AWS Certification – Analytics Services – Cheat Sheet

Data Pipeline

- orchestration service that helps define data-driven workflows to automate and schedule regular data movement and data processing activities
- integrates with on-premises and cloud-based storage systems
- allows scheduling, retry, and failure logic for the workflows

EMR

- is a web service that utilizes a hosted Hadoop framework running on the web-scale infrastructure of EC2 and S3
- launches all nodes for a given cluster in the same Availability Zone, which improves performance as it provides higher data access rate
- seamlessly supports Reserved, On-Demand and Spot Instances
- consists of Master Node for management and Slave nodes, which consists of Core nodes holding data and Task nodes for performing tasks only
- is fault tolerant for slave node failures and continues job execution if a slave node goes down
- does not automatically provision another node to take over failed slaves
- supports Persistent and Transient cluster types

    - Persistent which continue to run
    - Transient which terminates once the job steps are completed
- supports EMRFS which allows S3 to be used as a durable HA data storage

Kinesis

- enables real-time processing of streaming data at massive scale
- provides ordering of records, as well as the ability to read and/or replay records in the same order to multiple Kinesis applications
- data is replicated across three data centers within a region and preserved for 24 hours, by default and can be extended to 7 days
- streams can be scaled using multiple shards, based on the partition key, with each shard providing the capacity of 1MB/sec data input and 2MB/sec data output with 1000 PUT requests per second
- Kinesis vs SQS
    - real-time processing of streaming big data vs reliable, highly scalable hosted queue for storing messages
    - ordered records, as well as the ability to read and/or replay records in the same order vs no guarantee on data ordering (with the standard queues before the FIFO queue feature was released)
    - data storage up to 24 hours, extended to 7 days vs up to 4 days, can be configured from 1 minute to 14 days but cleared if deleted by the consumer

- supports multiple consumers vs single consumer at a time and requires multiple queues to deliver message to multiple consumers