Cut+Dry

AWS Services

2025-01-23 @ CSE

What you will learn?

1. Create your first account.
2. AWS general tools.
3. Most useful services.
4. Some design examples.
5. You will design a solution.

What is IaaS?

IaaS, or Infrastructure as a Service, is a cloud computing model that provides on-demand access to computing resources such as servers, storage, networking, and virtualization.

IaaS is attractive because acquiring computing resources to run applications or store data the traditional way requires time and capital.

Organizations must purchase equipment through procurement processes that can take months.

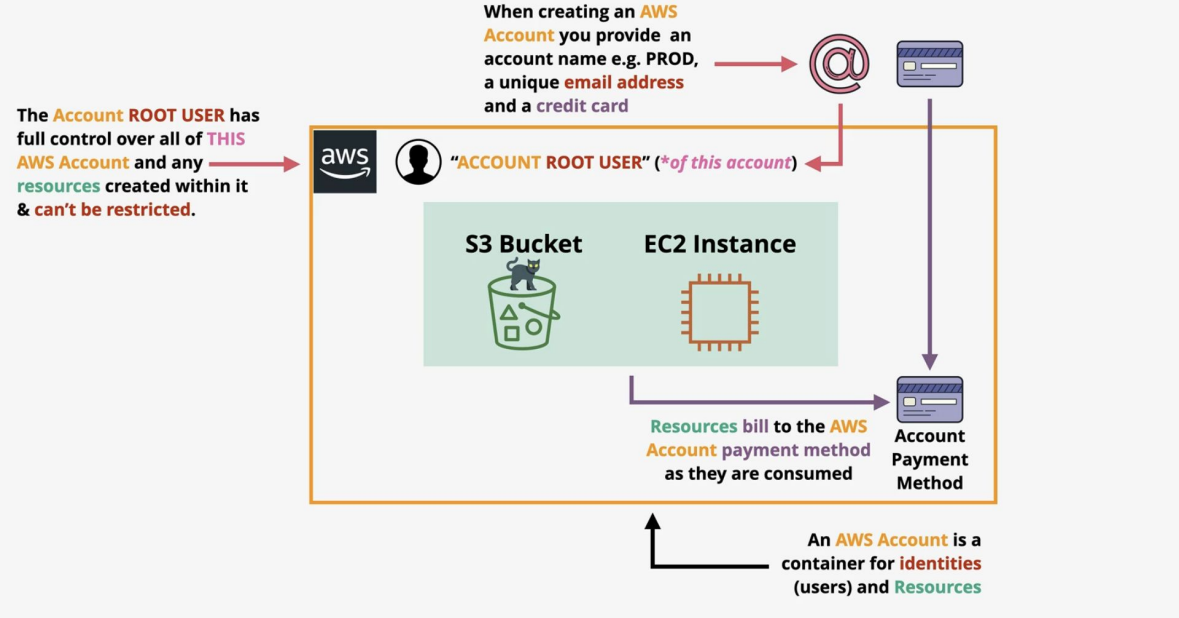
They must invest in physical spaces, typically specialized rooms with power and cooling.

And after deploying the systems, they need IT professionals to manage and maintain them.

Let’s setup your first account

- Need - email and a payment method
- Sign Up -> AWS Console - Signup
- Create two accounts - PROD & DEV
- Create IAM users
    - Dev account + root user -> has full access
    - IAM user to use daily -> grant permissions as required
- Add MFA
- Create a budget (Zero Budget) and always remember to cleanup
- Configure other billing preferences

AWS Account



learn.cantril.io

AWS CLI

The **AWS Command Line Interface (CLI)** is a tool that allows users to interact with AWS services via the command line or terminal, offering a scriptable and efficient way to manage cloud resources. Here are some key uses:

