Data Analytics based on Python & AWS

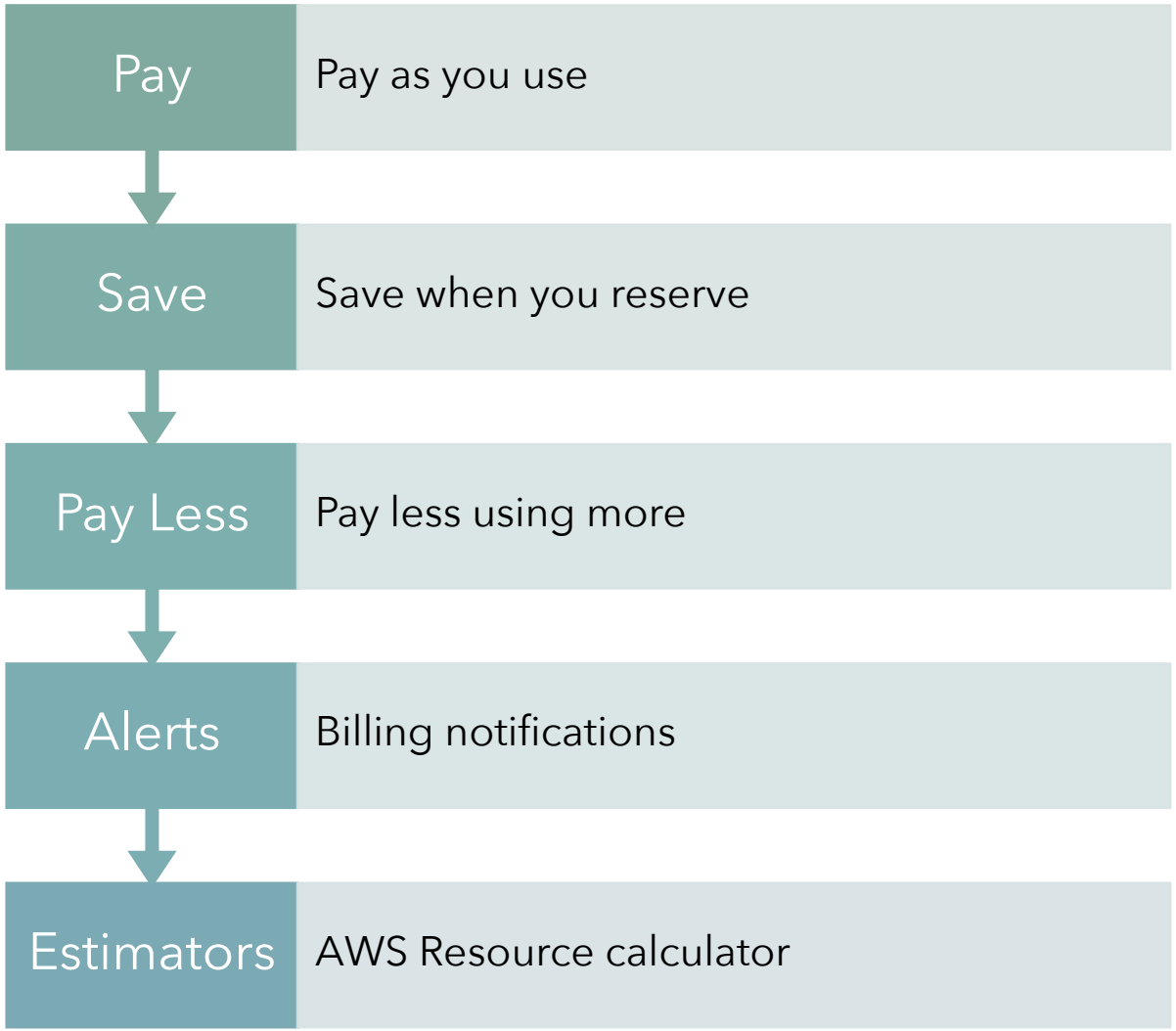
By Gezhi Cheng 8458204,
Haowei Lee 8460698,
Ziyi Liu 8467087,
Yang Lu 8382191,
Chaitanya Madduri 8459705
(sorted by first letter of last name)

Agenda

- Introduction
- Intro to Cloud Services
- Overview on AWS
- Demo of Lamba Functions
- Q&A Feedback

Why Cloud Services?

