AWS Big Data Training September 2018 - Useful slides

When datasets are **so large** that they are **difficult** to:

- Collect
- Store
- Organize
- Analyze
- Move
- Share

When does data become "big data"?

Dataset Size

The 3 V's

When the data **outgrows your ability to process it** due to the:

- Velocity
- Volume
- Variety

Consider these **four** things when determining the **right tool** for the job:

Data structure

Latency

Throughput

Access patterns

Collect

Near Real-time
Amazon Kinesis Firehose

Data Import
Amazon Import/Export Snowball

Message Queuing
Amazon SQS

Web/app Servers
Amazon EC2

Store

Object Storage
Amazon S3
Amazon Glacier

Near Real-time
Amazon Kinesis Streams

RDBMS
Amazon RDS

NoSQL
Amazon DynamoDB

Search
Amazon CloudSearch

Internet of Things (IoT)
Amazon IoT

Process and Analyze

Hadoop Ecosystem
Amazon EMR

Near Real-time
AWS Lambda
Amazon Kinesis Analytics

Data Warehousing
Amazon Redshift

Machine Learning
Amazon SageMaker

Elastic Search Analytics
Amazon Elasticsearch Service

Process and Move Data
AWS Data Pipeline
AWS Glue

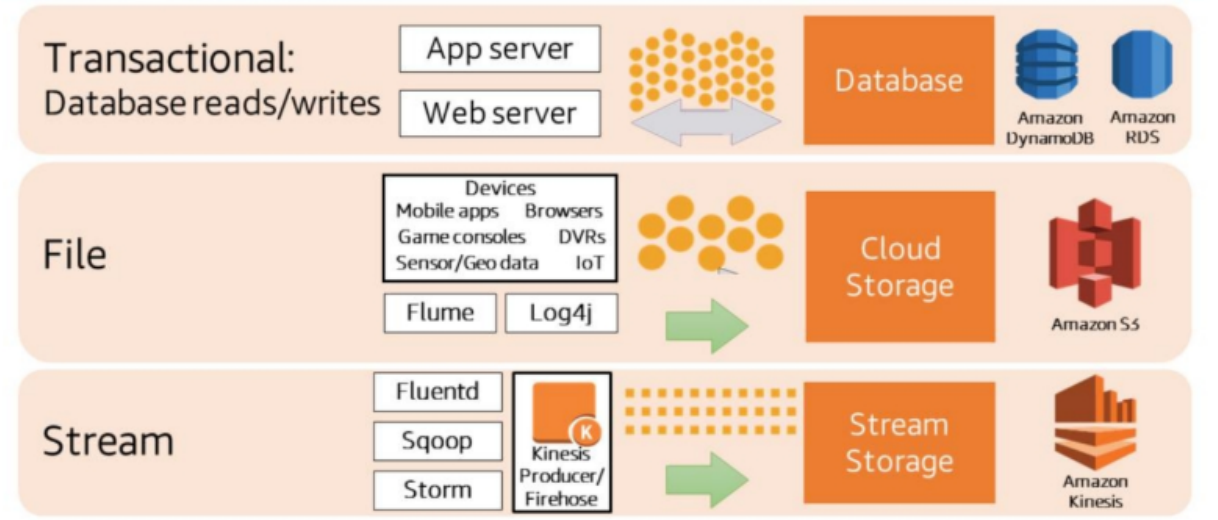
Ad Hoc Analytics
Amazon Athena

Visualize

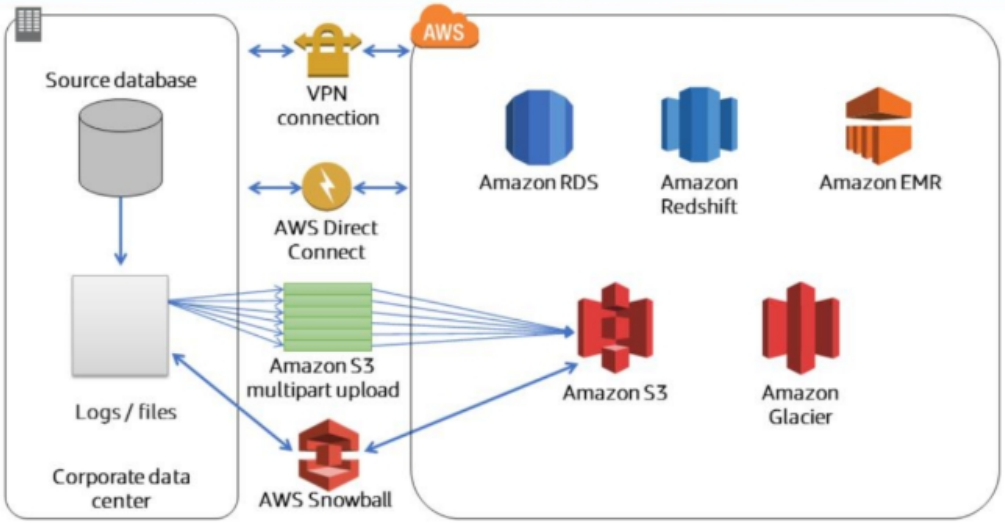
Business Intelligence and Data Visualization
Amazon QuickSight

