.

DynamoDB is supported by Amazon web services. It is more security for the data than others and the speed for loading is fast. It has many features for example, the free-text search will make the information searching more convenient[15]. Besides, it has high availability and flexible database schema.

Cloud Datastore is supported by Google cloud platform.it is fast and it has high availability and high scalability.

We will select the DynamoDB for storing because our client requires to use the Amazon web services as the platform. Besides, the DynamoDB has more features which is suitable for our product. For example, the free-text search allows to search the information easier and flexible schema make the schema easy to development. Furthermore, the cost for DynamoDB is very low.

9 Programming language for achieving Database functionality

There are many options for programming language such as java, python, php. Each programming language will have different features. Our goal for this part is to figure out the suitable language for our product and avoid using many different languages. Because the more languages we use the more mistake we will might have.we generate several criteria for evaluating different language such as APIs, testability,security and tool support .

language	API for inserting	API for updating	API for listing table	Testability	Security	DynamoDB support
Java	Yes	Yes	Yes	Testable	Secure	Yes
php	Yes	Yes	Yes	Testable	Normal	Yes
python	No	No	No	Testable	Secure	Yes

Java is the optimal choice for achieving Database functionality. Here are the reason as following. The document of Amazon DynamoDB provide the APIs for basic functionality in java such as inserting data, updating data and listing

table. These APIs make the database functionality achieving more easily. Besides, Java is testable and more secure than the php. Another point is most of tool for our product is using Java.If we use the language other than java, the process will become complex because we also need to figure out the translation way between java and that language. This will also make the test process become much more complex. therefore , java is the suitable language for achieving Database functionality.

10 THE VISUALIZATION TOOL USE TO DISPLAY THE DATA

There are various visualization tools could be using to make the data visualization. Each visualization tool has different features. Our goal for this part is to figure out the optimal choice of visualization tool for displaying the data. The visualization tool we choose to evaluate are Tableau,QuickSight and FusionChart. we generate several criteria for evaluating them such as platform, speed, features and cost.

Visualization Tool	Platform	Speed	Features	Cost
Tableau	Tableau Online	Fast	Allows cross database, beautiful design	Low
QuickSight	Amazon Web Services	Fast	High accessibility, Get answer fast, Easy share business insight, Smart Visualizations	Low
FusionChart	No platform	Fast	Controllable for chart	Normal

Tableau could generate the graph fast and the cost of it is not high. The QuickSight is supported by Amazon web services. It has high accessibility which allows access data from multiple source. Besides, it could share the business insight in security way.Another important feature for QuickSight is it could generate visualizations very fast for very large quantity of data.Furthermore, the cost of it is low[16]. FusionChart does not require the platform to support and it could generate the graph fast. The cost of Fusionchart is expensive than Tableau and QuickSight.

Comparing with Tableau and FusionChart, the QuickSight is the optimal choice for our product. The QuickSight is supported by Amazon web services which fits the requirement of platform. Besides, it has strong accessibility which allows to communicate with other data service on Amazon web services such as Amazon DynamoDB easily.Furthermore, it has fast speed for generate the large data set which fits our purpose. Therefore, the QuickSight will be the optimal option as the visualization tool use to display the data.

3.2 Technology Review Revision

In our technology review document, we would like to some tools for ingestion and parsing of unprocessed data such as Apache Spark and Amazon Kinesis, but eventually, we did not use them. The reason is in fact; we are not required to complete ingestion and parsing of data. The client has provided existing data. Thus we do not need to consider how to ingest data by ourselves.

We change the language for implementing database table. Instead of using Java APIs, we are using the Python SDK for implementing DynamoDB table. The reason is Java APIs only show how to implement the DynamoDB table but they don't have the steps for loading data from S3 to DynamoDB. In boto3, We find the specific steps for loading data from S3 to DynamoDB through EC2 instances. Furthermore, it is better to use the same language for loading data and implementing DynamoDB table. Therefore, we decide to use the Python language to do the implementation.

4 DESIGN

4.1 Original Document

Prototype Big Data Archive in a Public Cloud

Group 56: Pathfinder of Big Data

Zhi Jiang, Isaac T Chan, Zhaohensg Wang

CS 461: Senior Capstone Fall 2016

Oregon State University

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 which will effectively represent all kinds of behaviors of students for information technology. However, the data is very difficult to manage because it is collected from multiple sources and is hard to analyze. As a result, various technologies will be required for supporting this project. In this document, we will talk about these technologies based on different viewpoints such as context, information or algorithm. For each viewpoint, we will declare the design concerns for it. Besides, we will introduce our technology detail based on these viewpoints.

CONTENTS

1 INTRODUCTION

1.1 Purpose

The purpose of this document is to elaborate implementation of main technologies used in this product. We will introduce each technology from several design viewpoints such as interaction and structure. The intended audiences of this document includes the client, development group, and assessors of this product. The client can understand all details of compositions that the developer will implement in this product according to this document. On the other hand, the development group can implement special plans such testing plan based on attributes of these technologies. Assessors can evaluate technologies used in the final product and then compare them with this document to ensure final product matches these technologies.

1.2 Scope

In this document, technologies are separated into eight distinct pieces and each member will discuss their own pieces individually. Isaac Chan is in charge of these pieces: methods to measure performance metrics of database functionality, methods of database security, and methods of user interaction with the system. Zhi Jiang is in charge of these three pieces: S3, Amazon Kinesis and Data Pipeline. Zhaoheng Wang is in charge of DynamoDB table design, DynamoDB Functionality implementation, and Quicksight Visualization.

1.3 Summary

In the document, section one will give a brief introduction for the whole document. Section two and three declare reference terms for the document. From section four, we start introducing our own technologies based on different viewpoints context, information or algorithm. For each viewpoint,we will declare the design concerns for it. This document will be in much more detail than the technology review document.

2 REFERENCES

[1] "Elastic Compute Cloud (EC2) Cloud Server and Hosting AWS", *Amazon Web Services, Inc.*, 2016. [Online]. Available: https://aws.amazon.com/ec2/?nc1=h_ls. [Accessed: 30- Nov- 2016].

[2] *Amazon Simple Storage Service - Developer guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 3.

[3] *Amazon Simple Storage Service - Developer guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 196.

[4] *Amazon Kinesis Firehose - Developer guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 2.

[5] *Amazon Kinesis Analytics - Developer guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 43.

[6] *AWS Data Pipeline - Developer Guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 5.

[7] *AWS Data Pipeline - Developer Guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 148.

[8] *AWS Data Pipeline - Developer Guide*, 1st ed. Amazon Web Services, Inc, 2016, p. 136.

[9] "Amazon DynamoDB Developer Guide", *Amazon Web Services, Inc.*, 2016. [Online]. Available: http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/JavaDocumentAPIWorkingWithTables.html. [Accessed: 1- DEC- 2016].

[10] "Amazon DynamoDB Developer API for Querying Tables and Indexes", *Amazon Web Services, Inc.*, 2016. [Online]. Available: http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/QueryingJavaDocumentAPI.html. [Accessed: 1- DEC- 2016].

[11] "Amazon DynamoDB Developer API for Scanning Tables and Indexes", *Amazon Web Services, Inc.*, 2016. [Online]. Available: http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/ScanJavaDocumentAPI.html. [Accessed: 1- DEC- 2016].

[12] "Amazon QuickSight User Guide", *Amazon Web Services, Inc.*, 2016. [Online]. Available: http://docs.aws.amazon.com/quicksight/latest/user/getting-started-create-analysis-database.html. [Accessed: 1- DEC- 2016].

[13] "Amazon Cloudwatch PutMetricData API", *Amazon Web Services, Inc.*, 2016. [Online]. Available: http://docs.aws.amazon.com/AmazonCloudWatch/latest/APIReference/API_PutMetricData.html. [Accessed: 1- DEC- 2016]

[14] "How to collect RDS MySQL metrics", *John Matson*, 2016. [Online]. Available: https://www.datadoghq.com/blog/how-to-collect-rds-mysql-metrics/. [Accessed: 1- DEC- 2016]

[15] "Identity and Access Management (IAM)", *Amazon Web Services, Inc*, 2016. [Online]. Available: https://aws.amazon.com/iam/. [Accessed: 1- DEC- 2016]

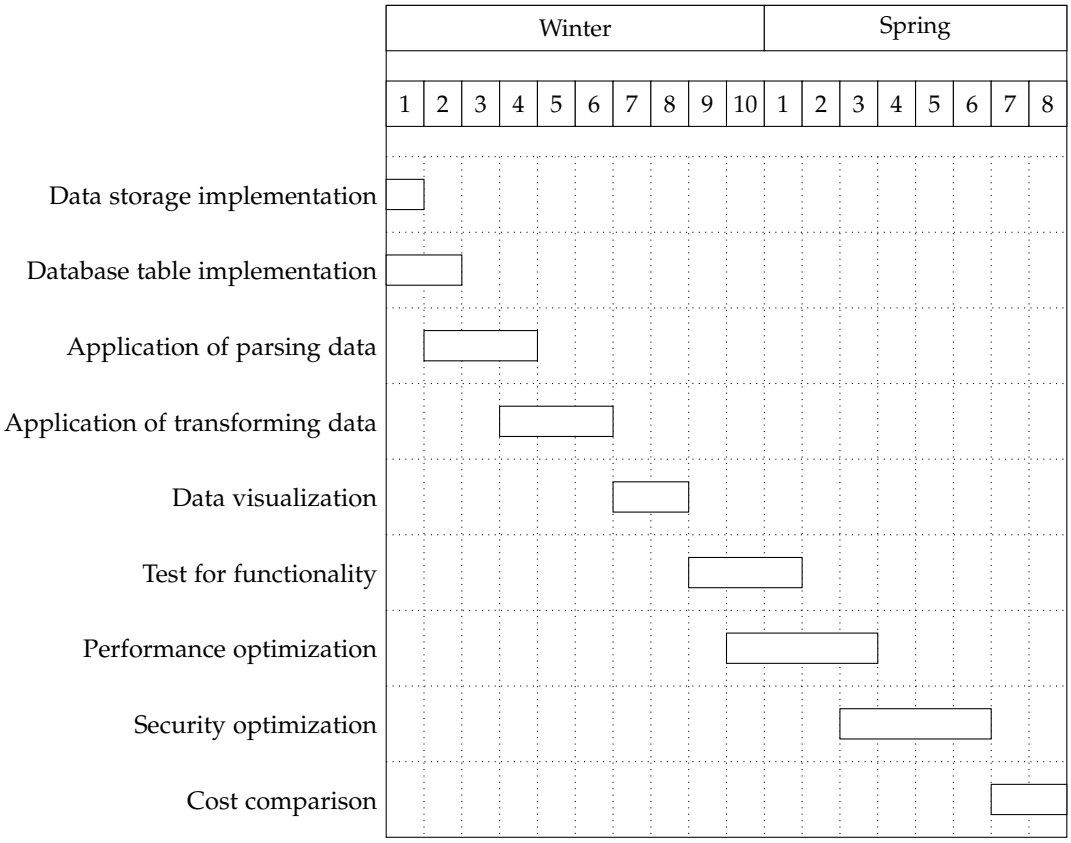
[16] "Amazon EMR", *Amazon Web Services, Inc*, 2016. [Online]. Available: https://aws.amazon.com/emr/?nc2=h_l3_al. [Accessed: 1- DEC- 2016]

[17] "Simple Linear Regression with Pure Python", *ActiveState Code*, 2014. [Online]. Available: http://code.activestate.com/recipes/578914-simple-linear-regression-with-pure-python/. [Accessed: 1- DEC- 2016]

3 GLOSSARY

Term	**Definition**
User	People who interact with our project
DB	Database
Nosql	Non-relational database
SDK	Software development kit
Schema	Database table
AWS	Amazon web service, a Platform as a service(PaaS) offered by Amazon
S3	Simple storage service provided by Amazon
Amazon Kinesis	A service used to process stream and log file data
SQL	a standard programming language used to access and process database or data storage
EC2	a web service that provides resizable compute capacity in the cloud[1]
EMR	A comprehensive tool that can manage data and provide analysis through conventional queries or machine learning technology

4 TIMELINE



5 AWS CLOUDWATCH

5.1 Context

Performance metrics for database functionality is an important element of the implementation. We will assess performance from typical database operations: data inserts, updates, and reads. After reviewing different technologies, AWS Cloudwatch is the utility we will use to perform these operations and measure the performance.

5.2 Viewpoint: Users

We will have one primary user for the implemented system: OSU staff who perform data analysis. From their viewpoint, performance is an incredibly important feature of the database. The data analyzer may be performing analysis on potentially extremely large data sets; if the speed of reading the data in the database is slow, it is not a successful implementation. We understand that slow is a vague term, but without a method of comparison it is impossible to judge the collected performance metrics as of now. We will meet with current data analyzers and our client to determine whether or not the performance meets expectations.

In order to accurately provide a model for realistic usage, we will need to measure performance of database operations using differing sizes of sample data.

5.3 Viewpoint: Scalability

Another large aspect of our database implementation is the ability for it to scale well in the future, with more and more data added. Although this will of course affect the user, scalability will also affect the database administrator.

