AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

STG319

Integrate serverless applications with AWS storage services

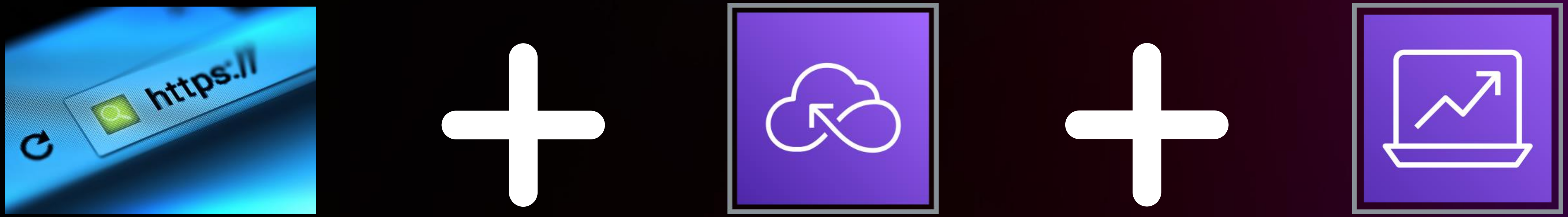
Brandon Dold

Software Development Engineer
Amazon Web Services

Rafael Koike

Principal Solutions Architect
Amazon Web Services

Serverless object detection application



Web Interface

REST API

Analytics

Workshop agenda

Amazon EFS + AWS Lambda

Module 1: Integrate a serverless API with Amazon EFS

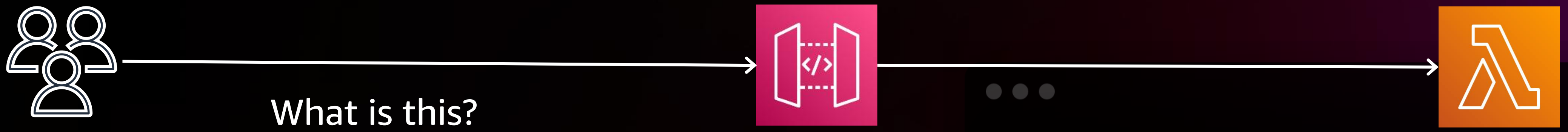
Web hosting on Amazon S3

Module 2: Build a serverless web application with Amazon S3

Amazon S3 as a data lake

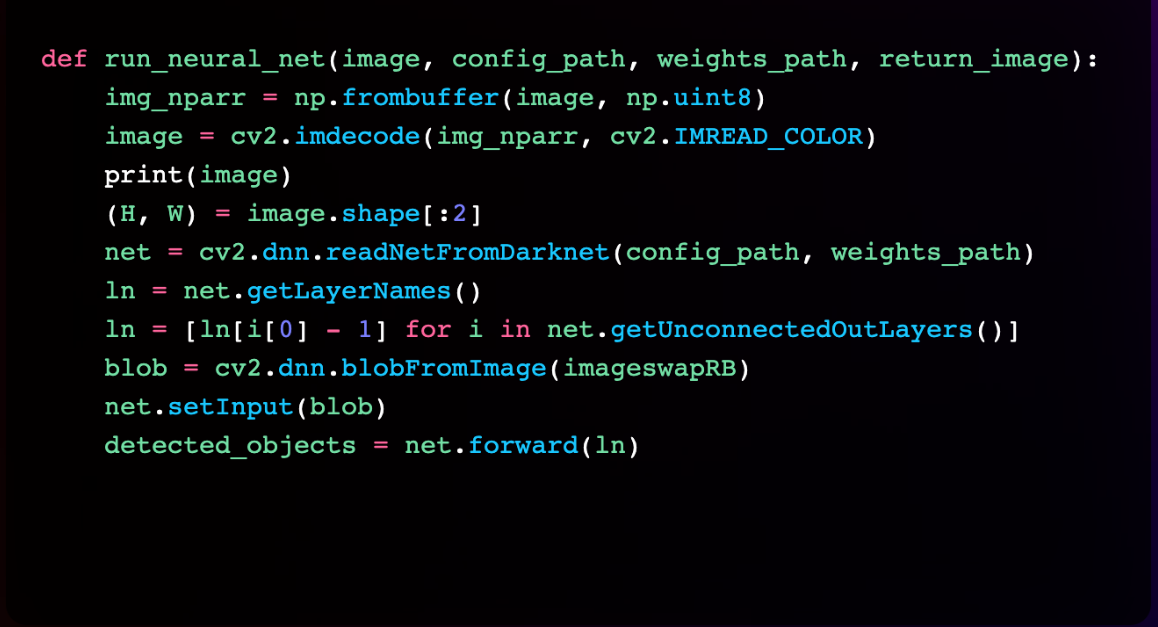
Module 3: Enhance application insights with an Amazon S3 data lake