1. Configuration

$aws configure

2. Resource Management

$aws s3 ls

3. Querying and Filtering Data

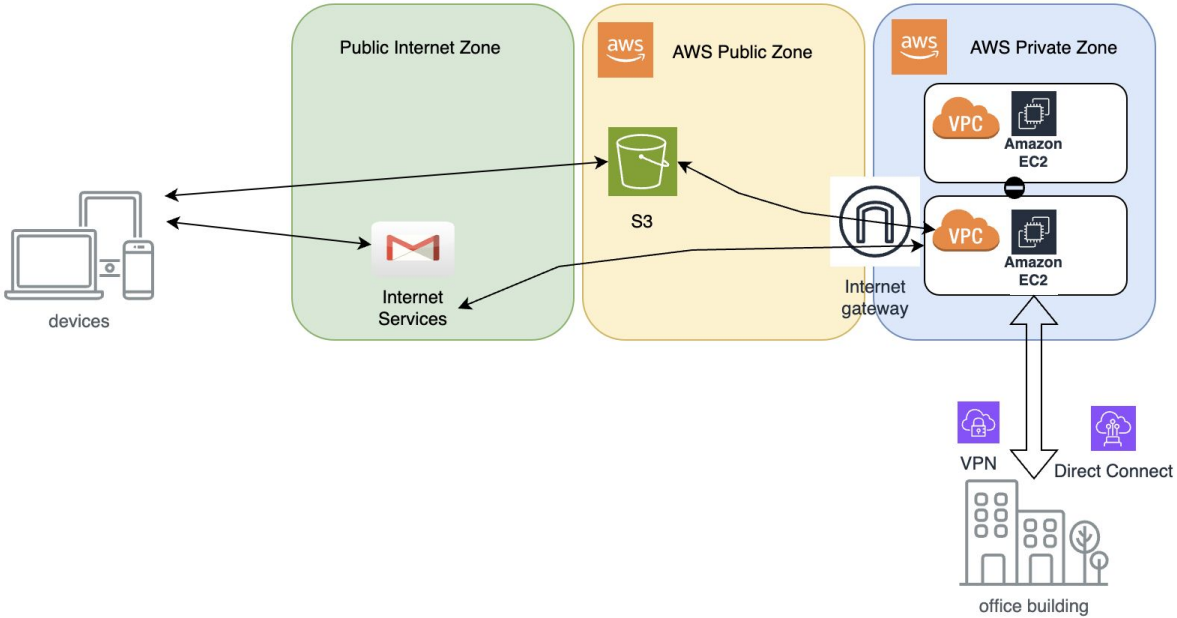
$aws ec2 describe-instances --filters "Name=instance-state-name,Values=running"

4. Automation with scripts

5. Cost Optimization and Monitoring

Public vs Private Services

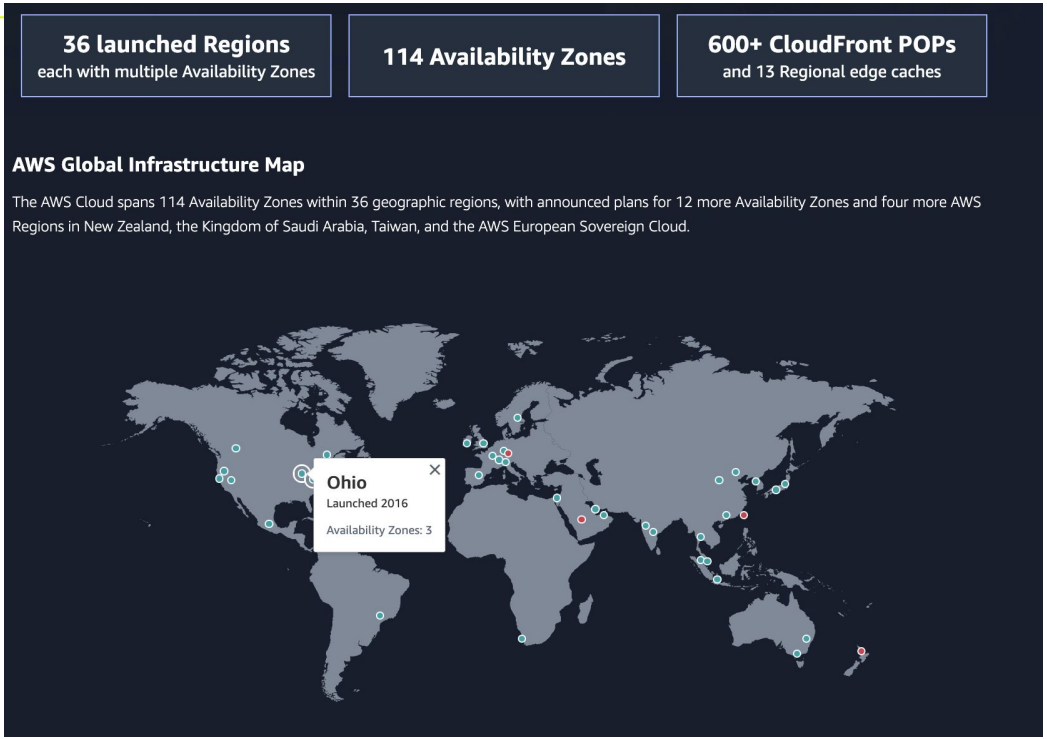
This is only for Networking perspective



AWS Global Infrastructure

1. **AWS Regions** - Full deployment of AWS Infrastructure
   - Geographical separation - Isolated Fault Domains
   - Geopolitical separation - Different governance
   - Location control - Performance
   - Each region has a **Region Code (ap-southeast-2) and Region Name (Asia Pacific (Sydney))**
2. Regions consist **Availability Zones(AZ)**
   - Logical resource separation / no clear visibility
   - For Resiliency
3. What is Service Resiliency?
   - Global Resilient -> IAM, Route53, S3, KMS, CloudFront
   - Region Resilient -> RDS
   - AZ Resilient -> EC2 etc more prone to failure