Elastic Search Analytics
Amazon Elasticsearch Service

Types of Data Ingestion

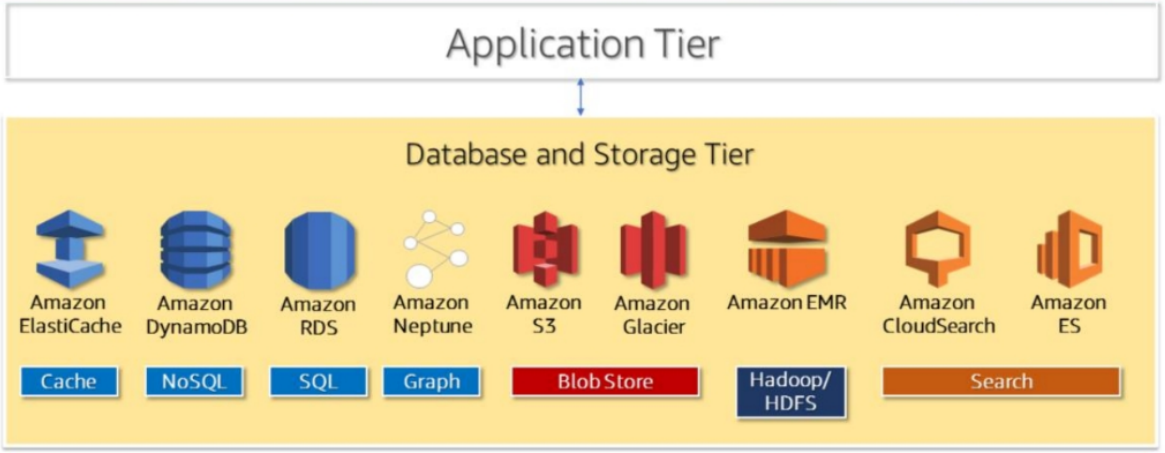


AWS Data Transfer Options



Storage Solutions With AWS

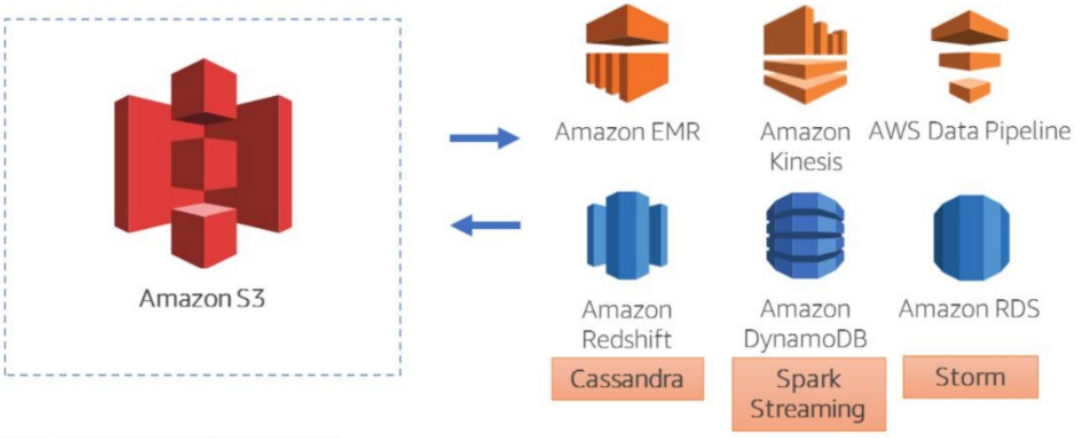
aws training and certification



Recommendation

aws training and certification

Aggregate all data in an S3 data lake surrounded by a collection of the right tools.



Data Store Considerations

aws training and certification

Structure of your data: **Fixed schema**, **JSON**, **key-value**, etc.

Access patterns: **Store** data in the format you will **access** it.

Data/access characteristics: **Hot? Warm? Cold?**

Cost: Is this solution the most **cost effective** for my anticipated usage?

Choosing a Data Store: Data Structure

aws training and certification

Data structure	What to use?
Fixed schema	SQL, NoSQL
Schema-free (JSON)	NoSQL
(Key, Value)	Cache, NoSQL
Graph	GraphDB

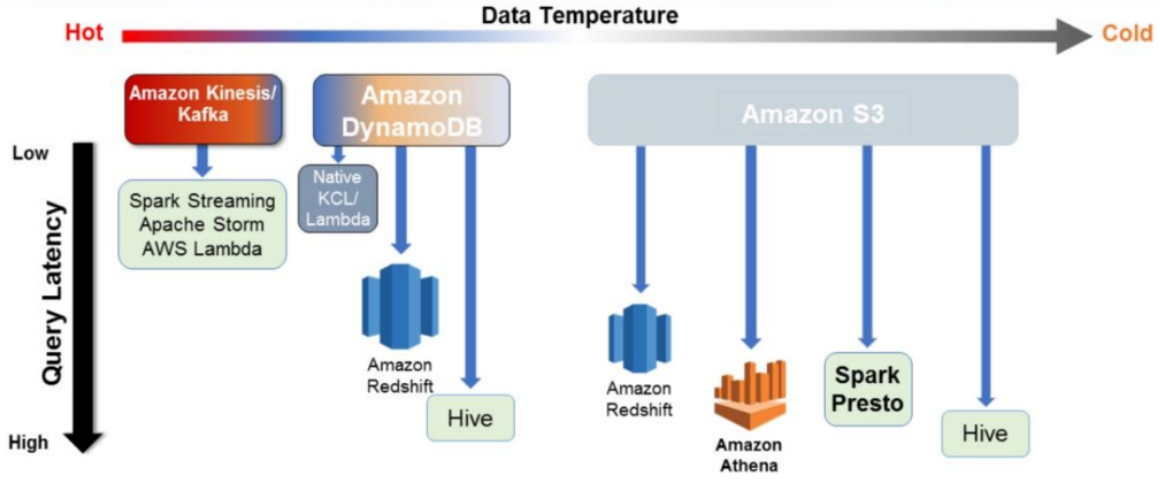
SQL vs. NoSQL

aws training and certification

	Traditional RDBMS	NoSQL
Data model	Fixed schema	Flexible schema
DB Transaction	Full ACID compliance	No ACID compliance
Performance	Optimized for storage	Optimized for compute
Scale	Scale vertically	Scale horizontally

Data Temperature vs. Processing speed

aws training and certification



Amazon Kinesis Data Firehose vs. Amazon Kinesis Data Streams

aws training and certification

Amazon Kinesis Streams

- Custom processing per incoming record
- Sub-1 second processing latency
- Choice of stream processing frameworks

Amazon Kinesis Firehose

- Zero administration
- Processing latency of 60 seconds or higher.
- Ability to use existing analytics tools based on Amazon S3, Amazon Redshift, and Amazon Elasticsearch Service

Amazon Kinesis

aws training and certification

Amazon Kinesis Data Streams
Collect and stream data for ordered, real-time processing