Module 1: A serverless API for ML-powered object detection



What is this?

/detectObjects



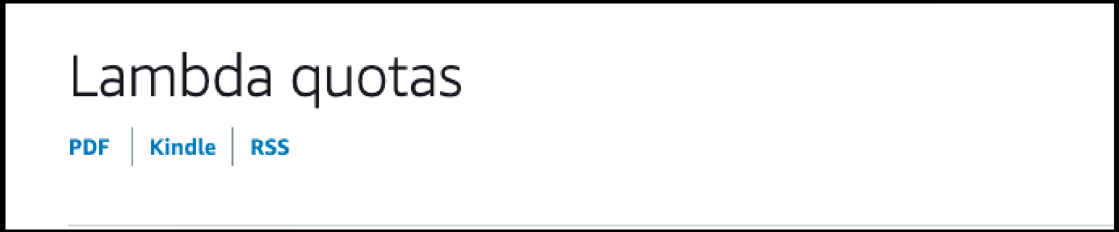
```
def run_neural_net(image, config_path, weights_path, return_image):
    img_nparr = np.frombuffer(image, np.uint8)
    image = cv2.imdecode(img_nparr, cv2.IMREAD_COLOR)
    print(image)
    (H, W) = image.shape[:2]
    net = cv2.dnn.readNetFromDarknet(config_path, weights_path)
    ln = net.getLayerNames()
    ln = [ln[i[0] - 1] for i in net.getUnconnectedOutLayers()]
    blob = cv2.dnn.blobFromImage(imageswapRB)
    net.setInput(blob)
    detected_objects = net.forward(ln)
```

More parameters = Higher accuracy

Model accuracy

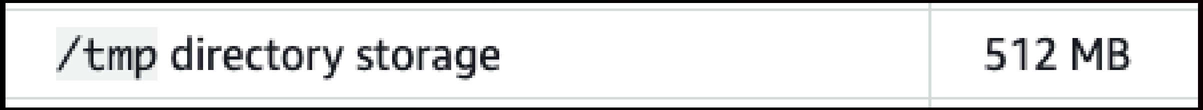
Model	File size	Mean average precision	# of parameters
Yolo tiny	~43 MB	23.7	~119
Yolo big	~240 MB	60.6	~682