AWS Global Infrastructure



AWS VPC

- VPC = A Virtual Network inside AWS
- Private and isolated unless you decide otherwise
- Two types
  - Default VPC
  - Custom VPC
- VPC considerations
  - What size should be the VPC
  - Try to predict future
  - VPC IP structure- subnets
- VPC router
  - Every vpc has a vpc router
  - Route traffic between subnets
- Internet Gateway (IGW)
  - Region resilient GW attached to a VPC
  - Runs within AWS Public Zone
  - Manage traffic between the VPC and the internet or AWS Public Zone

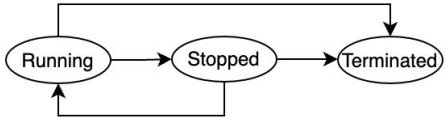
Amazon VPC

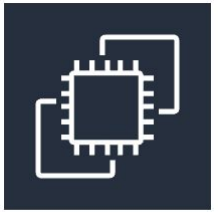
AWS IAM

- Handles Identity and Access Management for the AWS services
- **IAM Policies**
  - Policy documents with 1+ statements
  - Inline policy updates - update to each user/group
  - Managed policies - reusable,low management overhead
- **IAM Users**
  - Identity used for anything requiring long-term AWS access - Humans, Service Accounts
  - 5000 IAM users per account (need to consider in the design)
- **IAM Groups** - containers for users
- **IAM Roles -** provides required permissions that a service needs to interact with other AWS services

AWS EC2

- IaaS -> provides virtual machines => compute instances
- Private - used vpc networking
- AZ resilient -> instance fails if AZ fails
- Different instance sizes and capabilities
- On-demand billing
- Local on-host storage or Elastic block store (EBS)
- Instance state transition as below





- Can bootstrap EC2 using user data/ easy for automations
- Placement groups - provides infrastructure isolation (Spread/Partition)

AWS AMI

- Can be used to launch EC2 instance
- AWS or community provided
- Also can be taken from the Marketplace (including commercial software)
- Has regional unique id
- AMI Lifecycle
  - Launch
  - Configure - AMI Baking
  - Create new Image
  - Launch
- AMI can’t be edited
- Can be copied across regions

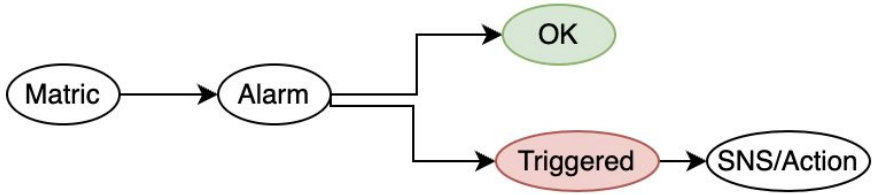
AMI

AWS S3

- Global storage platform - Regional resilient
- Public service, unlimited data and multi-user
- …any data type supported
- economical
- Accessible via UI/CLI/API/HTTP
- Store as objects in buckets (object storage- not file or block storage)
- Great for large scale data storage, distribution
- INPUT/OUTPUT for many AWS products
- Supports object versioning
- Support encryption - SSE-S3,SSE-C,SSE-KMS
- With life cycle events , can move to cheaper storages

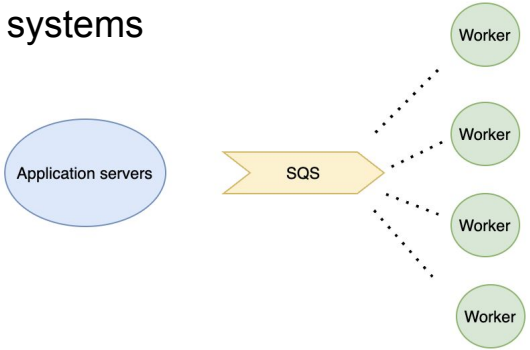
AWS CloudWatch

- Collects and manages operational data , logs
- Matrics - natively handled, some metrics available with CloudWatch Agent - ex EC2
- CloudWatch Logs
- CloudWatch Events - AWS services
- Alarms



AWS SQS

- Public, fully managed service
- Provides highly-available queues - two types -> Standard and FIFO
- Messages up to 256 kb in size- for large data links can be used
- Dead letter queues can be used for problem messages
- Auto scale groups can scale and lambdas can be invoked based on queue length
- Used to decouple systems



