Big Data AWS Tutorial

Agenda

01 Intro to Big Data on AWS

02 Data Collection
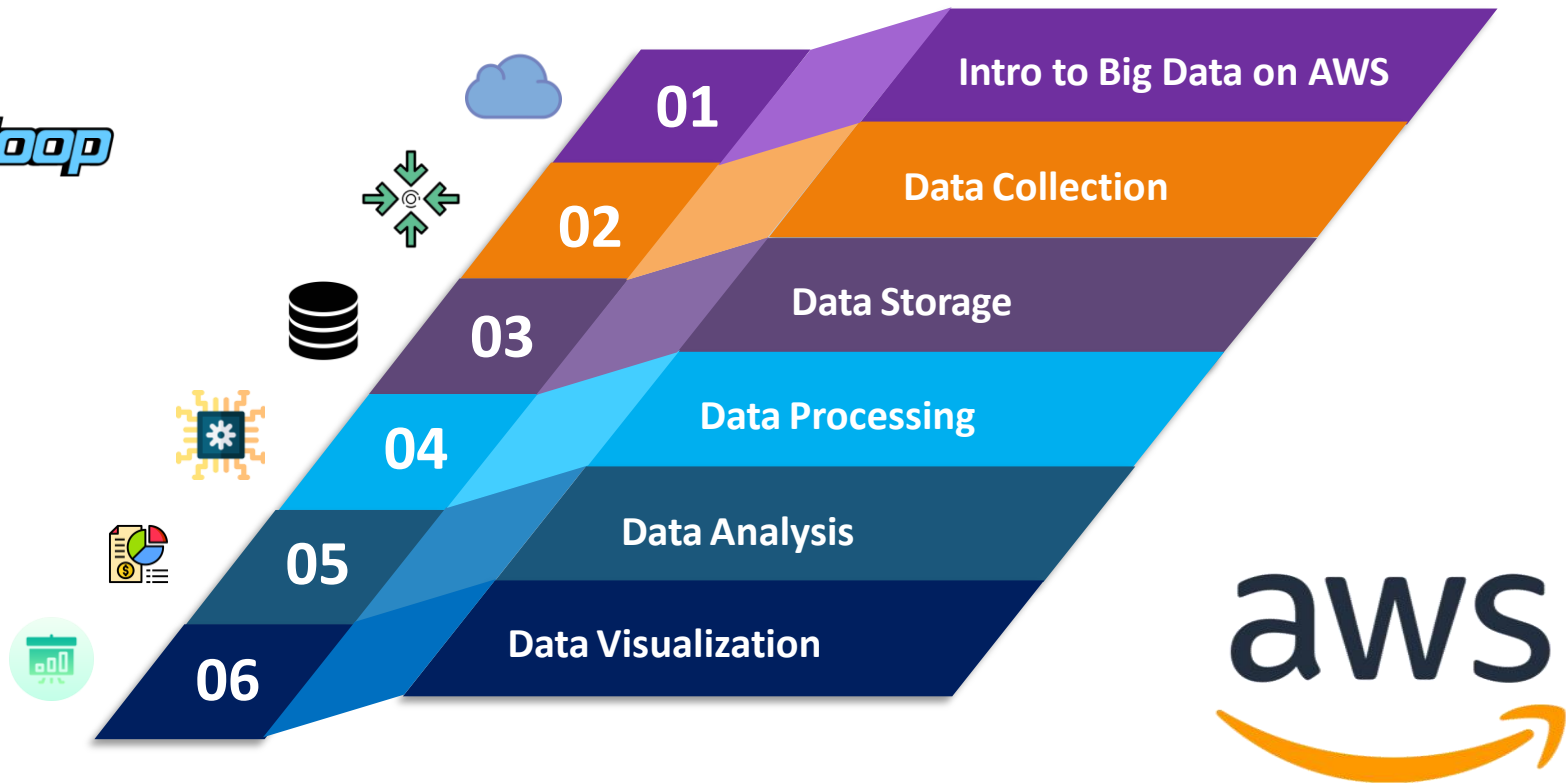
03 Data Storage

04 Data Processing

05 Data Analysis

06 Data Visualization

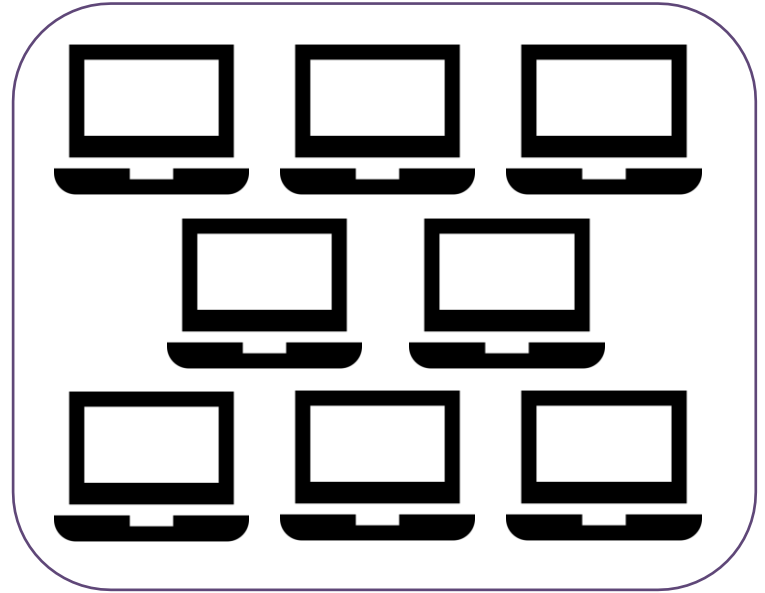# Introduction to Big Data on AWS

Introduction to Big Data on AWS

The traditional way of implementing Big Data

Remote Access

Client

A Big Data Cluster

Introduction to Big Data on AWS

Using AWS cloud services to implement Big Data

Abstract Visibility

Remote Access

Client

aws

AWS Big Data Services

AWS Servers

Introduction to Big Data on AWS

There are multiple advantages to using AWS for our Big Data needs

Fully managed

No hardware requirement

Pay only for what we use

Instant scale up or down

Integrates with other AWS services

Introduction to Big Data on AWS

