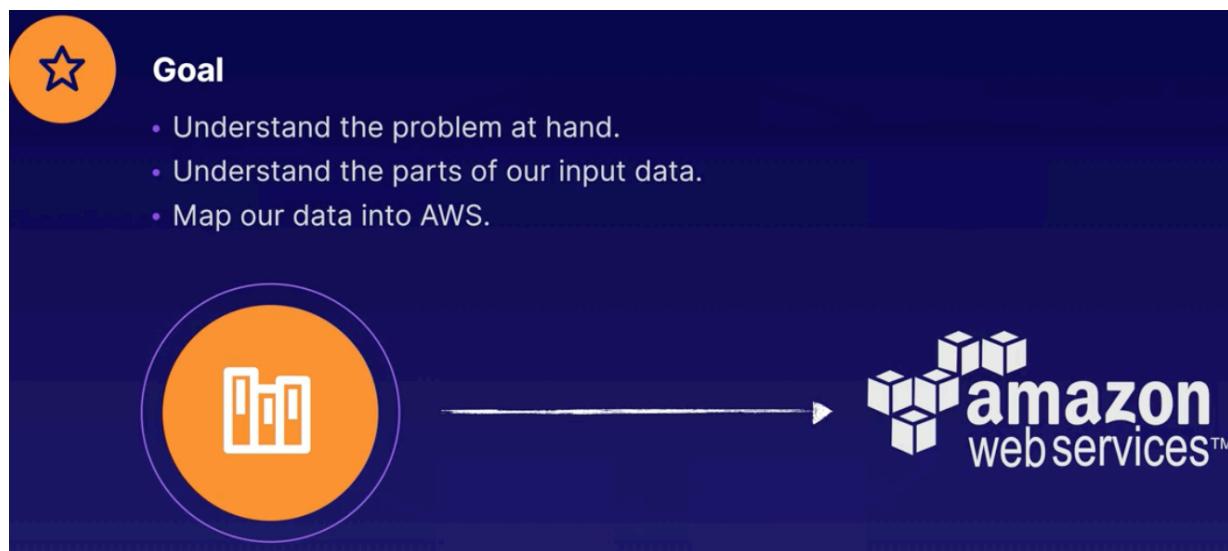
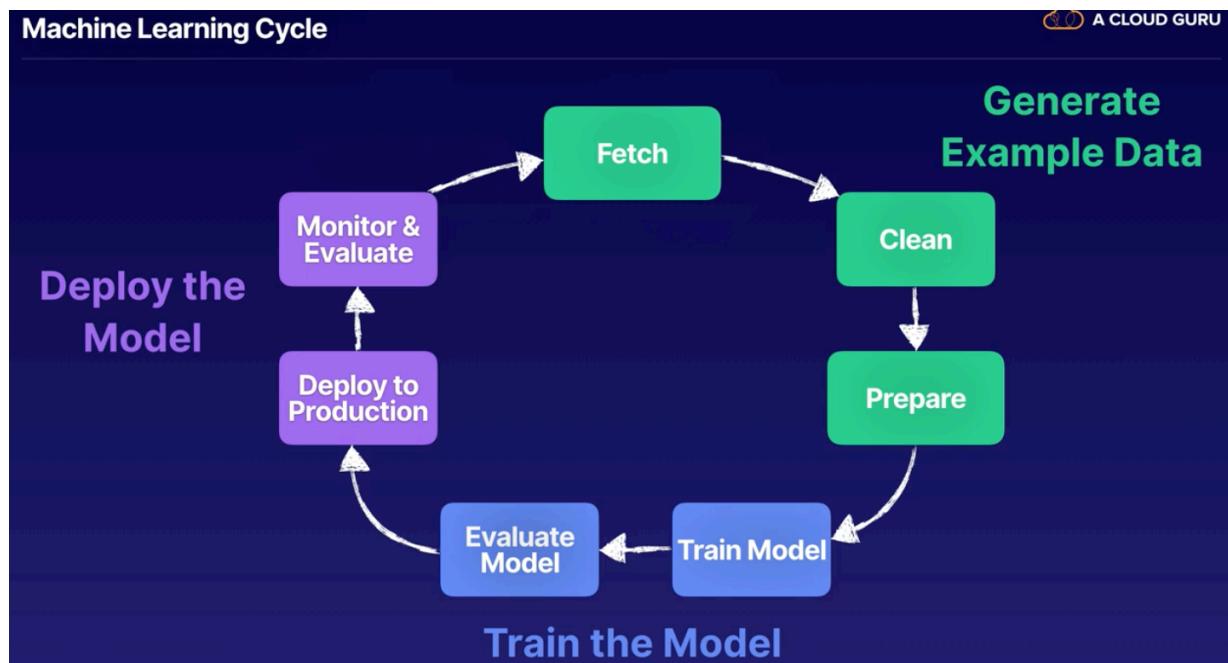