Amazon Kinesis Video Streams
Capture, process, and store video streams for analytics

Amazon Kinesis Data Analytics
Analyze data streams using SQL queries

Amazon Kinesis Data Firehose
Easily load massive volumes of data

Structured Data Stores: Amazon Redshift

aws training and certification

- Petabyte-scale database service
- Optimized for data warehousing
- Columnar storage
- Massively parallel processing architecture to parallelize and distribute SQL operations
- Run complex queries and analytics against structured data using SQL
- More on Amazon Redshift later in this course

Structured Data Stores: Amazon RDS

aws training and certification

- Relational database service
  - Supports MySQL, MariaDB, PostgreSQL, Oracle, Microsoft SQL Server, and Amazon Aurora
  - Automated patches and database backups
  - Multi-AZ deployment for high availability and failover
  - Standard storage and Provisioned IOPS for fast performance
  - Enables you to run complex queries using SQL
- Output Amazon EMR data directly to Amazon RDS using Apache Sqoop
  - Import and export data using MapReduce (map-only jobs)

Unstructured Data Stores: Amazon EMR (HDFS)

aws training and certification

- HDFS (local disk)
  - Storage-optimized instances provide ample disk space
- Data is local and not streamed from Amazon S3
  - Can copy data to Amazon S3 as needed
  - Data is lost when cluster is terminated
- Run queries against unstructured data using common processing frameworks such as Hive, Pig, and Spark
- Typically used for caching results from intermediate steps

Amazon Athena

aws training and certification

- Interactive query service
- Runs interactive SQL queries on Amazon S3 data
  - No need to load or aggregate data: "schema-on-read"
  - Cross-region queries are supported
- Supports ANSI SQL operators and functions
- No infrastructure or administration to create, manage, or scale (serverless)

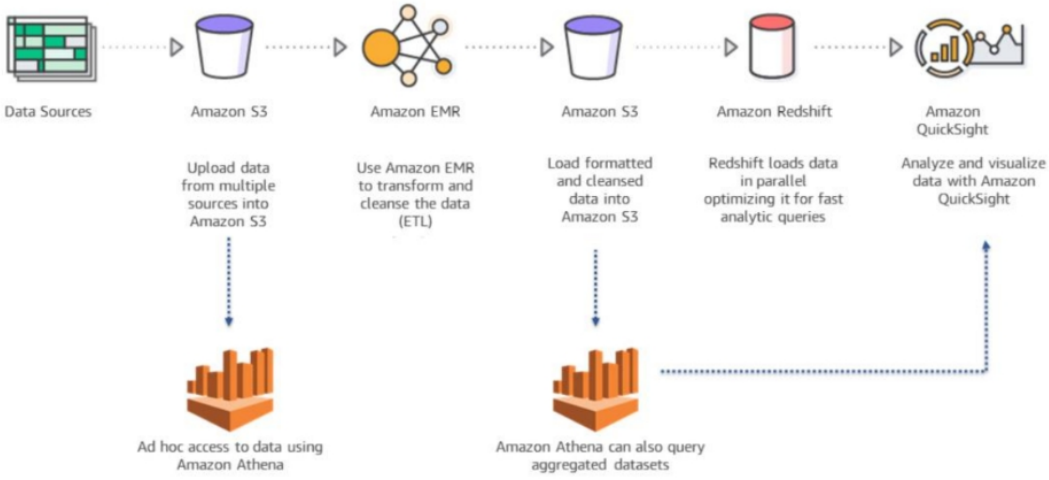
Amazon Athena

aws training and certification

	Amazon Athena
Use case	Ad hoc interactive queries
Scale/throughput	Automatic/No limits
AWS managed service	Yes, serverless
Storage	Amazon S3
Optimization	CSV, TSV, JSON, Parquet, ORC
Metadata	Athena Catalog Manager
BI tools support	Yes (JDBC)
Access controls	AWS IAM
UDF support	No

Data Lake Pipeline With Amazon Athena

aws training and certification



Amazon EMR

aws training and certification

- Managed cluster platform
- Run Hadoop, Spark, Presto and other applications
- Launch a cluster in minutes
- Deploy multiple clusters
- Resize a running cluster
- Cost-effectively process vast amounts of data
- Use HDFS and S3 file systems
- Automated install of common ecosystem projects by default, like Hive, Pig, Hue (Hadoop 2 only), and Spark, etc.