Data Collection

Data Collection

AWS Kinesis is a data streaming service

Input
Tractor sensors send data to Amazon Kinesis Data Streams

Amazon Kinesis Data Streams
Ingests and stores sensor data streams for processing

AWS Lambda
AWS Lambda is triggered and runs code to detect trends in sensor data, identify anomalies, and order replacements for faulty parts

Output
Replacement parts are delivered
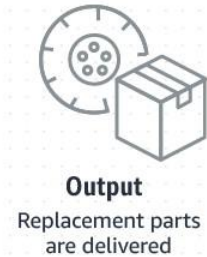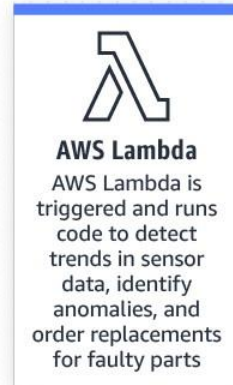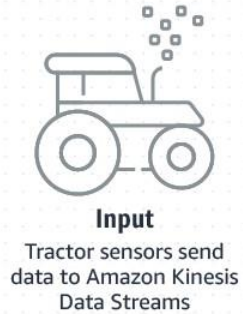
Data Storage

Data Storage

Amazon DynamoDB is a NoSQL database, which is fully managed by AWS to provide a fast and predictable performance. With DynamoDB, we can create database tables that can store and retrieve any amount of data
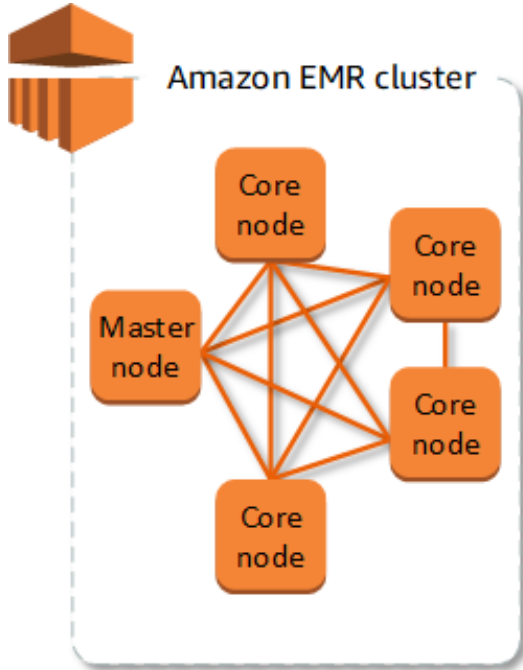