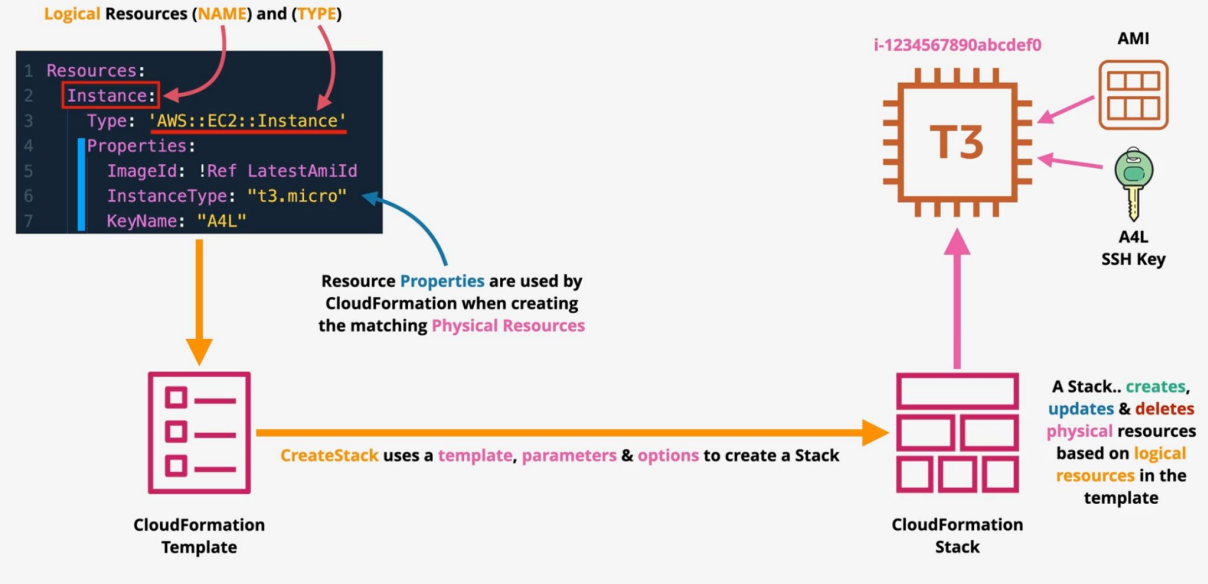
AWS SNS

- **Message Distribution**: AWS Simple Notification Service (SNS) enables pub/sub messaging for delivering messages to multiple subscribers.
- **Protocol Support**: Supports multiple delivery protocols, including HTTP/S, email, SMS, and AWS Lambda.
- **Event-driven Architecture**: Integrates seamlessly with AWS services like S3, EC2, and Lambda for event notifications.
- **High Scalability**: Automatically scales to handle millions of messages per second.
- **Reliability**: Provides durable message delivery with retry logic and message archiving using Dead Letter Queues.
- **Access Control**: Ensures secure messaging with IAM policies and encryption options using AWS KMS.
- **Cost-effective**: Pay-as-you-go pricing model with no upfront costs or minimum fees.

AWS CloudFormation

- **Infrastructure as Code**: AWS CloudFormation enables defining and provisioning infrastructure using templates.
- **Simplified Management**: Manages AWS resources as a single stack for easier deployment and updates.
- **Declarative Syntax**: Uses YAML or JSON to define resources and configurations.
- **Automated Scaling**: Automatically provisions and adjusts resources as defined in the templates.
- **Drift Detection**: Identifies changes in resources that deviate from the original template.
- **Cross-region Support**: Allows managing resources across multiple AWS regions using StackSets.
- **Cost Management**: Provides visibility into resources, helping track and optimize costs effectively.