Performance for inserting and updating data within the database is directly related to scalability, and whether or not the performance is heavily affected by the amount of stored data.

We will use large sets of sample data to load into the database. These large sizes should again, differ in size in order to model an estimation for future database performance.

5.4 Implementation

In AWS Cloudwatch, we will utilize the built-in PutMetricData API to input custom metrics for Cloudwatch to monitor[13]. This API works by passing measurements of interest into the API and the measurements are saved to a location within AWS for us to view.

Here is a code example of checking CPU utilization, using the Cloudwatch command line interface. As shown, it displays the namespace, as well as the runtime. After collecting these types of metrics using differing sizes of data, we can provide an accurate model and estimate projected metrics with extremely large data sets.

```
mon-get-stats CPUUtilization
    --namespace="AWS/RDS"
    --dimensions="DBInstanceIdentifier=instance-name"
    --statistics Maximum
    --start-time 2015-09-29T00:00:00
    --end-time 2015-09-29T00:05:00
```

Listing 1: Cloudwatch monitoring example[14]

6 Database Security

6.1 Context

With methods database security, we hope to improve the security of our database by restricting access and preventing malicious utilization of data. Our best options for database security is AWSs user authentication policies, encrypting sensitive data fields, and using sufficient input validation to avoid injection attacks.

6.2 Viewpoint: Users

From a data analyzer viewpoint, database security is a concern, in order to prevent unintended data loss, either from malicious users or unfortunate accidental injection attacks. Data loss, if unrecognized while performing analysis, can result in incorrect conclusions drawn from incomplete data. Additionally, waiting for backups to be restored or missing data to be re-inserted takes up time and hinders work.

6.3 Adminstrators

From an administration perspective, database security is a concern to reduce the possibility of malicious user access. Administrators will carefully restrict the number of authenticated users using AWSs user authentication policies. Most of the data inside the database is OSUs system users, with identifying fields such as student IDs. If a malicious user

were to gain access to the database, it would be a massive privacy breach for them to have access to identifying features of such high volume. Administrators will encrypt these identifying fields to prevent identification of the data.

Elements of course include the database and subsequent data. Additionally, after we assess the necessity of including backups, possibly backups.

6.4 Implementation

AWSs user authentications, also known as AWS IAM (Identity and Access Management), is quite robust. Administrators can create and manage AWS users and groups, while setting permissions to restrict access to certain AWS resources[15]. For example, administrators can set modularize groups to only have access to data ingestion, processing, or database for analysis. This prevents a large number of users to have access to the whole system and increases liability and responsibility for the users.

As we implement the database, the user-identifying fields within the data is already encrypted to protect us as developers and reduce liability for our client. This encryption will remain within the database and future data inserts will be encrypted as according to the administrators.

Finally, minimizing possibility of injection attacks. There is inherent protection when user authentication is done properly, because authenticated users will not be as likely to perform malicious operations. However, we will be implementing our NoSQL database with AWS DynamoDB, which inherently prevents injection attacks by not allowing multiple operations within one command.

7 User Interaction

7.1 Context

Users will need to be able to interact with the system. They will need to use different utilities for different interactions, such as monitoring resources, monitoring performance, managing data, and performing different methods of analysis. As discussed previously regarding performance metrics, the utility of choice for monitoring is AWS Cloudwatch. For the different methods of analysis, we chose AWS EMR as the most well-rounded utility for comprehensive coverage of analysis techniques.

7.2 Viewpoint: Data analyzers

From the viewpoint of a data analyzer, they will be of course interested mainly in performing analysis on the data within the database. It is their choice what utility they use, but as the developers we would like to ensure the database implementation is acceptable. We chose AWS EMR as the utility we will test our database with. From EMR, we can perform basic tests to ensure data analysis is possible.

AWS EMR is a managed Hadoop framework in a command line interface[16]. Scripts and code can be run from AWS EMR. As developers, we will be running basic data analysis code from EMR against the implemented database to ensure analysis is possible.

7.3 Viewpoint: Administrators

From an administrative perspective, they will be interacting with the system with monitoring and management utilities. Because monitoring was covered in-depth in the performance metrics section, this section will mainly be on the management utility. Database management includes operations such as inserting new or updating data within the database. This can be achieved using our database implementation utility, AWS DynamoDB. See section 10, DynamoDB for more details.

7.4 Implementation

Provided is a sample python script to perform a linear regression analysis on a set of data. Obviously as of now, the analysis is purely hypothetical, considering we do not currently have a source of data, nor we do know how the data will be analyzed. This sample code is only a proof of concept that scripts such as these can be run within AWS EMR to ensure analysis is possible with our database implementation.

```
def fit(X, Y):

    def mean(Xs):
        return sum(Xs) / len(Xs)

    m_X = mean(X)
    m_Y = mean(Y)

    def std(Xs, m):
        normalizer = len(Xs) - 1
        return math.sqrt(sum((pow(x - m, 2) for x in Xs)) / normalizer)

    def pearson_r(Xs, Ys):
        sum_xy = 0
        sum_sq_v_x = 0
        sum_sq_v_y = 0

        for (x, y) in zip(Xs, Ys):
            var_x = x - m_X
            var_y = y - m_Y
            sum_xy += var_x * var_y
            sum_sq_v_x += pow(var_x, 2)
            sum_sq_v_y += pow(var_y, 2)
        return sum_xy / math.sqrt(sum_sq_v_x * sum_sq_v_y)

    r = pearson_r(X, Y)
    b = r * (std(Y, m_Y) / std(X, m_X))
```

```
A = m_Y - b * m_X

def line(x):
    return b * x + A
return line
```

Listing 2: Sample EMR linear regression analysis example[17]

8 S3

8.1 Context

The purpose of S3, as a data storage, is used to store all files, and these files include unformatted and uncleaned data and results integrated by analytics tool. S3 gains any types of data and format data from user and then move them to other services in Amazon platform.

8.2 Composition

Buckets: bucket is a basic container used to store objects in S3[2].
Objects: Objects are the fundamental entities stored in S3, and it consists of object data and metadata. The role of metadata is to store set of name-value pairs that describe the object[2].
Keys: each object has the unique identifier and which is called key. In order to ensure uniqueness of each object, the combination of bucket, key and version ID is used to identify each object in S3[2].

8.3 Interaction

S3 need to upload data to analytics and send data to database, thus S3 will interact with data pipeline, and then complete data transforming between different compute and storage services according application on data pipeline. On the other hand, the developer needs to program some functions for S3, thus S3 will interact with AWS Lambda. The role of AWS Lambda is to run code for all backed services in this platform.

8.4 Algorithm

The AWS SDK supports several programing languages to develop S3 by programming. The programing languages cover Java, .NET, Ruby, Python and PHP. On the other AWS SDK also provide many API for these programing languages. Take an instance with Java, the developer can utilize low-level API to implement create, update, and delete operations that apply to buckets and objects in S3[3].

9 Amazon Kinesis

9.1 Context

Parsing data like log file and streaming can ensure consistency and normalization of data which will be moved to database. Thus, the service Amazon Kinesis provides is complete the basic integrating and parsing for these data.

9.2 Composition

Amazon Kinesis Firehose: the key concept of Amazon Kinesis Firehose is to create a delivery stream and then the delivery stream is used to receive data from user. Eventually, this delivery stream will be processed by Amazon Kinesis Analytics as the underlying entity of Firehose[4].
Amazon Kinesis Analytics: Amazon Kinesis Analytics support user use a series of SQL statements to complete multiple operations for stream. On the other hand, the developer can also develop own application to satisfy some special requirements.
Amazon Kinesis Streams: the developer can create a stream through this composition.

9.3 Dependency

The service is not completely independent from others. The reason is it needs data from external source such as S3, so the precondition of using this service is developer must have existing data.

9.4 Interaction

Amazon Kinesis must interact with other entities. There are two reasons. One is Amazon Kinesis needs to receive data from other services. Second reason is the results should be exported to other services after data parsed by Amazon Kinesis Analytics.

9.5 Algorithm

The core concept of Amazon Kinesis Analytics is to create one or more applications for completing multiple tasks. The programming language is SQL in this composition, and it provides enough functions to solve some general problems.

Pre-pocessing streams: The purpose of pre-pocessing streams is to normalize data and avoid errors appear before some applications do high-level processing, because sometime streaming source contains many kinds of extra conditions. For example, if the streaming source includes multiple record types, which means developers need to integrate all data from these two record types. In this condition, developers can create two additional in-application streams to store these two record types separately. And the filter the rows from original source based on record type and insert them in the newly created streams using pumps[5].

Most frequently occurring values: seeking most frequently is also a usual operation for data analysis. For example, according to our clients requirements, developers need to find which printer are most commonly used by students, so in the other word, the application must be able to calculate the frequently occurring values of printer column in the table. The function TOP_K_ITEMS_TUMBLING provided by AWS can effectively deal with this kind of problem. Developers still can set some extra conditions or constraints such as finding the top three most frequently printer in library.

Date anomalies: client needs to ensure everything is normal, so developers need to detect data anomalies on a stream in order resolve any potential issues. The AWS also provide a function to complete this kind of task, and which is RANDOM_CUT_FOREST. Take an instance, if client needs to find condition of delay time for certain webpage, developer can assign an anomaly score in this function, and then output results including all of anomaly delay time.

10 DATA PIPELINE

10.1 Context

The product relates to a number of different compute and storage services, thus function of AWS Data Pipeline is to help user efficiently and massively move data among these services.

10.2 Composition

Data node: entity of data source in the pipeline, and attributes of it consist of name, locations, and formats[6].

Activity: activity represent methods to transform data such as moving data from one location to another.

Schedule: each activity has own schedule for operating data

Resources: entity that implement activities when they are scheduled

10.3 Dependency

Data pipeline must depend on other services because the main goal of it is to transform data from other data sources. On the other hand, the activities are extensible, so the developer is allowed to run custom scripts to implement more combinations.

10.4 Interaction

AWS provides management console to create a data pipeline directly. The developer can set data source and data destination, and then the new data pipeline will automatically make connection between these two locations.

10.5 Structure

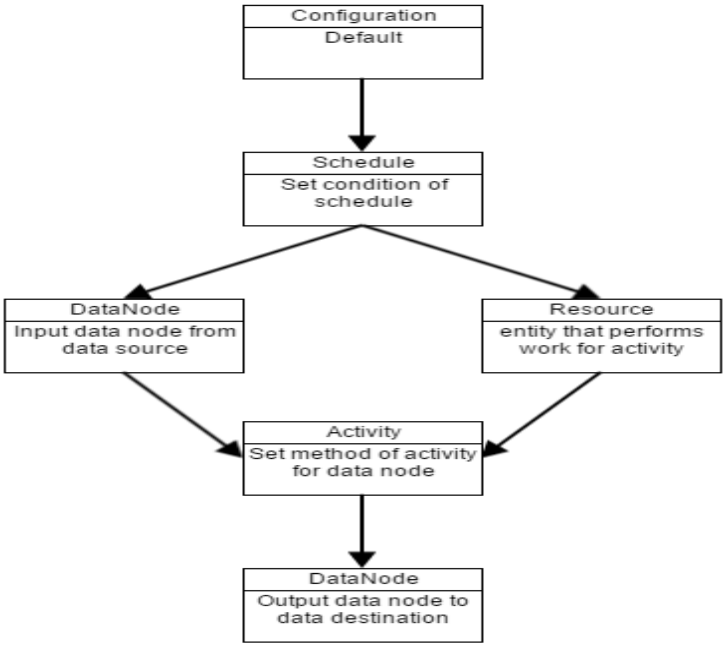


Figure 1: Complete structure of Data Pipeline

10.6 Algorithm

In this product, the locations are S3 and DynamoDB for data pipeline, thus all of algorithms must be associated to these two services. AWS Data pipeline supports **S3DataNode** and **DynamoDBDataNode**. The object example of **3dataNode** is

```
{
    "id" : "OutputData",
    "type" : "S3DataNode",
    "schedule" : { "ref" : "CopyPeriod" },
    "filePath" : "s3://myBucket/#{@scheduledStartTime}.csv"
}
```

Listing 3: S3 Data Node example[7]

And the object example of DynamoDBDataNode is

```
{
    "id" : "MyDynamoDBTable",
    "type" : "DynamoDBDataNode",
    "schedule" : { "ref" : "CopyPeriod" },
    "tableName" : "adEvents",
    "precondition" : { "ref" : "Ready" }
}
```

Listing 4: DynamoDB Data Node example[8]

In term of algorithm of activity, AWS Data pipeline provides several general activities to accommodate common scenarios. These activities include CopyActivity, HiveActivity, PigActivity, etc.

Resource is used to make these activities work on data node. The AWS Data Pipeline supports two kinds of resources: EC2 and EMR. The concept of Ec2Resource is to use EC2 instance to perform the activity but EmrCluster is to use EMR cluster to perform the activity.