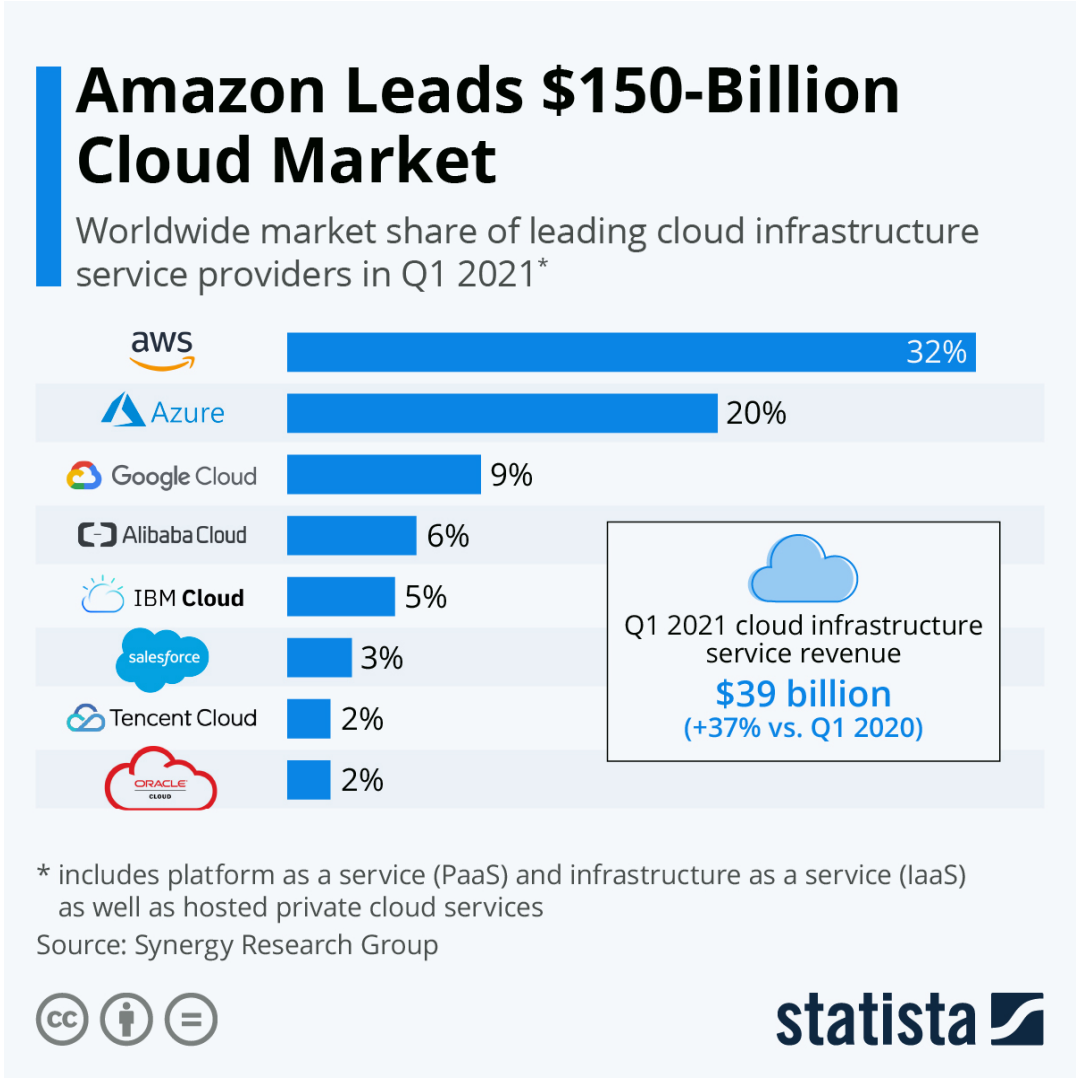
Pricing


Pay
Pay as you use
Save
Save when you reserve
Pay Less
Pay less using more
Alerts
Billing notifications
Estimators
AWS Resource calculator

How to choose a Cloud Service ?

