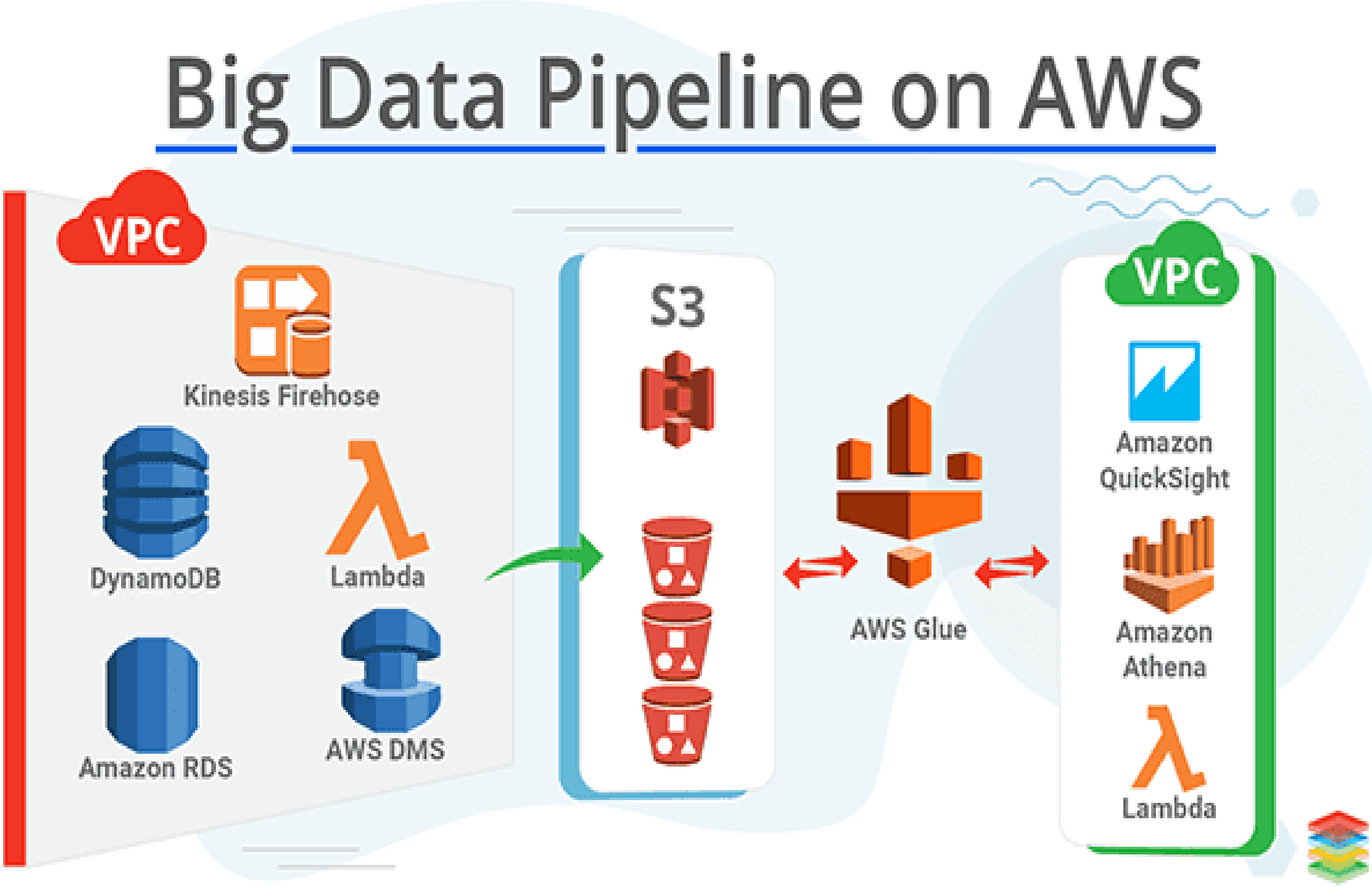
Data Pipeline Design Example

Contents:

- Problem Statement
- Understanding the DP ecosystem
- Solution Architecture

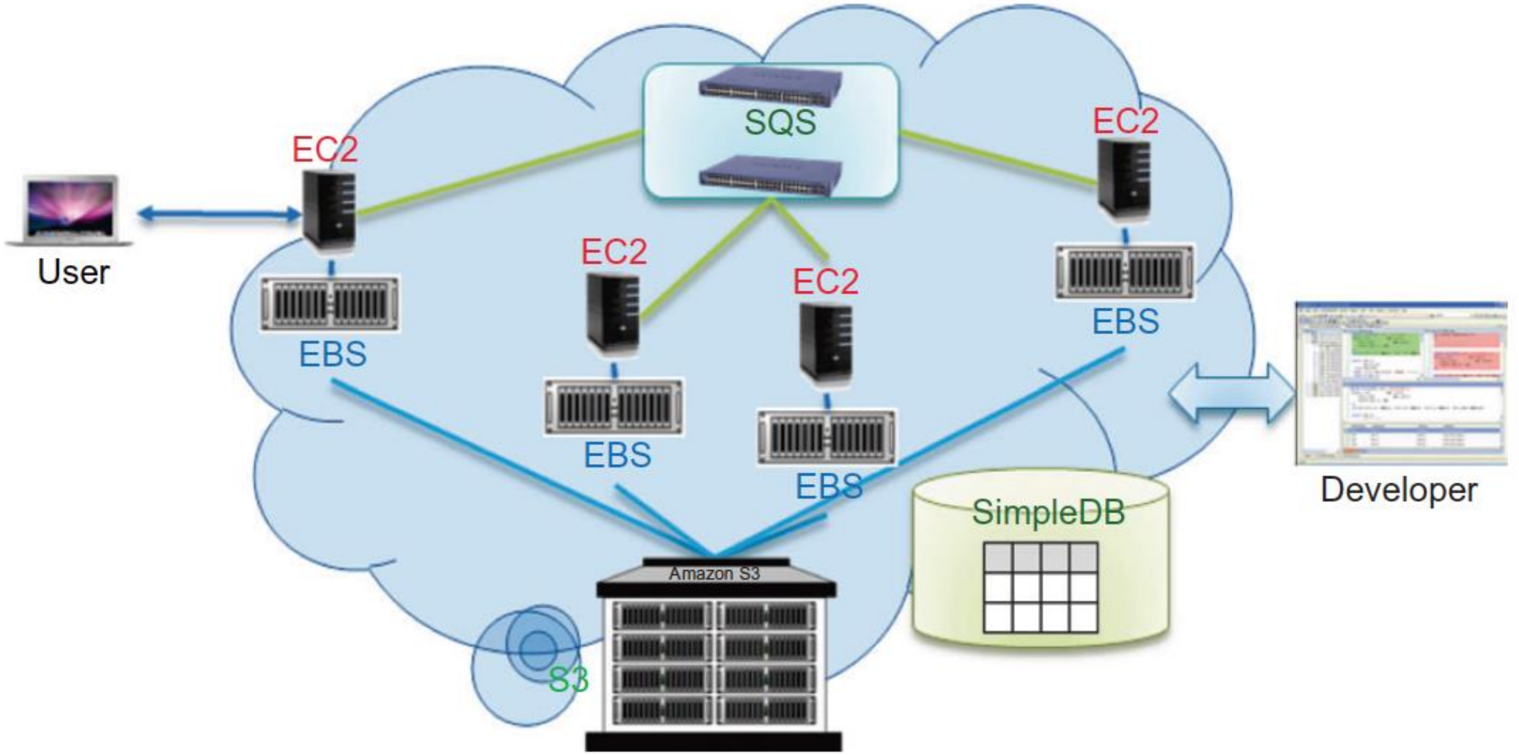
Not for your Quiz! ☺

Recapitulation



Slides 2, 4 to 7
Courtesy:
amazon.com

AWS Cloud – Some frequently used terminology!



EC2 – Elastic Compute Cloud (a virtual server); SQS – Simple Queue Service (for exchanging msgs between s/w components); EBS – Elastic Block Store (ability to store the data in blocks); S3 – Simple Storage Service (data storage via a web service interface)

Ref: Parts of the material presented in *slides #3 – 8* can be found in amazon.com

AWS Cloud – Some remarks

Amazon AWS - Compute service categories

- General Purpose Instances
- Computer Optimized Instances
- Memory Optimized Instances
- Accelerated Computing Instances
- Storage Optimized Instances
- Dense Storage Instances

Instance types comprise varying combinations of - CPU, memory, storage, and networking capacity;

Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

AWS Cloud – Some remarks

Amazon AWS – Storage service categories

- Amazon Simple Storage Service (Amazon S3)
- Amazon Glacier
- Amazon Elastic File System (Amazon EFS)
- Amazon Elastic Block Store (Amazon EBS)
- Amazon EC2 Instance Storage.
- AWS Storage Gateway.
- AWS Snowball.
- Amazon CloudFront.

AWS Cloud – Some remarks

Quick note on AWS S3! Most supported storage platform

AWS S3 is an object storage model that is built to store and retrieve any amount of data from anywhere - *websites, mobile apps, corporate applications, and data from IoT sensors or devices;*

Very well-suited for hosting web content that requires bandwidth along with high demand;

S3 is also used to host entire static websites and storage for images, videos, and client-side scripts in formats such as JavaScript.

AWS Cloud – Some remarks

Quick note on AWS S3!

Durability of AWS S3! Runs upon the world's largest global cloud infrastructure, and was built from the bottom-up fashion to deliver a customer promise of 99.999999999% durability;

Availability of AWS S3! Data is *automatically distributed across a minimum of three physical facilities that are geographically separated within an AWS Region*, and also automatically replicates data to any other AWS Region;

AWS Cloud – Some remarks

Quick note on AWS S3!

