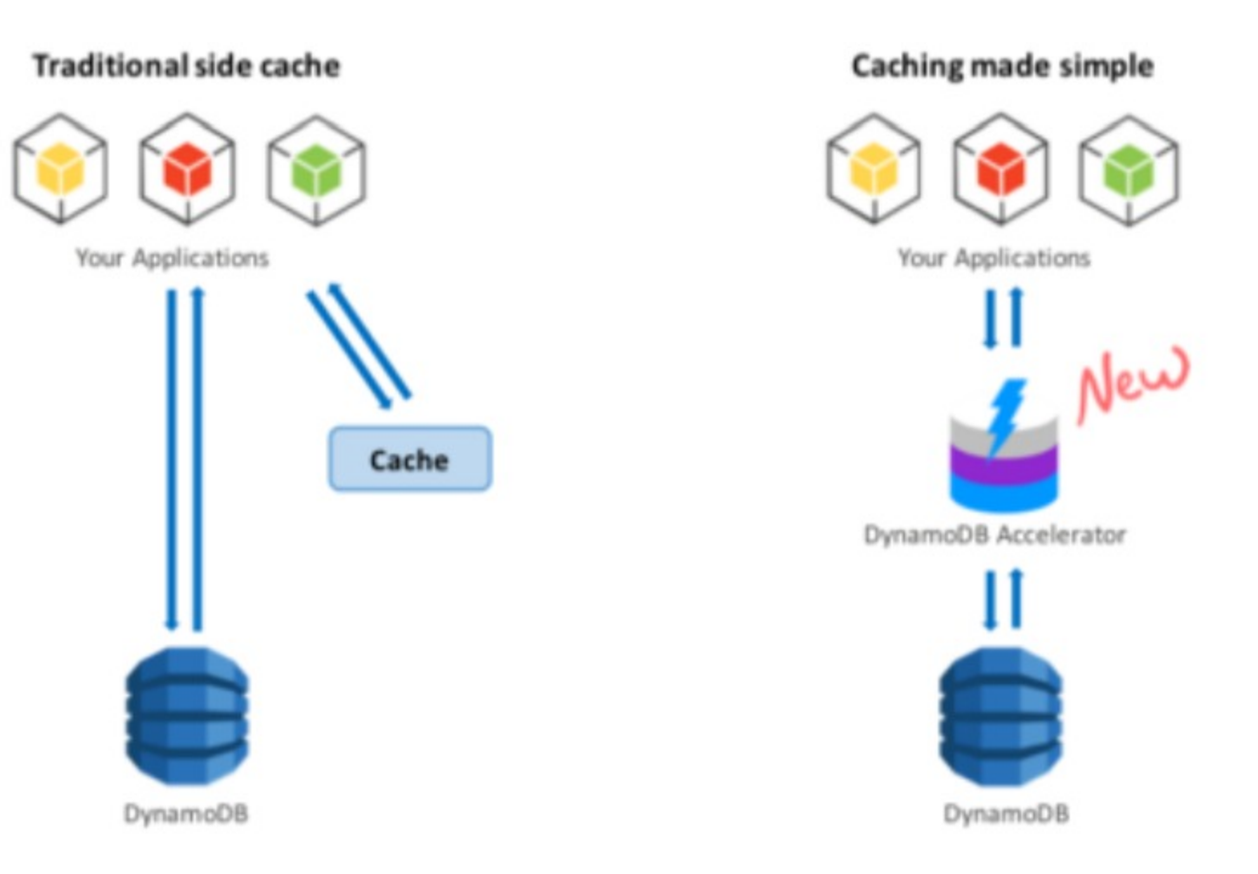
AWS Big Data Specialty Exam Tips and Tricks

05 Dec 2018

If you're planning on taking the AWS Big Data Specialty exam, I've compiled a quick list of tips that you may want to remember headed into the exam.