amazon

DynamoDB

Data Storage

Amazon Glacier is designed for long-term data storage for cold data that is not needed to be retrieved very frequently and quickly. It is also a backup service, which is used to back up the infrequently accessed data in S3

amazon
S3 Glacier

Data Processing

Data Processing

Amazon EMR cluster

- Master node
- Core node
- Core node
- Core node
- Core node

Amazon EMR is a Big Data cluster hosted through EC2 instances

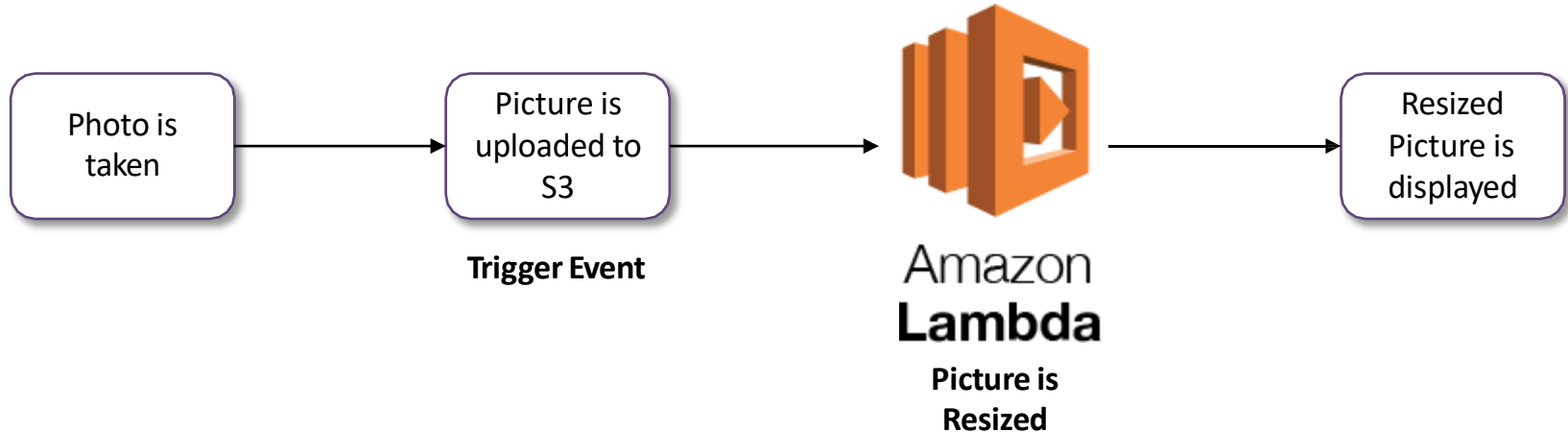Begin Processing → Step 1 COMPLETED → Step 2 RUNNING → Step 3 PENDING → Step N... PENDING → End Processing

EMR Services cease after the processing ends

Data Processing
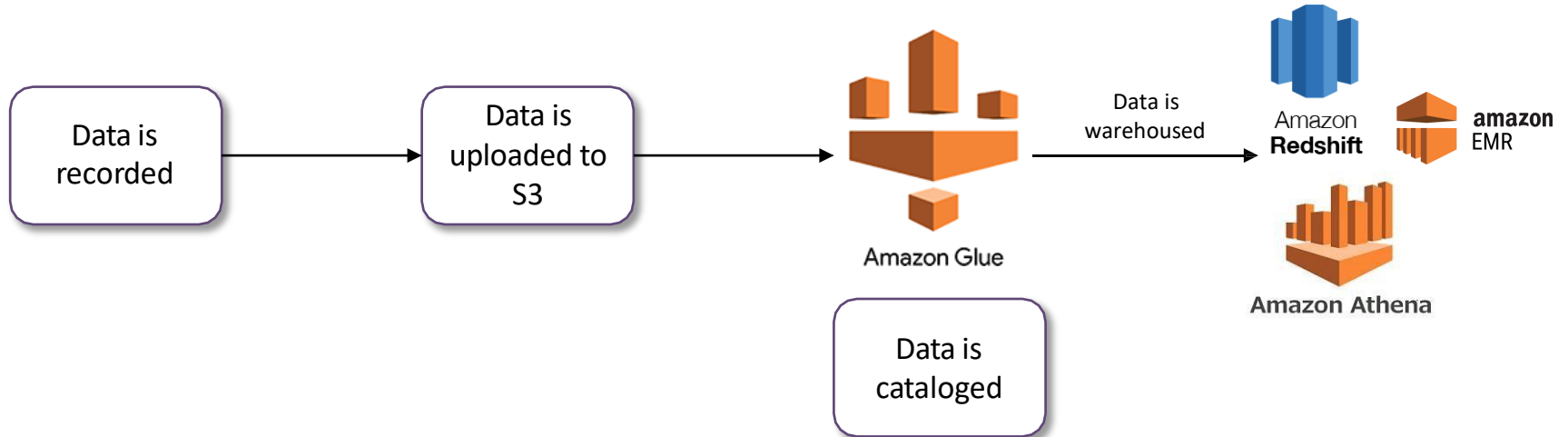
AWS Lambda is a serverless, event-driven, and fully managed computing solution where we pay only for the computation we perform