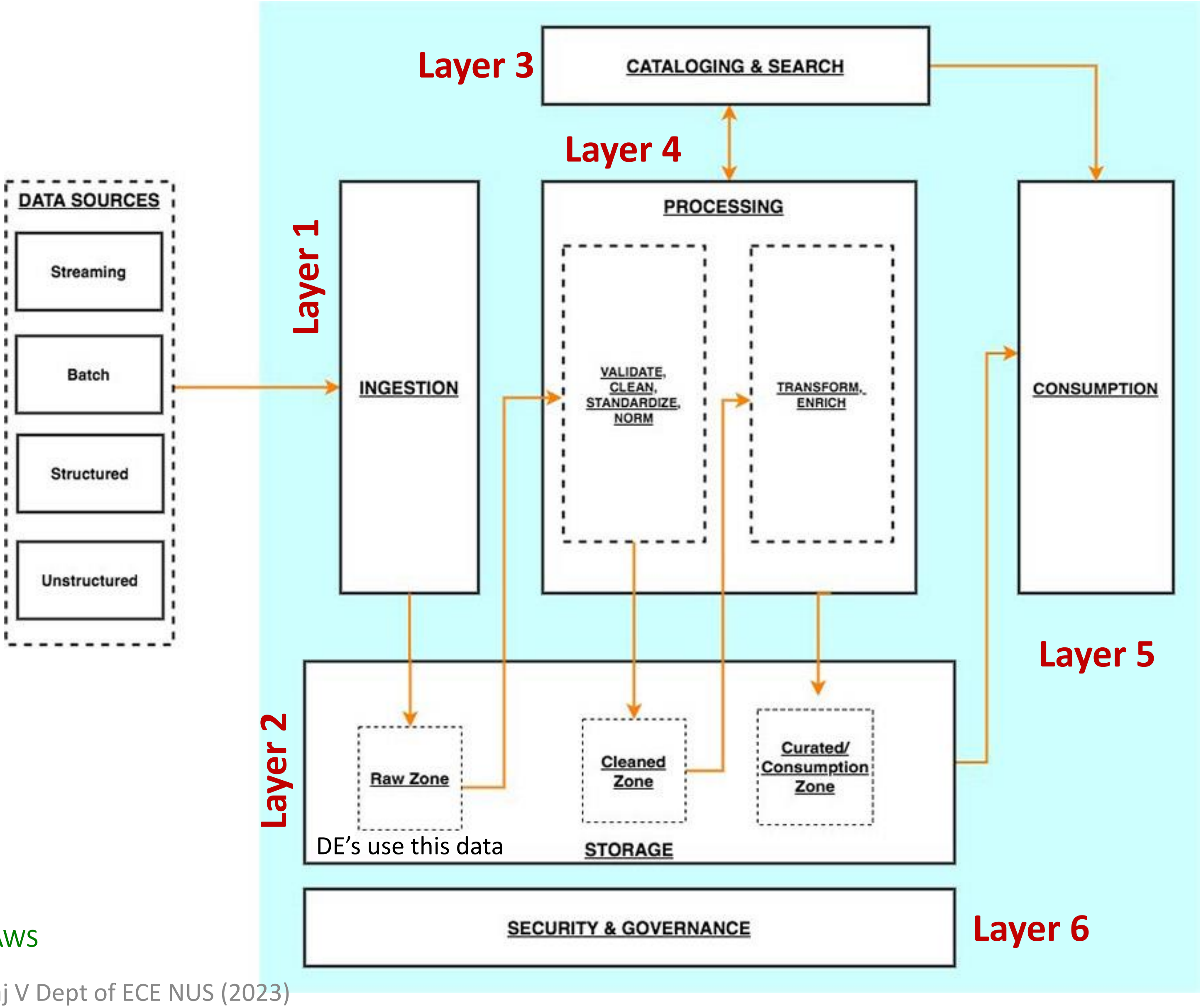
Security on AWS – a highly secure storage service!

AWS S3 supports three different forms of encryption, including server-side-encryption and client-side-encryption.

Supports usual access management polices;

Facilitates adding an optional layer of security by enabling *Multi-Factor Authentication (MFA)* for object operations.

Overview of the DP Architecture - Data lake centric analytics architecture



Courtesy: AWS

Understanding the DP Components

Six logical layers

- **Ingestion Layer:**

  - Responsible for bringing data into the data lake.
  - Has the ability to connect to internal and external data sources over a variety of protocols.
  - Both batch and streaming data are ingested into the storage layer.
  - Responsible for delivering ingested data to a diverse set of targets in the data storage layer (including the object store, databases, and warehouses)

- **Storage Layer:**

  - Responsible for providing durable, scalable, secure, and cost-effective components to store vast quantities of data;
  - Raw zone - Storing data as it is; DE often uses this zone!
  - Cleaned zone – Validate, clean, standardization operations; Original format preserved; DE & DS interact by referring to this data!
  - Curated Zone – Transformed & enriched data!

Understanding the DP Components

- Catalogue & Search Layer:

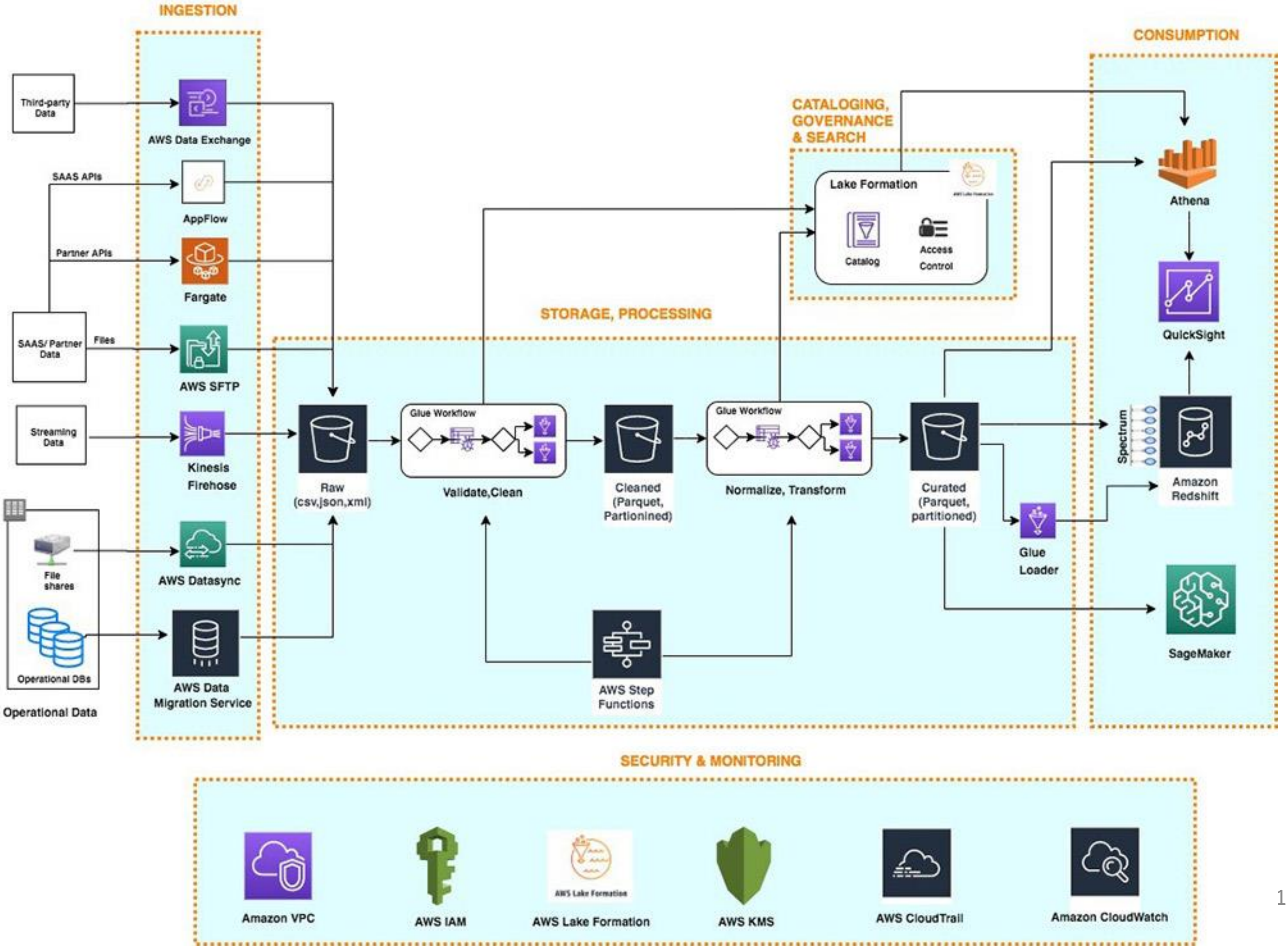
  - Responsible for storing metadata about datasets hosted in the storage layer
  - Has the ability to track schema and the granular partitioning of dataset information in the lake;
  - Supports mechanisms to track versions to keep track of changes to the metadata

- Processing Layer – Our most of the ETL here!

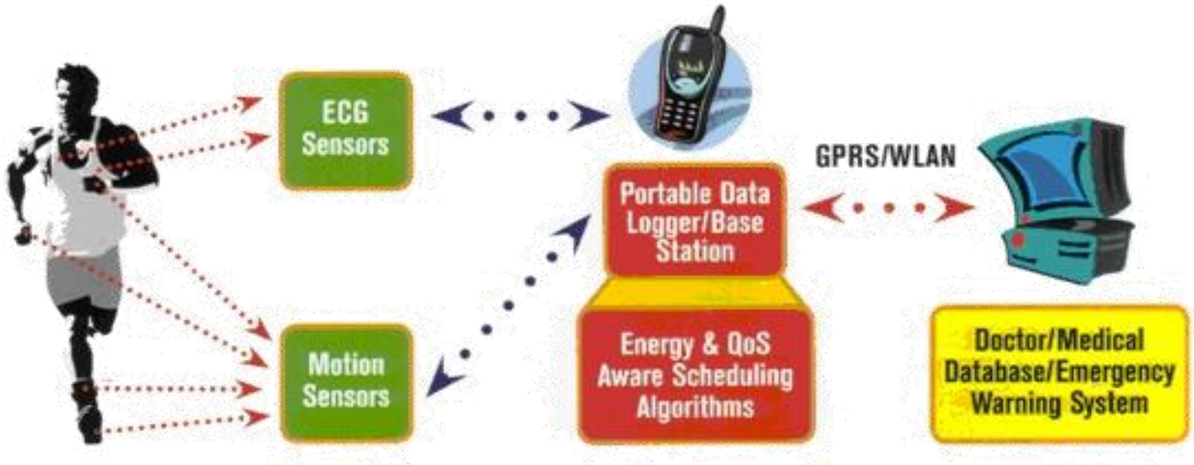
- Consumption Layer - Responsible for providing scalable and performant tools to gain insights – All kinds of analytics support is provided – SQL data processing, batch analytics, BI dashboards, reporting, and AI/ML. Integrates with the data lake's storage, cataloging, and security layers.

- Security Layer - Responsible for protecting the data in the storage layer and processing resources in all other layers; Provides mechanisms for access control, encryption, network protection, usage monitoring, and auditing; Monitors activities of all components in other layers and generates a detailed audit trail.