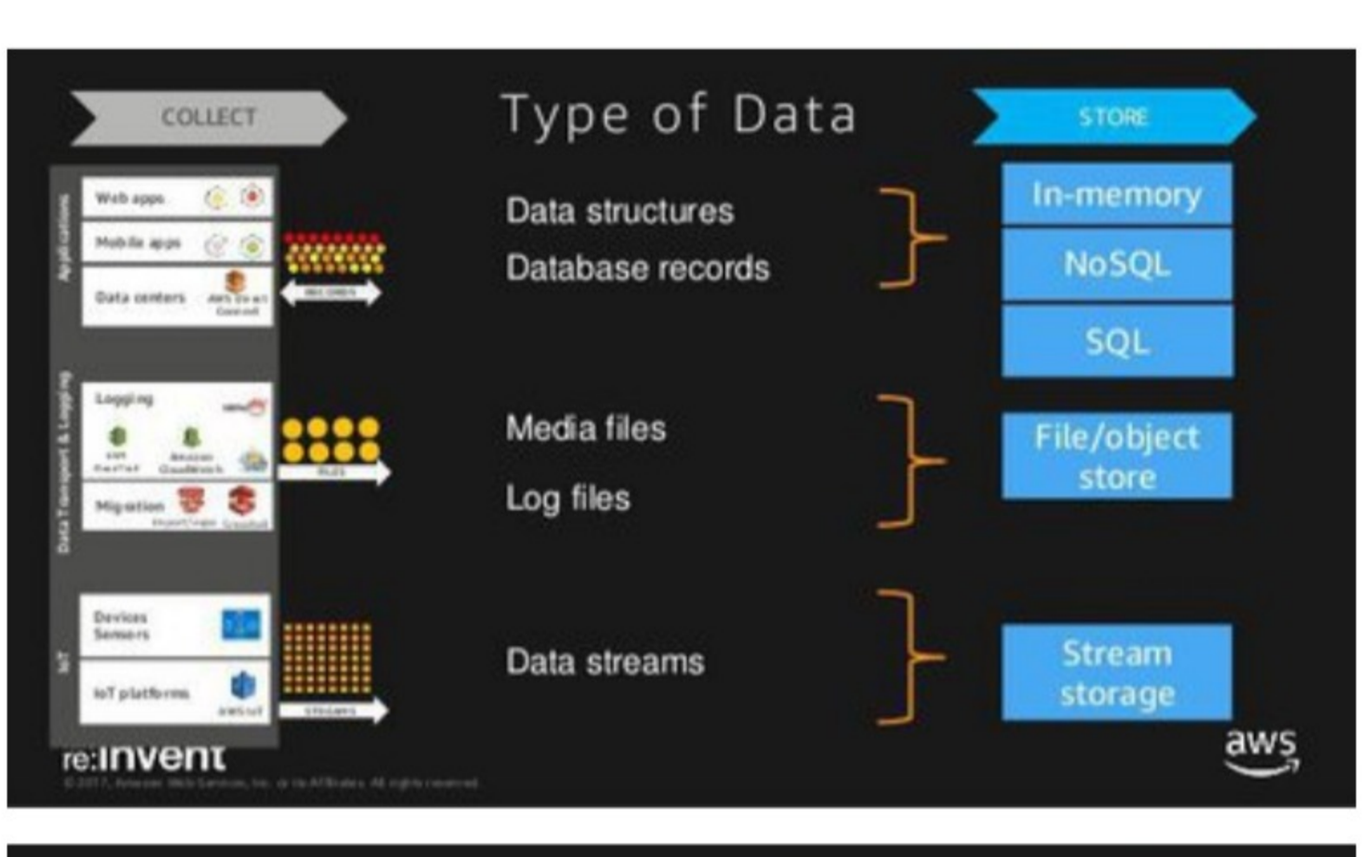
I passed the exam on December 6, 2018 with a score of 76%. **In my opinion, this exam is more difficult than the AWS Solutions Architect Pro!**

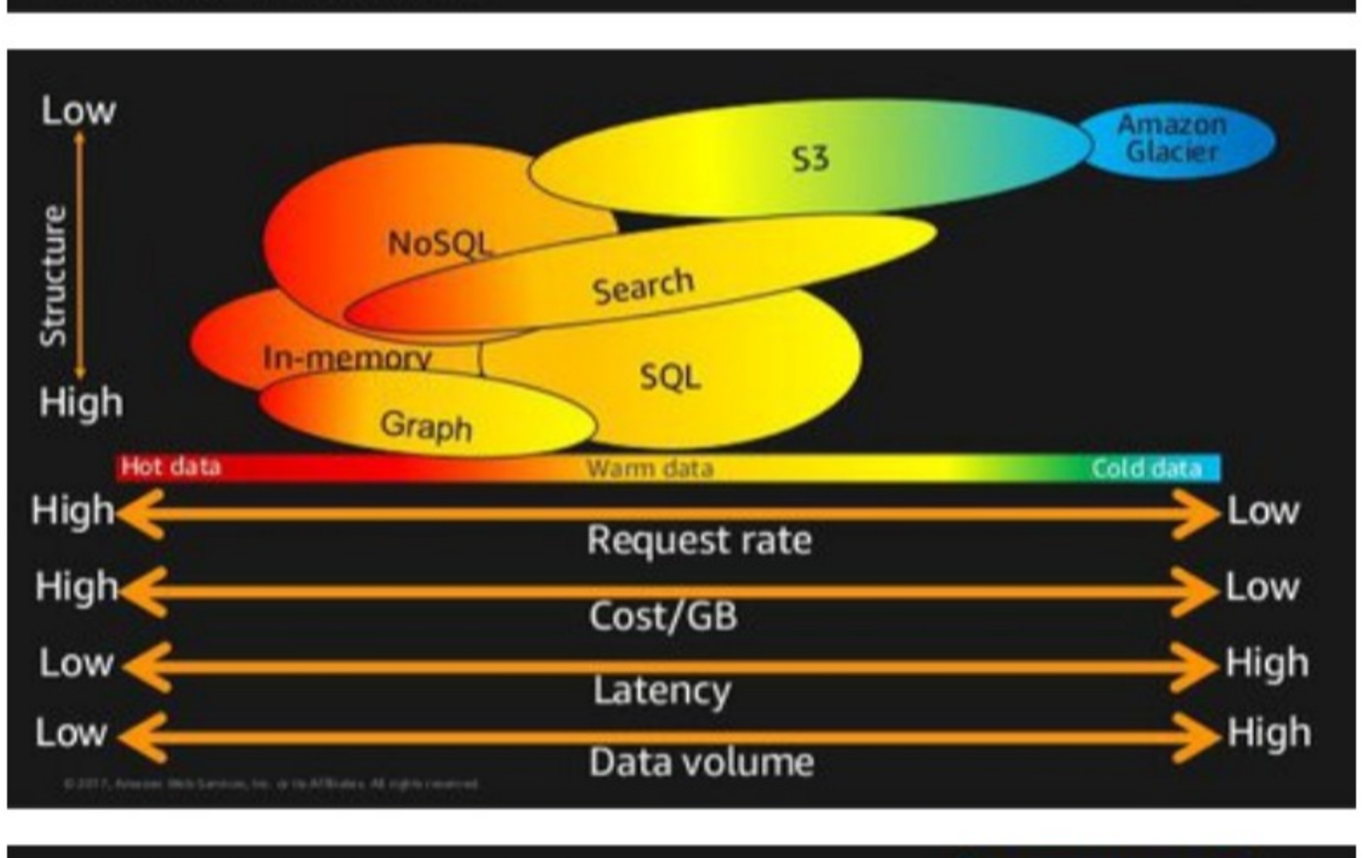
- **You really, really need to understand Redshift distribution strategies. Here are some things to remember:**
  - **Automatic Distribution**: The default option, Redshift automatically manages your distribution strategy for you, shifting from an initial `ALL` strategy (for smaller tables) to `EVEN` distribution (for larger tables). **Note:** Redshift will *not* automatically switch back from `EVEN` to `ALL`.
  - **Even Distribution**: With the `EVEN` distribution, the leader node distributes rows equally across all slices. This is appropriate for tables that do not participate in joining.
  - **Key Distribution**: With the `KEY` distribution, rows are distributed according to a selected column. Tables that share common join keys are physically co-located for performance.
  - **All Distribution**: A copy of the *entire* data set is stored on *each* node. This slows down inserting, updating, and querying. This distribution method is only appropriate for small or rarely-updated data sets.
- You need to know the DynamoDB partition sizing formula by heart: `(Desired RCU/3000 RCU) + (Desired WCU/1000 RCU) = # of partitions needed`
- AWS Machine Learning does not support unsupervised learning - you will need Apache Spark or Spark MLLib for real-time anomaly detection.
- AWS IoT accepts four forms of identity verification: X.509 certificates, IAM users/roles, Cognito identities, and Federated identities. *"Typically, AWS IoT devices use X.509 certificates, while mobile applications use Amazon Cognito identities. Web and desktop applications use IAM or federated identities. CLI commands use IAM."*
- In the context of evaluating a Redshift query plan, `DS_DIST_NONE` and `DS_DIST_ALL_NONE` are good. They indicate that no distribution was required for that step because all of the joins are co-located.
- `DS_DIST_INNER` means that the step will probably have a relatively high cost because the inner table is being redistributed to the nodes. `DS_DIST_ALL_INNER`, `DS_BCAST_INNER` and `DS_DIST_BOTH` are not good. (Source)
- You must disable cross-region snapshots for Redshift before changing the encryption type of the cluster.
- Amazon recommends allocating **three** dedicated master nodes for each production ElasticSearch domain.
- Read up on **DAX** and DynamoDB.

