DSCI 5350 – Big Data Analytics

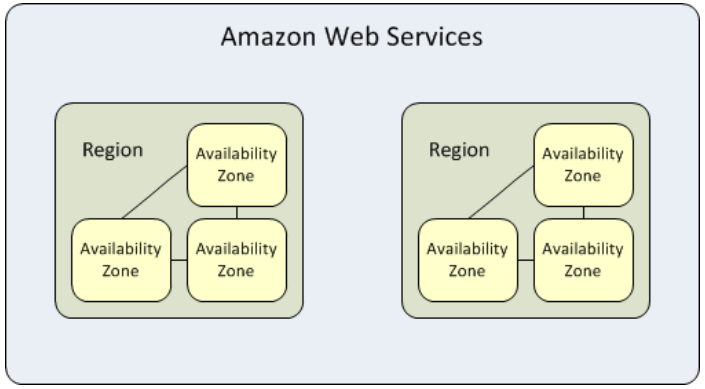
Lecture x – AWS Basics

Kashif Saeed

AWS – Regions and Availability Zones (AZ)

Regions & AZ

- ✓Region is a geographic area
- ✓Regions are completely isolated from one another
- ✓Resources in AWS are tied to a region
- ✓A region has several Availability Zones
- ✓Each Availability Zone is isolated, but the Availability Zones in a Region are connected through low-latency links
- ✓Availability zones consist of one or more data centers, each with redundant power, networking, and connectivity, housed in separate facilities.



AWS Regions

Code	Name
us-east-1	US East (N. Virginia)
us-east-2	US East (Ohio)
us-west-1	US West (N. California)
us-west-2	US West (Oregon)
ca-central-1	Canada (Central)
eu-central-1	EU (Frankfurt)
eu-west-1	EU (Ireland)
eu-west-2	EU (London)
eu-west-3	EU (Paris)
eu-north-1	EU (Stockholm)
ap-east-1	Asia Pacific (Hong Kong)
ap-northeast-1	Asia Pacific (Tokyo)
ap-northeast-2	Asia Pacific (Seoul)
ap-northeast-3	Asia Pacific (Osaka-Local)
ap-southeast-1	Asia Pacific (Singapore)
ap-southeast-2	Asia Pacific (Sydney)
ap-south-1	Asia Pacific (Mumbai)
sa-east-1	South America (São Paulo)

AWS Terminologies – EC2

EC2 (Elastic Compute Cloud)

- ✓Service to create a VM in cloud
- ✓You only pay for the capacity you use
  - ○dedicated host is an exception
- ✓Options:
  - ○On-demand → pay for usage, no contract, no upfront payment
  - ○Reserved → 1-3 year contract to reserve a server
  - ○Spot → allows you to bid a price
  - ○Dedicated host → Physical dedicated server

AWS Terminologies – Instance Store

Instance Store (EC2)

- ✓Provides temporary block-level storage for your instance
- ✓Data persists for the lifetime of the EC2 instance
  - oData will not persist if the instance stops or terminates
- ✓This storage is located on disks that are physically attached to the host computer
- ✓Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content
- ✓Not ideal for data that you'd like to permanently store

AWS Terminologies – S3

S3 (Simple Storage Service)

- ✓Object based storage with NO storage maximum
- ✓99.9% availability
- ✓Files are stored in buckets (folders)
- ✓Buckets have a universal namespace & each bucket has a unique URL
- ✓Bucket URL: https://s3-<region>/amazonaws.com/<bucketname>
- ✓Storage Classes:
  - S3 Standard
  - S3 IA (infrequently accessed) – lower cost
  - S3 One Zone IA – even lower cost (stored in 1 zone; can't be used for HA)
  - Glacier – lowest cost; used for archival
- ✓Use Cases: Backup & Restore, DR, Data Archival, Data Lakes

AWS Terminologies – EBS

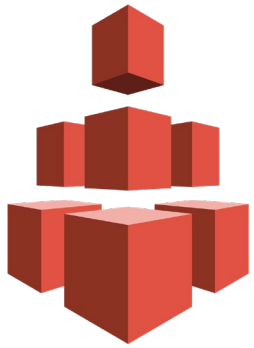
EBS (Elastic Block Storage)

- ✓Persistent block storage volumes for EC2 instance
- ✓Network attached storage that can be mapped to multiple EC2 instances
- ✓Unlike S3, you must have an EC2 instance to access EBS
- ✓Each Amazon EBS volume is automatically replicated within its Availability Zone for HA
- ✓EC2 instance and its EBS volume are in the same AZ
- ✓Use Cases:
  - Storage for Relational and NoSQL databases
  - Storage for streaming applications
  - Storage for data warehouses