Detailed DP Flow Diagram



Hospital Management System on AWS – Design of a Data Pipeline – A Case Study



ECG Sensors

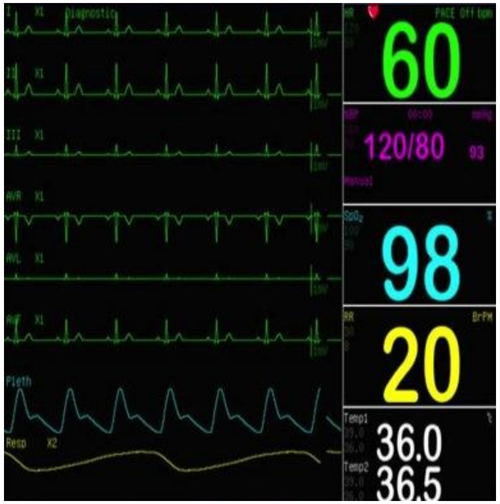
Motion Sensors

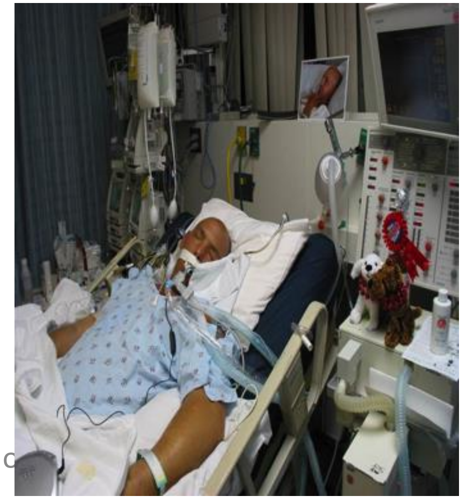
Portable Data Logger/Base Station

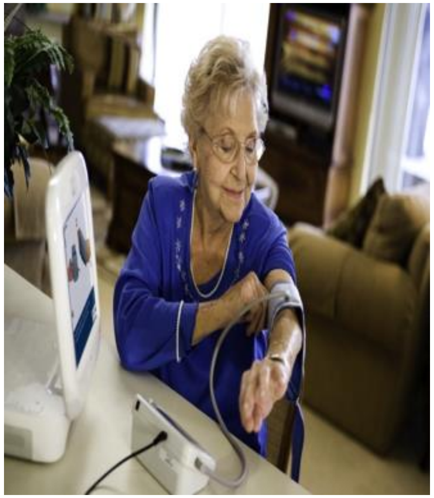
Energy & QoS Aware Scheduling Algorithms

GPRS/WLAN

Doctor/Medical Database/Emergency Warning System







Problem Statement

Hospital Management System (HMS) wants to build a real-time analytics for supporting and detecting patients with high-risk cardiac issues. They want to host their processing system on a cloud based infrastructure, such as AWS which can support both Batch and Real-time streaming analytics. This is to facilitate a peer-to-peer service facility for medical practitioners as well as patients. Patients will receive alerts from doctors, if need arises.

Problem Description

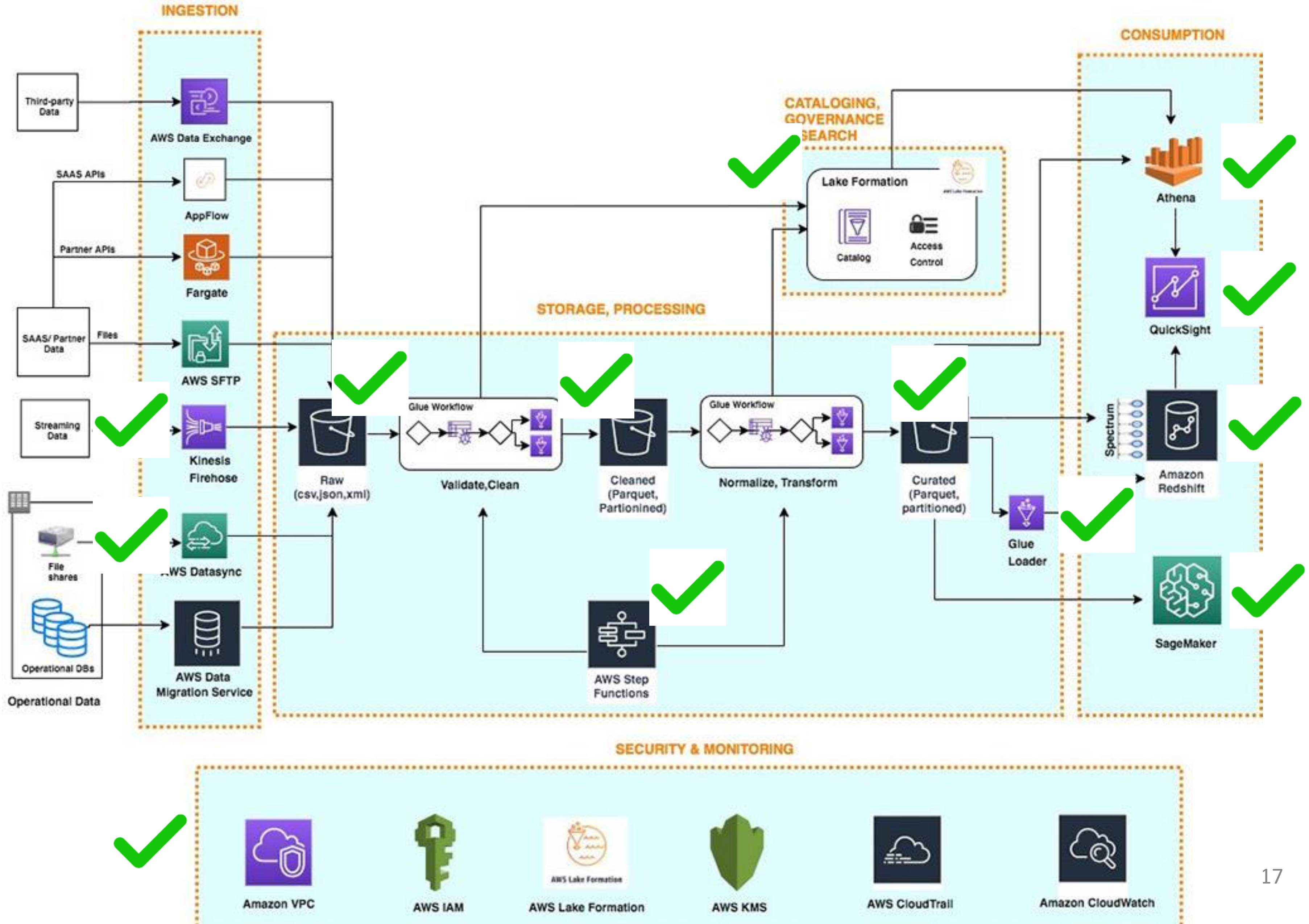
- Patients data is stored on SQL DBs on-premises (hospitals) for security reasons. Patients data are streamed into the Cloud using IoT sensors, through his/her mobile apps, and wearables.
- It is required to fetch on-premises data and fuse it with real-time streaming data to interpret and complete the analytics required to extract the current condition of the patient.
- Raw data must be available at any time. Since the data is large, they need to be compressed and stored to minimize cost.
- Upon detecting any anomalies, the analytics unit should alert the medical practitioner as well as the patients.