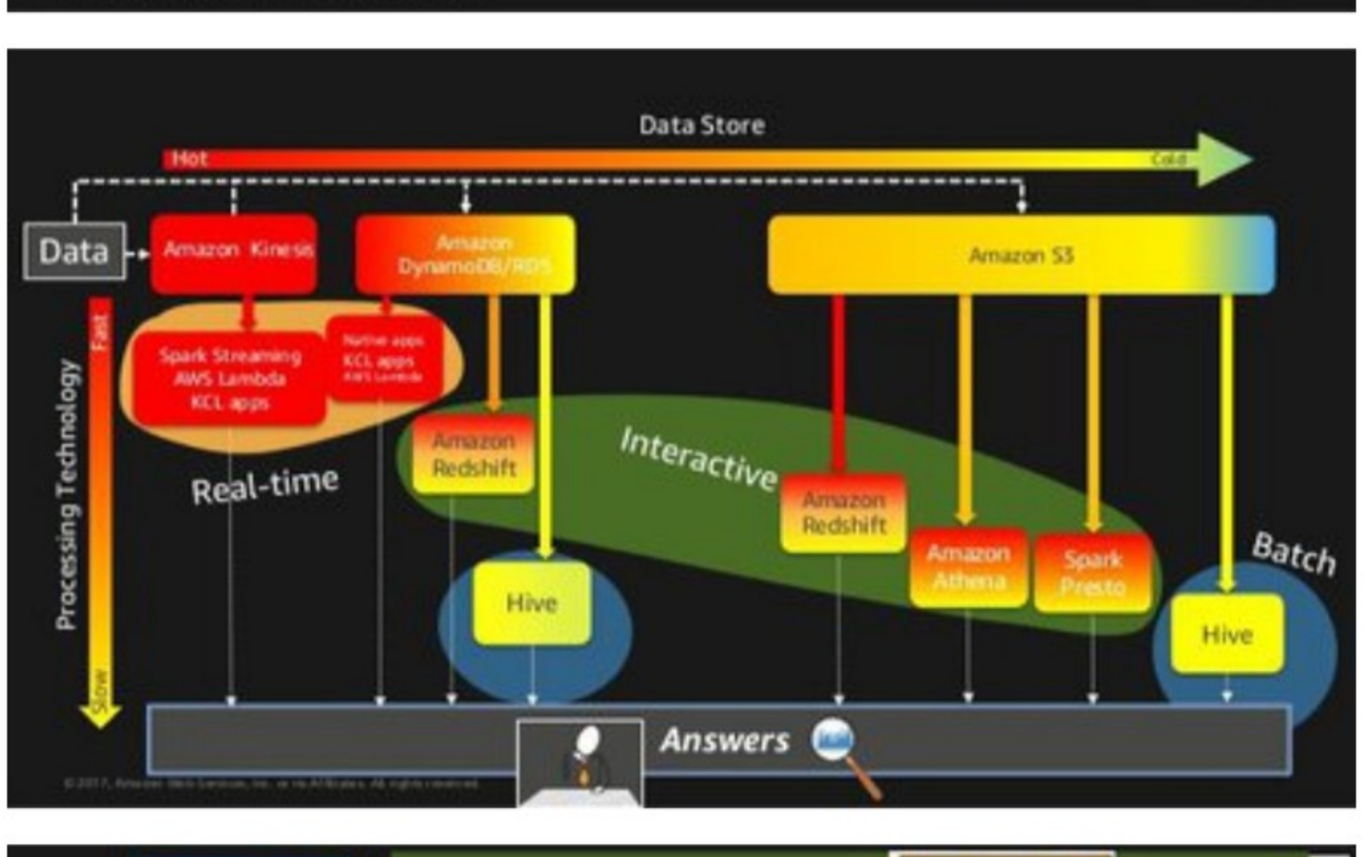


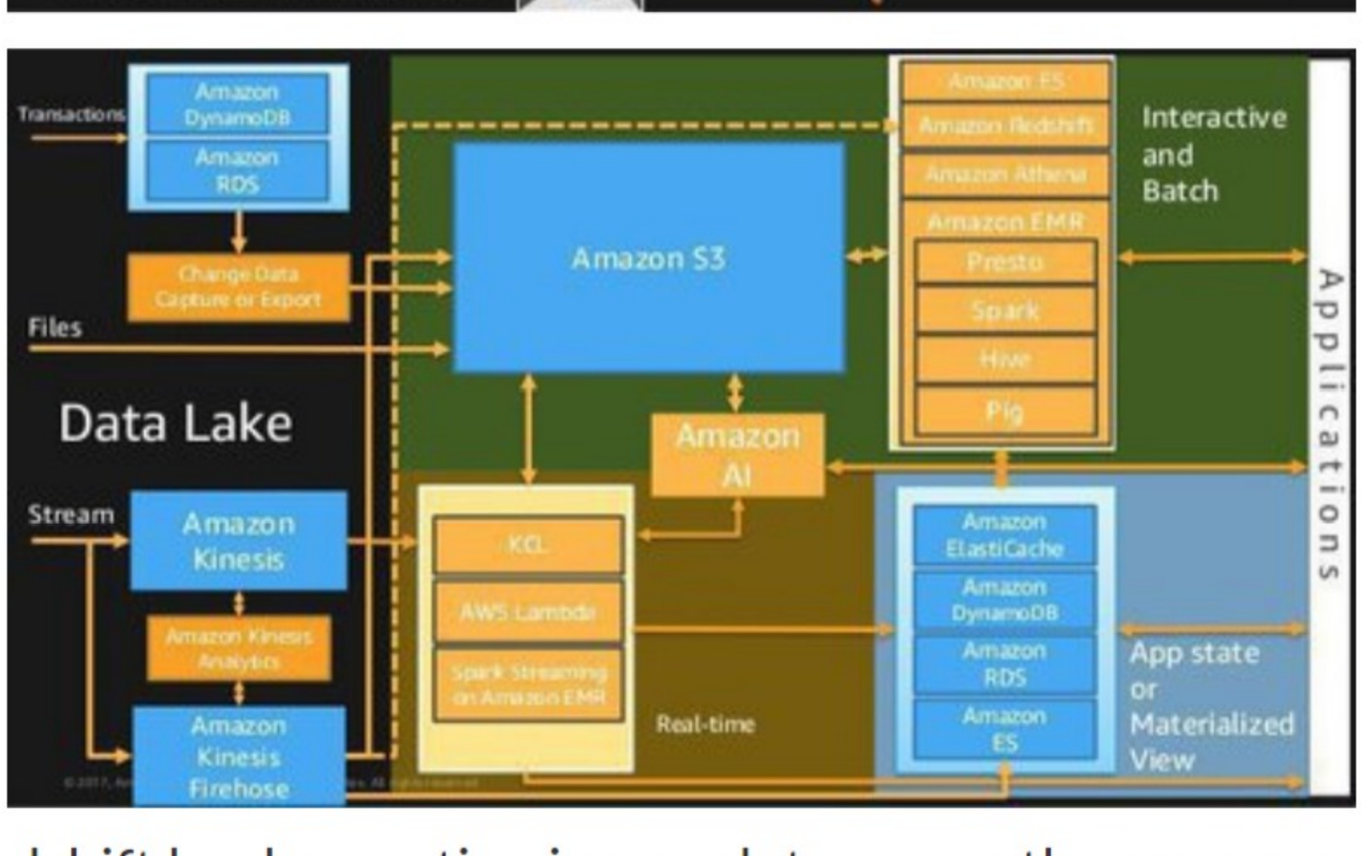
In most cases, the DynamoDB response times can be measured in single-digit milliseconds. However, for use cases that require response times in microseconds, DynamoDB Accelerator (DAX) delivers fast response times for accessing eventually consistent data.