File system limitations



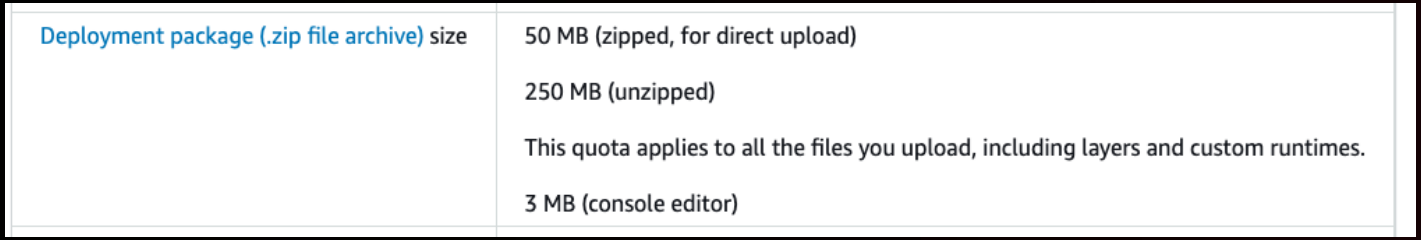
Lambda quotas

PDF | Kindle | RSS



/tmp directory storage	512 MB

Ephemeral, for example, must be downloaded for each invocation ✗



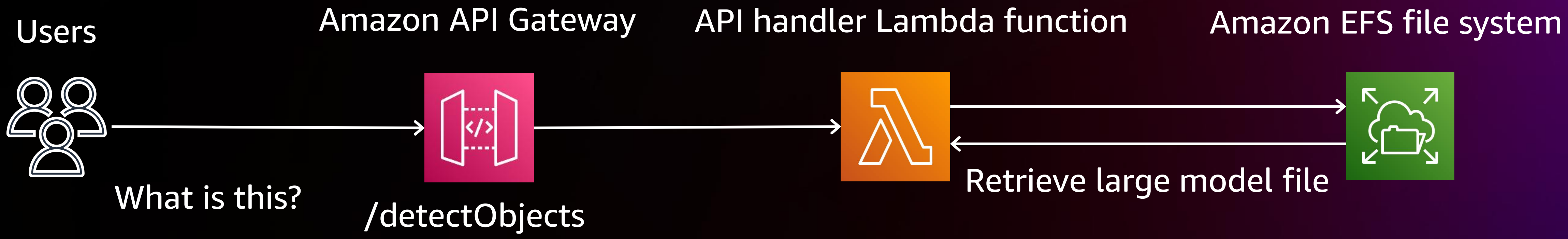
Deployment package (.zip file archive) size	50 MB (zipped, for direct upload) 250 MB (unzipped) This quota applies to all the files you upload, including layers and custom runtimes. 3 MB (console editor)

Model size + required code exceeds limit ✗

The solution

Amazon EFS + AWS Lambda

How it works


Users

Amazon API Gateway

API handler Lambda function

Amazon EFS file system

What is this?