AWS CloudFormation



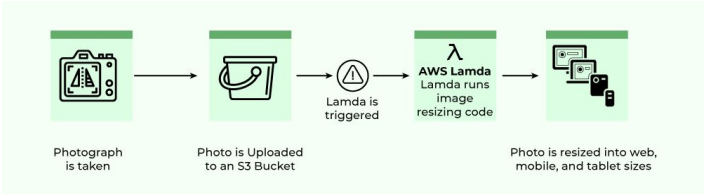
learn.cantril.io

AWS Route53

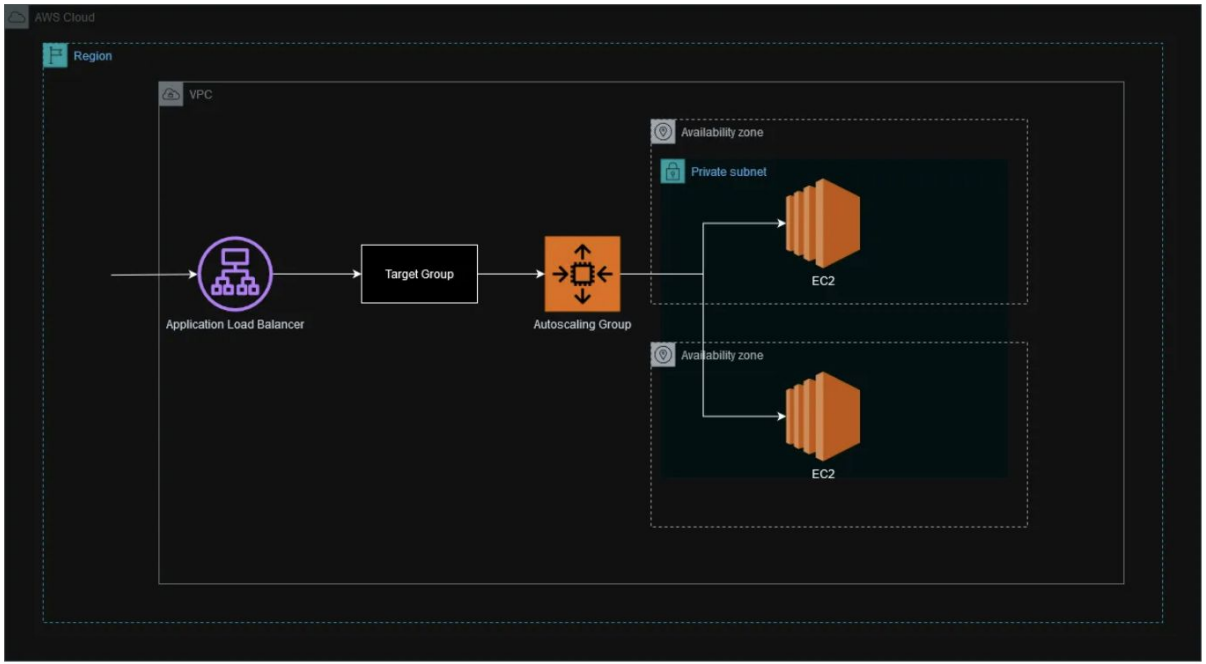
- **Global DNS Service**: Amazon Route 53 provides scalable and reliable Domain Name System (DNS) services.
- **Domain Registration**: Supports domain registration and management for a wide range of TLDs.
- **Routing Policies**: Offers various routing options, including latency-based, geolocation, and weighted routing.
- **Health Checks**: Monitors endpoint health and automatically routes traffic to healthy resources.
- **High Availability**: Ensures low-latency and fault-tolerant DNS resolution with a global infrastructure.
- **Integration**: Seamlessly integrates with AWS services like CloudFront, S3, and Elastic Load Balancer.
- **Scalable and Secure**: Designed to handle large query volumes with DDoS protection via AWS Shield.

AWS Lambda

- Function-as-a-service (FaaS)
- A key part of serverless architecture
- Short running and focused
- Highly scalable and fully managed service
- Runs a lambda function - a piece of code
- Runtime can be provided and functions are loaded in a runtime environment - zipped 50mb and unzipped 250mb
- Billed for the duration that a function runs
- Runs for 15 minutes and then timeout