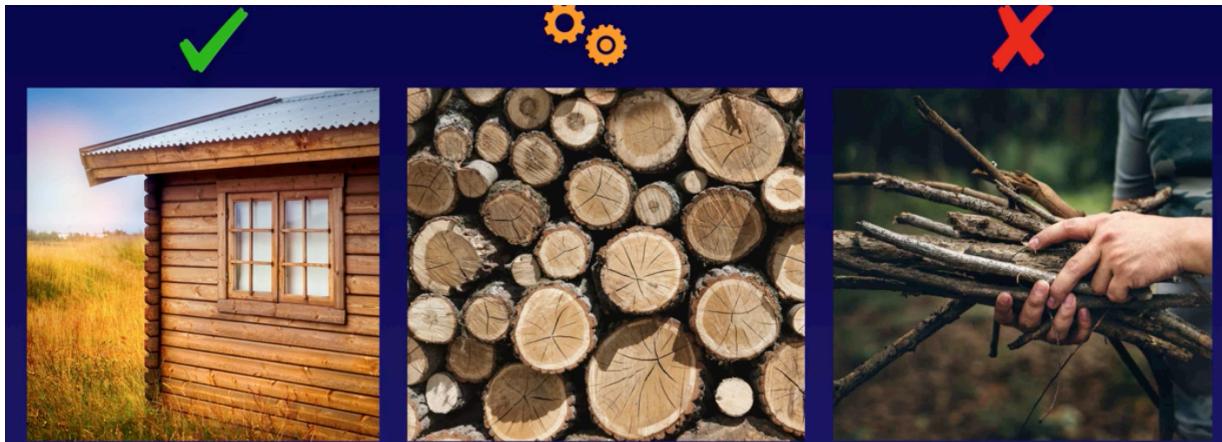
## CloudGuru - 1 - Data collection

<https://acloud.guru/course/aws-certified-machine-learning-specialty/learn/bc7319b8-616e-f972-ec23-d2b49728a81d/c068feca-3ae9-ce9b-0b5a-a7eca7f6f895/watch>



Data mostly comes unstructured, not prepared,...

A model would not infer anything good from noisy/crapy data



## Good Data

DATA A CLOUD

Traits of good data	Traits of bad data	Why
Large datasets.	Small datasets (less than 100 rows).	Generally, more data means better model training.
Precise attribute types, feature rich.	Useless attributes, not needed for solving problem at hand.	Models need to train on important features.
Complete fields, no missing values.	Missing values, null fields.	Models can skew results when data points are missing.
Values are consistent.	Inconsistent values.	Models like clean and consistent data.
Solid distribution of outcomes.	Lots of positive outcome, few negative outcomes.	Models cannot learn with skewed distributions of outcomes.
Fair sampling	Biased sampling	Models will skew results with biased data.

How much data do we need?

**You should have at least 10 times as many data points as the total number of features.**

DATA Terminology

*dataset = input data = training/testing data*

*column = attribute = feature*

Column/Attribute

ID	Name	Evil
1	Luke	0
2	Leia	0
3	Han	0
4	Vadar	1

*row = observation = sample = data point*

Row/Observation →

ID	Name	Evil
1	Luke	0
2	Leia	0
3	Han	0
4	Vadar	1

Other formats: JSON, CSV,...

ID	Name	Evil
1	Luke	0
2	Leia	0
3	Han	0
4	Vadar	1

```
{
  "Characters": [
    {
      "ID": 1,
      "Name": "Luke",
      "Evil": 0
    },
    {
      "ID": 2,
      "Name": "Leia",
      "Evil": 0
    },
    ...
  ]
}
```

ID, Name, Evil  
1, Luke, 0  
2, Leia, 0  
3, Han, 0  
4, Vadar, 1

Also images, video, audio,... not just text



Structured data: SQL,...  
It's how we would love our data



## Structured Data

Structured data has a defined schema and a schema is the information needed to interpret the data, including attribute names and their assigned data types.

But majority of data is unstructured, with no defined schema  
=> PDF, images, video, audio,.. log files



## Unstructured Data

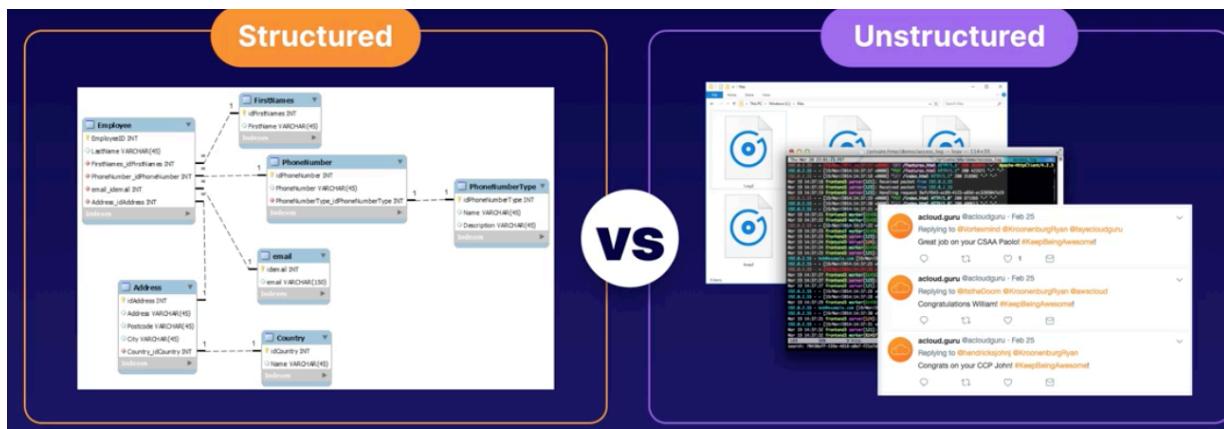
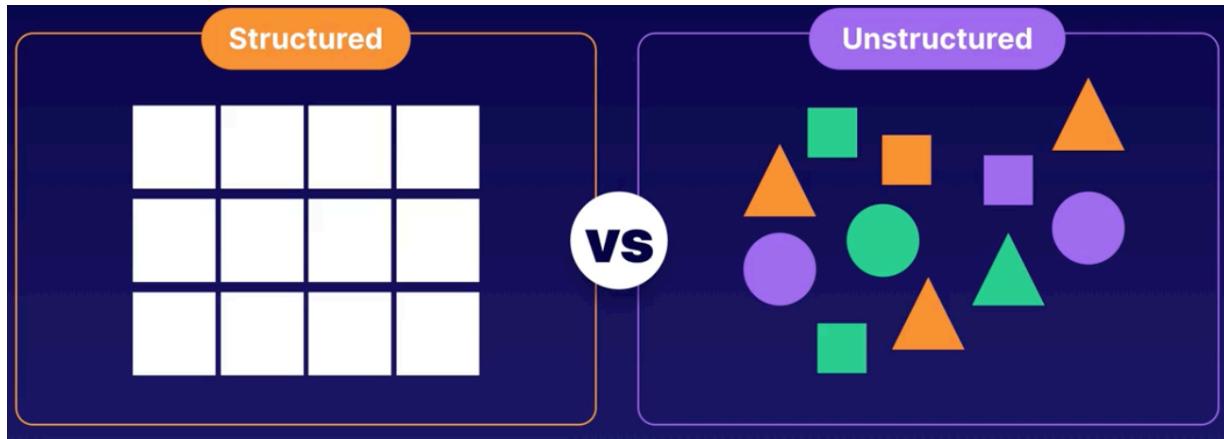
Unstructured data has no defined schema or structural properties.  
Makes up majority of data collected.