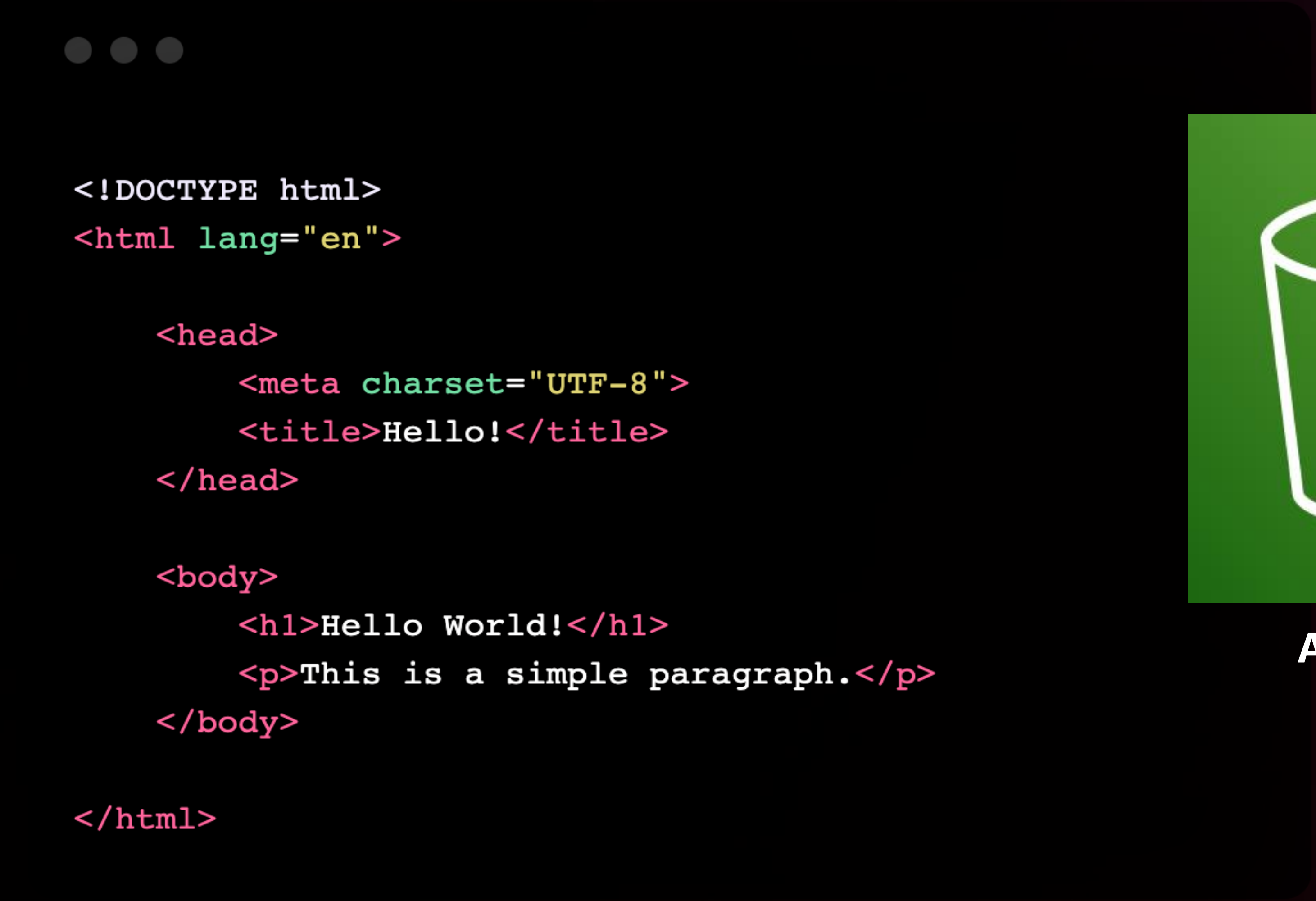
/detectObjects

Retrieve large model file

- Mounts file system when execution environment is prepared
- Minimal latency
- If Lambda is warm, mount is already available
- Scales to 25,000 concurrent connections

https://s12d.com/stg319

Module 2: Serverless web application with Amazon S3



```
<!DOCTYPE html>
<html lang="en">

    <head>
        <meta charset="UTF-8">
        <title>Hello!</title>
    </head>

    <body>
        <h1>Hello World!</h1>
        <p>This is a simple paragraph.</p>
    </body>

</html>
```

Amazon S3

Traditional web server

Install OS

Updates

Security patches

Configure networking

Configure firewall

Secure the OS

NGINX

APACHE
HTTP SERVER PROJECT

Setup requires specific domain knowledge . . .

. . . and takes hours or even days to complete

Writing code

✓

Provisioning server

✗

Web applications are distinguished based on where the code is run

Modern web applications perform logic in a web browser and communicate with backend cloud resources, such as databases, through web APIs

A web application consists of a set of HTML, CSS, and JavaScript files

These files can be hosted by Amazon S3 and served to users with no need to provision or maintain a single server