- DynamoDB: Remember to use "write sharding" to allow writes to be distributed evenly across partitions. There are two methods for this: **Random Suffixes** and **Calculated Suffixes**.
- A Kinesis data stream retains records for 24 hours by default, but this can be extended to **168 hours** using the `IncreaseStreamRetentionPeriod` operation.
- To make your data searchable in CloudSearch, you need to format it in **JSON** or **XML**.
- You can use popular BI tools like Excel, MicroStrategy, QlikView, and Tableau with EMR to explore and visualize your data. Many of these tools require an ODBC (Open Database Connectivity) or JDBC (Java Database Connectivity) driver. (Source)
- **Hue** is the web interface for an EMR cluster.
- The following are some fantastic slides from the invaluable AWS re:Invent 2017: Big Data Architectural Patterns and Best Practices on AWS (ABD201) talk. Watch it two or three times.

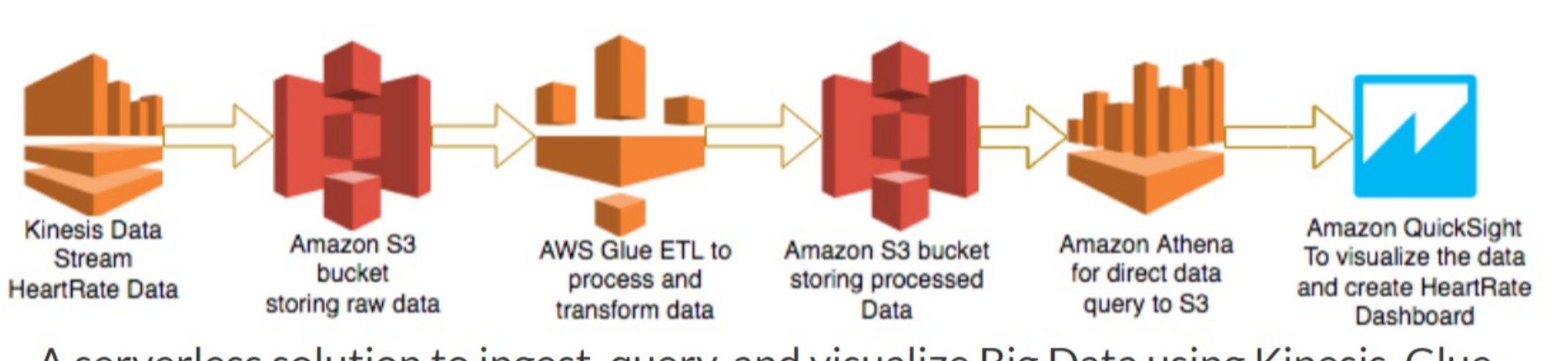






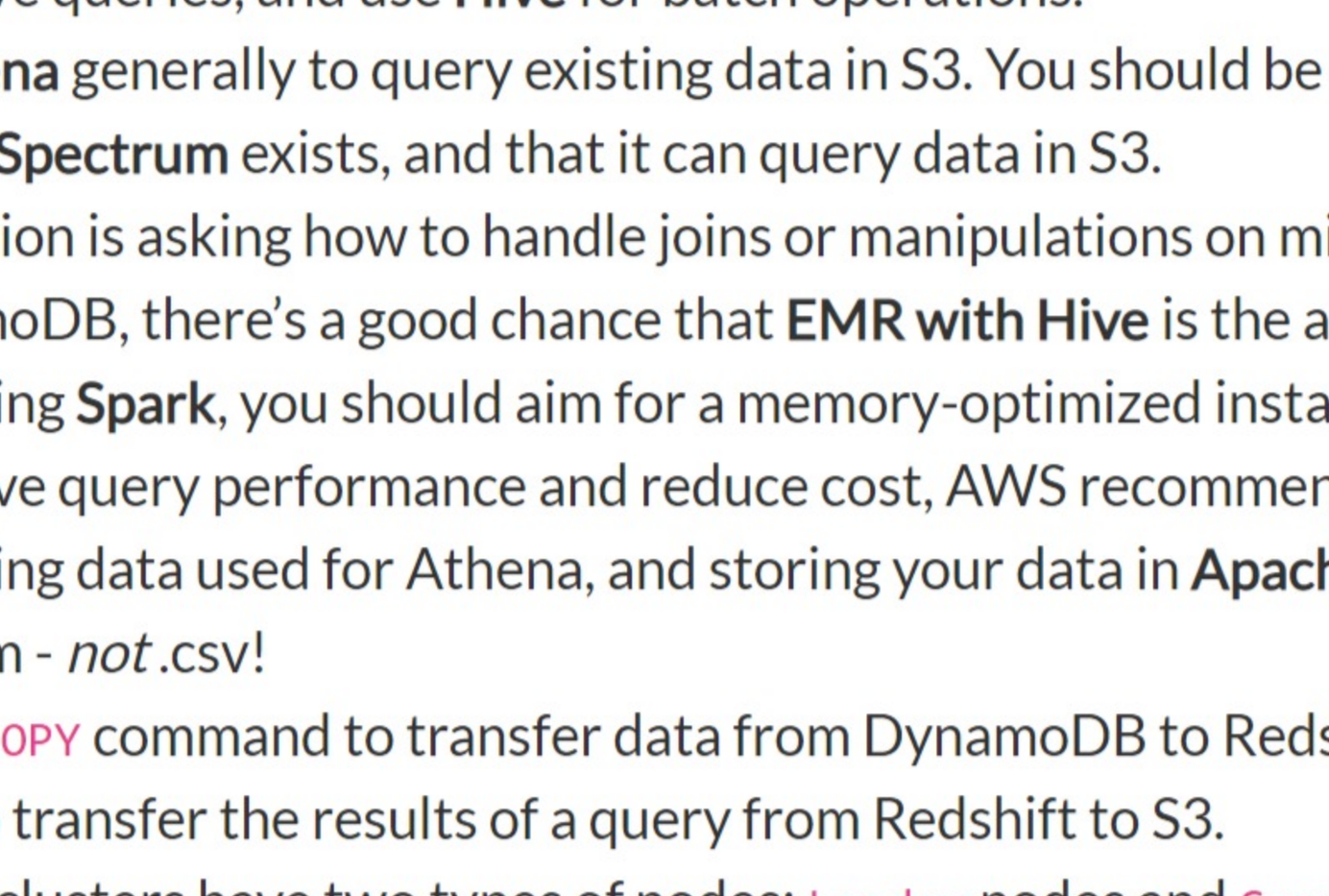


- After a Redshift load operation is complete, query the `STL_LOAD_COMMITS` table to verify that the expected files were loaded.
- **Mahout** is a machine learning library with tools for clustering, classification, and several types of recommenders, including tools to calculate most-similar items or build item recommendations for users. Use it to carry out Machine Learning work on top of Hadoop.