Some other data is semi-structured  
NoSQL, JSON, XML, JSON



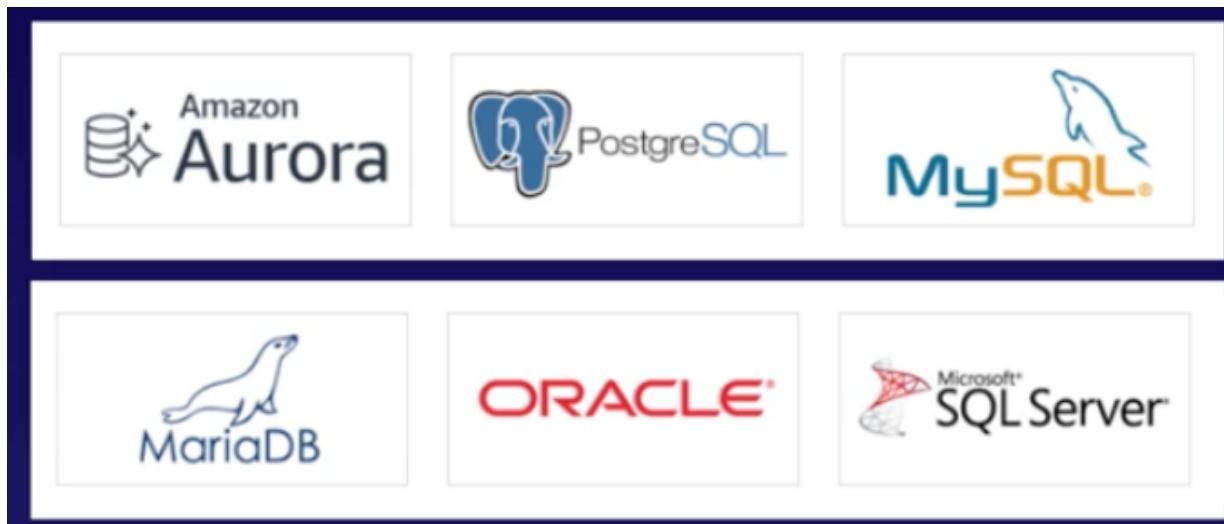
## Semi-structured Data

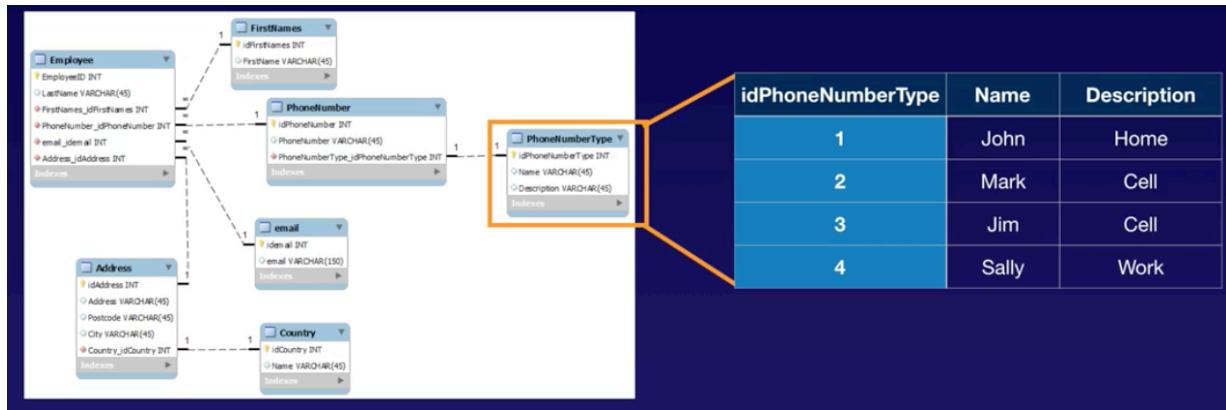
Semi-structured data is to unstructured for relational data, but has some organizational structure. Usually in the form of CSV, JSON, or XML.