11 DynamoDB

11.1 Overview

The processed data will be store into DynamoDB table. So there will be a step for modeling data structure (design table). After that,there is a step for creating table, unloading data and implement operations such as updating. The Information viewpoint will be use to design the table. Algorithm viewpoint will be use to implement different operation in database.

11.2 Information viewpoint

The Information viewpoint will be use to modeling data structure (design the tables).

Design concern

The major concern will be modeling the data structure of NoSQL database in a clear way. It is very important because

it will be easy to create tables if the modeling part is clear. Another concern will be how to convert data into schema in NoSQL database. In relational database, the ER-diagram is usually used to convert data into schema. In the key-value type of NoSQL database, the ER-diagram is not suitable to represent the table. However, the ER diagram can still represent the relationship between data and we can convert it into NoSQL database.

ER-diagram to NoSQL data model

we will first figure out the entities and the attributes of each entities. Then, we need to find the relationship (one to one, one to many or many to many) between the entities. After that, we will design the ER diagram and convert it into NoSQL table model. For instance, an ER-diagram have four entities: user, terminal, OS and printer. The relationships are: user use different terminal to access the printers and the terminal have various operating system.The figure 2 is an example of ER-diagram to NoSQL database table

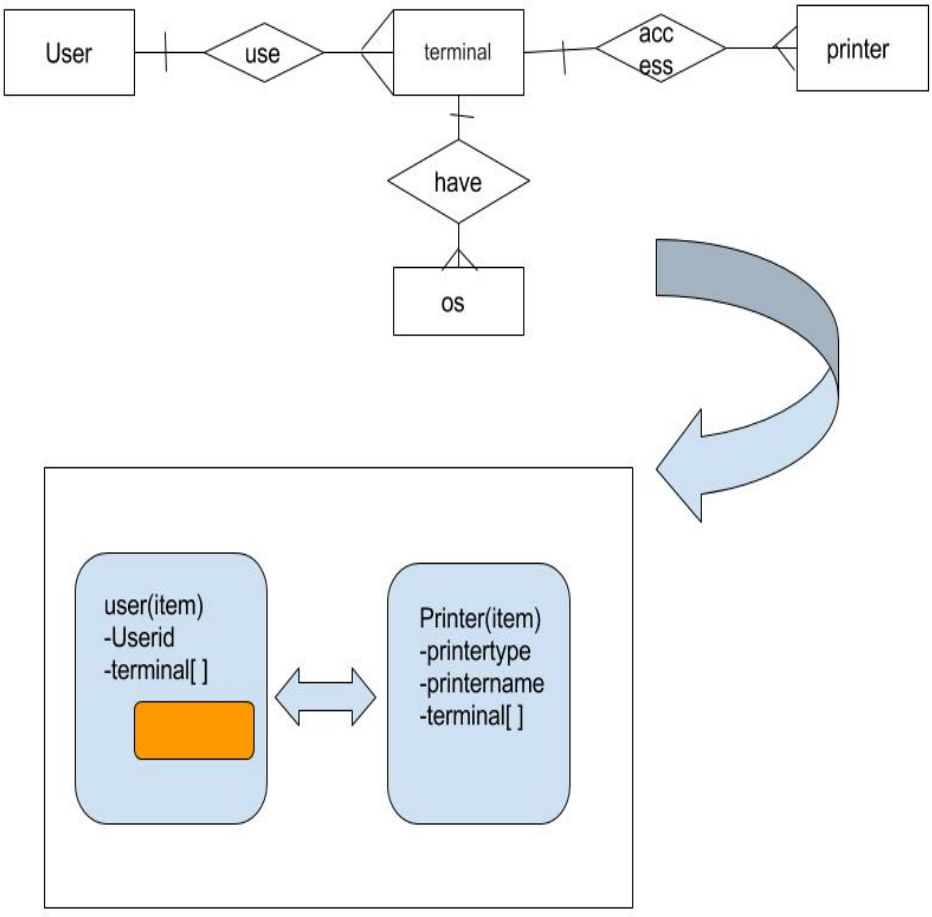


Figure 2: ER-diagram to NoSQL database table

11.3 Algorithm viewpoint

Algorithm viewpoint will be use to implement different operation such as create table, upload data, update table, delete table, list table, query table and scan table.

Design concerns

The concern for this part will be how to design the algorithm for operations. Fortunately, the Amazon provides basic APIs for the operations. So we can adapt it from the Amazon document API.

APIS for operation

(1) Create Tables

We will use the AWS SDK for creating tables. The Amazon provides the Java document API for creating tables.These API will help us to create the database schema that we need.The API will first create a dynamoDB class. After that it declares a table request which contains the table name, key of table, the definitions of attributes and capacity.

Here is the API[9] :

```
DynamoDB dynamoDB = new DynamoDB(new Client( new Provider()));
ArrayList [Attribute] attribute= new ArrayList[Attribute] ();
attribute.add(new Attribute().withAttributeName(”Id”).withAttributeType(”N”));
ArrayList[Element] key = new ArrayList[Element] ();
key.add(new Element().withAttributeName(”Id”).withKeyType(KeyType.HASH));
CreateTableRequest request = new CreateTableRequest()
.withTableName(table)
.....
.withProvisionedThroughput(new ProvisionedThroughput()
.withReadCapacityUnits(size)
.withWriteCapacityUnits(size));
Table table = dynamoDB.createTable(request);
table.waitForActive();
```

(2) Upload Data

We will use the AWS SDK for uploading data into schema. The Amazon provides the Java document API for uploading data.We will use these API to upload the data into the table that we create. The API will first create dynamoDB class.After that it declares a new item which contains the primary key, string set and the numbers for uploading.

Here is the API[9]:

```
Table table = dynamoDB.getTable(Name);
try
Item item = new Item()
.withPrimaryKey(Key)
.withString(string)
.withNumber(number)
```

.withBoolean()

....

.withString();

table.putItem();

(3) Update Table

We will use the AWS SDK for updating table. The Amazon provides the Java document API for updating table.We will use these API to update the data in the table that we create.The API will first create table class. After that it declares a new class called provisionedthroughput which contains the information for updating.

Here is the API[9]:

DynamoDB dynamoDB = new DynamoDB(new Client(new Provider()));

Table table = dynamoDB.getTable(table);

ProvisionedThroughput provisionedThroughput = new ProvisionedThroughput() .withReadCapacityUnits(size)

.withWriteCapacityUnits(size);

table.updateTable(provisionedThroughput);

table.waitForActive();

(4) Delete Table

We will use the AWS SDK for deleting table. The Amazon provides the Java document API for deleting table. We will use these API to deleting table. The API will create table class. After that, it deletes the table which is the table need to delete.

Here is the API[9]:

DynamoDB dynamoDB = new DynamoDB(new Client(new Provider()));

Table table = dynamoDB.getTable(table);

table.delete();

table.waitForDelete();

(5) List table

We will use the AWS SDK for listing table. The Amazon provides the Java document API for listing table.We will use these API to list the data in the table that we create.The API will first create DynamoDB class. After that it will execute the list table method for listing table.

Here is the API[9]:

DynamoDB dynamoDB = new DynamoDB(new Client(new Provider()));

TableCollection[Result] tables = dynamoDB.listTables();

Iterator[Table] iterator = tables.iterator();

while (iterator.hasNext())

Table table = iteratornext();

System.out.println(table.getTableName());

(6) Query table

We will use the AWS SDK for query the table. The Amazon provides the Java document API for querying the table.We will use these API to query the table that we create. The API will first create DynamoDB class.Then, it will create a table class which use to represent the table for retrieving. After that, it will use the query to retrieving the data.

Here is the API[10]:

```
DynamoDB dynamoDB = new DynamoDB( new Client(new Provider()));
Table table = dynamoDB.getTable(table);
QuerySpec spec = new QuerySpec()
.withValueMap(new ValueMap()
ItemCollection[Query] items = table.query(spec);
Iterator[Item] iterator = items.iterator();
Item item = null;
while (iterator.hasNext())
item = iterator.next();
System.out.println(item.toJSONPretty());
```

(7) Scan Table

We will use the AWS SDK for scan the table.The Amazon provides the Java document API for scan the table.We will use these API to scan the table in order to read the data which store in the table. The API will create a class called Client. Then, it create a class which use to provide the information of table for scanning.

Here is the API for Scan Table[11]:

```
AmazonDynamoDBClient client = new Client( new Provider());
ScanRequest scanRequest = new ScanRequest()
.withTableName(table);
ScanResult result = client.scan(Request);
for (Map[String, Attribute] item : result.getItems())
printItem(item);
```

12 QUICKSIGHT

12.1 Overview

The data store in the DynamoDB will be used to create an analyze and a visual on Amazon visualization tool called Quicksight. In this part, Information viewpoint will be used.

12.2 Information viewpoint

The information viewpoint can be used to achieve visualization part. Basically, we will use the Quicksight as the visualization tool to represent the analyzing result.

Design concern

The major concern for this part will be how to generate a visual on the Quicksight. Fortunately, the Amazon provides tutorial on the Quicksight document. Therefore, we can a follow the instruction step by step from the document.

Step for creating a visual

Base on the Quicksight document, there will be several general step for creating a visual. The first step is to create a data source in your database. These data source are very important for Quicksight because it will require these data source for creating data set and do the analyzing. After that, the Quicksight needs to connect with the data source. In this step, if the database is on the Amazon web services the data source will be easily to connect. However, if the data source is not on the Amazon web services it requires port, database access information to connect the data source in database. The detail information is on the Amazon Quicksight document. The next step is to create the data set according to data source and do an analyzing. The last step is to create a visual according to the data set. There are different types of visual on the Quicksight and we can choose one to generate the suitable graph for the analyzing result.[12]

The figure 3 is the general process:

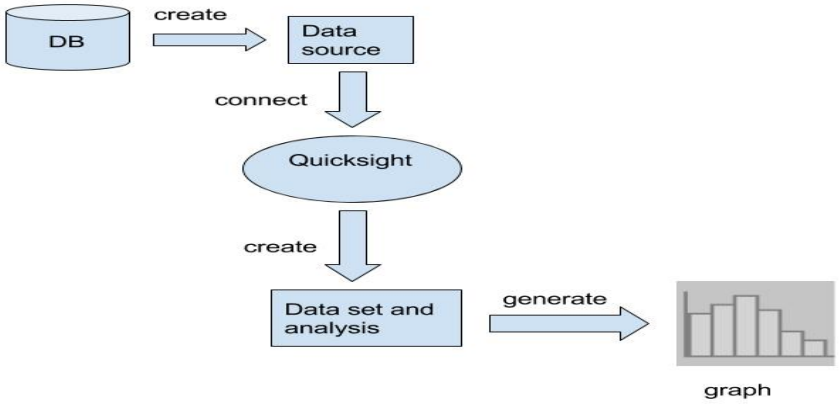


Figure 3: General process

4.2 Design Revision

We changed Amazon Kinesis to AWS SDK. We found Amazon Kinesis is a useful tool for the data stream, but data our client provided is formatted. On the other hand, for some reasons, we were authorized completely to use AWS EMR, so we were not able to use some services. Python is a good choice for us because we are more familiar with it. The AWS provides a package of Python, and which includes many useful functions to complete our requirements such as downloading files from data storage or creating table into database.

On the other hand, we didnt use Data Pipeline. The reason we chose this tool is we would like to implement automation. The advantage of Data Pipeline is it can make a schedule for some specific tasks, so this point satisfies our needs. Eventually, we gave up this tool because when we used Python, we found it can also implement automation for the entire project.

We discuss the process with our data analyst for loading data and we find that a graph of the workflow will better represent the steps for loading new data. Because of this, we add the workflow for loading the new data from S3 into the DynamoDB.Furthermore, we also make some changes for implementing DynamoDB table. Instead of using Java APIs, we use the boto3 which is the Python SDK to implement the table in DynamoDB.The reason is the boto3 provides specific steps for loading data from S3 into DynamoDB by using EC2 instances.

5 Weekly Blogs

5.1 Fall Term

5.1.1 Week 3

Zhi Jiang

Progress since last week:

Last week we completed the first assignment problem statement. The assignment effectively made me understand the entire condition of this project. This assignment helped me quickly collect all kinds of information about this project. On the other hand, I still learned two kinds of markup languages, and which are LaTex and Markdown. These markup languages will be useful tool for I working this project.

Any problems you encountered:

The most important problem we encountered is database. According to client's comments, we need to evaluate some ways to implement NoSQL database, but actually we are not familiar with database. I spent a lot of time on searching materials about NoSQL database when I modified our problem statement. Another problem is we should have more effective communication with the client. The Client told us it would be a good idea to let him know of deadlines, so we should ensure information symmetry for client.

Plans for the coming week:

We will keep going to focus on database in next week. I will try to send email to my instructor in CS 340 and ask him some questions about NoSQL and way of implementation of NoSQL. We will find more details in this project, in the other words, we will do a complete analysis for it.

Isaac T Chan

Progress since last week:

Last week we wrote the problem statement and sent it and received it from the client. There was only one revision necessary which was good, meaning we understand the problem that he presented fairly accurately.

Any problems you encountered:

We somehow got a bit behind on deadlines and sent it to the client too late. However he was very good about reading and signing it so we were lucky in that respect. We will be much more prompt in the future.