1. Popularity – Competency
2. Cost incurred
3. Services Provides & Locations
4. Free Tier Services

Cloud Market Share



(Source: https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/)

Machine Type Vs Cost

Machine Type	AWS	Azure	GCP
Smallest Instance / Basic Configuration 2 CPU + 8 GB RAM + 256GB HDD	US$69 per month.	US$70/month.	US$52/month.
Largest Instance	3.84 TB of RAM and 128 vCPUs will cost you around US$3.97/hour.	3.89 TB of RAM and 128 vCPUs. It costs around US$6.79/hour.	largest instance that includes 3.75 TB of RAM and 160 vCPUs. It will cost you around US$5.32/hour.

(Souce: https://intellipaat.com/blog/aws-vs-azure-vs-google-cloud/)

Services provided & Locations

Cloud Service	AWS	Microsoft Azure	Google Cloud
Total Services	~222	~200+	90
Data centres	84 Availability Zones within 26 geographic regions	54 Azure regions available in 140 countries.	88 Zones and 200+ Countries
Always free Services (always)	~33	~25	~20+

AWS Challenges

Understanding the services and usage

Billing Can be Confusing

Limiting of resources by region

AWS Free Tier

STORAGE

Free Tier — 12 MONTHS FREE

Amazon S3

5 GB

of standard storage

Secure, durable, and scalable object storage infrastructure.

- **5 GB** of Standard Storage
- **20,000 Get Requests**
- **2,000 Put Requests**

COMPUTE

Free Tier — ALWAYS FREE

AWS Lambda

1 Million

free requests per month

Compute service that runs your code in response to events and automatically manages the compute resources.

- **1,000,000** free requests per month
- **Up to 3.2 million seconds** of compute time per month

- A service that can run, manage, share and employ code.
- Traditionally:
- For managing and sharing codes: Github, GitLab, Bitbucket……
- For employing Python codes: run from local machine.

What are we looking for? (Objectives)

- A web service that provides secure, resizable compute capacity in the cloud
- Allows you to obtain and configure capacity with minimal friction
- Provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment
- Security, CPU & Memory utilization …
- Compute tier, Optimized Memory, General purpose

EC2

(Elastic Compute Cloud)

AWS Lambda

- A compute service that lets you run code without provisioning or managing servers ("Serverless")
- Code
- Lambda function (output:ARS)
- Memory(128GB/2056GB)

Decision Tree

(Source: https://servian.dev/choosing-a-suitable-aws-compute-product-a-decision-tree-1dc46caef824)

Dimension	Lambda	EC2
Cost	Charged by event	Charged by time
Setup & Management	Easier, low flexibility, low maintenance	Pick by yourself, high flexibility, high maintenance
Performance	Maximum 3 GB memory, significant timeout constraints of 15min	Maximum limit is 1,000 concurrent executions, infinite running time
Security	Most of the onus on the AWS side	Full control over system-level security
Integration	API Gateway, SNS...	Self-setup

(Source: https://lumigo.io/blog/aws-lambda-vs-ec2/)

Why Lambda?

Which One Should We Choose?

Characteristics of Lambda

- **Event-driven applications**
- **Infrequently-accessed applications or scheduled jobs**