Where is data stored:

- structured => transactional DB when're changes re rolled back if a transaction is not successful





### Other data store: Data Warehouse

Collects from many sources and congrats them together, and make them avail to BI tools

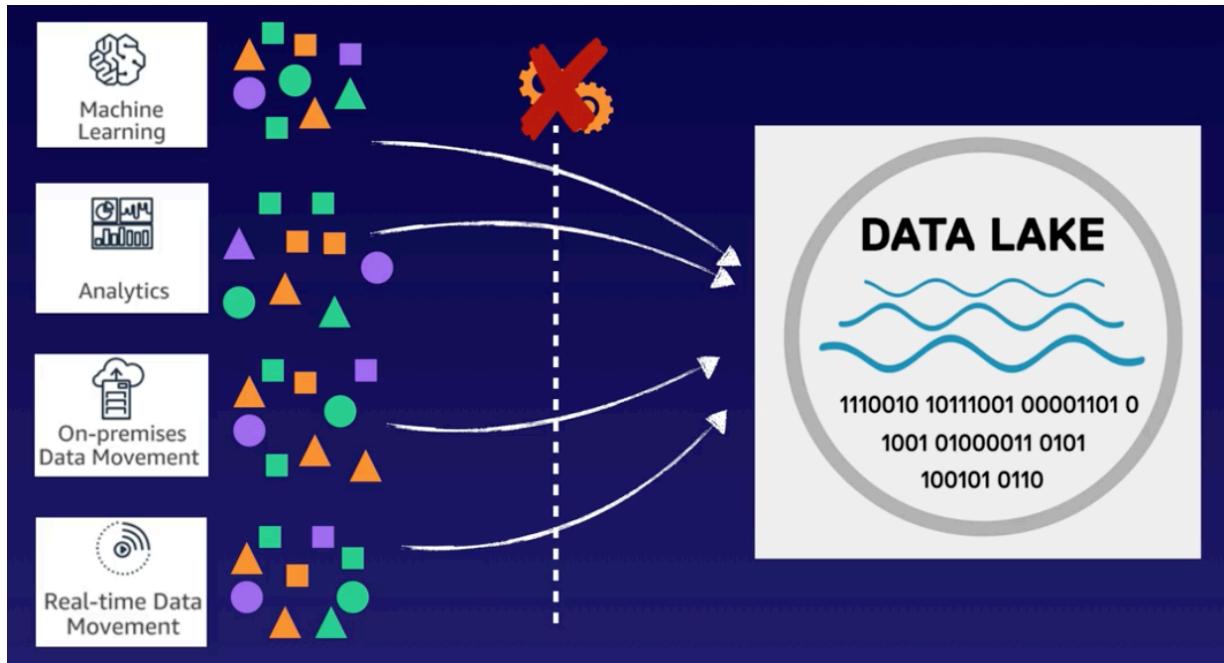
=> PROCESSING done before storing

Can store petabytes/heabytes of data and allow us to analyze more easily



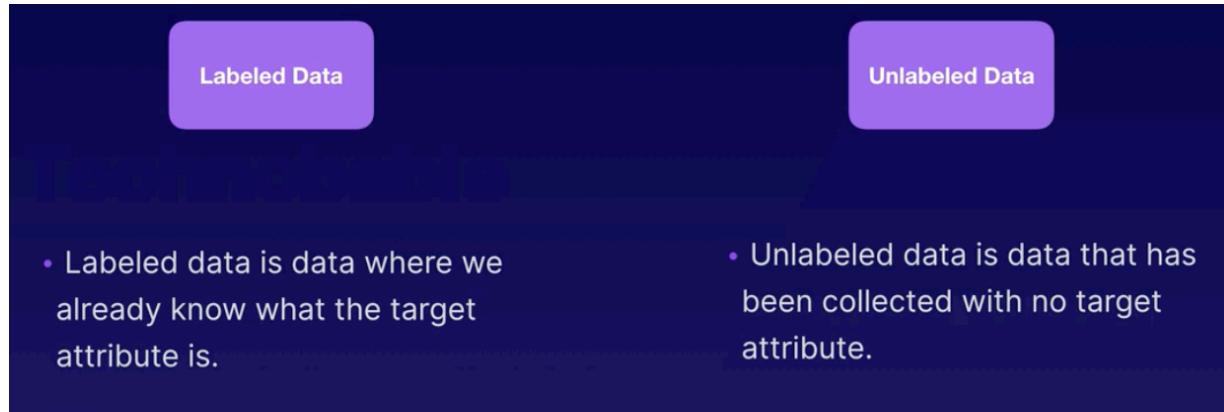
Lastly, **Data Lakes** store VAST amount of data

There is **NOT processing** done. One stop shop for dumping our data (for historical data or data we are not sure what to do with but that is important)



<b>Databases</b>	<ul style="list-style-type: none"> <li>• Traditional relational databases</li> <li>• Transactional</li> <li>• Strict defined schema</li> </ul>
<b>Data Warehouses</b>	<ul style="list-style-type: none"> <li>• Processing done on import (schema-on-write)</li> <li>• Data is classified/stored with user in mind</li> <li>• Ready to use with BI tools (query and analysis)</li> </ul>
<b>Data Lakes</b>	<ul style="list-style-type: none"> <li>• Processing done on export (schema-on-read)</li> <li>• Many different sources and formats</li> <li>• Raw data may not be ready for use</li> </ul>

Labeled/Unlabeled data



Example of labelled day on email with what is spam or not

Spam is the label (or target attribute)