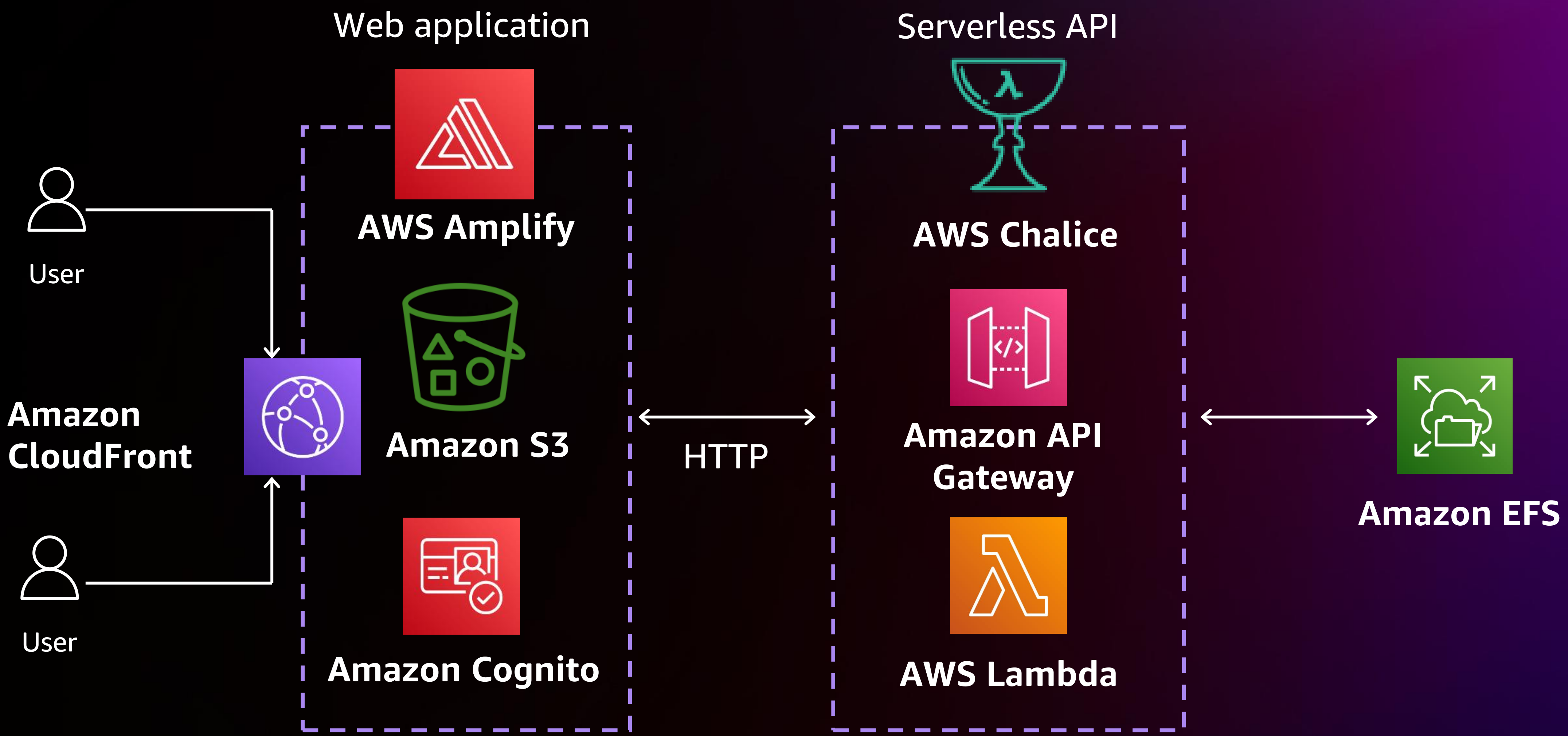
Can scale to handle enterprise-level traffic with no additional work