Four Best Practices

aws training and certification

- Aggregate data size
- Control data aggregation size
- Choose an appropriate compression algorithm
- Data partitioning

Best Practice 4: Data Partitioning

aws training and certification

Log processing:

Data processed once per hour

Data access pattern based on time

Partition based on day and hour

/data/logs/YYYY/MM/DD/HH/logfiles for this given hour

/data/logs/2016/01/01/02/logfile1
.../logfile2
.../logfile3

/data/logs/2016/01/01/03/logfile1
.../logfile2
.../logfile3

Three primary things to consider when determining how to partition your data:

1. Data type (time series)
2. Data processing frequency (per hour, per day, etc.)
3. Data access and query pattern (query on time vs. query on geo location, etc.)

High-Level Hadoop Programming Frameworks

aws training and certification

Apache Hive

aws training and certification

Description:

Open-source, SQL-like data warehouse solution that runs on top of Hadoop

Audience:

SQL developers and data scientists looking for a familiar programming model to prototype and run jobs

Pros	Cons
Parity with SQL language (new skills are not needed)	Does not give direct access to low-level MapReduce features
JDBC and ODBC drivers are available to access data and submit jobs from compliant tools	Limited data transformation capabilities
Supports data partitions in Amazon S3 to reduce scope of data analysis	
Inherits linear scalability of Hadoop	
Supports user defined functions	
Can join two large tables (100M+ rows) with minimal optimization and automatic retries on failure	

Presto

aws training and certification

Description:

Open-source, distributed, in-memory SQL query engine

Audience:

Data scientists and analysts who need quick interactive query responses

Pros	Cons
SQL-based (new skills not needed)	Requires careful optimization when joining two large tables
Interactive queries are optimized for lowest response times (ms to minutes)	Not optimal for batch processing jobs: Limited by amount of memory available Limited fault tolerance; failed jobs must be re-run
JDBC and ODBC-compatible middleware (Prestogres) available to submit jobs from compliant tools	
Supports user defined functions	
Can query data from multiple stores (Hive, Cassandra, relational databases, proprietary data stores, etc.)	

Apache Pig

aws training and certification

Description:

Simple, high-level, textual data flow language (Pig Latin)

Audience:

Script developers who need intuitive way to prototype and develop jobs

Pros	Cons
Intuitive, procedural syntax for coding of multiple, complex transformations run in sequence	May require new programming skills
Compiled into MapReduce jobs	
Supports unstructured and semi-structured data, such as weblogs and clickstreams	
Supports user-defined functions in Piggybank	

Comparing Latency and SQL Compatibility

aws training and certification

	Query Latency	SQL Compatibility
Hive	High (MapReduce)	✓
	Medium (Tez)	
Pig	High (MapReduce)	✗
	Medium (Tez)	
Spark	Low	✓ (SparkSQL)
Presto	Low	✓

Advantages of Apache Spark

aws training and certification

- Allows in-memory data mining and querying big datasets at fast speeds
- Resilient distributed datasets (RDDs)
- Processing engine can be run in-memory or from disk
  - Both can produce results faster than Hive on MapReduce
  - Up to 10x faster than MapReduce when using on-disk storage