**Label/Target**

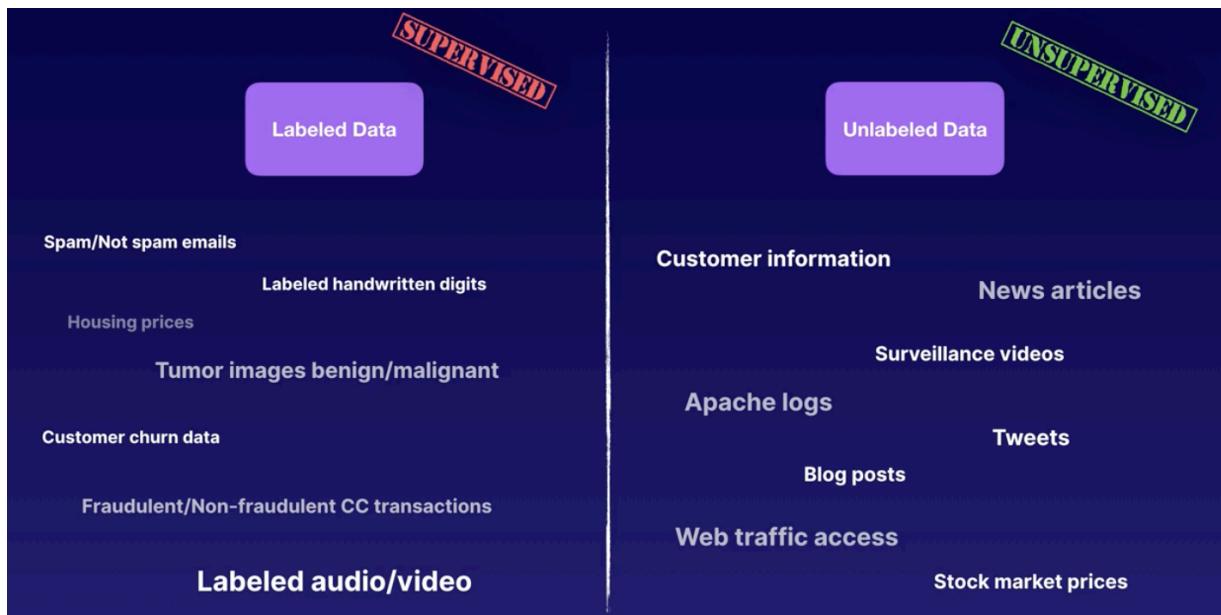
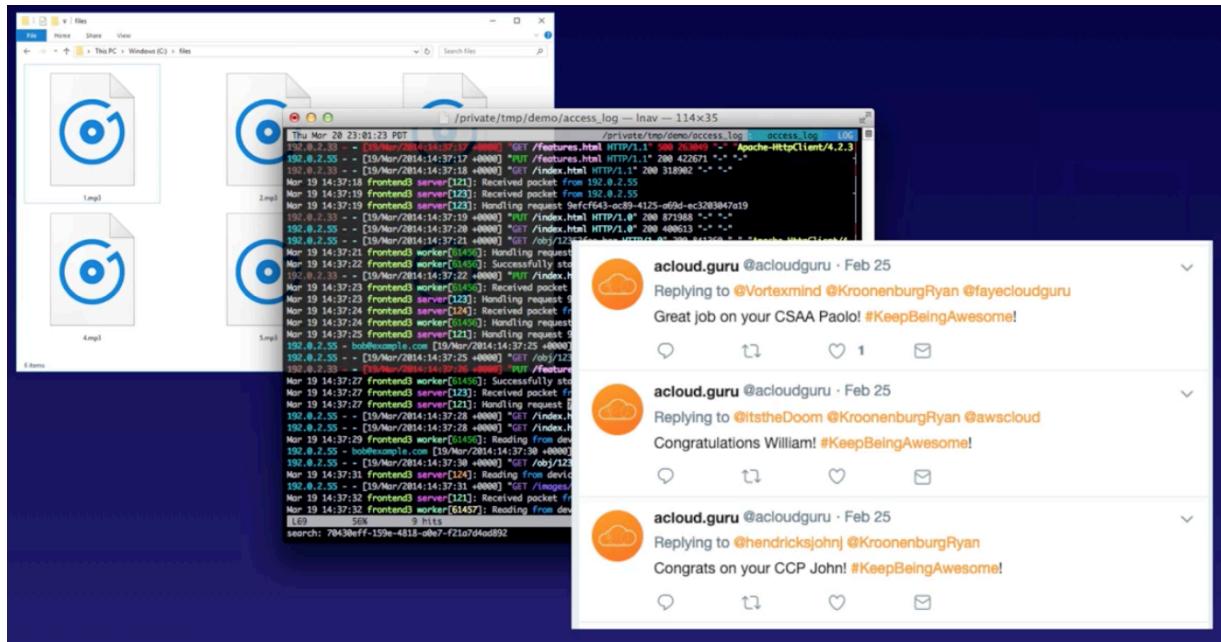
ID	From Domain	Body	...	Spam
1	gov	Hi ...	...	F
2	com	Welcome...	...	F
3	com	Hey ...	...	F
4	net	Click here...	...	T
5	org	New...	...	F
6	com	Free gift...	...	T

**Variables/Features/Attributes**

ID	From Domain	Body	...	Spam
1	gov	Hi ...	...	F
2	com	Welcome..	...	F
3	com	Hey ...	...	F
4	net	Click here...	...	T
5	org	New...	...	F
6	com	Free gift...	...	T

Unlabelled data: social media streams, log files, video, audio,...

There is no answer or tag associated to them



Categorical vs Continuous

## Categorical

- Categorical features are values that are associated with a group.
- Qualitative
- Discrete

## Continuous

- Continuous features are values that are expressed as measurable number.
- Quantitative
- $\infty$

Top to remember:



Group/Category				
ID	Height	Weight	...	Breed
1	26	6	...	Cairn Terrier
2	82	45	...	Airedale Terrier
3	38	9	...	Fox Terrier
4	40	20	...	Bull Terrier
5	36	6.5	...	Fox Terrier
6	55	20	...	Airedale Terrier
7	55	30	...	Bull Terrier
8	31	7.5	...	Cairn Terrier

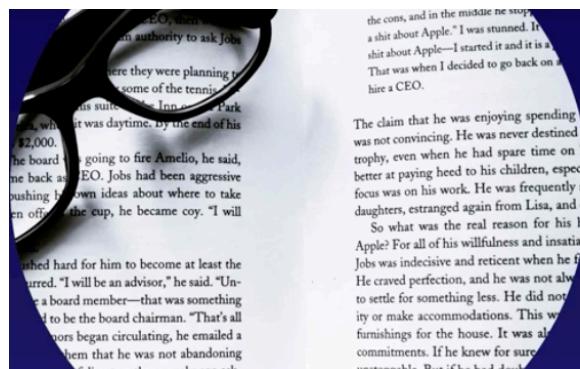


Group/Category				
ID	From Domain	Body	...	Spam
1	gov	Hi ...	...	F
2	com	Welcome...	...	F
3	com	Hey ...	...	F
4	net	Click here...	...	T
5	org	New...	...	F
6	com	Free gift...	...	T

If the attribute falls into a group or category, the attribute is categorical.