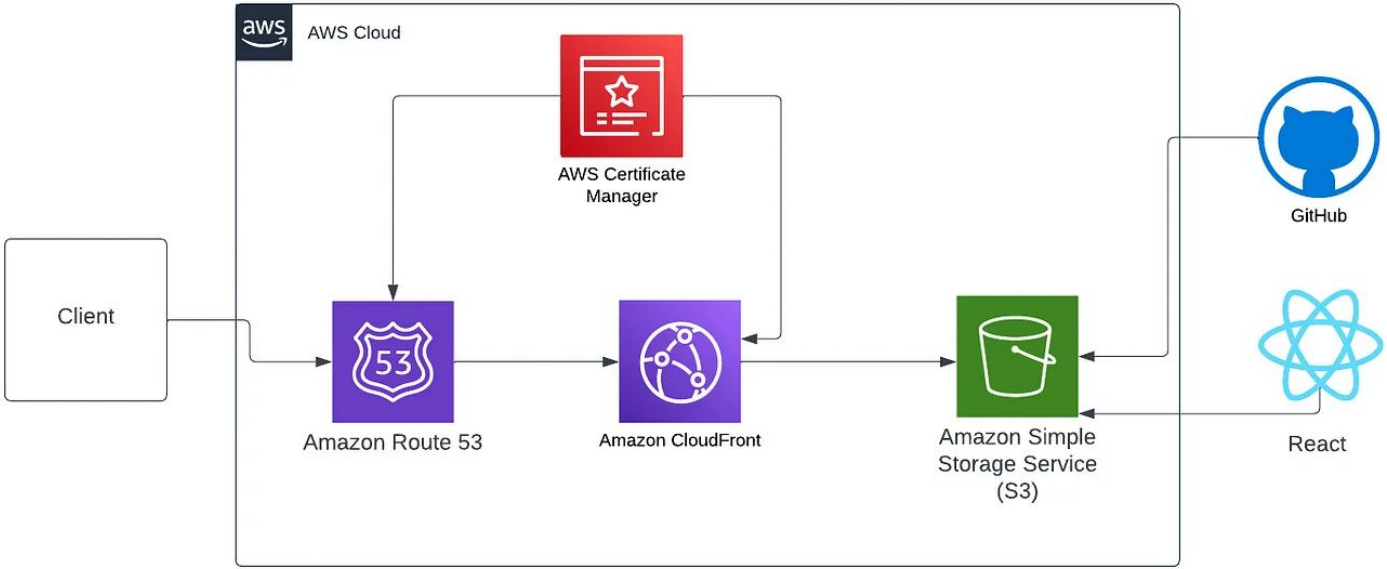
AWS Auto Scaling Group + Load Balancer



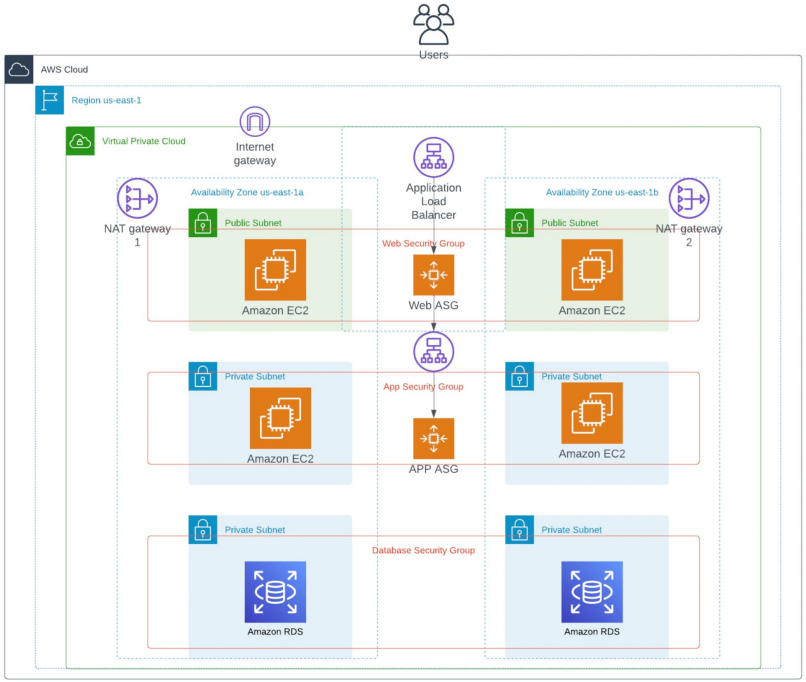
AWS Auto Scaling Group + Load Balancer

- Scaling policies
  - Manual
  - Scheduled
  - Dynamic
    - Simple scaling - cpu above 50% -> add 1
    - Stepped scaling
    - Target tracking
- Load Balancer types
  - Application Load Balancer (ALB)
  - Network Load Balancer (NLB)
  - Gateway Load Balancer (GWLB)