Take advantage of AWS Amplify for fully managed hosting

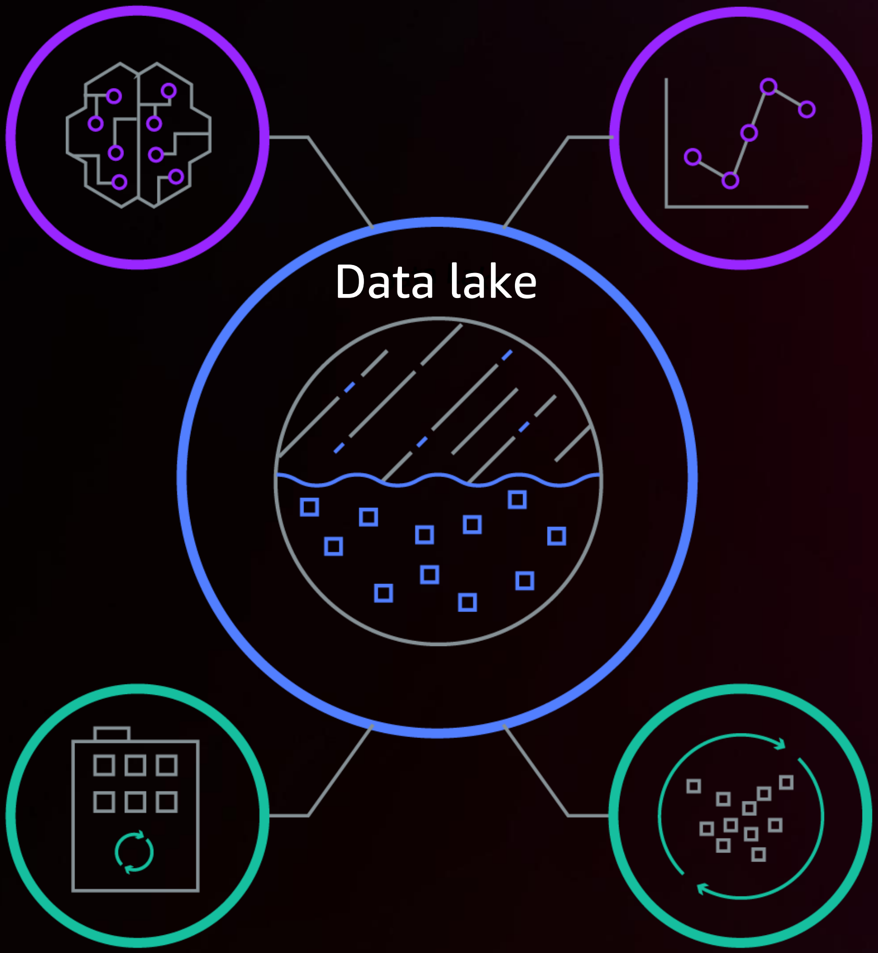
Amazon S3

Instead of hours or days to deploy a web application . . .

deploy your web application in minutes


Web application
Serverless API
User
User
Amazon CloudFront
AWS Amplify
Amazon S3
Amazon Cognito
HTTP
AWS Chalice
Amazon API Gateway
AWS Lambda
Amazon EFS

Module 3: Amazon S3 data lake



Machine learning

Analytics

Data lake

On-premises data movement

Real-time data movement

Traditional data analytics resources

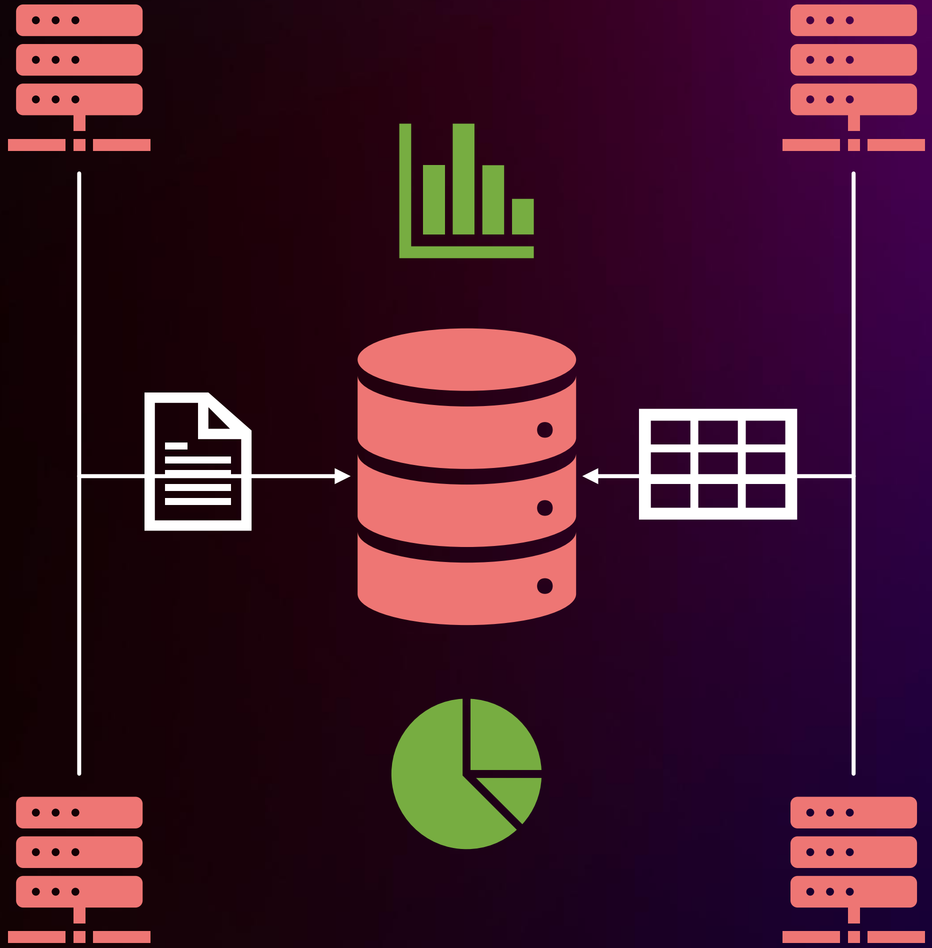
Can have similar issues to those described earlier, such as OS/software configuration, network/security configuration