Photo is taken → Picture is uploaded to S3 → → Resized Picture is displayed

Trigger Event

Picture is Resized

Data Processing

AWS Glue is a fully managed ETL (Extract, Load, and Transform) service that helps users prepare their data for analysis

Data is recorded → Data is uploaded to S3 → Amazon Glue → Data is warehoused → Amazon Redshift, amazon EMR, Amazon Athena

Data is cataloged

Data Analysis

Data Analysis

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud that allows users to deal with as much as a petabyte of data

Data Analysis

Amazon SageMaker is a fully managed Machine Learning service that provides an easy platform for users to train and build models and an environment to deploy them
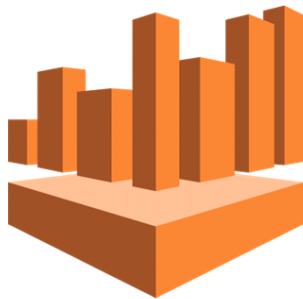
Data Analysis

Amazon Athena is a serverless, interactive query service that makes it easy to analyze data in Amazon S3, using basic SQL

Amazon Athena

Data Analysis

Amazon Elasticsearch Service is a fully managed, cost-effective AWS service that helps us deploy, secure, and run Elasticsearch and scale it
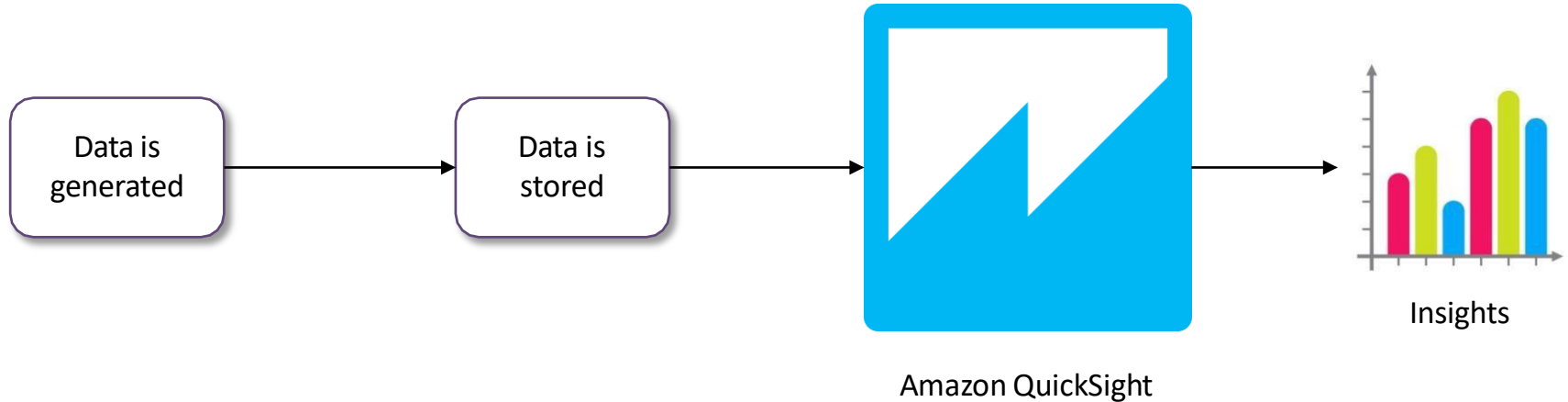
Data Visualization

Data Visualization and Security

Amazon QuickSight is a fast and interactive visualization tool that helps users perform analysis and derive insights from data

Data is generated → Data is stored → Amazon QuickSight → Insights

+919030485102

rganesh0203@gmail.com

https://topmate.io/rganesh_0203