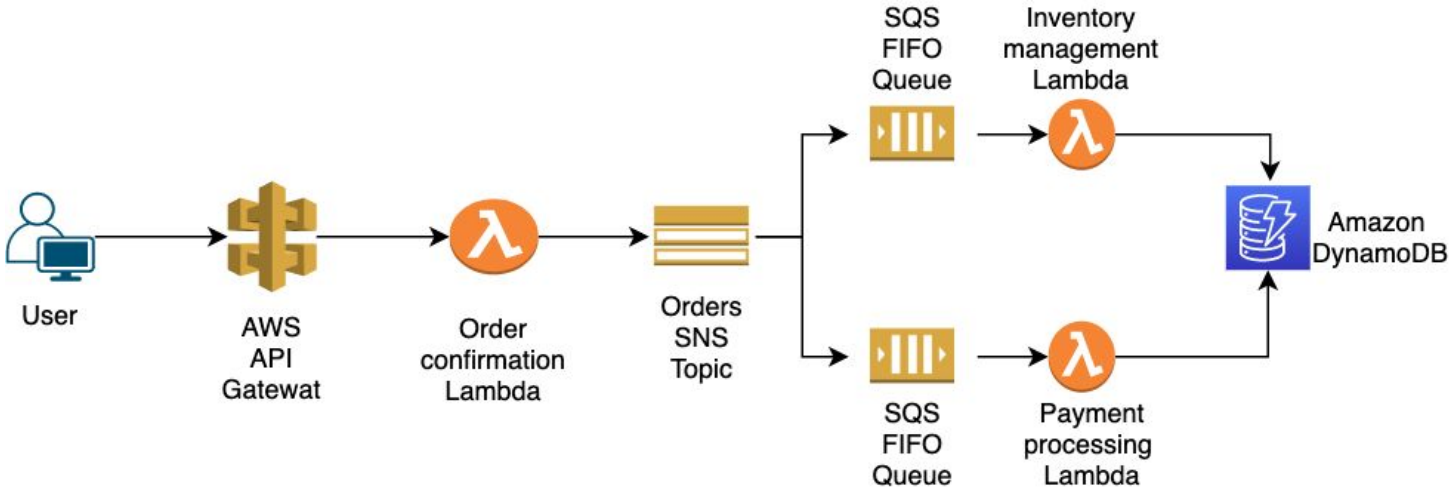
Design #1 - Static site hosting



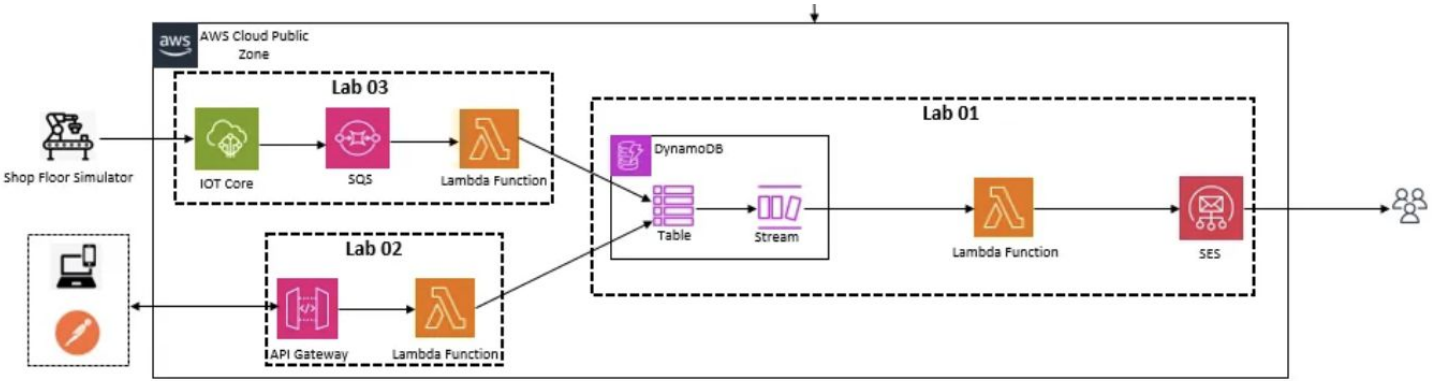
Design #2 - HA 3-tier web application



Design #3 - Highly scalable event based design



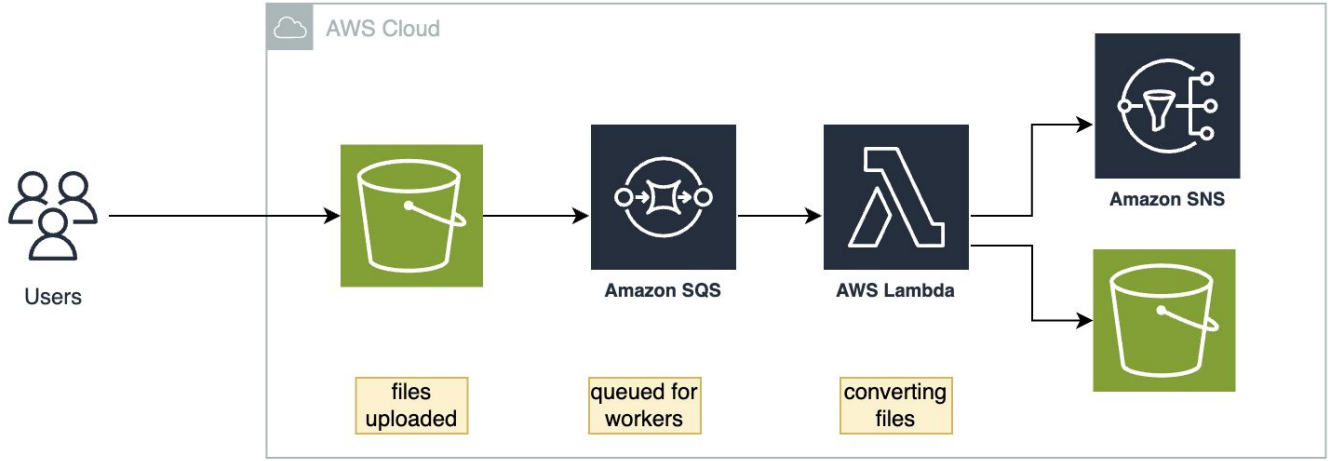
Design #4 - Serverless and scalable design