Plans for the coming week:

I don't know the difference between relational and non-relational databases and need to learn. We will probably want to fill in the gaps in our personal knowledge and start thinking about a plan to implement.

Zhaoheng Wang

The process we finish last week:

We finish our problem statement part. So I figure out the general process of project about what should we do and how

to evaluate whether our project meets the requirements.However,it is not specific enough.

Any problems you encountered:

Our project requires to use the amazon web service as an platform. However, I never use it before and I am not familiar with this platform. Besides, I am still confused about the concept of NoSQL database.

Plans for the coming week:

Firstly, I would like to search more information about this platform about how to use the AWS. Secondly, our project will focus on collecting data, organizing and analyzing data and implement the database to store the data. Therefore, I should figure out how to implement a NoSQL database on AWS.Furthermore,I might search some videos about what is the difference between NoSQL and SQL.

5.1.2 Week 4

Zhi Jiang

Progress since last week:

In the last week, we have done three things. First one is we fixed some minor mistakes in problem statement according to instructor's feedback. Secondly, we sent our modified document to client and made an appointment with him in next week. Third thing is we met with our TA and talked somethings about requirements documents. As for me, I reviewed some commands of Git because I think I am not familiar to this tool.

Any problems encountered:

The most important problem we encountered should be we still have a lot of problems about our project, in the other word, we need more details such as implementation of database and types of sample data.

Plans for the coming week:

We will focus on requirements document in next week. As I mentioned above, we made an appointment with our client in next week because we need more specific information about project. After meeting with client, we will start to do requirements document. We will read related material and understand how to write correctly requirements. For instance, we need to know differences between functional requirements and non-functional requirements.

Isaac T Chan

Progress since last week:

There was not a lot of progress this last week. Because of the extended revision deadline for the problem statement, we were on hold until we received feedback for it. When we did receive feedback, it required very minimal corrections before we sent it again to our client. We also met with our TA for the first time, made plans for meeting with our client, and have a plan for the requirements document, due next week.

Any problems encountered:

There were no problems we encountered. I did worry a little about the condition of our problem statement because we

had only a short time to fix problems and get it approved and signed by our client, but was relieved to see that the feedback was positive.

Plans for the coming week:

This next week will be busy. The requirements document seems to need a lot of work. I am not that experienced in LaTeX - some time will be spent on the IEEE formatting. Additionally, we hope to meet with our client as early in the week as possible, so we can clarify some details and be ready to write an excellent requirements document.

Zhaoheng Wang

Progress since last week:

In the last week, we revise our problem statement base on the feedback. After that, we send it to our client to check whether fits the requirement about the project. Besides, we meet our TA on Friday. we go though the general process about future meeting and we also ask several question about requirement document such as how detail should our requirement document should be. Furthermore, I do some research about NOSQL database which required by our project. it seems the biggest difference between SQL and NOSQL is SQL database uses the table and the NOSQL uses the document. The NOSQL is more focus on availability and performance.

Any problems encountered:

For the requirements document, the lecture shows five general parts about what we need to write such as functionality, security, response time. However, i could not give a clear answer for several parts such as security. For example, we are going to use the database on AWS. Usually, there will be some disaster recovery for the database and we might figure out how to do it.

Plans for the coming week:

we are going to have a meeting with our client and figure out how to organize our requirements document. Additionally,I am not familiar with latex style so i would like do some research on it. Furthermore, if I have enough time after finishing the requirements document and the research about latex, I will focus on searching the information about how to use Amazon Web Services.

5.1.3 Week 5

Zhi Jiang

Progress since last week:

In this week, we finished the final draft of problem statement and started to write software requirement specification. This document needs more details about requirements, thus we met our TA again and then we gained many information. In process of writing document, we feel benefits of requirement specification, because we need to completely understand what we do in project, and provide some specific plans to each part, otherwise we cannot complete this document very well.

Any problems encountered:

The third part of document ”the specific requirements” is main problem we encountered, because we still have many

questions about this part such as interfaces and functional requirements. In fact, according to client's introduction, our project is not very complex and the main part is database, so we will consider which section need to be ignored in this document.

Plans for the coming week:

The document is due next week, so the primary task for us is complete this document bases on feedback of TA and client. Probably we need to search more material about database, and find useful information for specific requirements.

Isaac T Chan

Progress since last week:

This week we completed our rough draft of the requirements document. We realized early that there were some requirement details that we didn't know; we met with our client early on to define them. Then we wrote the rough draft and gave it to our client to check the content.

Any problems encountered:

We still need to put the document into IEEE format. Also, there was some confusion regarding the "specific requirements" section of the document. It seems like we've covered all the requirements already in the previous section. This is likely an organizational issue and we will need to figure it out.

Plans for the coming week:

Luckily we have another week to turn in the final draft. We will need to put it into IEEE format. Regarding the missing section, we may need to meet with an instructor or TA to reorganize our document. Finally we need to finish it before Friday so we will have enough time to verify requirements one last time with our client, get it signed, and turn it in.

Zhaoheng Wang

Progress since last week:

This week we have finished our rough draft of requirements document and we send it to the client in order to get some feedback. After that, we revised it base on the feedback form client.Then,we send it to our TA in order to get more feedback.Besides, I have do some research about the idea between cloud platform ,PaaS and IaaS.The cloud platform is using the cloud computing technology. 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 dont need to worry about managing cloud infrastructure

Any problems encountered:

we have some question about the specific requirement part in requirements document. Because, we are not clearly figure out all the specific functions we need to use yet.

Plans for the coming week:

Next week, we will do some revision after getting the feedback from TA. Then, send it to our client to check whether fit

his requirement. Besides, I would like to do some research about some analyze tool such as tableau because our project need to have some basic reporting and analyzing functionality.

5.1.4 Week 6

Zhi Jiang

Progress since last week:

In this week, we talked grade of problem statement with Kirsten, and she gave us some useful suggestions about writing structure. We will do appropriate adjustment for our problem statement through her comments. Secondly, we finished the requirements document but our TA thinks it is not good enough. He thinks we need more stuffs in ”Specific requirements” section. I still learned some knowledge about technology when I was doing the requirements document. I download the document of Dynamo DB for developer and read it. I find that it has provided some APIs of basic operations of NoSQL Database. These APIs will make our tasks more convenient.

Any problems encountered:

The problem we need to solve are mistakes in the requirements document. We have to change our format of Latex and we also need to provide more details for functional requirements according to TA feedback. So it means we need to do more research about NoSQL database and completely understand its functions.

Plans for the coming week:

We will modify our requirements document as soon as possible and return it to client and TA. We still need to learn more thing about NoSQL.

Isaac T Chan

Progress since last week:

This week we received our graded problem statement back and finished up the final draft of our requirements document. We made an appointment with Kirsten to talk about our problem statement grade and the ”specific requirements” section of our requirements document. We plan to reorganize and resubmit our problem statement, because the content was good but the structure was poorly designed.

Any problems encountered:

Our requirements document is likely not in ”final draft” shape. We received feedback from our client, who thinks it is good, and also our TA, who thinks it is not so good. Due to the delay of feedback we may need to get an extension and turn it in at a later date.

Plans for the coming week:

Probably will need to finish up the requirements document and restructure the problem statement. Plus any new assignments.

Zhaoheng Wang

Progress since last week:

In this week, we make an appointment with Kirsten about our problem statement and get useful feedback from her. we find that our problem statement need to organize the information more clearly for the readers. Besides, we finish our requirements document and get some feedback from our TA. Furthermore, i do some research about the whole process again and i have a more clearly idea about how this whole process goes.

Any problems encountered:

Our requirements document still have some problems: 1. The syntax for latex need to change 2. The content for specific requirements part should be more specific 3. The format for specific requirement part need to change 4. Adding a schedule for the requirements

Plans for the coming week:

We need to revise our Requirements document and problem statement again. Besides, I would like do more research about loading data from different source and the algorithm for searching specific data in a file. i believe this will help a lot for the detail design part.

5.1.5 Week 7

Zhi Jiang

Progress since last week:

In this week, the first thing is that we completed the final draft of requirements document. In this process, we add two more sections ”Hardware interface” and ”Software interface”. We added more details for functional requirements. Eventually we created our schedule bases on functional requirements. On the other hand, I started to prepare for writing technology review. I have found many materials about each piece I will deal with so they will be useful for me to complete this assignment.

Any problems encountered:

Actually, when I read those materials about my pieces, I was frustrated, because there are many concepts I don’t know. It is hard to understand relationship between some technologies, so I believe that this is the biggest problem I encountered.

Plans for the coming week:

In next week, we will start to do technology review, so it means we must understand completely all pieces in our product, so probably we need to find more materials or ask corresponding questions to TA or instructor.

Isaac T Chan

Progress since last week:

We completed the final draft of our requirements document. We added functional requirements, software system attributes, and added more details for performance metrics. We also added a rough schedule based on our functional requirements. There is a rough outline of our technology review.

Any problems encountered:

In researching for the technology review, I found that with the information we have right now about the data and the system, it is difficult to envision the final product. I think we need to get a sample of data to work with, or all of our design and reviews are just speculation and our final product could deviate far from the preliminary design.

Plans for the coming week:

We plan to finish our technology review and start the design document. We will need more details on the design document and may need intermediate meetings with our TA to make sure we're doing it correctly.

Zhaoheng Wang

Progress since last week:

In this week, we revise our Requirements document again base on our TA's feedback. We first change the functional requirement part and make this part more detail. Besides, we add performance requirement part and system attribute part which let reader understand our system more clearly. Furthermore, we make a schedule in gantt chart which give a clearly schedule for our project.

Any problems encountered:

For the technology review, i am not familiar with how to make the data suitable before visualization so i need to do more research about this part. Besides, i also need to figure out what kind of tools we might use for visualization.

Plans for the coming week:

we might start our detail design next week. Therefore, it is better to ask an appointment with our client and data analyzer. we need to check our thought with our client and make sure it fits our client's requirement. For example, we need to ask for some sample data from our client so we could figure out the detail process for loading data. Besides, we also need to check our thought with our user(data analyzer) and figure out what types of data we need to create for loading into the analyze tool.

5.1.6 Week 8

Zhi Jiang

Progress since last week:

We completed technology review and submitted it on Monday. And then we started to do the design document. In this document, we need to focus on how to implement technologies we discussed in the last document. When we met our TA today, we asked some question about our technologies such as what's the difference between data storage and database. TA given us useful suggestions and explanation. And he also provided some basic ideas of big data,so these are good reminders for us to consider the whole product.

Any problems encountered:

The problem is also how we implement these technologies. We need to do more research base on suggestions of TA

because we are not familiar with some technologies like the filesystem.

Plans for the coming week:

As for next week, probably we will meet with our client and talk something about sample data. According to TA's suggestion, we should get sample data as soon as possible because we need to design our product depends on characteristics of these data. On the other hand, we will make a video for our product.

Isaac T Chan

Progress since last week:

This week we finished our technology review. Not a lot was discovered because we had previously researched technologies we would use, and our research affirmed those technologies. Additionally, I think we found that, because our client requires us to use AWS, AWS actually offers a lot of the technologies we need. This is very helpful because it reduces the complexity of our implementation.

Any problems encountered:

Our problems remain the same as last week. We still need a sample of the data to determine specific implementation details and how to design the database.

Plans for the coming week:

I plan on restructuring our problem statement next week and resubmitting it. I also plan on completing a portion of the design document. I think that we have fallen behind schedule slightly and I think the progress report presentation will take more time than we think. As it's already the end of week 8, we have a lot of work ahead.

Zhaoheng Wang

Progress since last week:

on this week, i finish our technology plan which is going to figure out the optimal options for achieving the requirements. This document is a necessary step for our design documents and it is very important. In this document, i choose three pieces of technologies: the first one is to figure out which is the optimal option of the storage way for dealing with processed data. In this part, i compare the other storage way with database. After that,I compare the sql and nosql database. Finally, I compare three types of nosql databases.because of the research, I benefit a lot and get more clearly understanding for the features of database than before.

Any problems encountered:

i am going to start our design document however there are several things we have to check with our client: 1 i need to get some sample data first because our first step is to parse the file. 2 i need to figure out the information we need to store. 3 For the visualization part, i need to check with data analyzer because he is our user.

Plans for the coming week:

we are planning to have an appointment with our client next week and figure out the questions above.Besides, there are many research for our project. it is very important to get a clear scope of our project before we start to design it.

furthermore, we should start to think about the progress report on next week.

5.1.7 Week 9

Zhi Jiang

Progress since last week:

In this week, we met with our client on Monday. The main thing we talked is about sample data. We are doing the design plan for entire product, so sample data can let us understand how to design each part. Another important thing was we completed the revision of problem statement and submitted to Winters. As for myself, we found more information about solution I chose for my parts. I read some developer guide provided by AWS and known some basic idea for building Hadoop framework.

Any problems encountered:

The problem we encountered is we have not got sample data from our client, so it will affect our plan, because sample data is very important for us and all of tasks we will do must depend on sample data.