Continuous				
ID	Bedrooms	Square Feet	...	Price
1	2	1250	...	\$125,000
2	3	1600	...	\$165,000
3	1	975	...	\$98,000
4	2	1400	...	\$145,000
5	4	2500	...	\$200,000
6	1	1075	...	\$105,000

If you can place the attribute value on a number line, the attribute is continuous.



the cons, and in the middle he says, "I'm going to fire Amelio," he said. "I have the authority to ask Jobs where they were planning to go. I asked him if he had some of the tennis clubs in his suit. He said, 'Inn at the Park'—which was daytime. By the end of his \$2,000.

The board was going to fire Amelio, he said, he back as CEO. Jobs had been aggressive pushing his own ideas about where to take the company. At the end of the cup, he became coy. "I will

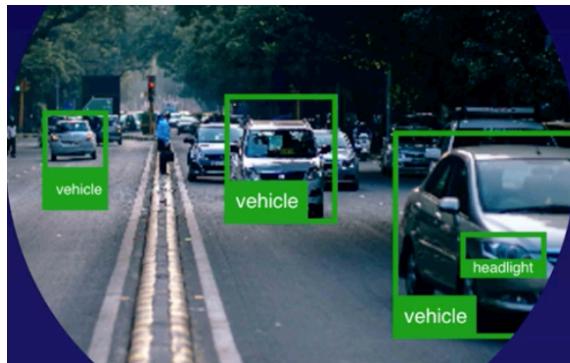
pushed hard for him to become at least the chairman. "I will be an advisor," he said. "Unless I'm a board member—that was something I wanted to be the board chairman. "That's all I wanted to do." Rumors began circulating, he emailed a friend, telling them that he was not abandoning Apple. But if he had doubts,

the claim that he was enjoying spending time with his wife Laurene and their two daughters, estranged again from Lisa, and on a shirt about Apple." I was stunned. It was not convincing. He was never destined for a trophy, even when he had spare time on his hands, he was better at paying heed to his children, especially his focus was on his work. He was frequently absent from his family.

So what was the real reason for his love of Apple? For all of his willfulness and insatiable desire for perfection, he was indecisive and reticent when he faced challenges. He craved perfection, and he was not always willing to settle for something less. He did not like to compromise or make accommodations. This was evident in his furnishings for the house. It was always a commitment. If he knew for sure that he would be unapproachable. But if he had doubts,

## Text Data (Corpus Data)

These are datasets collected from text. Used in Natural Language Processing (NLP), speech recognition, text to speech, and more.



## Ground Truth

Ground truth datasets refers to factual data that has been observed or measured. This data has successfully been labeled and can be trusted as "truth" data.

SameMaker Ground Truth - tool to enable label data

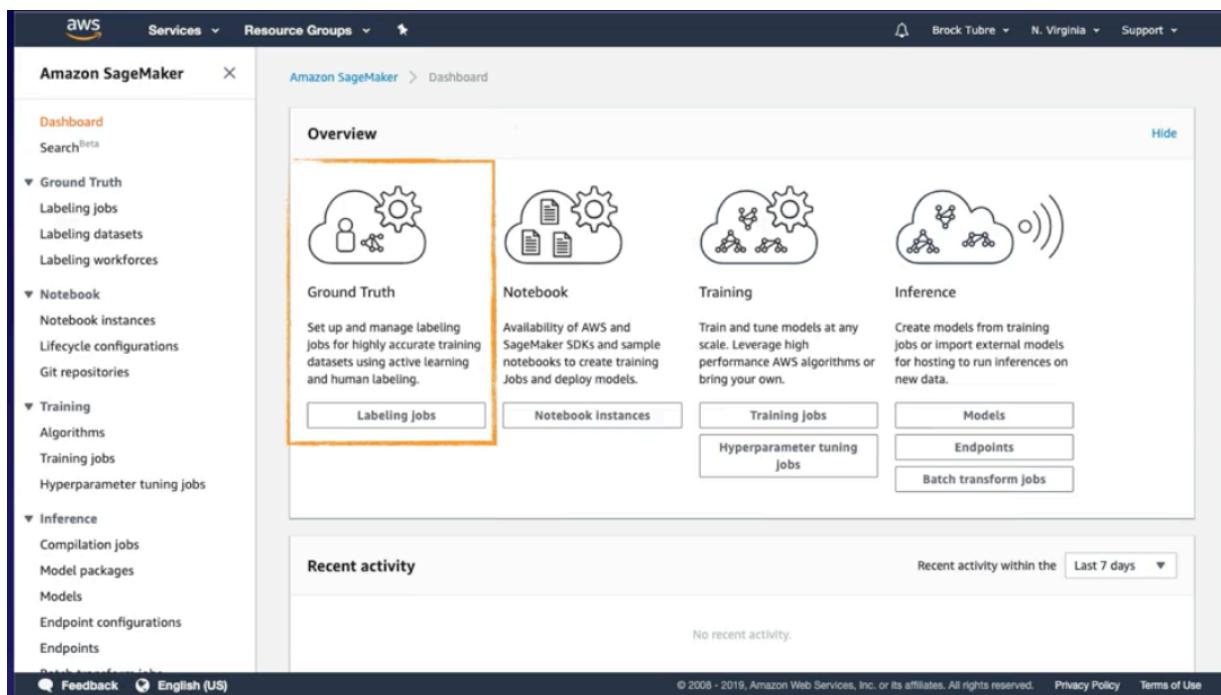
### Amazon SageMaker Ground Truth

Tool that helps build ground truth datasets by allowing different types of tagging/labeling processes.