Design #5

User drops a csv file (~100MB). We need to convert that file into JSON and upload the processed file to S3. After processing, admin needs to be alerted to proceed with that JSON file.

Design #5


AWS Cloud
Users
Amazon SQS
AWS Lambda
Amazon SNS
files uploaded
queued for workers
converting files

Q&A

Cut+Dry

Part 2 - AI/ML Services

2025-01-23 @ CSE

Amazon Forecast

- Forecasting for time-series data
- Retails demand,supply chain, staffing, energy, server capacity, web traffic
- Import historical and related data
- Understand what is normal
- Output = forecast and forecast explainability
- Supports web console , APIs, has SDKs

Amazon Fraud Detector

- Fully managed fraud detection service
- New account creations, payments,guest checkouts
- Can upload historical data
- Can configure model types
- Online fraud detection - account creation
- Transaction fraud - transaction history, identify suspect payments
- Account takeover - Identify phishing or another social based attacks

Amazon Sagemaker

- Fully managed Machine Learning (ML) service
- Data labeling and preparation,Model training and tuning with built-in algorithms or custom code,Model deployment with managed endpoints for real-time or batch inference
- Automatic Model Tuning - Use SageMaker to automatically find optimal learning rates or batch sizes for your ML model.
- Managed Infrastructure with Scalability - Elastic scaling during training and inference,Distributed training across multiple GPUs/instances
- Integration with Other AWS Services

Amazon Textract

- Detect and analyse text contained in input documents
- Input = JPEG,PNG,PDF or TIFF
- Output = Extracted text, structure and analysis
- Most documents -> synchronous , real time
- Large documents (Big PDFs) -> asynchronous
- Pay for usage
- Document analysis
- Receipt analysis
- Identity documents analysis

Amazon Transcribe

- Automatic speech recognition (ASR) service
- Input = Audio, Output = Text
- Language customisations, Filters for privacy, audience appropriate language, speaker identification
- Customer vocabularies and language models
- Pay as you go- per second of transcribed audio

Amazon Translate

- Text translate service.. ML based
- Translate text from native language to other languages
- Auto detection source language
- Encode reads source => semantic representation (meaning)
- Decoder reads meaning => write target language
- Attention mechanisms ensure ‘meaning’ is translated properly

Amazon Polly

- Converts text into “life-like” speech
- Text (language) => Speech - No translation
- Much more human/natural sounding but more complex
- Output formats - Mp3,Ogg,Vorbis,PCM
- Additional control over how polly generated speech
  - Pronunciations
  - Whispering
  - Emphasis
- Speech synthesis markup language (SSML)

Amazon Rekognition

- Deep learning image and video analysis
- Identify objects, people, text, activities,content moderations, face detection, face analysis, face comparison etc
- Per image or per minute (video) pricing
- Integrated with other applications and event driven
- Can even analyse live video streams - kinesis video streams

Amazon Comprehend

- Natural Language Processing (NLP)
- Input = Documents
- Output = entities, phrases, languages, PII, sentiments
- Pre-trained models or Custom
- Real-time analysis for small workloads
- Async jobs for large workloads
- Console & CLI… interactive or use APIs to build into applications

Amazon Kendra

- Intelligent search service
- Designed to mimic interacting with a human expert
- Supports wide range of question types
- Factoid - who, what, where
- Descriptive
- Kendra helps determine Intent
- Index - searchable data organized in an efficient way
- Kendra connects to data sources and indexes from that location, synchronise with the index based on a schedule
- Integrates with other AWS services IAM, Identity Center

Amazon Lex

- Text or voice conversational interfaces
- Powers alexa services
- Automatic speech recognition (ASR) - speech to text
- Natural language understanding (NLU) - Intent
- Build understanding into your application
- Scaling, Integrations, Quick deployment, Pay as you go pricing
- Chatbots,Voice Assistant , Q&A Bots
- Can fulfill the intent … lamda integrations

Thank you !