Often proprietary systems

Intricate and involves many components

Requires data model to be defined in advance

Data must be structured in a tabular format, for instance



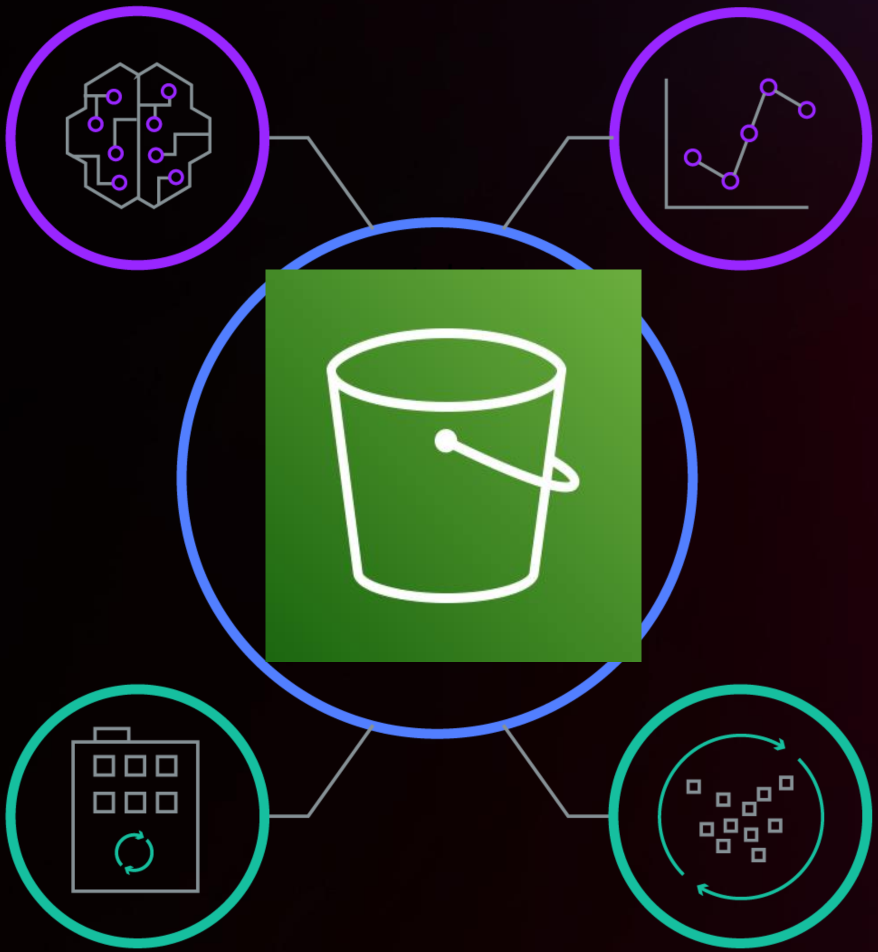
Some use cases do require that sort of data analysis architecture . . .

but it’s not needed for analyzing unstructured or semi-structured ML data

Central repository for all data, structured or unstructured

Schema written at the time of analysis (schema-on-read)