Plans for the coming week:

The primary is to get sample data for comping week and then complete design plan as soon as possible. I got grade for my technology review, and I will probably talk feedback with Winters in next week.

Isaac T Chan

Progress since last week:

We met with our client early in the week and requested a set of sample data. This is very important for the design document because our database structure depends on how the sample data is set up. There are many components that are reliant on the sample data, such as whether or not we need to cleanse, parse, or otherwise process the data before it can be stored in the database. We also restructured our problem statement and sent it to Dr. Winters.

Any problems encountered:

We are waiting for the set of sample data from our client. Due to the break, we don't know when to expect it but as soon as it does we can inspect it and base some design decisions off of it.

Plans for the coming week:

Hopefully we receive the set of data as soon as possible. Then we can organize and collaborate on the design document. Additionally at some point we should write the progress report and begin the presentation.

Zhaoheng Wang

Progress since last week:

This week we make an appointment with our client and ask for a sample data. The sample data is very important for design document because if we get the sample data then we will figure out the data type we should parse for our project. Besides, the sample data will give a good example about the information we should store into database. I also

watch some video about how dynamodb works however it is not very helpful.

Any problems encountered:

The problem is i still confuse about the way that dynamodb store and there are several questions i need to figure out: 1. Does dynamodb requires er diagram? 2. How quicksight works on aws? 3. How to achieve the functionality for database? another thing is we don't get our sample and we could not start our design part.

Plans for the coming week:

Once we get the sample data, it is time to start our design document. Furthermore, we might also start to think about the process report and think about how to organize it.

5.1.8 Week 10

Zhi Jiang

Progress since last week:

In this week, we gained the sample data from our client and then completed the design document. According to this document, we understand more details about our technologies. We also made a new timeline in this document because there are some changes which are different from previous documents.

Any problems encountered:

The most difficult problem is also about our technologies. We still need to spend more time to learn how to use these technologies. I think we can start to use sample data to attempt some basic functions for tools on AWS.

Plans for the coming week:

We plan to complete the last assignment on Sunday. And we still need to submit the hard copy with signature as soon as possible.

Isaac T Chan

Progress since last week:

This week we completed our design document. In the beginning of the week we received a sample of data from our client. This data was in a CSV format, which should be simple to load into the database. During our work with the design document, we learned much more about the technologies we will implement. After seeing the format of the data and what types of fields it contains, it made me more confident in my technology choices because I had no conflicts with the data.

Any problems encountered:

I think that we are still in hypothetical terms when designing our project. The problems we will encounter will be small nuances in the technologies we choose. As of now, I don't think any of us have issues with our technologies or the sample data, but the issues we may have are very difficult to identify without actually implementing the project. Now that we have sample data, we can probably begin to work with it on AWS and experiment, but we will need credentials

given to us by our client.

Plans for the coming week:

We plan to complete the progress report on Sunday, 12/4. Also we submitted an unsigned version of the design document and will need to submit the signed version as soon as it is returned to us.

Zhaoheng Wang

Progress since last week:

This week, we receive our sample data and finish our preliminary design document.In this document, we introduce our technology based on various such as context viewpoint,algorithm viewpoint and information viewpoint.This document gives us more clear visual for our project.Besides,I have done some personal research about the process for designing NoSQL table.Usually, the relational database table will design by ER-diagram but the ER-diagram is not suitable for represent NoSQL database table.However, the ER-diagram can convert to the NoSQL database table model.Therefore, we can create the ER-diagram first and convert it into the NoSQL database table model.

Any problems encountered:

The sample data provided by data analyzer is in csv file type which is already readable by using excel.In this file, it contains the data such as userid,time,terminal type, OS and others.Therefore, there will be two things I need to check with data analyzer.The first thing is whether the future data is the same type with the sample data. The second thing is whether the sample data contains all the information we need to store.

Plans for the coming week:

In the next week, we will finish our process document for this term. Besides, I might start to think the specific step about how to implement this project.

5.2 Winter Term

5.2.1 Week 2

Zhi Jiang

Progress since last week:

In this last week, we met with our TA on Wednesday and we talked something about our project. Actually, we have not yet got the account for Amazon Web Services, because our client explained that they cannot access to the cost and billing alters so they have to solve this problem before we gain account. As for myself, I read all documents we completed in this week because I want to review more details about our project.

Any problems encountered:

Obviously, the problem we encountered is the AWS account, we hope we can get it as soon as possible.

Plans for the coming week:

We are planning to connect our client with the account. On the other hand, we should start to work on our project bases

on the schedule we made before.

Isaac T Chan

Progress since last week:

The week before last, we contacted our client for details regarding our AWS account access. We received a reply last week, explaining the difficulties our client is having with billing. We sent another email explaining the urgency and how we are behind schedule.

Any problems encountered:

We are now 2 weeks behind schedule and need to accelerate as soon as we get access to AWS and the data.

Plans for the coming week:

Hopefully we receive AWS account access and will begin development as soon as we do.

Zhaoheng Wang

Progress since last week:

In this week, we schedule our group meeting with our TA and we talk about our current process of our project. Unfortunately, we don't get the AWS account access and the data from our client.As a result, some important step can not start yet. we have send an email with our client to check about this problem.

Any problems encountered:

we are waiting for client's email and personally i would like to do some research about this project.

Plans for the coming week:

we don't get the account for AWS which is necessary for our project. Besides, i am still unsure for some of the detail for the project.

5.2.2 Week 3

Zhi Jiang

Progress since last week:

The good news for us is we get the access to AWS in this week, so which means we can start to work on our project formally. My task is to transform data from data storage to database in this project, so I read some developer guides for tools I probably use.

Any problems encountered:

I think the problem is I am not familiar with these tools. In my task, I will use a tool call EMR, and it is a little bit complex, so I think I should spend more time to understand it.

Plans for the coming week:

I will start to work on my task in this weekend, and I will continue reading material for them. If I have any questions I will send email to TA or technical adviser of our client. Finally, I hope I can gain some progress before the next weekend.

Isaac T Chan

Progress since last week:

We got access to AWS this week and have emailed our client for access to the specific tools we require. I reviewed the documents for a refresher for the tasks I'm assigned.

Any problems encountered:

I don't have any issues aside from delayed access to AWS. However, I find that most of my assigned tasks are after the majority of the work (alpha release will work without my contributions) so I want to work with my group on their tasks so I can keep up with the progress.

Plans for the coming week:

I want to try loading sample data into the database. It seems that our knowledge about NoSQL tables is lacking and I think that trying to load the data in will help my overall understanding.

Zhaoheng Wang

Progress since last week:

In this week, we schedule an appointment with our client on Tuesday. After that, we get the access account for our project.On Wednesday, we meet our TA and we talk about the current progress for our project. On Thursday,we send the email for asking the permission for the software such as S3, Dynamodb and quicksight.

Any problems encountered:

Now, we don't have the problem yet seems we just have the permission for access the software.

Plans for the coming week:

For next week, we would like to start our project as soon as possible. It seems our progress is behind the schedule.

5.2.3 Week 4

Zhi Jiang

Progress since last week:

In this week, the first thing is we have gained tools we need from our client. We started to discuss a specific plan for next step. And we will start to modify our design document because actually some tools and solutions are different from what we planned before.

Any problems encountered:

The problem for me is I cannot create a cluster successfully because we lack a tool called IAM. I have reported this issue

to our client on Friday. I hope it can be solved as soon as possible, otherwise, we could not do next step.

Plans for the coming week:
We will show some results for our TA in next week, Which is we will import data from S3 to Dynamo. This is the very important step in the whole project because all of the unprocessed files will be stored in S3.

Isaac T Chan
Progress since last week:
This week we have obtained access to almost every AWS utility we will need. We tested some of them out and realized that our understanding of the utilities was flawed as we were writing the preliminary design documents. We have come up with a plan to tackle the rest of the project.

Any problems encountered:
No problems really, in fact this week we realized we aren't as far behind in the schedule as we thought so things are going well.

Plans for the coming week:
Our group plans to have a working meeting next week. We will load some data into DynamoDB using a python script executed on an EMR or EC2 cluster, then demonstrate this to our TA. Additionally, I'm keeping note of the changes that I'll need to make to our design document before the end of the term.

Zhaoheng Wang
Progress since last week:
On this week, we have the group discussion on Wednesday and we discuss our project current process and the next step for our project. after that, we meet with our TA to talk about our project and the future plan. This week, we all access the AWS server in order to get familiar with our project and after trying some software, we find we might make some changes for our project.

Any problems encountered:
We need to do some preparation for loading data and we are catching up our schedule.

Plans for the coming week:
For the following week, our group plan to load the data from S3 into the DynamoDB by using EMR. We plan to load some sample data into the DynamoDB. If the sample data works, the data we load in the future should be working. After that, we will show it with TA to get some feedback.

5.2.4 Week 5

Zhi Jiang
Progress since last week:

In this week, TA canceled the meeting, so we did not meet with her on Monday. Another thing is that I completed the wired assignment. I still introduced our project to my partner.

Any problems encountered:

There is no any problem so far.

Plans for the coming week:

I will finish midterm progress report in next week.

Isaac T Chan

Progress since last week:

This week we submitted our poster for printing after we got it client/instructor approved. I continued to work with the data analyst to connect to EMR and gained access. After gaining access, I ran the code provided, after making modifications for it to work within the AWS environment. None of this needed to be documented, as it was a proof of feasibility for our client's benefit and to ensure our project is finished.

Any problems encountered:

Haven't received usage estimates.

Plans for the coming week:

I still haven't received an estimate for future usage so I can't complete the cost comparison. I'll plan on sending a follow-up email for those numbers if I don't receive them by Monday. I think next week the midterm progress report is due, so I'll do that next week.

Zhaoheng Wang

Progress since last week:

On this week, I introduce the project of my group with the classmate that instructor assigns for the WRIED card assignment. We have a short discussion with our projects. The project of him is about the mobile app which is designing for guiding the tourists. The users could easily access to map without internet connection on this app. It is very useful and it is cheap for the users. Besides, he also introduces the feature of this app. For example, there will be a website which is designing for uploading videos and feedback for the viewpoints. Another thing I do this week is to update the images for workflow and make it clearly.

Any problems encountered:

Currently, There is no any problem for me.

Plans for the coming week:

For the next week, the midterm report will come soon. So I plan to start to writing the midterm report. Besides, I would like to introduce my project again with the partner who assigned for WRIED card assignment. Because I think some of the point I introduce is not very clear.

5.2.5 Week 6

Zhi Jiang

Progress since last week:

In this week, we have completed the midterm progress presentation and report. In presentation, we showed off our project at some alpha level functionalities such as importing data and visualization of data. On the other hand, we have revised some documents includes Requirement Document and Technology Review. Finally, we created OneNote and organized all materials and documents on OneNote.

Any problems encountered:

In this week, I did not encounter any problems.

Plans for the coming week:

In this week, we have completed the midterm progress presentation and report. In presentation, we showed off our project at some alpha level functionalities such as importing data and visualization of data. On the other hand, we have revised some documents includes Requirement Document and Technology Review. Finally, we created OneNote and organized all materials and documents on OneNote.

Isaac T Chan

Progress since last week:

This week, we scrambled to complete all the requirements for the midterm progress report. This included meeting to record our presentation, revising documents, and completing the OneNote. In the OneNote, we added an organizational scheme for our poster.

Any problems encountered:

Not really. Probably just need to talk to our client and data analyzer to be ready to implement my portions.

Plans for the coming week:

During my review of my midterm progress, I realized that I hadn't done much personally regarding the project due to our task division. Actually there are still a few more weeks before I'm scheduled to begin my portions (have to begin after we have a working solution). But there's many requirements I need to work on my tasks, including getting data from our client and consulting with the data analyzer to see what kinds of analysis we should try.

Zhaoheng Wang

Progress since last week:

This week,we first revise our review document and the design document. After that,we finish our presentation slide on Monday night and we schedule a group appointment with each other on Tuesday in order to finish our presentation. Our video is around 20 mins. it is a little short in our project. As a result, we check it with TA on Wednesday.After that, we start our progress report on Wednesday night. After finishing the progress report, we need to make a simple structure of the poster.

Any problems encountered:

1. Check with data analysis whether there is an efficient way for organizing table 2. Figure out which sample data can be ignored

Plans for the coming week:

There still have some improvement for these parts in our project. Firstly, the way I choose for design the data structure may not be the optimal choice. Therefore, I would like to check with the data analyzer whether it is the efficient way or not. Secondly, the graph which made by QuickSight needs to improve because some of sample data needs to be filter. Furthermore, the QuickSight seems not handle the complex analysis and our client doesnt give specific topic that we need to analysis. As a result, I would like to check with the data analyzer what kind of analysis we need to do

5.2.6 Week 7

Zhi Jiang

Progress since last week:

In this week, I started to improve my portion. On the other hand, we send design document to our client for signature, because we did some changes on it.

Any problems encountered:

We did not get the signature on time, so I just hope we can get it as soon as possible.

Plans for the coming week:

In the next week, I think we should have an appointment with our client because we should report to our client what we did in several weeks ago specifically. I will continue to improve my portion and test it.