AWS Terminologies – EFS

EFS (Elastic File System)

- ✓Shared storage that can work with thousands of EC2 instances
- ✓Pay for usage; no upfront cost
- ✓Amazon EFS is a regional service storing data within and across multiple Availability Zones (AZs) for high availability and durability
- ✓EBS is only available in a particular region; you can share files between regions on multiple EFS instances
- ✓Use cases: Container Storage, Content Management, Analytics



AWS Storage Comparison: https://cloud.netapp.com/blog/ebs-efs-amazons3-best-cloud-storage-system

AWS Terminologies – RDS

RDS (Relational Database Service)

- ✓A service to set up, operate, and scale a relational database in the cloud
- ✓Automates hardware provisioning, database setup, patching and backups
- ✓Allows you to choose from six database engines: Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and SQL Server
- ✓Use Cases: Operational databases like web/mobile/ecommerce backend databases

AWS Storage Comparison: https://cloud.netapp.com/blog/ebs-efs-amazons3-best-cloud-storage-system

AWS Terminologies – Redshift

Redshift

- ✓Amazon Redshift is a fast, scalable data warehouse running on columnar storage on high-performance disk
- ✓Allows you to query your Data lake
  - oAmazon Redshift extends your data warehouse to your data lake to help you gain unique insights that you could not get by querying independent data silos
  - oYou can directly query open data formats stored in Amazon S3 with Redshift Spectrum, a feature of Redshift, without the need for unnecessary data movement
  - oThis enables you to analyze data across your data warehouse and data lake, together, with a single service

AWS Terminologies – DynamoDB

DynamoDB

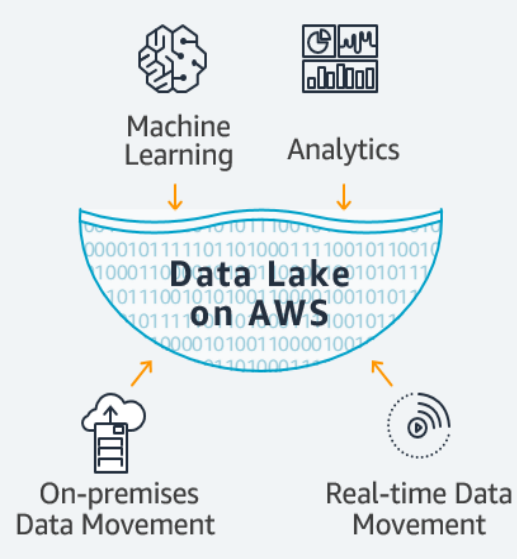
- ✓Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale
- ✓Ideal for the backend of mobile or web applications with really high traffic
- ✓DynamoDB supports ACID transactions to enable you to build business-critical applications at scale
- ✓Serverless: With DynamoDB, there are no servers to provision, patch, or manage and no software to install, maintain, or operate

Querying is not that easy in NoSQL databases – if you're not dealing with the transaction load like Amazon, Uber, Redfin, etc., you can live with RDS

Building a Data Lake in AWS

Data Lakes and Analytics in AWS

- AWS provides all the tools necessary to build and maintain a data lake in AWS for analytics
- Analytics and ML services are available to use the data in the data lake
- AWS also provides tools for moving On-premises and real-time data into the data lake
- Customers like NASDAQ, Zillow, Yelp, iRobot, and FINRA use AWS for their Analytics work load
- In the next few slides we will focus on the data movement, storage, and Analytics/ML services related to a data lake solution in AWS



Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

1. Data Movement to AWS Data Lake

On-premises data movement

- ✓AWS Direct Connect – a dedicated network connection between your data center and AWS
- ✓AWS Snowball – a storage device that can be used to transfer Petabytes of data into AWS if you'd like to avoid internet transfers using Direct Connect
- ✓AWS Snowmobile – a data center in a truck that can handle Petabytes to Exabytes of data
- ✓Storage Gateway – a service that allows on-premises applications to write directly to S3 by showing virtual hard drives of S3 in your Gateway VM

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

1. Data Movement to AWS Data Lake

Real-time (streaming) data movement

- ✓AWS Kinesis – allows you to ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data
  - Real-time: Kinesis is a service that allows you to ingest, buffer, and process streaming data in real-time
  - Fully Managed: Amazon Kinesis is fully managed and runs your streaming applications without requiring you to manage any infrastructure
  - Scalable: Kinesis is seamlessly scalable to any amount of streaming data

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

More about Amazon Kinesis

Amazon Kinesis capabilities

Kinesis Video Streams

Capture, process, and store video streams

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing.