Problem Description (Cont’d)

- The analytics must constantly produce the required insights via visualization tools.
- Further the dashboard must facilitate doctors to query any past 1 year data to know about that patient's history.
- For DEs and DSs there must be a provision to search the required data to perform analytics. So data describing the data, *referred to as meta-data* must be stored.
- Finally, the patients data must be stored on the Cloud for at least 3 years and then needs to be archived.

As a Data Engineer, you are expected to design a Data Pipeline architecture (solution architect) that meets the above specs!

Solution ?


INGESTION
Third-party Data
AWS Data Exchange
SAAS APIs
AppFlow
Partner APIs
Fargate
SAAS/ Partner Data
Files
AWS SFTP
Streaming Data
Kinesis Firehose
File shares
AWS Datasync
Operational DBs
AWS Data Migration Service
Operational Data
CATALOGING, GOVERNANCE SEARCH
Lake Formation
Catalog
Access Control
STORAGE, PROCESSING
Raw (csv,json,xml)
Glue Workflow
Validate,Clean
Cleaned (Parquet, Partionined)
Glue Workflow
Normalize, Transform
Curated (Parquet, partitioned)
Glue Loader
AWS Step Functions
CONSUMPTION
Athena
QuickSight
Spectrum
Amazon Redshift
SageMaker
SECURITY & MONITORING
Amazon VPC
AWS IAM
AWS Lake Formation
AWS KMS
AWS CloudTrail
Amazon CloudWatch

What about the requirement - *The patients data must be stored on the Cloud for at least 3 years and then needs to be archived ?*

- One of the S3 instances is called Glacier - provides storage for data archiving and backup; Can be configured to trigger backup based on user's criteria;