- When preparing to use a Lambda/Kinesis combination, make sure to optimize your Lambda memory and batch size, and adjust the number of shards used by the Kinesis streams.
- Triggers do **not** exist in Redshift.
- Amazon **Kinesis Aggregators** is a Java framework that enables the automatic creation of real-time aggregated time series data from Kinesis streams. (Source)
- You can't encrypt an existing DynamoDB table. You need to create a new, encrypted table and transfer your data over. (Source)
- **Presto** is a fast SQL query engine designed for interactive analytic queries over large datasets from multiple sources. (Source)
- Redshift creates the following log types:
  - **Connection log** — logs authentication attempts, and connections and disconnections.
  - **User log** — logs information about changes to database user definitions.
  - **User activity log** — logs each query before it is run on the database.
- Kinesis Data Firehose can send records to S3, Redshift, or Elasticsearch. It **cannot** send records to DynamoDB. (Source)



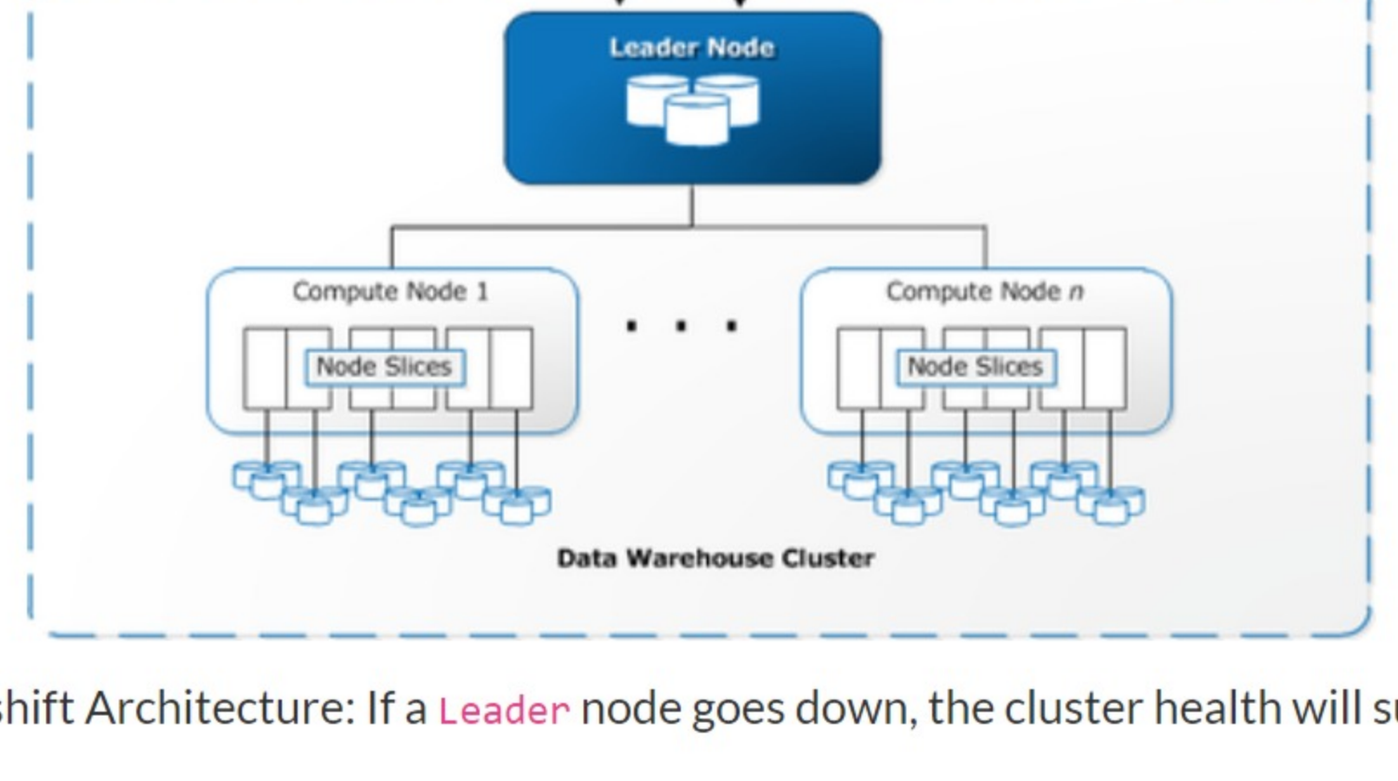
A serverless solution to ingest, query, and visualize Big Data using Kinesis, Glue, Athena, and QuickSight.

