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

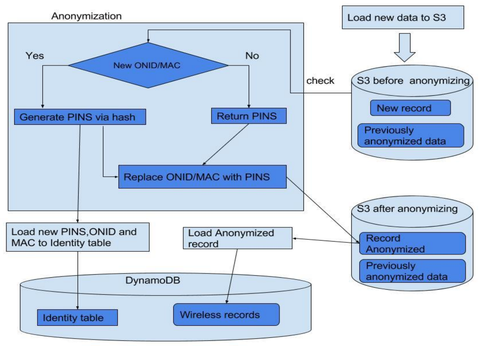
Programming Language

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

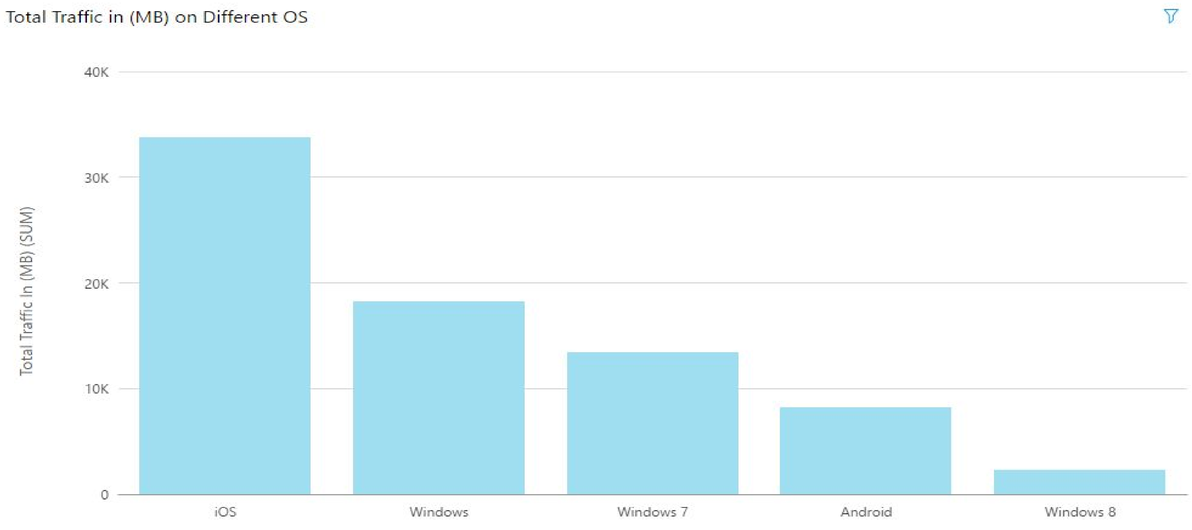
Workflow (loading new data)

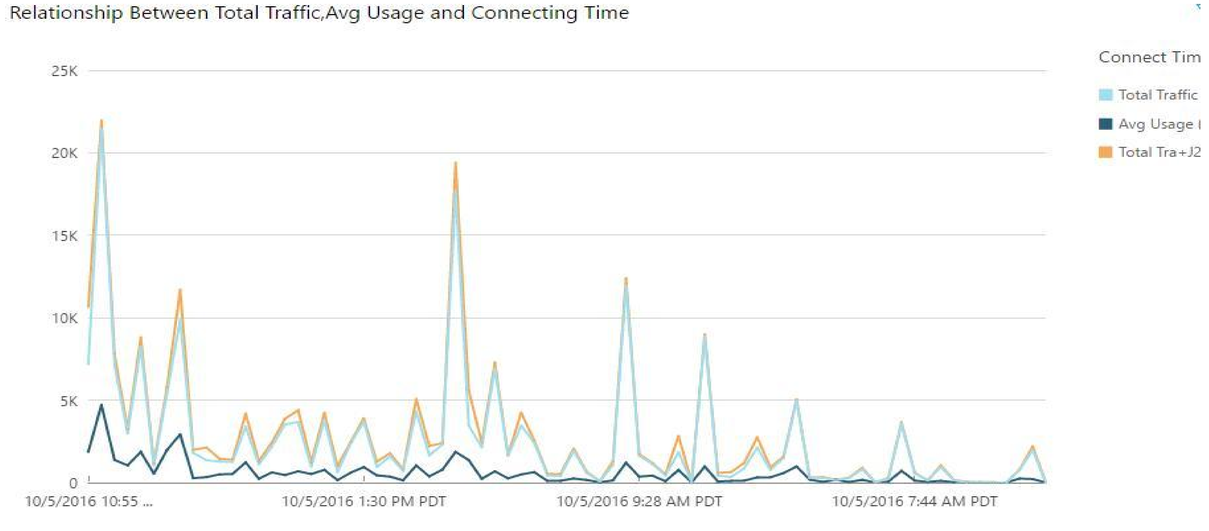
The workflow of loading new data into DynamoDB starts with S3. At first, the system stores the new record in S3. After that, it checks whether the identifier for the device or person is new. If the unique identifiers are new, a new system identifier is generated based on a hash of the original device 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