Data Hub

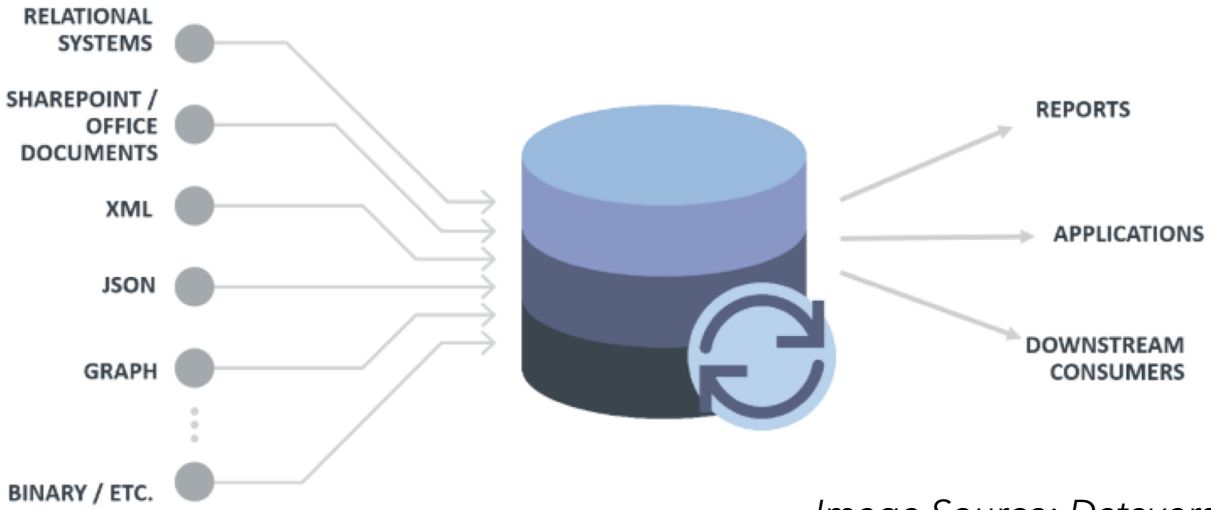


Image Source: Dataversity

Definition:

A modern data storage system

Consolidate and store enterprise-wide data

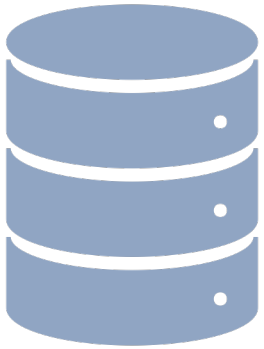
Allows companies to push data into other systems for further analysis

How does Data Hub work?

- Transfer data to the data hub repository
- Centralize and Standardize data
- Analyze data
- Consume data

Why Data Hub?

Connect all data touchpoints and make the data available at a central location (data integration)



Data Hub

Amazon Redshift

- The most widely used, petabyte-scale data warehouse service in the cloud
- Automatically patch and back up customers data warehouse
- Use replication and continuous backups to enhance availability and improve data durability
- Automatically recover from component and node failures

Why Redshift

- Significantly lower the cost and operational overhead
- With Redshift Spectrum, easily to analyze large amounts of data in its native format without requiring customers to load the data

Data Hub

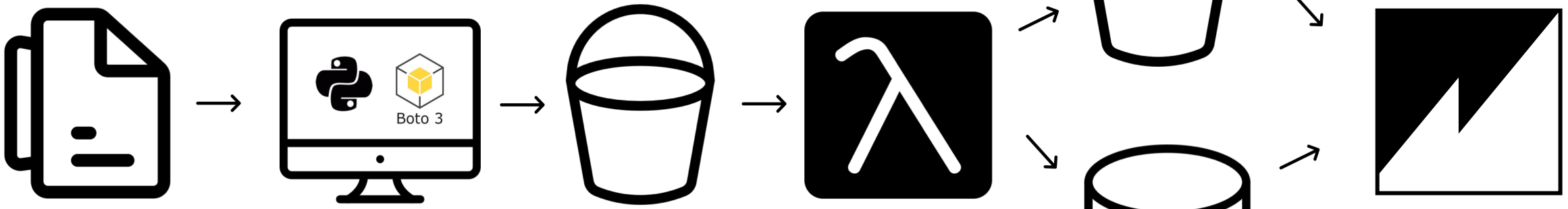
Amazon Simple Storage Service (S3)

- The largest and most performant object storage service for structured and unstructured data
- Customers can use native AWS services to run big data analytics, artificial intelligence (AI), machine learning (ML), high-performance computing (HPC) and media data processing applications

Why S3?