- Use **Spark** for general purpose Amazon EMR operations, use **Presto** for interactive queries, and use **Hive** for batch operations.
- Use **Athena** generally to query existing data in S3. You should be aware that **Redshift Spectrum** exists, and that it can query data in S3.
- If a question is asking how to handle joins or manipulations on millions of rows in DynamoDB, there's a good chance that **EMR with Hive** is the answer.
- When using **Spark**, you should aim for a memory-optimized instance type.
- To improve query performance and reduce cost, AWS recommends partitioning data used for Athena, and storing your data in **Apache Parquet** or **ORC** form - *not* .csv!
- Use the `COPY` command to transfer data from DynamoDB to Redshift. Use `UNLOAD` to transfer the results of a query from Redshift to S3.
- Redshift clusters have two types of nodes: `Leader` nodes and `Compute` nodes.

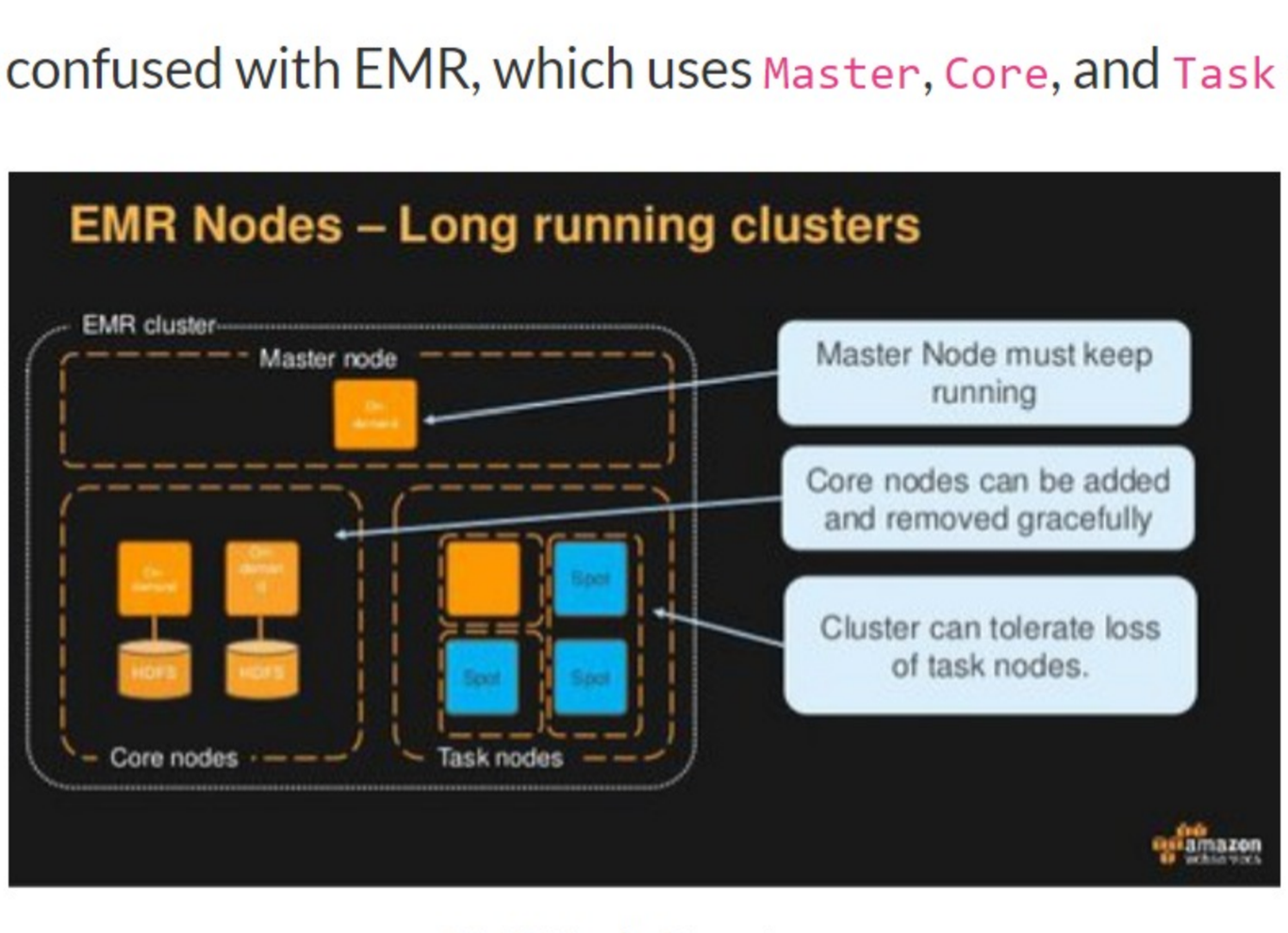


Redshift Architecture: If a `Leader` node goes down, the cluster health will suffer.

- Not to be confused with EMR, which uses `Master`, `Core`, and `Task` nodes.