Isaac T Chan

Progress since last week:

This week we sent our updated design document to our client for approval and signature.

Any problems encountered:

We haven't yet received a signature for the design document from our client and need to submit it by today.

Plans for the coming week:

This coming week I need to contact our client and data analyzer for approval on the strategies I want to employ for data analysis and performance testing. This involves getting larger sets of sample data than we have received so far so there might be some issues regarding how much data they are willing to give us.

Zhaoheng Wang

Progress since last week:

On this week, we meet with our TA on Wednesday to talk about the current process of our project and we send our

design document to our client for asking the signature.

Any problems encountered:

we don't get the signature for the new design document and we will send an email to TA. Hopefully, we could get the signature as soon as possible.

Plans for the coming week:

For next week, we are planing to have an appointment with our client in order to check whether we are on the correct position. Besides, we might optimize our project. For example, the table can be optimized by creating three separate tables. By doing this way, some important data can be store more security.

5.2.7 Week 8

Zhi Jiang

Progress since last week:

First of all, I continued to improve my portion, but I did not complete it. Secondly, we got the signature from the client for the design document. On the other hand, we met with our data analyzer, and she gave us some suggestions about this project. On important thing is we should create three tables when we import data from S3 to dynamoDB.

Any problems encountered:

The problem I encountered is I did not complete function to create a table automatically. In this library, it has fixed format to create a table, so I was not able to use loop statement for setting name of each attribute.

Plans for the coming week:

In next week, we will try our best to implement the solution based on data analyzer's suggestion. On the other hand, we still have a class meeting with our instructor, so probably we should discuss how to show our project in class.

Isaac T Chan

Progress since last week:

This week we received our signed design document from our client. He included a suggestion to meet with the data analyzer, which we were already planning to do. After meeting with her on Wednesday, we got several suggestions about the pipeline from S3 to the database, and creating two other tables with identifiable data (MAC address, ONID) with a hashed unique key to tie them back to the main data. This is something we will implement with fake ONID and MAC address, due to the sensitivity of the data.

Any problems encountered:

No problems this week.

Plans for the coming week:

The data analyzer gave us some suggestions that we plan to implement, including hashed user identifiers and also

replicating data to check database performance. She will send us some code to test data analysis as well. Finally we have a class this week on Thursday, so our group should meet and prepare a little bit before presenting in that class.

Zhaoheng Wang

Progress since last week:

On the Tuesday night of this week, we get signature and the feedback about our design document from our client. On Wednesday, we have an appointment with our TA to discuss our current process. On Thursday, we schedule an appointment with the data analyzer and she gives several suggestion for our project. For example, we would have three table and two of them will be link table which contains Onid,MAC. She also mentions that we would use pipeline to export the data from database.

Any problems encountered:

Currently, we don't have the problem on this week.

Plans for the coming week:

On the next week, since the group number of us is 56 we will have the on class group meeting on Thursday.

5.2.8 Week 9

Zhi Jiang

Progress since last week:

In this week, we learned some experiences about Expo in class on Thursday.

Any problems encountered:

No problems in this week.

Plans for the coming week:

As for next week, we have a lot of things. We will go to test your program by real data. On the other hand, we need to complete the progress video and progress report, and finally, we have to finish the poster.

Isaac T Chan

Progress since last week:

This week we met with the TA on Wednesday and discussed the deliverables for the coming week. Then we had class on Thursday and learned more about how to conduct ourselves during expo.

Any problems encountered:

Aside from requiring real data, nothing. Unless stressing about the million things due in this class counts as a problem.

Plans for the coming week:

During our meeting with our TA we talked about the current state of the project and I realized that for us to use real data in the expo, even with hashing, we have to have at least one group member able to access to real data. I'll probably

send an email to our client/data analyzer and see if we have to meet in person. We also have a meeting on Wednesday of next week to record the final report presentation.

Zhaoheng Wang

Progress since last week:

On the Tuesday night of this week, we change the table organization base on the suggestion from data analyzer after that we reload the sample data again. On Wednesday, we have an appointment with our TA to discuss our current process and the meeting on Thursday morning. On Thursday, we discuss our project with other group in class time.

Any problems encountered:

Currently, i have some question about the progress report which is due on the final week.

Plans for the coming week:

on the next week, we schedule an appointment with group members on Wednesday 3pm to record the video for our project. After that, we are going to discuss the poster for our project since the poster is due on the Friday of week 10.

5.2.9 Week 10

Zhi Jiang

Progress since last week:

We met with our client in this week. We discussed his questions about midterm progress report, also he given us some suggestions for this project. We got a new workflow diagram; this is very important for us because we can improve functions of our project according to this diagram. On the other hand, we still completed the presentation for winter term progress.

Any problems encountered:

No problems in this week.

Plans for the coming week:

As for the coming week, we would like complete winter term progress report as soon as possible.

Isaac T Chan

Progress since last week:

This week, our group met to complete the recording of our presentation. Additionally, we had a meeting with our client to update him on our progress and get a signature for our week 6 progress report. We received a workflow diagram, which was their expectations for our progress. It aligned quite well with our solution, and we plan to update our design document with this workflow.

Any problems encountered:

There was some miscommunication between our client and us. He had thought that our midterm progress report was

the final report, so some of the content didn't make sense with that assumption. We clarified this in our meeting.

Plans for the coming week:

The next week I plan to complete my individual progress report.

Zhaoheng Wang

Progress since last week:

On monday, we send the midterm progress report to our client. On Wednesday, we schedule a group meeting from 3pm to 5pm in order to record the video for our project. On Thursday, I schedule an appointment with our instructor to discuss about the final report. After that, i start to write the final report and the poster on Thursday. The poster is due on Friday this week.

Any problems encountered:

Currently,we don't get the signature for our project. Hopefully, we could get it as soon as possible.

Plans for the coming week:

For the coming week, i would like to revise my design document again because we get the suggestion from the data analyzer.

5.3 Spring Term

5.3.1 Week 1

Zhi Jiang

Progress since last week:

There is no any progress.

Any problems encountered:

There is no any problem.

Plans for the coming week:

We need to complete our poster as soon as possible. We still need to do extra credits with other group.

Isaac T Chan

Progress since last week:

No progress was made since the last week was Spring Break. I evaluated our position in the project to make sure I haven't forgotten anything.

Any problems encountered:

No problems this week.

Plans for the coming week:

We have a poster due soon and should get some feedback on the drafts we turned in over winter term. Also, I need to send an email to the data analyst to obtain a script so we can ensure the usability of our implementation.

Zhaoheng Wang

Progress since last week:

On this week, i add a workflow for the design document base on the feedback which get from our client. After that, our group schedule a time for meeting with TA every week.

Any problems encountered:

Currently we have no problem for our project.

Plans for the coming week:

For the next week, we will start to revise our poster since the poster is due very soon. Besides, we will also do some preparation for the Expo.

5.3.2 Week 2

Zhi Jiang

Progress since last week:

We met with our TA on Monday, and she gave us some guidance about what we should do in this term. On Friday, we still did extra credit with another group in Winters' office. We got some feedback and suggestion about poster and presentation.

Any problems encountered:

There is no any problem so far.

Plans for the coming week:

As for next week, we will finish the draft of our poster based on feedback from Winters and another group. We will still send it to our client for feedback.

Isaac T Chan

Progress since last week:

This week we met with our TA on Monday to talk about things we need to turn in, the current state of our project, and the plan for the rest of the term. Then, on Friday we had a meeting with Dr. Winters and another group to review our poster and practice presentations.

Any problems encountered:

Not really but we should be in better contact with our client this term. We'll send him our poster draft and see what he thinks.

Plans for the coming week:

We received some feedback regarding our poster and will incorporate it over the weekend. We have another draft due on Monday, and will turn it in with our changes then. Additionally we'd like to send it to our client too see what he thinks. I also should send an email to the data analyst to obtain the rudimentary analysis script.

Zhaoheng Wang

Progress since last week:

On this week, we have the group meeting with TA on Monday to discuss the current process for our project. On Friday, we have the discussion for our poster and project with other group in order to get some feedback for our poster. After that, we will revise our poster as soon as possible because the second draft is due on next Monday.

Any problems encountered:

For the next week, we will finish revising our poster base on the feedback from other group. After that, we will send it to our instructor and our client in order to get more useful feedback. Once we get the feedback, we will change our poster. Besides, we might practice more for how to represent our project in expo.

Plans for the coming week:

Currently we have no problem for our project but we are not sure whether the poster fits our client's requirements. Therefore, we will send it to our client and ask the feedback from him.

5.3.3 Week 3

Zhi Jiang

Progress since last week:

In this week, we talked about our current state with our TA and we also get a code used to test our project from the data analyst.

Any problems encountered:

we are waiting for feedback for our poster from the client.

Plans for the coming week:

As for next week, we will go to test our project by using test code. On the other hand, if we get the feedback from the client, we still modify our poster again.

Isaac T Chan

Progress since last week:

On Monday, we met with our TA to discuss coming deadlines and the current state of our project. We also sent an email to the data analyst to obtain code in order to test the usability of our implemented database. Additionally, we submitted our second poster draft and also sent a copy to our client. On Wednesday, we had class. Later in the week, we received the code from the data analyst.

Any problems encountered:

We haven't received poster feedback from our client for the second draft.

Plans for the coming week:

The code we received from our data analyst will need to be ran on AWS. I plan on testing this next week. Additionally we will revise our poster, probably along with a meeting with our client. We need to clarify the deliverable "price model with locally hosted hardware" to make sure we meet the expectations of our client and professors, which we can do at our client meeting. Also, we need to remember to get written approval for using the "OSU Information Services" logo on our poster. Finally, I need to submit my model release form.

Zhaoheng Wang

Progress since last week:

On this week, we have the group meeting with TA on Monday to discuss the current process for our project.After that, we send an email for asking the testing code from our data analyst since she will provide the test code for our project.Besides, we send our poster to our client in order to get more feedback. On Wednesday, I personally schedule an appointment with Kirsten to ask some question about the poster and the midterm report requirement.

Any problems encountered:

Currently, we don't get the feedback from our client so we can not start to revise our poster.

Plans for the coming week:

For the next week, we will revise our poster since it is due on May 1th. Besides, we might test our project on next week.

5.3.4 Week 4

Zhi Jiang

Progress since last week:

In this week, our TA checked the status of the project on Monday. The second thing is we got the feedback about our poster from the client. According to his feedback, we revised our poster and then we got the signature from him.

Any problems encountered:

There is no any problem so far.

Plans for the coming week:

We will go to finish our entire project as soon as possible. On the other hand, we need to do document assignments such as progress reports.

Isaac T Chan

Progress since last week:

This week we received feedback for our poster from our client and revised it. Then we got a signature for both the use of the logo and approval for our poster. I worked with our client and data analyst to get access to EMR to prove feasibility

and also to assess client expectations for a cost comparison.

Any problems encountered:

AWS permissions weren't granted so the proof of feasibility must wait until I can access EMR. Additionally I don't yet have an estimate for future usage, so for the cost comparison I have to wait until those numbers will be given to me before I can do that cost comparison. The comparison won't have any code involved, and will simply be a document. I'll put it in GitHub after the code freeze, as soon as those numbers are given to me.

Plans for the coming week:

I wasn't able to get access to EMR and need to in the coming week, as well as prove feasibility by running a sample analysis script within EMR against our database. The proof of feasibility won't need to go in GitHub before the code freeze so we're fine on that account. I also know what I need to do for the cost comparison and will do that as soon as possible.

Zhaoheng Wang

Progress since last week:

On this Monday, our group schedule an appointment with our TA. Zhi Jiang and I go through the steps how to load data and create the visualization on QuickSight with our TA. We check out the first two requirements on the project. However, Issac is still working on the third part and he is not coming to the meeting. On Tuesday, our client email back the feedback for our poster. So we revise the poster based on the feedback and send back for asking a signature. On Wednesday, we get the signature for our poster.

Any problems encountered:

Currently, we get the permission from our client for the logo on the poster. Jiang Zhi and I have finished our parts but we are not sure whether Issac finishes his part or not.

Plans for the coming week:

Since the Expo will come soon, personally I will do some practice for how to represent our project. Besides, most of my job are finished on the last term. Therefore, I would like to figure out whether my job can be optimized or not.

5.3.5 Week 5

Zhi Jiang

Progress since last week:

In this week, TA canceled the meeting, so we did not meet with her on Monday. Another thing is that I completed the wired assignment. I still introduced our project to my partner.

Any problems encountered:

There is no any problem so far.

Plans for the coming week:

I will finish midterm progress report in next week.

Isaac T Chan

Progress since last week:

This week we submitted our poster for printing after we got it client/instructor approved. I continued to work with the data analyst to connect to EMR and gained access. After gaining access, I ran the code provided, after making modifications for it to work within the AWS environment. None of this needed to be documented, as it was a proof of feasibility for our client's benefit and to ensure our project is finished.

Any problems encountered:

Haven't received usage estimates.