- 99.999999999% data durability
- Automatically create and store copies of all uploaded S3 objects

Pipeline



- Files
- Raw data

Scripting
- Local Machine
- Lambda
- EC2

S3
Bucket
(storage)

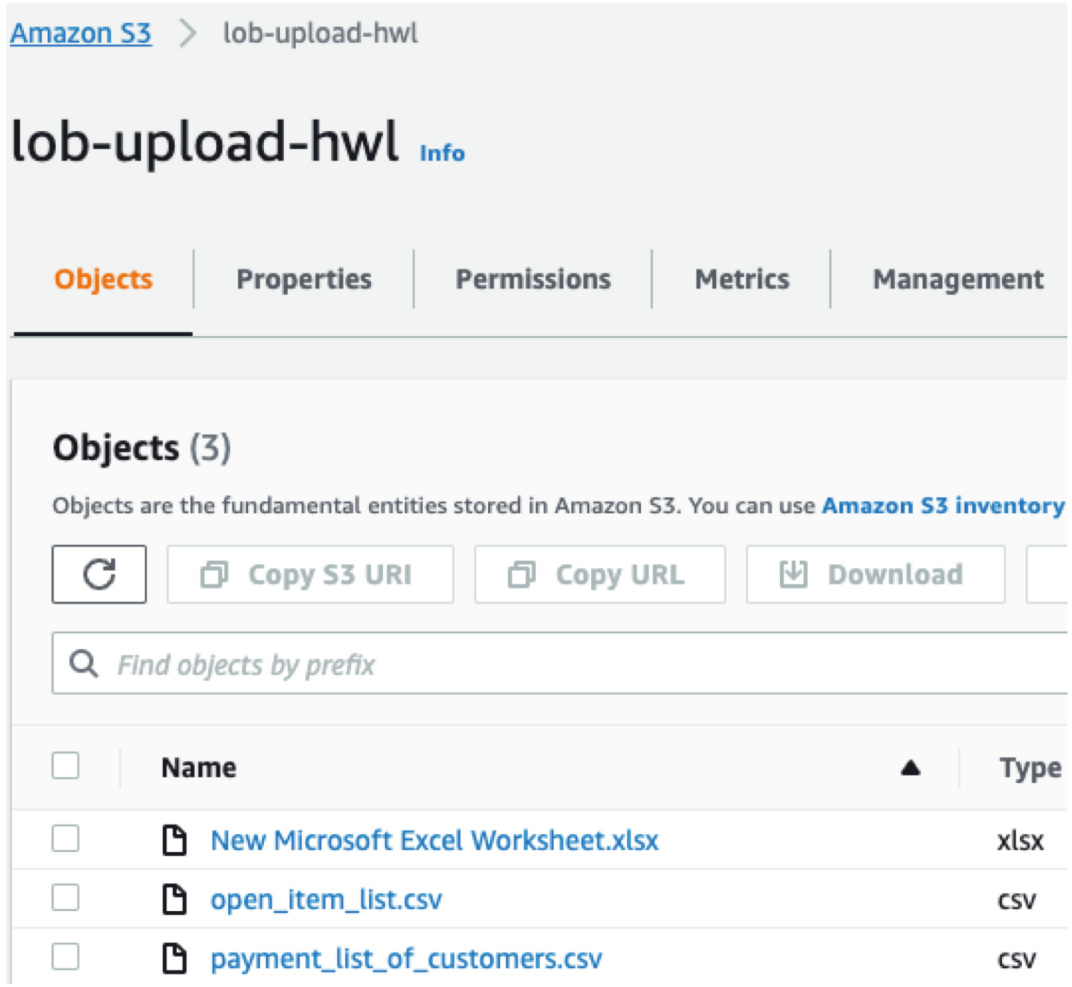
Lambda
Function
(execution)

S3
Bucket
(short run)

Database
(long run)
- Redshift
- DynamoDB

QuickSight
(Visualisation)