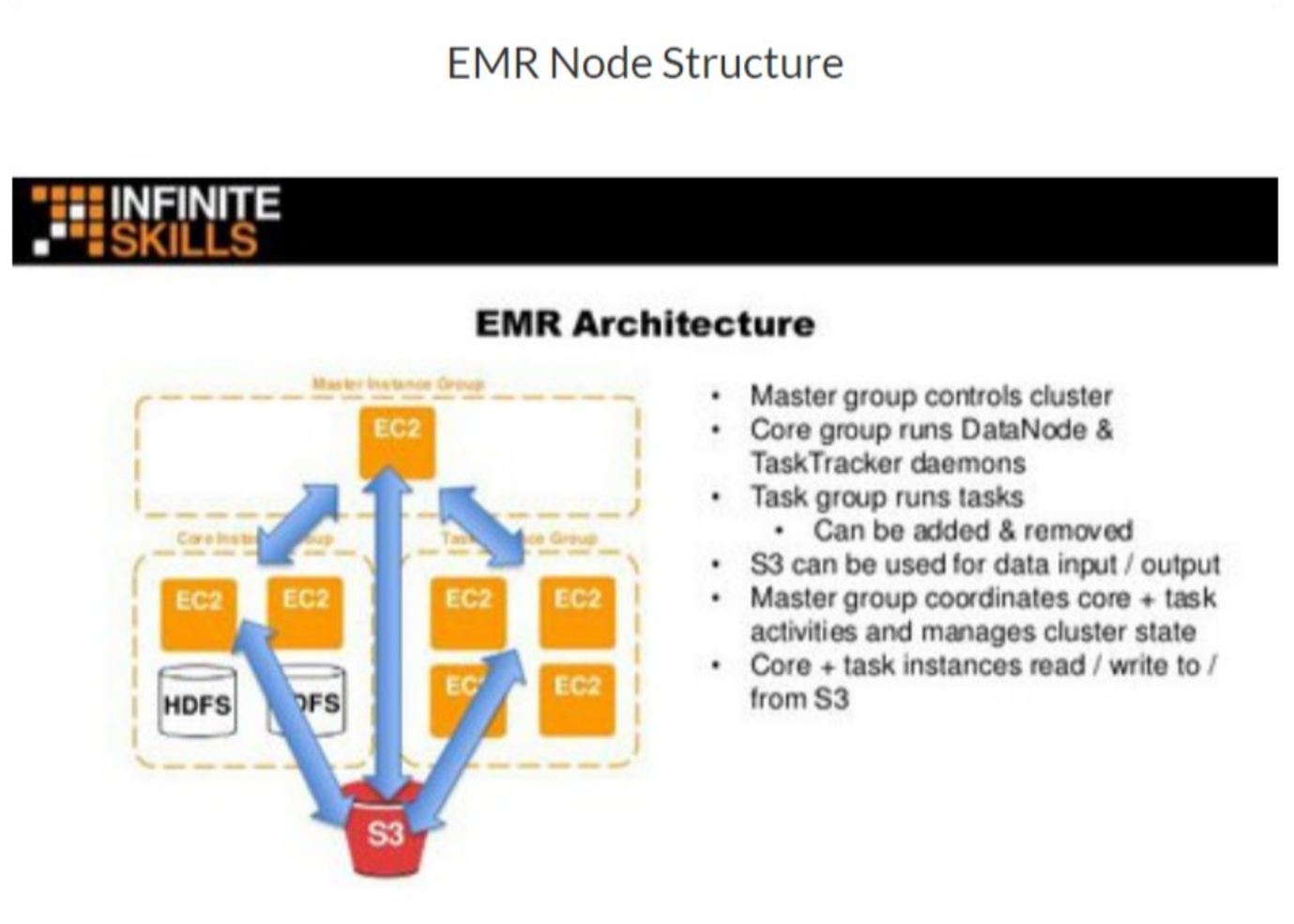


EMR Node Structure



EMR Architecture: EMR stores log files on the `Master` node by default.

- Use ElasticSearch to analyze data stream updates from other services, such as Kinesis Streams and DynamoDB.
- Amazon Schema Conversion Tool is sometimes referred to as SCT.



Review EMR security and encryption

Training Materials I Used

- ACloudGuru - AWS Big Data Specialty Course
- WhizLabs Practice Exams (None of these questions appear on the exam, but they do get you valuable practice with AWS-style questions)

Whitepapers I Read

- Comparing the Use of Amazon DynamoDB and Apache HBase for NoSQL
- Lambda Architecture for Batch and Stream Processing
- Enterprise Data Warehousing on AWS

Other Links

- Top 8 Best Practices for High-Performance ETL Processing Using Amazon Redshift
- Glacier Expedited, Standard, and Bulk Retrieval Types
- Redshift Engineering's Advanced Table Design Playbook
- QuickSight FAQ
- Server-Side Encryption for Kinesis Data Streams
- Amazon EMR - Submit a Streaming Step
- Redshift - Choosing the Best Sort Key
- Capturing Table Activity with DynamoDB Streams
- Redshift: Choosing the best Distribution model
- Visualize AWS Cloudtrail Logs Using AWS Glue and Amazon QuickSight
- Redshift: Vacuuming Tables
- Analyze Apache Parquet optimized data using Kinesis Data Firehose, Athena, and Redshift
- Building a Binary Classification Model with Machine Learning and Redshift
- AWS IoT Authentication
- Use Business Intelligence Tools with EMR
- JOIN Redshift AND RDS PostgreSQL WITH dblink
- Ensuring Consistency When Using S3 and Elastic MapReduce for ETL Workflows

Related Posts

Accessing Arbitrary Paths via String Dot Notation in JavaScript
10 May 2020

COVID-19 - Eye of the Storm
28 Apr 2020

Creating a React/Redux JupyterLab Extension
24 Apr 2020

COVID-19 - Calm Before the Storm
31 Mar 2020