Supports different types of analytics

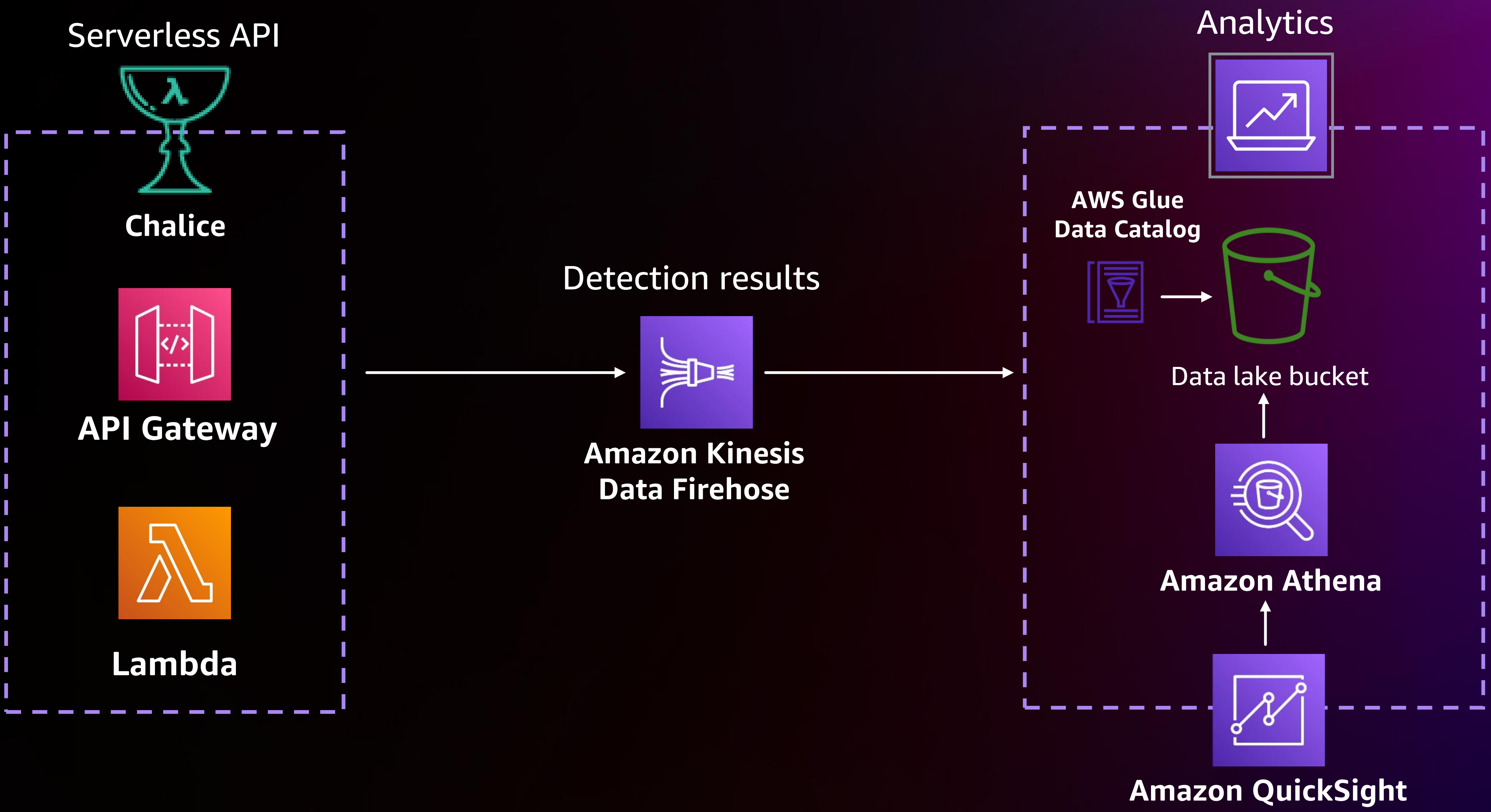


Amazon S3 data lake

Automatically scales

No need to carefully define data models

Not a single server to manage


Serverless API
Chalice
API Gateway
Lambda
Detection results
Amazon Kinesis Data Firehose
Analytics
AWS Glue Data Catalog
Data lake bucket
Amazon Athena
Amazon QuickSight

Thank you!

Brandon Dold
GitHub: brandold

Rafael Koike
koiker@amazon.com

Please complete the session survey in the **mobile app**