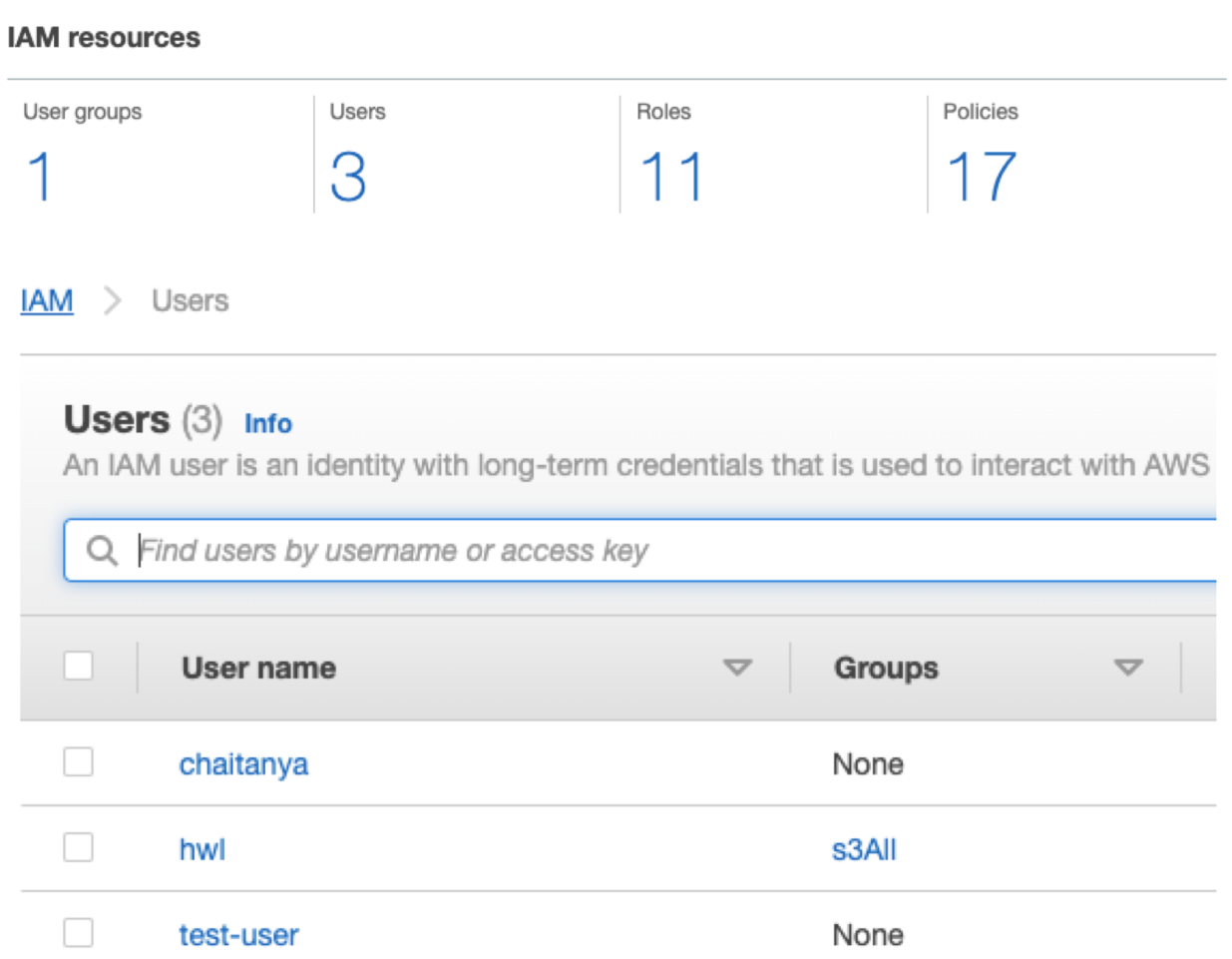
Demo



S3
(Simple Storage Service)

To create a bucket for **file storage**

- **Buckets** -> Folders
- **Objects** -> Files
- Flow
  Create -> *Upload*



IAM
(Identity and Access Management)

To set **permission** to the aws resources, such as S3

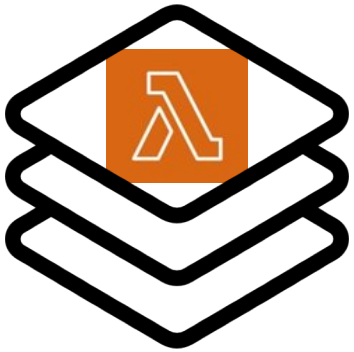
- User Groups
  - long-term credentials
- Roles
  - permissions for specific validation / short durations
- Policies
  - Permission Definition