Learn more »

Kinesis Data Streams

Capture, process, and store data streams

Amazon Kinesis Data Streams is a scalable and durable real-time data streaming service that can continuously capture gigabytes of data per second from hundreds of thousands of sources.

Learn more »

Kinesis Data Firehose

Load data streams into AWS data stores

Amazon Kinesis Data Firehose is the easiest way to capture, transform, and load data streams into AWS data stores for near real-time analytics with existing business intelligence tools.

Learn more »

Kinesis Data Analytics

Analyze data streams with SQL or Java

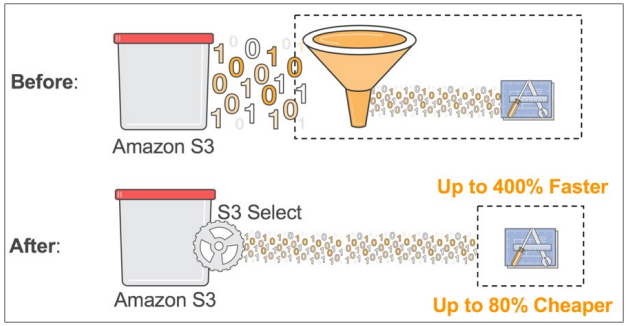
Amazon Kinesis Data Analytics is the easiest way to process data streams in real time with SQL or Java without having to learn new programming languages or processing frameworks.

Learn more »

Source: https://aws.amazon.com/kinesis/?nc2=h_m1

2. Data Storage for AWS Data Lake

- AWS uses S3 and Amazon Glacier for storing data in the data lake
- Amazon S3 – object storage for more frequently queried data
  - ✓Is secure, highly scalable, durable object storage with millisecond latency for data access
  - ✓Can store structured, unstructured, or semi-structured data
  - ✓S3 Select, a feature that enables applications to retrieve only a subset of data from an object by using simple SQL expressions, reduces response times up to 400%



Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

2. Data Storage for AWS Data Lake

- Amazon Glacier – extremely low cost storage for long-term backup and archival
  - ✓Is secure, highly scalable, durable object storage with minutes latency for data access
  - ✓Costs as little as $0.004 per gigabyte per month
  - ✓Glacier Select reads and retrieves only the data needed
- AWS Glue – a fully managed extract, transform, and load (ETL) service
  - ✓You simply point AWS Glue to your data stored on AWS, and AWS Glue discovers your data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog
  - ✓Once cataloged, your data is immediately searchable, queryable, and available for ETL

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

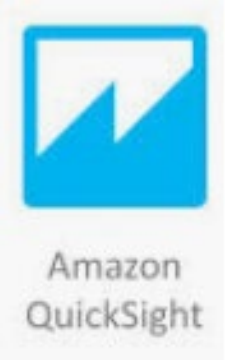
3. Analytics in AWS Data Lake

- Amazon Athena – interactive analysis using SQL
  - ✓Standard SQL for data in S3 and Glacier
  - ✓Serverless: no infrastructure to setup or manage
  - ✓You only pay for the queries you run
- Amazon EMR– big data processing
  - ✓Managed service for big data processing using the Spark and Hadoop frameworks
  - ✓supports 19 different open-source projects including Hadoop, Spark, HBase, and Presto, with managed EMR Notebooks for data engineering, data science development, and collaboration
- Amazon Redshift – data warehousing solution on your data in the data lake

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

3. Analytics in AWS Data Lake

- Amazon Kinesis – real-time analytics
  - ✓enable you to process, and analyze data as it arrives in your data lake, and respond in real-time instead of having to wait until all your data is collected before the processing can begin
- Amazon QuickSight– for dashboards and visualizations
  - ✓provides you a fast, cloud-powered business analytics service, that that makes it easy to build stunning visualizations and rich dashboards that can be accessed from any browser or mobile device



Source: https://aws.amazon.com/big-data/datalakes-and-analytics/

4. Machine Learning in AWS Data Lake

- AWS Deep Learning AMIs – ML and DL framework for Data Scientists
  - ✓ allows you to launch Amazon EC2 instances pre-installed with popular deep learning frameworks and interfaces such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, and Keras
- Amazon SageMaker – for developers to consume ML
  - ✓a platform service that makes the entire process of building, training, and deploying ML models easy by providing everything you need to connect to your training data, select, and optimize the best algorithm and framework, and deploy your model on auto-scaling clusters of Amazon EC2

Source: https://aws.amazon.com/big-data/datalakes-and-analytics/