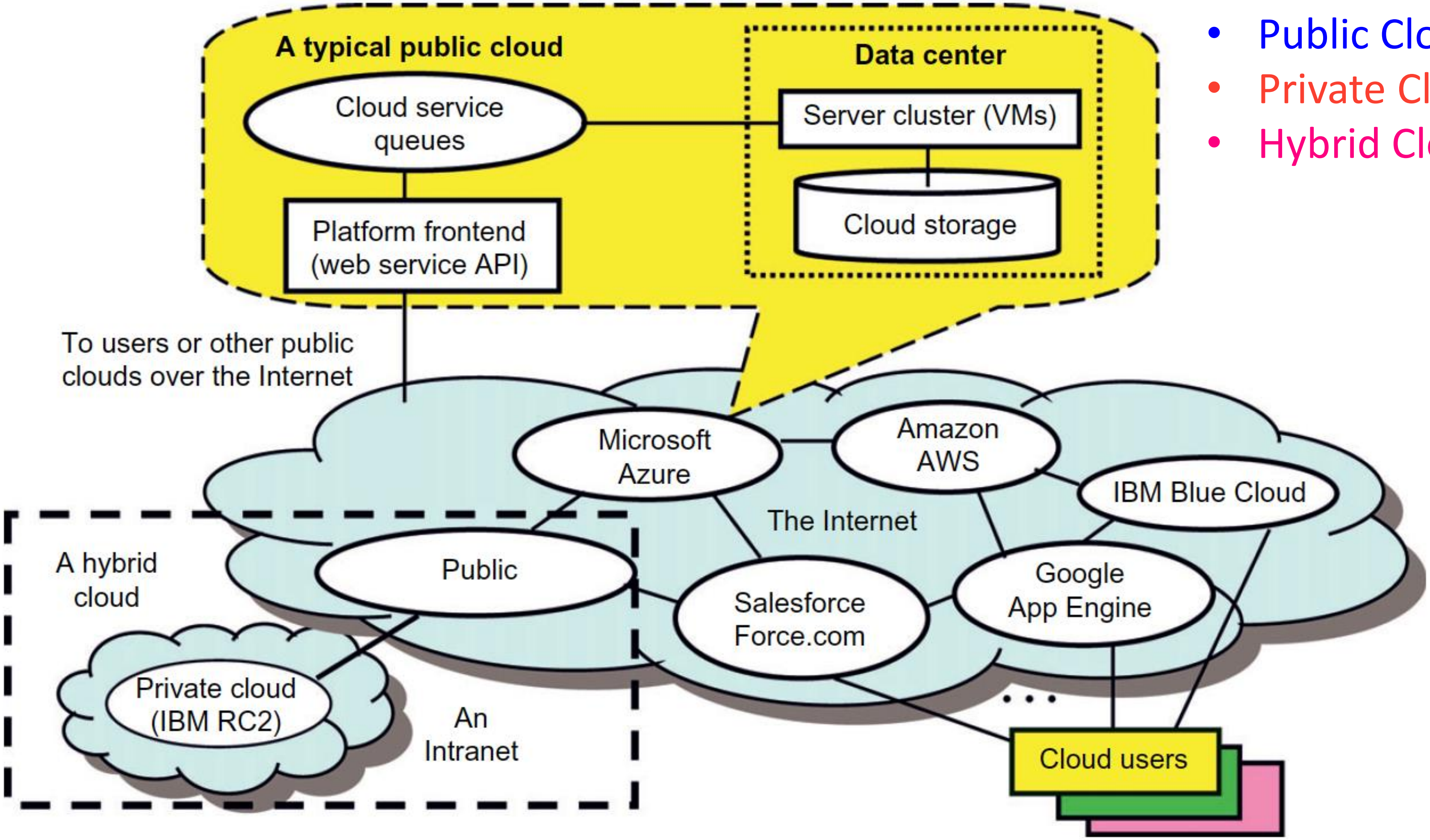
What about the requirement - *Upon detecting any anomalies, the analytics unit should alert….. ?*

- One of the key features of *AWS Kinesis* is a two-way communication protocol; This is also available with Apache Kafka service; This is a configuration based step;

What about the Step Function? This corresponds to an orderly execution of functions (using lambda service) that can be scheduled while performing ETL steps;

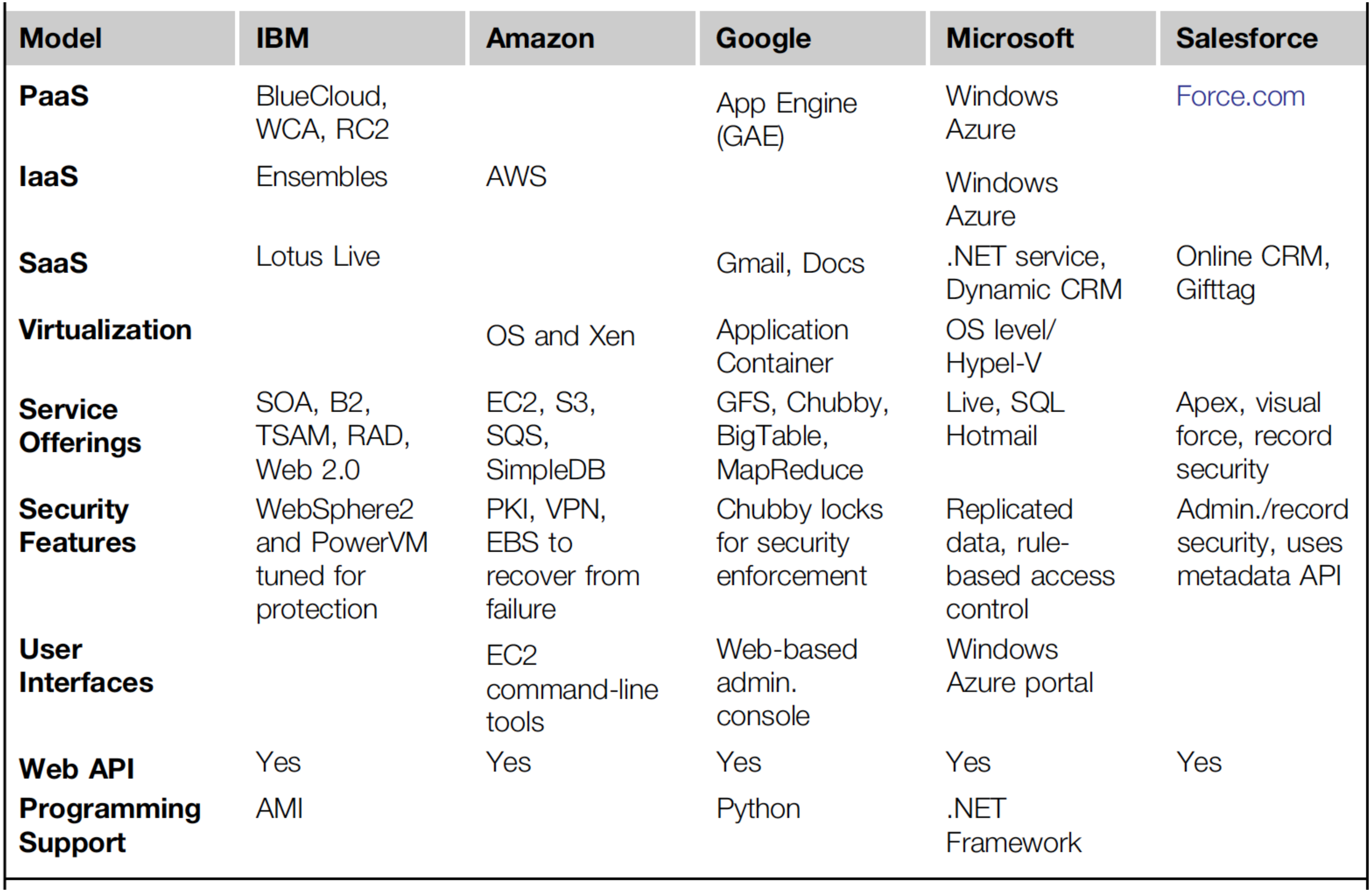
Thank you!