Spark Programming Model

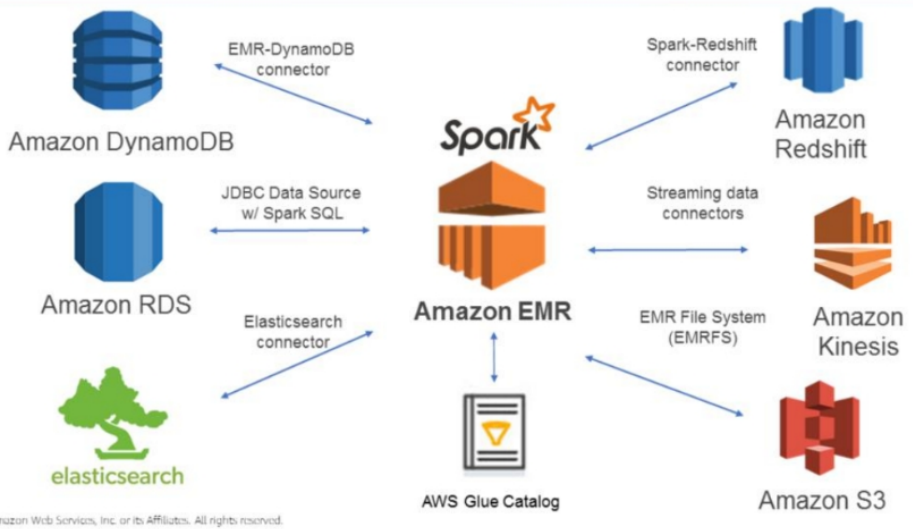
aws training and certification

Resilient distributed datasets (RDDs):

- Are read-only distributed collections of objects that are cached in memory across cluster nodes.
- Allow apps to keep working sets in memory for efficient reuse.
- Retain attractive properties of MapReduce: Fault tolerance, data locality, and scalability.

Storage Options

aws training and certification



How Is a Data Warehouse Different From a Database?

aws training and certification

Traditional databases are designed for transactional workloads.

In transactional processing, operations are processed in series, and each one relies on its previous operation in order to complete the entire job.

How Is a Data Warehouse Different From a Database?

aws training and certification

Running analytics on a traditional database can prevent business users from using that database.

While an analytics job is running, business users may be locked out and have to wait for the job to be completed before being able to access the database again.

Using a data warehouse prevents analytics jobs and transactional jobs from competing for resources.

How Is a Data Warehouse Different From a Database?

aws training and certification

Data warehouses can run on top of traditional databases and typically use tables with duplicated data.

This reduces the need for intensive operations that require disk I/O.

Additionally, analytic operations can run separately from transactional requests, on discrete resources.

Amazon Redshift

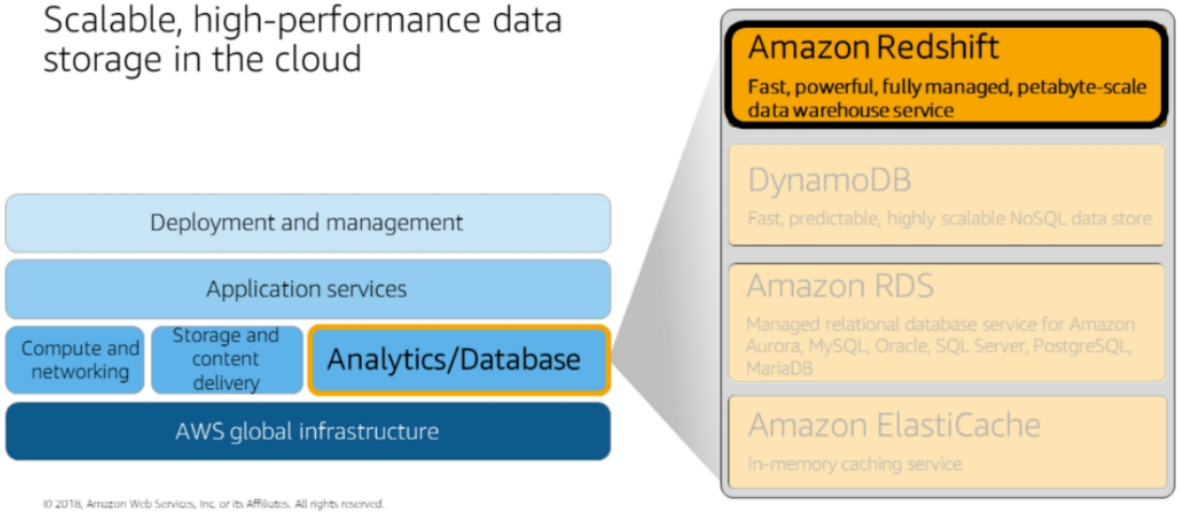
aws training and certification

- Relational data warehouse
- Massively parallel; petabyte scale
- Fully managed
- HDD and SSD platforms
- $1,000/TB/year; starts at $0.25/hour