Plans for the coming week:

I still haven't received an estimate for future usage so I can't complete the cost comparison. I'll plan on sending a follow-up email for those numbers if I don't receive them by Monday. I think next week the midterm progress report is due, so I'll do that next week.

Zhaoheng Wang

Progress since last week:

On this week, I introduce the project of my group with the classmate that instructor assigns for the WRIED card assignment. We have a short discussion with our projects. The project of him is about the mobile app which is designing for guiding the tourists. The users could easily access to map without internet connection on this app. It is very useful and it is cheap for the users. Besides, he also introduces the feature of this app. For example, there will be a website which is designing for uploading videos and feedback for the viewpoints. Another thing I do this week is to update the images for workflow and make it clearly.

Any problems encountered:

For the next week, the midterm report will come soon. So I plan to start to writing the midterm report. Besides, I would like to introduce my project again with the partner who assigned for WRIED card assignment. Because I think some of the point I introduce is not very clear.

Plans for the coming week:

Currently, There is no any problem for me.

5.3.6 Week 6

Zhi Jiang

Progress since last week:

We went to complete the midterm progress in this week. Today we completed the video and then we planned to finish

the midterm progress document on the weekend.

Any problems encountered:

There is no any problem so far.

Plans for the coming week:

We will complete midterm progress document as soon as possible, and we will go to do expo!

Isaac T Chan

Progress since last week:

This week we focused on our midterm progress report. We met Friday afternoon to record the video and will complete the document over the weekend. I also sent a follow-up email to our client for an estimation so I can complete the cost comparison requirement.

Any problems encountered

Still haven't been able to start the cost comparison requirement.

Plans for the coming week:

The coming week we will submit our progress report on Monday. The rest of the week will be preparation for Expo on Friday.

Zhaoheng Wang

Progress since last week:

In this week, we schedule an appointment with our TA to discuss current process for our project. We also ask some question about the midterm report and the midterm video. We are working on the slide for midterm report on Thursday and finished it on Friday. On Friday morning, I personally finish my midterm report and do some practice for the expo. Besides, our group are recording the midterm video on Friday afternoon.

Any problems encountered:

Currently, I finished my job which is designing the data structure for the DynamoDB and the workflow for loading new data. Besides, I finished the rudimentary analysis on QuickSight. However, I think it is better to check the current situation with our client and make sure our project are in the correct direction. Thus, our group plan to schedule an appointment with client.

Plans for the coming week:

In next week, we plan to schedule an appointment with our client to talk about our project. We will check our project with the data analyst since the data analyst is the only user for our project. Besides, the expo is coming on Friday next week. Therefore, I would like to do more preparation for the expo.

5.3.7 Week 7

Zhi Jiang

Progress since last week:

On Monday, we completed midterm progress and sent them to instructors and TA. On Friday, were done Expo well. We introduced our project to other people and answered their question in the Expo.

Any problems encountered:

I don't have any problems.

Plans for the coming week:

I think in the week we will discuss our project with our client. We would like to know does he have any opinion on our work.

Isaac T Chan

Progress since last week:

This week we submitted our midterm progress report and presented our poster at Expo.

Any problems encountered:

I have a final deliverable for our client - the cost analysis. I'll complete this next week. We might also meet with him and see if he'd like us to do any further work in the next couple weeks.

Plans for the coming week:

No issues.

Zhaoheng Wang

Progress since last week:

On this week, our group first submit our midterm report and the midterm video for the project. On Wednesday class, we get much information on how to prepare for the Expo on Friday. On Friday, our group attends in the Expo in KEC. Each of the group members introduces our general ideas of our project to the people and answer their questions. For example, one of them is interesting where are these big data analysis used for. I answer that these big data can be used to study the student behavior. For instance, we can analyze the average usage of the network and report the data. These data will help the network manager to figure out when is the best time to maintain the server.

Any problems encountered:

Currently, I have finished my part and I have no question.

Plans for the coming week:

On Expo, our group meets with our client and we have discussed our project.For the next week, our group will probably meet again with our client and check the requirements for the project.

5.3.8 Week 8

Zhi Jiang

In the fall term, I chose this project because I was interested in big data, but I did not understand it. Through these three terms, I have a preliminary understanding of big data and cloud computing, which is a pretty good experience for me. If I was to redo the project, I will tell myself that I should spend more time learning the knowledge involved in the project. When we did technology review document, I mentioned that we would use a tool called "Lambda," but when we were implementing the project, we did not use this tool. Thus this mistake was caused by my lack of knowledge of these tools.

Throughout the process, the skills I have learned are about the use of cloud computing services and the design of large data workflow. Also, I also learned to be more skilled in using GitHub and Latex. I did not intend to work on big data or cloud computing later, but the project made me aware of the importance of the data.

If I were the client for this project, I would be satisfied with the work, because I am well conscious of my efforts for this project. If this project were continued next year, I think a user interface should be a good choice for this project. The user interface can provide visualization of analysis results to students.

Expo is an excellent experience for me. I was very happy to explain to others what we have done. Throughout the process, I found myself not very good at introducing our projects in simple language. Not everyone knows the technology, so Expo teaches me how to give a clear account of those who know nothing about the technology.

Isaac T Chan

Retrospect:

For me, this project did not have a focus on development and has been an exercise in writing and documentation, with occasional collaboration and scheduling. So a pretty standard software engineering experience. If I were to redo the project from fall term, I would spend more time on development in fall term, and less in winter. The project was structured in that there was initial requirements and design in fall term. However, without access to sample data and AWS, it was difficult to determine the exact structure of the project. When we started development in the winter, it quickly deviated from the design document and the technologies we said we would use. I've learned a lot about writing documents in LaTeX and making nice charts. I think that a lot of the documentation skills could be utilized in the future. Unfortunately, the overload of documentation necessary in this project has taught me a healthy distaste of documentation. I enjoy working within AWS but the scope of the project was not ideal for me. It was targeted at people with zero experience in AWS and was a bit trivial for me. However, our client was very happy with our final outcomes and found it to be a very nice proof of concept for future infrastructure at OSU. I think this project could easily be carried on for next year and be actually implemented into a workable pipeline, processing real data as it comes in. However, this might not be feasible as the data cannot be seen by students.

Expo Experience

Expo was a marathon of finding different ways to describe something I have already described many times before. I

found that our project did not generate a lot of interest with the majority of expo viewers. We never had to describe our project to children, or high schoolers. Instead, we did get some industry representatives. I was looking forward to being challenged by their questions, but found that most of them, once they understood the outcomes and processes, found it interesting and then left. I think that if we were able to implement our solution faster, and transfer it to the data analysts sooner to get actual results instead of hypothetical ones, we could potentially have generated more interest in the crowd and have us be more engaged in the expo.

Zhaoheng Wang

Redo the project

If I can redo the project from Fall term, I would like to tell myself research more information about the project such as workflow and tools. Besides, I would also tell myself to ask more questions to understand the whole project. It is important to have some background knowledge for the whole project. Otherwise, it might cost much time to figure out. Take my project as an example, the project my group work on is to deal with the big data on the Amazon web service. At first, I have no idea about how to start the project. Therefore, I search online about the big data analysis. However, there are many ways to do that so it is hard to decide which is the best way for dealing with the big data. Besides, I am not familiar with Amazon web service. Thus, I can only guess the step before starting this project. Because of this, we spend a lot of time on asking the permission for using the tool.

Skill learned I think the most useful skill I learned will be how to solve a problem in the short time. I think it is very important skill for most of the computer science students. With the technology developing, there will be lots of problems in our daily life. Therefore, this skill become more and more important to us. Besides, another important skill I learned teamworking. I believe it is also important skill because most of project requires to work in a team. Thus, a good teamworking makes the work more efficient. I believe these two skills help me to work better in the future work.

Project like and dislike

I think the whole project is very interesting. Currently, there are many places require the big data analysis. For example, there are more people shopping in Amazon on the festival. The Amazon company needs the big data analysis to figure out whether the server can work normally if there are many people shopping online in the same time. The part I dont like will be researching part, there are many different tools can be used for the project and it is very hard to decide which one is the optimal choice. For example, we can use the Nosql or sql database as the storage way. However, the Nosql will be easy to load data but it is hard to search data. The sql database will be hard for loading the data but easily to search the data.

Teammates, if I am the client and if continued next year

I think I am benefits a lot from working with the other two teammates. I believe the most important skill I learn from the teammates is how to work efficient.If I am the client, I think I will be satisfied with the work because the project fits all the requirements and each group members communicates with the client frequently.Currently, the data our project use simple data which doesnt real ONID and MAC address. However, if the project were to be continued, the real data will be stored and the whole project will be applying into the real situation.

Expo experience

The expo is very interesting to me. At first, I think there might be only few people come to talk about our project because this project is designing for the specific users such as companies. However, I find there are many people come to talk about this project. Sometimes, the people talk with our project ask some questions which is very useful. For example, I remember of the them ask me why we need to do the big data analysis. But some of them are asking very strange question. For instance, how to relate the big data with music tools. It is hard to answer that question. In a word, the Expo is very interesting and it is a way to demo our project which is very useful.

COLLEGE OF ENGINEERING

Electrical Engineering & Computer Science

Prototype Big Data Archive in a Public Cloud

What is OSU big IT data?

Data from usage of information technology can provide statistics on student computer ownership and behavior. The purpose of this project is to create a workflow to unify, organize, and securely store, that data in the Amazon Web Services (AWS) cloud, perform basic analysis, and visualize results.

Our Work

1. Method of ingestion and management of sample data onto Amazon's cloud platform
2. Rudimentary data analysis, reporting, and visualization
3. Provide a cost-value comparison between the Amazon cloud solution and locally-hosted hardware

Our Tools

- Amazon Web Services (AWS)
- Simple Storage Service (S3)
- Elastic Compute Cloud (EC2)
- DynamoDB
- QuickSight

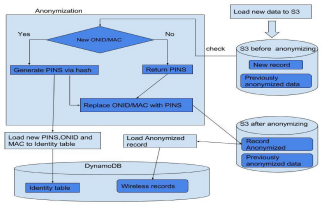
Programming Language

- **Python** script is used to import data from S3 to DynamoDB. The library Boto3 provides useful methods such as connecting S3 and DynamoDB.

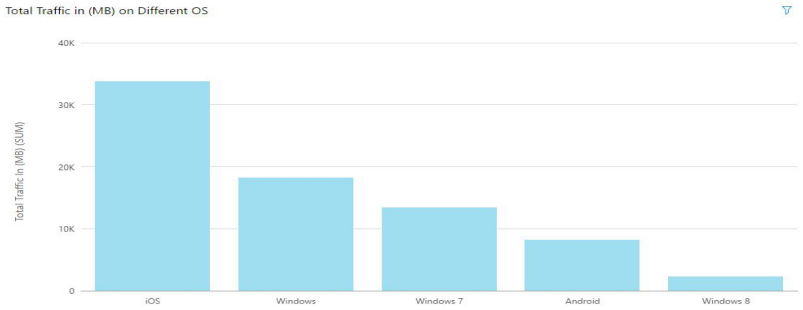
Workflow (loading new data)

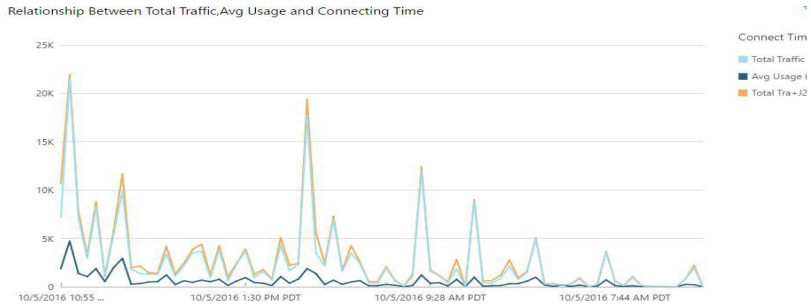
The workflow of loading new data into DynamoDB starts with S3. At first, the system stores the new record in S3. After that, it checks whether the identifier for the device or person is new. If the unique identifiers are new, a new system identifier is generated based on a hash of the original device and student identifiers. The original identifiers and the new identifier are then added to a restricted access, DynamoDB Identity table. In the record, the original identifiers are replaced with the new system identifier, and the anonymized record is loaded into a DynamoDB data table. If the unique identifiers are not new, they are replaced with the system identifier already available in the Identity table, and the record is added to the data table.

Workflow diagram



Results





Our team:

Zhaoheng Wang
wangzhao@oregonstate.edu
Zhi Jiang
jiangzh@oregonstate.edu
Isaac Chan
chani@oregonstate.edu

Client:

David Barber
David.Barber@oregonstate.edu

Oregon State UNIVERSITY
Information Services

Oregon State UNIVERSITY

7 PROJECT DOCUMENTATION

7.1 Project Overview

The purpose of this project is to implement a work flow of big data. The following digram shows loading new data into DynamoDB starts with S3. The new record is stored in S3 firstly. When it is imported to DynamoDB, we have to check whether the identifier for the device or person is new. If the unique identifiers are new, a new system identifier is generated based on a hash of the original device and student identifiers. The original device and student identifiers are loaded to Identity table. On the other hand, the original identifiers are replaced with the new system identifier, and the anonymized record is loaded into a DynamoDB data table. If the unique identifiers are not new, they are replaced with the system identifier already available in the Identity table, and the record is added to the data table.