Easily create labeled data.



Amazon SageMaker  
Ground Truth



AWS Services Resource Groups Brock Tuber N. Virginia Support

Amazon SageMaker > Dashboard

**Overview**

- Ground Truth**  
Set up and manage labeling jobs for highly accurate training datasets using active learning and human labeling.  
[Labeling jobs](#)
- Notebook**  
Availability of AWS and SageMaker SDKs and sample notebooks to create training Jobs and deploy models.  
[Notebook Instances](#)
- Training**  
Train and tune models at any scale. Leverage high performance AWS algorithms or bring your own.  
[Training jobs](#)  
[Hyperparameter tuning jobs](#)
- Inference**  
Create models from training jobs or import external models for hosting to run inferences on new data.  
[Models](#)  
[Endpoints](#)  
[Batch transform jobs](#)

**Recent activity**

No recent activity.

Recent activity within the

Feedback English (US)

© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use



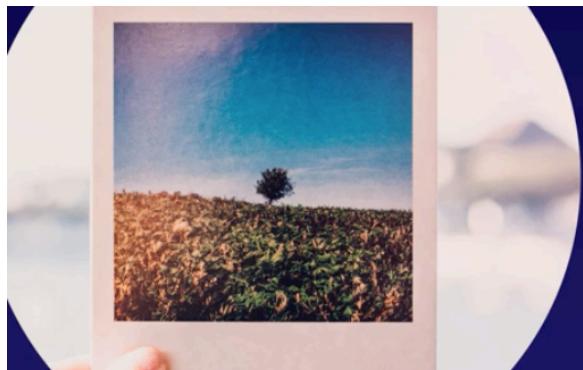
Which sport is being played?



Select an option

- Soccer
- Swimming
- Baseball
- Cricket

Submit

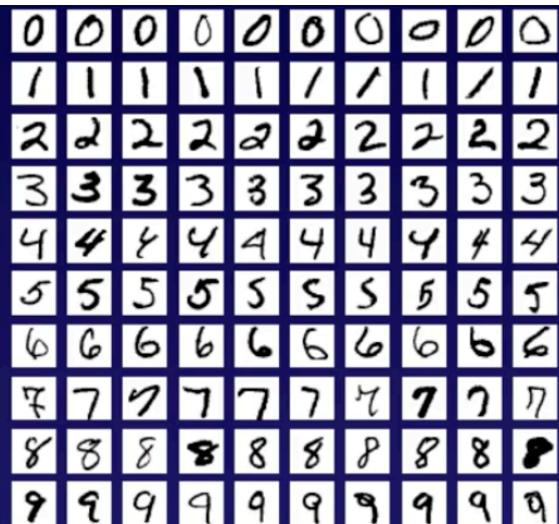


## Image Data

Image data refers to datasets with tagged images.

Example of Tagged image data





MNIST data



Image Net

### Dataset examples

Dataset Type	Example cases	Format
Image Data	Facial recognition, action recognition, object detection, handwriting and character recognition.	images, videos
Text Data	Reviews, news articles, messages, twitter and tweets, dialogs.	text, csv
Sound Data	Speech, music, other sounds.	mp3, text
Signal Data	Electrical signals, motion-tracking, chemical compounds.	text

Dataset Type	Example cases	Format
Physical Data	high-energy physics, systems, astronomy, earth science.	text
Biological Data	human, animal, plants, microbes.	text
Multi-variable Data	financial, weather, census, transit, internet, games.	csv, text

### AWS Data Stores and services to store data

# AWS Data Stores

S3

**Amazon Simple Storage Service (S3)**



Unlimited data storage that provides object based storage for any type of data.



**S3**



Go to place for storing Machine Learning data.

## What is S3?

- Files can be from 0 bytes to 5 TB.
- There is unlimited storage.
- Files are stored into buckets (similar to folders).
- S3 is a universal namespace. That is, names must be unique globally.
- <https://s3-us-east-1.amazonaws.com/machinelearningdata>