Cloud Computing Platforms



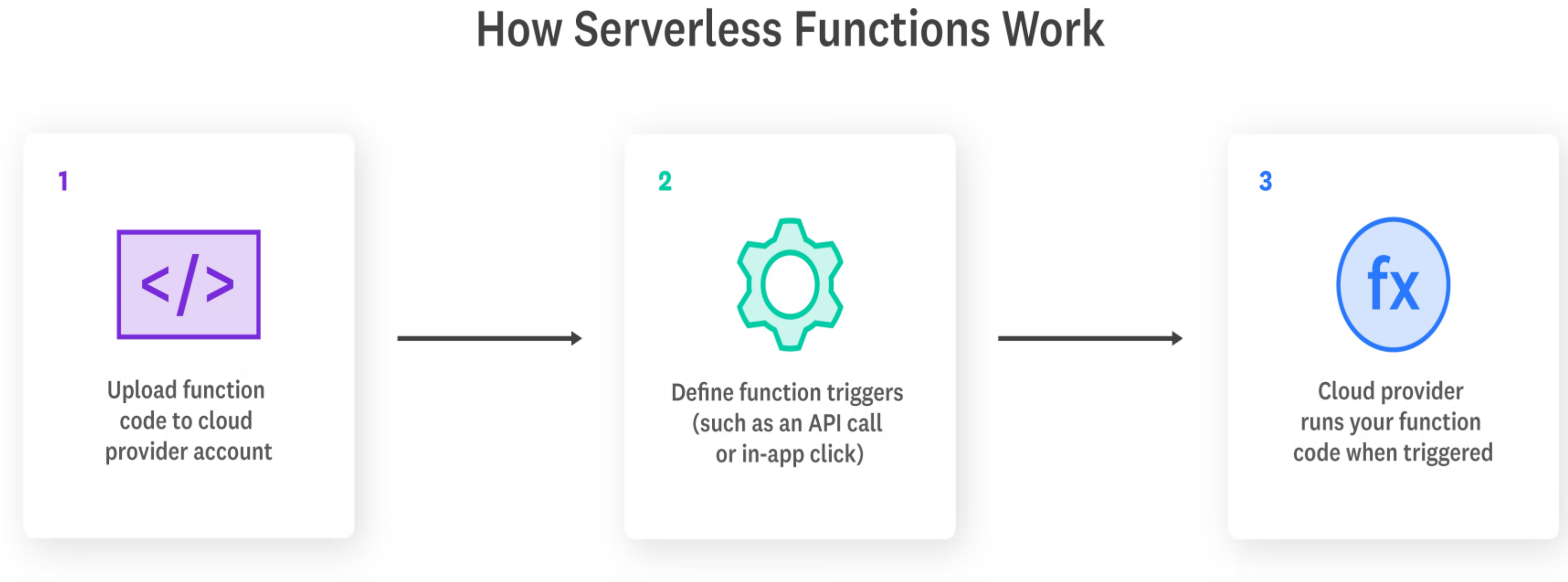
- Public Clouds
- Private Clouds
- Hybrid Clouds

Sample Cloud platforms & their service offerings



Model	IBM	Amazon	Google	Microsoft	Salesforce
PaaS	BlueCloud, WCA, RC2		App Engine (GAE)	Windows Azure	Force.com
IaaS	Ensembles	AWS		Windows Azure	
SaaS	Lotus Live		Gmail, Docs	.NET service, Dynamic CRM	Online CRM, Gifttag
Virtualization		OS and Xen	Application Container	OS level/ Hypel-V	
Service Offerings	SOA, B2, TSAM, RAD, Web 2.0	EC2, S3, SQS, SimpleDB	GFS, Chubby, BigTable, MapReduce	Live, SQL Hotmail	Apex, visual force, record security
Security Features	WebSphere2 and PowerVM tuned for protection	PKI, VPN, EBS to recover from failure	Chubby locks for security enforcement	Replicated data, rule-based access control	Admin./record security, uses metadata API
User Interfaces		EC2 command-line tools	Web-based admin. console	Windows Azure portal	
Web API	Yes	Yes	Yes	Yes	Yes
Programming Support	AMI		Python	.NET Framework	

ANNEX - Serverless Architecture - Data lake centric analytics architecture


How Serverless Functions Work
1
</>
Upload function code to cloud provider account
2
Define function triggers (such as an API call or in-app click)
3
fx
Cloud provider runs your function code when triggered

Source: Datadog

Most of the components used in a DP adopt a serverless functional style as shown above.

Courtesy: datadoghq.com