**Lecture Notes for AWS Solution Architect**

**MODULE-7**

**Amazon EMR**

Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto. Amazon EMR makes it easy to set up, operate, and scale your big data environments by automating time-consuming tasks like provisioning capacity and tuning clusters. With EMR you can run petabyte-scale analysis at less than half of the cost of traditional on-premises solutions and over 3x faster than standard Apache Spark. You can run workloads on Amazon EC2 instances, on Amazon Elastic Kubernetes Service (EKS) clusters, or on-premises using EMR on AWS Outposts.

**Benefits**

**Easy to use**

Analysts, data engineers, and data scientists can use EMR Notebooks to collaborate and interactively explore, process, and visualize data. Simply specify the version of EMR applications and type of compute you want to use. EMR takes care of provisioning, configuring, and tuning clusters so that you can focus on running analytics.

**Low cost**

EMR pricing is simple and predictable: You pay a per-instance rate for every second used, with a one-minute minimum charge. You can launch a 10-node EMR cluster for as little as $0.15 per hour. You can save 50-80% on the cost of the instances by selecting Amazon EC2 Spot for transient workloads and Reserved Instances for long-running workloads.

**Elastic**

Unlike the rigid infrastructure of on-premises clusters, EMR decouples compute and storage, giving you the ability to scale each independently and take advantage of the tiered storage of Amazon S3. With EMR, you can provision one, hundreds, or thousands of compute instances or containers to process data at any scale. The number of instances can be increased or decreased automatically using Auto Scaling (which manages cluster sizes based on utilization) and you only pay for what you use.

**Reliable**

Spend less time tuning and monitoring your cluster. EMR is tuned for the cloud and constantly monitors your cluster — retrying failed tasks and automatically replacing poorly performing instances. Clusters are highly available and automatically failover in the event of a node failure. EMR provides the latest stable open source software releases, so you don’t have to manage updates and bug fixes, which leads to fewer issues and less effort to maintain your environment.

**Secure**

EMR automatically configures EC2 firewall settings, controlling network access to instances and launches clusters in an Amazon Virtual Private Cloud (VPC). Server-side encryption or client-side encryption can be used with the AWS Key Management Service or your own customer-managed keys. EMR makes it easy to enable other encryption options, like in-transit and at-rest encryption, and strong authentication with Kerberos. You can use AWS Lake Formation or Apache Ranger to apply fine-grained data access controls for databases, tables, and columns.

**Flexible**

You have complete control over your EMR clusters and your individual EMR jobs. You can launch EMR clusters with custom Amazon Linux AMIs and easily configure the clusters using scripts to install additional third party software packages. EMR enables you to reconfigure applications on running clusters on the fly without the need to relaunch clusters. Also, you can customize the execution environment for individual jobs by specifying the libraries and runtime dependencies in a Docker container and submit them with your job.

**Deployment options**

**Amazon EMR on Amazon EC2**

You can deploy EMR on Amazon EC2 and take advantage of On-Demand, Reserved, and Spot Instances. EMR manages provisioning, management, and scaling of the EC2 instances. AWS offers more instance options than any other cloud provider, allowing you to choose the instance that gives you the best performance or cost for your workload.

**Amazon EMR on Amazon EKS**

You can use EMR to run Apache Spark jobs on demand on Amazon Elastic Kubernetes Service (EKS), without needing to provision EMR clusters, to improve resource utilization and simplify infrastructure management. Amazon EKS gives you the flexibility to start, run, and scale Kubernetes applications in the AWS cloud or on-premises. With Amazon EMR on EKS, you can share compute and memory resources across all of your applications and use a single set of Kubernetes tools to centrally monitor and manage your infrastructure.

**Amazon EMR on AWS Outposts**

Amazon EMR is available on AWS Outposts, allowing you to set up, deploy, manage, and scale EMR in your on-premises environments, just as you would in the cloud. AWS Outposts brings AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility.

**Use cases**

**Machine learning**

Use EMR's built-in machine learning tools, including Apache Spark MLlib, TensorFlow, and Apache MXNet for scalable machine learning algorithms, and use custom AMIs and bootstrap actions to easily add your preferred libraries and tools to create your own predictive analytics toolset.

**Extract, transform, load (ETL)**

EMR can be used to quickly and cost-effectively perform data transformation workloads (ETL) such as sort, aggregate, and join on large datasets.

**Clickstream analysis**

Analyze clickstream data from Amazon S3 using Apache Spark and Apache Hive to segment users, understand user preferences, and deliver more effective ads.

**Real-time streaming**

Analyze events from Apache Kafka, Amazon Kinesis, or other streaming data sources in real-time with Apache Spark Streaming and Apache Flink to create long-running, highly available, and fault-tolerant streaming data pipelines on EMR. Persist transformed data sets to S3 or HDFS and insights to Amazon Elasticsearch Service.

**Interactive analytics**

EMR Notebooks provide a managed analytic environment based on open-source Jupyter that allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analyses.

**Genomics**

EMR can be used to process vast amounts of genomic data and other large scientific data sets quickly and efficiently. Researchers can access genomic data hosted for free on AWS.

**Amazon ElastiCache**

Amazon ElastiCache allows you to seamlessly set up, run, and scale popular open-source compatible in-memory data stores in the cloud. Build data-intensive apps or boost the performance of your existing databases by retrieving data from high throughput and low latency in-memory data stores. Amazon ElastiCache is a popular choice for real-time use cases like Caching, Session Stores, Gaming, Geospatial Services, Real-Time Analytics, and Queuing.

Amazon ElastiCache offers fully managed Redis, voted the most loved database by developers in the Stack Overflow 2020 Developer Survey, and Memcached for your most demanding applications that require sub-millisecond response times.

**Benefits**

**Extreme performance**

Amazon ElastiCache works as an in-memory data store and cache to support the most demanding applications requiring sub-millisecond response times. By utilizing an end-to-end optimized stack running on customer dedicated nodes, Amazon ElastiCache provides secure, blazing fast performance.

**Fully managed**

You no longer need to perform management tasks such as hardware provisioning, software patching, setup, configuration, monitoring, failure recovery, and backups. ElastiCache continuously monitors your clusters to keep your workloads up and running so that you can focus on higher value application development.

**Scalable**

Amazon ElastiCache can scale-out, scale-in, and scale-up to meet fluctuating application demands. Write and memory scaling is supported with sharding. Replicas provide read scaling.



**Amazon ElastiCache Engines**

1. **Amazon ElastiCache for Redis**

Amazon ElastiCache for Redis is a blazing fast in-memory data store that provides sub-millisecond latency to power internet-scale real-time applications. Built on open-source Redis and compatible with the Redis APIs, ElastiCache for Redis works with your Redis clients and uses the open Redis data format to store your data. Your self-managed Redis applications can work seamlessly with ElastiCache for Redis without any code changes. ElastiCache for Redis combines the speed, simplicity, and versatility of open-source Redis with manageability, security, and scalability from Amazon to power the most demanding real-time applications in Gaming, Ad-Tech, E-Commerce, Healthcare, Financial Services, and IoT.

1. **Amazon ElastiCache for Memcached**

Amazon ElastiCache for Memcached is a Memcached-compatible in-memory key-value store service that can be used as a cache or a data store. It delivers the performance, ease-of-use, and simplicity of Memcached. ElastiCache for Memcached is fully managed, scalable, and secure - making it an ideal candidate for use cases where frequently accessed data must be in-memory. It is a popular choice for use cases such as Web, Mobile Apps, Gaming, Ad-Tech, and E-Commerce.