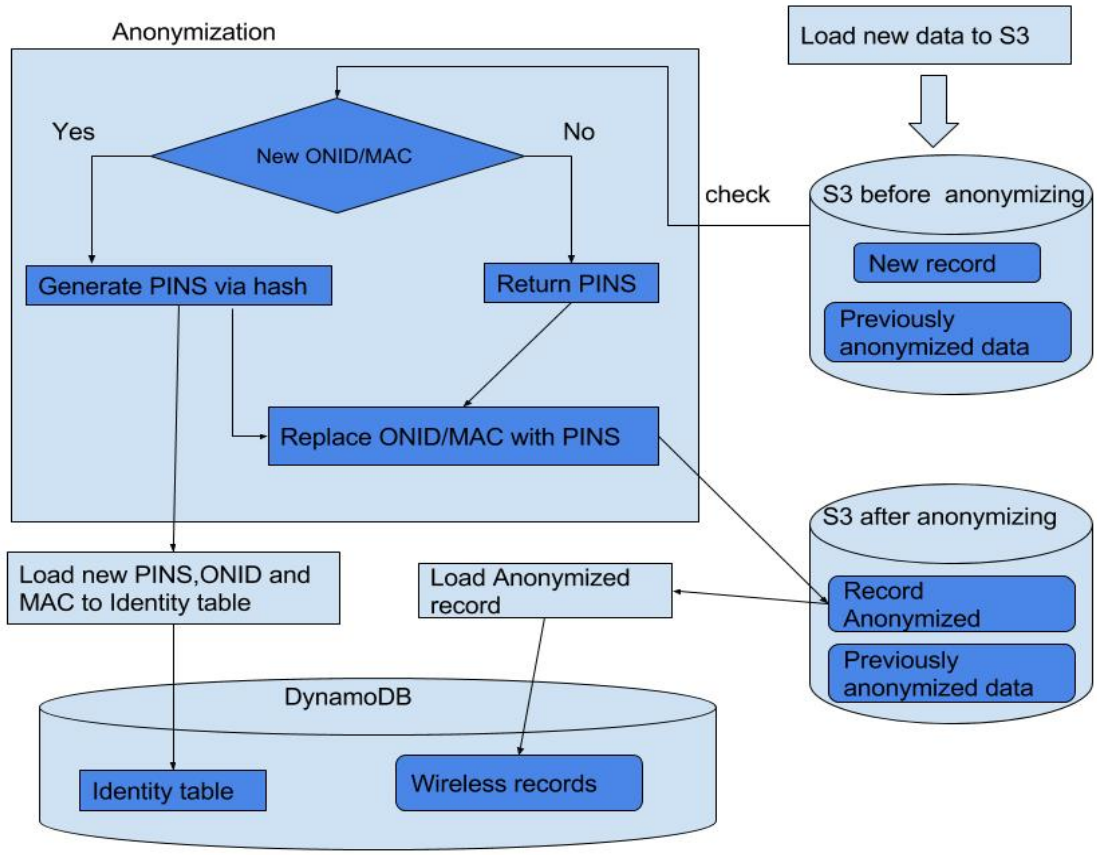


Figure 1: The workflow of importing new record

The workflow of the visualization starts from S3. After the DynamoDB table results have been stored back into S3, the data will be stored as CSV files in S3. These CSV files are important to the QuickSight because it requires the CSV files to generate the data set on the QuickSight console. After the data set have been created on the QuickSight console, the visualization will be easily generated. The figure 2 is the workflow of creating visualization on QuickSight.

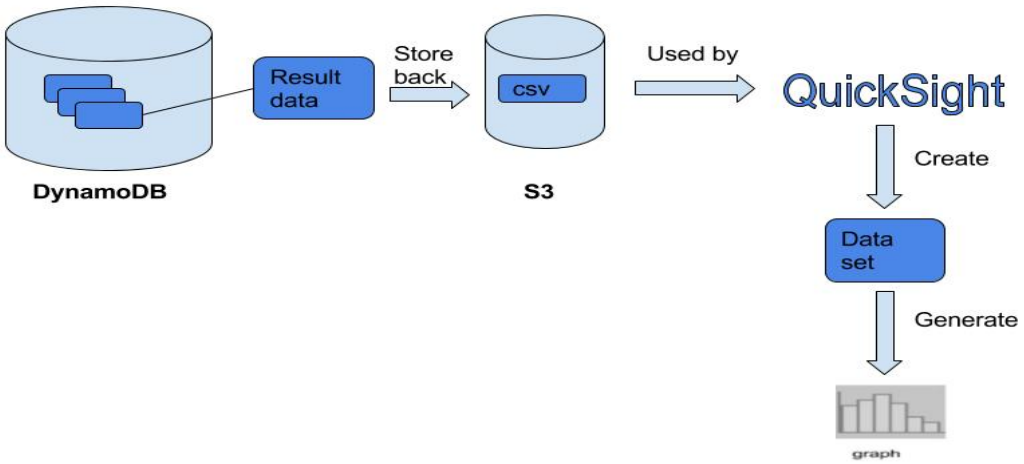


Figure 2: The workflow of creating visualization on QuickSight

7.2 Operating instructions

The project is working on the Amazon Web services, so it requires you must have a AWS account and this account has permission to access our project.

Load data from S3 into DynamoDB:

- Connecting into the EC2 instance, run the python file `create.py` for creating table
- After creating tables, run the `import.py` for loading data login in the DynamoDB
- The command you need to type is `python import.py [input bucket name] [input file name] [output bucket name]`. The input bucket is the path of input file, and output bucket is the path of anoymized record.
- Check the result on the console Stop the EC2 instance on EC2 console

Save the result back to the S3:

- Anonymized record have been created after run the import.py
- This record is stored in a certain bucket

Create visualization on the QuickSight

- Open the QuickSight console on AWS and Click New analysis button
- Choose New data set and click upload a file and choose the csv file which export from the ”Records” table
- Using the data set and select the conlumn which need to evalaute
- Select the rows that need to evaluate and choose the visual types
- The visual will be created on right side

8 The Approach to Learn New Technology

8.1 Useful Websites

As for this project, The first helpful website is the Amazon Web Services introduction page. From this website, we learned the general steps for big data analysis. Another helpful website we find is the Boto3 Document, Boto3 is the Python SDK for Amazon Web Services. In this project, we are using the file system s3 to store the sample data and the DynamoDB for storing the processed data. The QuickSight will be using for doing the analysis. In general, the project will start with parsing sample data from S3. After that, the sample data will be processed and store into the DynamoDB. When we start coding, we use Python to load sample data from S3 the data into DynamoDB. Therefore, the Boto3 website helps us figure out how to load the sample data from S3 into DynamoDB by Python.

8.2 Helpful People

The first helpful people on campus is our client David. He helps us on writing document. He gives useful feedback to us. So we can revise the document based on these feedback. The data analyst is also helpful. She helps us about permission for using the tool on AWS. Finally, our instructors and TA are very helpful. If we have problem with our project, we will ask them for help.

i

9 Personal Statement

9.1 Zhi Jiang

In the past three terms, I learned a lot of things from this project. From a technical point perspective, I learned and understood some basic workflow and framework of big data. I still learned some valuable tools of cloud computing. In our project, we mainly used Amazon Web Services. This cloud platform provides many useful services to complete all kinds of operation for big data. For example, S3 is used to store all original data as a data storage; DynamoDB is a kind of NoSQL database; EC2 is a virtual machine to connect other services and operate data on them. Moreover, according to this project, I learned the advantage of cloud computing on big data. For instance, cloud computing platform can avoid large capital expenditure on upgrades and hardware. On the other hand, it can also improve cost efficiency by more closely matching your cost pattern to your demand pattern. The purpose of the client setting up this project is he would like to compare the cost of cloud platform and local hardware. On the other hand, most cloud providers are extremely reliable in providing their services such as AWS. The security of data and project maintenance can be guaranteed significantly.

The most thing I did in project work is research. In our project, we must use several services on AWS. I have never used them before, so I started to do research for those services. I read developer guide and learned their basic operation method hence doing research is a very helpful point in project work. In fact, a programmer will always face problems he or she has never seen, so the research is the best approach to solving it. Moreover, I am glad that I had carefully considered the alternative solution. For some reasons, I had to give up the prior solution in our original plan and chose the alternative solution. Eventually, we still completed my portion by this alternative solution. In the whole developing process, the accident may occur at any time. A good alternative can help developer deal with emergencies and ensure that processes are running smoothly.

I still learned many things from project management. I believe that the schedule should be the most important component in the project management. As we know, there are many uncontrollable factors in the development process such as new requirements. If you have an unreasonable and chaotic schedule, any uncontrollable factors will make the influence on your development. On the contrary, the perfect plan can let you know what you did, what you should do right now, and what you left. You can understand the status of your project.

Teamwork is a necessary basis for success. By getting along with my teammates in these three terms, I found many benefits from collaboration. First of all, teamwork can increase the development efficiency, but the precondition is every member of this team should do their duty. If someone is lazy on the team, it will affect the performance of the entire team. Secondly, teamwork can more efficiently produce the best idea. A team comes up a more effective solution than one person dealing with on the same problem because each member has different knowledge and perspective. Meanwhile, I still learned that it was important to respect ideas of everyone because members have different habits and work styles. When we have different opinions with each other, the most appropriate way is to calmly discuss rather than quarrel. Thirdly, I feel I can learn more knowledge in teamwork because everyone has some skills that you dont know such as

programming language or an editing tool.

If I could do it all over, I believe user interface should be a good choice for this project. After visualization of analysis, the user interface can show the results to students. On the other hand, students are probably able to enter their person information to know their data in the campus.

9.2 Zhaoheng Wang

The technical information I learned is the process of big data analysis on Amazon Web Services. The project for my group is to deal with the big data on Amazon Web Services. It is brand-new project for me. It is difficult to figure out how to start this project without any background knowledge. Fortunately, the homepage of Amazon web Service gives the general steps about how to deal with the big data on Amazon Web Services. In general, the big data will be generated from multiple sources and the data file might be unreadable. Therefore, the first step is called parse data. The purpose of this step is to format the data file and make it readable. After that, the parsed data needs to store and organize the storage place such as database. After the data has been organized, the visualization graph for the data will be created for analyzing.This is the technical information I learned from the project.

The non-technical information and project work I learned is how to figure out a brand-new question. I believe it is impossible to solve every problem in the future work just based on the current knowledge I have. Therefore, the steps for solving a brand-new question will be a very important skill to have. In this project, I started with discussing with client. After discussion, I summarized the purpose and requirements for this project. Then, I started to search keywords such as AWS, big data, analysis and many others for this topic. According to searching information, I narrowed down the requirements and designed the general workflow for the whole project. After that, I discussed the project with my teammates and client to ensure that I was going the right direction. Personally, I think this skill is one of the most important skill I learned from this project.

As for the project management and team working, I have learned that good communication will make team work more efficient. Through good communication, I work with my group better. For example, one of my teammates has experience working on big data analysis. When I had some question about the project, I asked his advice. Another skill I learned from team is to manage time and make the schedule for the project. As for my group, our personal schedule conflicted with each other most of time. Because of this, we made the group schedule for the project and this schedule helped us to finish the requirements on time.

If I could do it all over, I would like to research more information about the project such as workflow and tools because it is very helpful to have more background knowledge before starting a project. Without any background knowledge, I spent more time on searching than I would have liked. Take my project as an example, the project my group worked on is to deal with the big data on the Amazon web service. At first, I had no idea about how to start the project. Therefore, I searched online about the big data analysis. However, there are many ways to do that so it is hard to decide which is the best way for dealing with the big data. Besides, I was not familiar with Amazon web service. Thus, I could only guess the step before starting this project. Because of this, the process for asking the permission for using the tool was complicated.

9.3 Isaac Chan

I learned a bit more about different databases. I was previously aware of relational databases, specifically the more widely known AWS offering, Redshift. NoSQL with DynamoDB was new to me. However, I found that DynamoDB seemed to just a lazier version of Redshift because there was no upfront schema preparation, which might cause problems further on. Most of the other things, including workflow design, AWS utilities, I was already familiar with.

I think the most valuable non-technical information I learned was to constantly communicate with a client. I'm not sure if it was the balance of other classes or what, but during winter term, our communication with our client slipped a lot. We only met once or twice the entire term and he probably had no idea how we were progressing with his project.

In this project, it was very important to communicate with my group. We met a lot and were in constant communication, which helped us deliver on time and keep the schedule moving. I found it very helpful to develop a rough schedule and divide tasks early on. This meant that each of us was accountable for a portion and we progressed at a satisfactory rate.

If I could do it over again, I think I would have finished the project earlier on. The project itself was not very complex, but I felt like we were tied down by the structure of the class and how we were "supposed" to progress with it. Because of this, the project work was stretched out over winter and part of spring term, and it felt like we weren't doing a lot.