AWS Database Services

aws training and certification

Scalable, high-performance data storage in the cloud



Amazon Redshift Architecture

aws training and certification

Leader Node

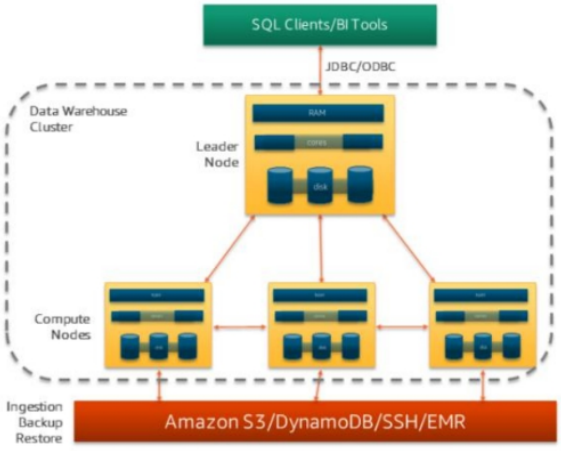
- SQL endpoint
- Stores metadata
- Coordinates query execution

Compute Nodes

- Local, columnar storage
- Execute queries in parallel
- Load, back up, restore via Amazon S3; load from Amazon DynamoDB or Amazon EMR or SSH

Two hardware platforms

- Optimized for data processing
- dc1: SSD; scale from 160 GB to 326 TB
- ds2: HDD; scale from 2 TB to 2 PB



Choose the Right Tools for Processing/Analytics

aws training and certification

	Amazon Redshift	Amazon EMR			Amazon Athena
		Presto	Spark	Hive	
Use case	Optimized for data warehousing	Interactive query	General purpose (iterative ML etc.	Batch	Ad hoc interactive queries
Scale/throughput	~Nodes	~Nodes			Automatic/No limits
AWS managed service	Yes	Yes			Yes, serverless
Storage	Local storage	Amazon S3, HDFS			Amazon S3
Optimization	Columnar storage, data compression, and zone maps	Framework-dependent			CSV, TSV, JSON, Parquet, ORC
Metadata	Amazon Redshift managed	Hive Meta-store			Athena Catalog Manager
BI tools support	Yes (JDBC/ODBC)	Yes (JDBCS/ODBC and Custom)			Yes (JDBC)
Access controls	Users, groups, and access controls	Integration with LDAP			AWS IAM
UDF support	Yes (Scalar)	Yes (Scalar)			No

Amazon EMR vs. Amazon Redshift

aws training and certification

Choose Amazon EMR when you need to:

- Batch-process large amounts of unstructured or semi-structured data.
- Perform extract/transform/load operations on big data.

Choose Amazon Redshift when you need to:

- Query structured data using SQL (for visualization).
- Store large quantities of structured data for rapid access.

Apache Hive vs. Amazon Redshift

aws training and certification

- Apache Hive cannot replace the full data warehousing capabilities of Amazon Redshift.
- Amazon Redshift is a structured data store solution; data may require transformation before loading.
- Hive queries on Amazon EMR leverage MapReduce (by default):
  - Performance impact for complex SQL operations
  - Complicated joins across several tables are not recommended

Presto vs. Amazon Redshift

aws training and certification

- Presto is not a true data warehouse solution, but a distributed SQL query engine.
- Presto requires extra data storage solutions:
  - HDFS (local store) or Amazon S3 is commonly used.
  - Some customers, such as Netflix, leverage Amazon S3 for their DW storage and use Presto/Hadoop for their queries/analytics instead.

Cost Considerations for Amazon EMR

aws training and certification

Hourly charges include:

- Amazon EC2 instance type
- Amazon EMR upcharge

Amazon EC2 instance hours are charged for master, core, and task nodes.

Amazon EMR charges vary with instance type.

- Amazon S3, DynamoDB, and Amazon RDS have additional costs.
- Data transfer out is also a cost.