Lambda Function

Trigger

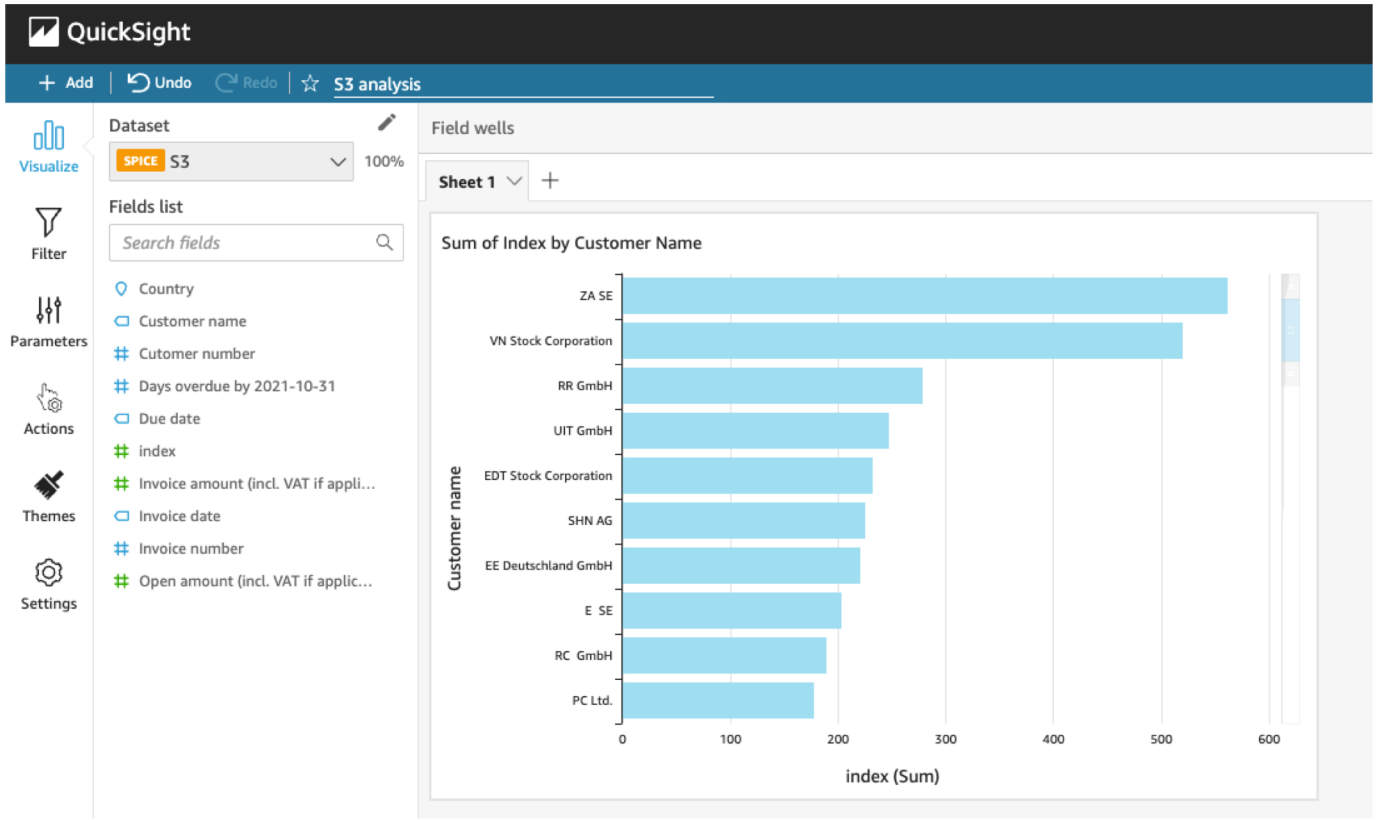
Layers

pandas



Lambda Function

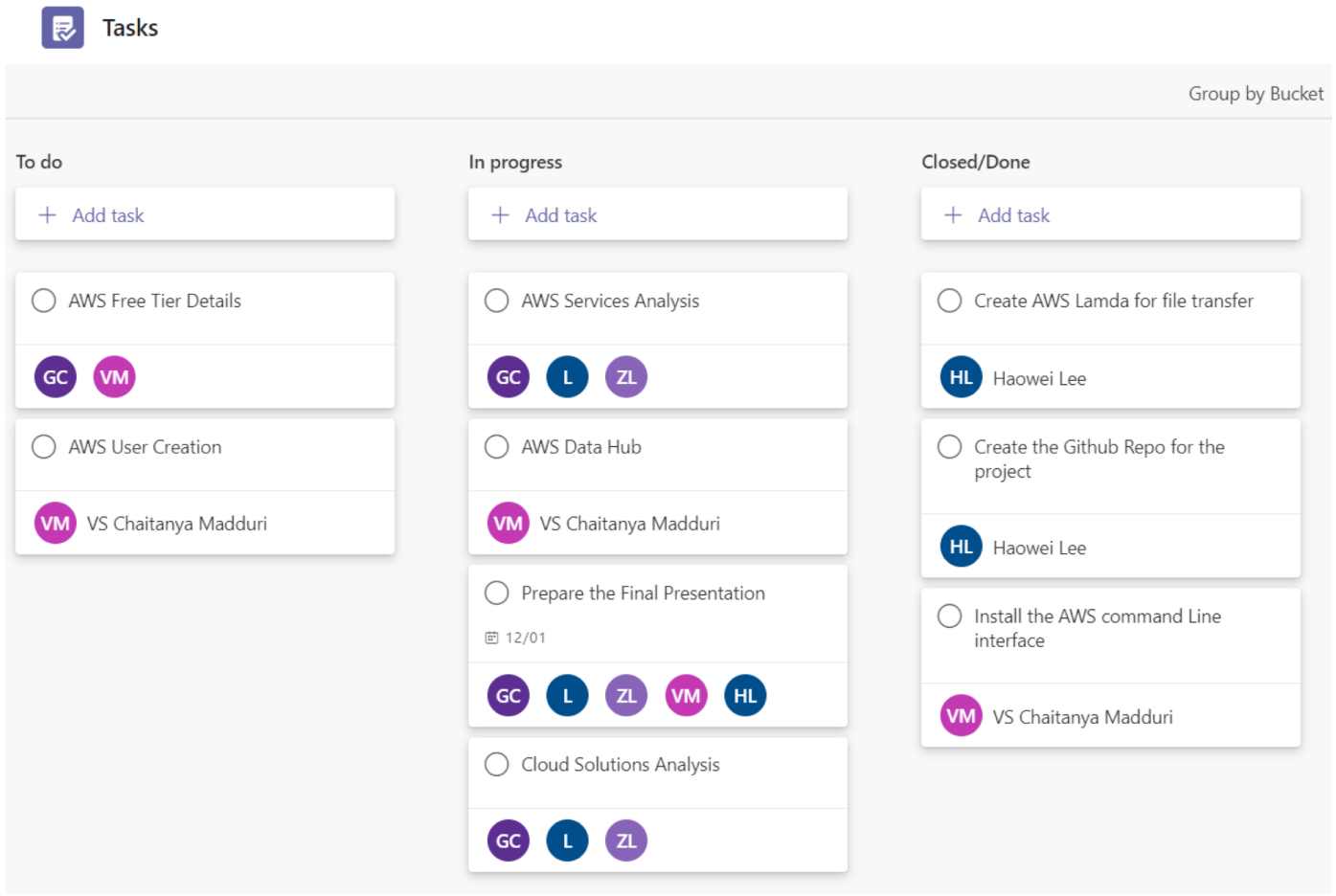
- Trigger
  - Source of events
- Layers
  - ensure packages dependency
  - avoid repeated tasks
  - decrease starting time of lambda function
- Flow
  Create -> *Test* -> *Deploy*



QuickSight

A BI tool for

- Visualizing data
- Business analytics
- Demo



Project Management

- Used the Teams Task planner for project management.

6. Q&A

Appendix

- https://geomotiv.com/blog/best-cloud-services-comparison/
- Amazon Free Tier - https://aws.amazon.com/free
- https://sados.com/blog/aws-benefits-and-drawbacks/
- Github::
  - *https://github.com/leehaowei/team5-lob-mads*

Thank You