A Multi-Tenancy Cloud-Native Digital Library Platform

Yinlin Chen, Jim Tuttle, William A. Ingram

{ylchen, jim.tuttle, waingram}@vt.edu

Information Technologies and Services
Virginia Tech Libraries
Agenda

• Cloud-native concept
• Virginia Tech Digital Library Platform (VTDLP)
• Design strategy
• Architecture overview
• Implementation overview
• VTL experiences
Cloud-native Concept

- Entire infrastructure is deployed in the Cloud (AWS)
- Platform is composed of a suite of microservices and managed services
- Focus on the business logic and workflow
- Utilize the advantages provided by the Cloud
New services to Digital Library Platform
- ID Minting service, Access Service, Metadata service, ...

Migrating legacy services to Digital Library Platform
- IAWA, VTechWork, ...
VT DLP Overview

Preservation staging
- VtechWork ETDs
- IAWA Images
- SW Virginia
- Others

Storage
- Amazon S3
- APTrust

Presentation
- IAWA
- BeyondVT
- Others

Batch Metadata Service
ID Minting Service
Metadata Service
Resolution Service
Serialization Service
Other Services
Design Strategy

• Cloud native (AWS ecosystem)
• Microservice/SOA (AWS lambda)
• Serverless (AWS managed services)
• CI/CD Pipeline
• Caching as much as possible
  – Static files
  – Lambda functions
• Automation as much as possible
  – Infrastructure as code
  – No manual provisioning or managing servers
Software stacks

- React
- AWS Amplify
- AWS AppSync
- Node.js
- Python

Web App

Microservice (AWS Lambda)
Preservation Pipeline

- Checksums
- Fixity
- Virus Scan
- AWS S3
- APTrust
- PREMIS

BagIt → Apache Airflow
Lambda Example – Metadata file

1. File upload to S3
2. S3 triggers a Lambda function
3. Lambda function parses file content and inserts/updates record in the DynamoDB
1. Data modifications in DynamoDB will trigger a Lambda function.
2. Lambda function captures changes and updates Amazon ES.
The International Archive of Women in Architecture

• A level 0 compliant image server using Amazon S3 and Amazon CloudFront
• Tiles images, manifest JSON files, and etc.
• Terabytes of scan images to be processed
Image processing workflow

Amazon S3

Raw images

Amazon EC2

Amazon Elastic File System

AWS Lambda

CloudWatch

Rule

Batch Job – image set 1

Batch Job – image set 2

Batch Job – image set 3

Batch Job – image set N

Tiles & Manifest

AWS Batch

Amazon S3
Batch job - IIIF_S3 Docker

- Command
- Parameters
- Environment variables
- vCPUs
- Memory

AWS Batch

Amazon S3

Tiles & Manifest

IIIF

Amazon Elastic File System
CI/CD with AWS

1. Developers
2. GitHub
3. AWS CodeBuild
4. Amazon S3
5. AWS CodePipeline
6. AWS Lambda
7. Amazon API Gateway
8. AWS CloudFormation
Cloud benefit - Backup examples

- **S3**
  - Amazon S3 is 99.999999999% durability and 99.99% availability.
  - On average, may lose one of 10,000 objects every 10 million years or so.
  - Cross-region replication

- **DynamoDB**
  - Point-in-time recovery (Last 35 days)
  - On-Demand Backup (Stored in S3)

- **ElasticSearch**
  - Daily snapshots (Last 14 days)
  - On-Demand Backup (Stored in S3)
VTL Experiences

• Entire development team is AWS certified

• One AWS Certification Subject Matter Expert (SME)
• AWS trainings and conferences
• Thinking and implementing new ideas the Cloud way
Q & A

Thank You!