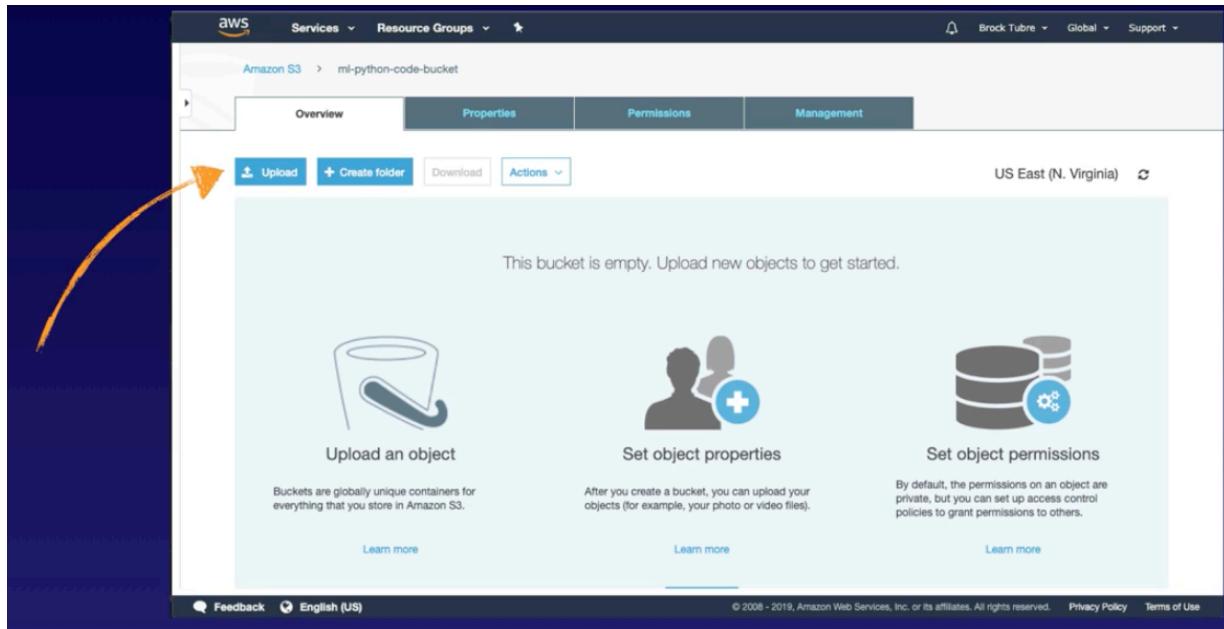
Sept 2020: virtual hosted style coming

## What is S3?

- Files can be from 0 bytes to 5 TB.
- There is unlimited storage.
- Files are stored into buckets (similar to folders).
- S3 is a universal namespace. That is, names must be unique globally.
- <https://s3-us-east-1.amazonaws.com/machinelearningdata>
- <https://machinelearning.s3.amazonaws.com/>

How to upload:

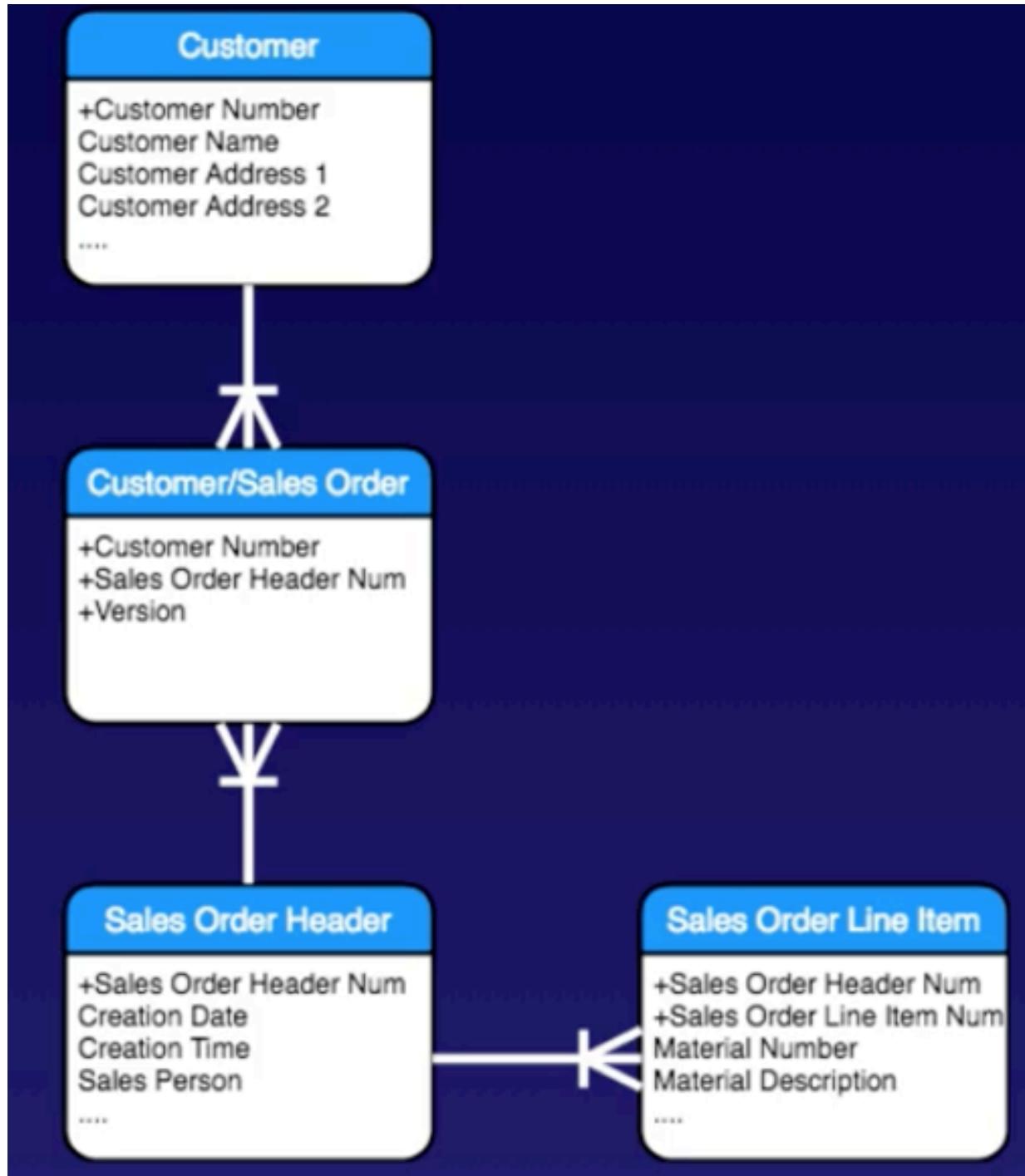
- via console



- via SDK
- via CLI

## RDS: Relational DB Service





The screenshot shows the AWS RDS 'Create database' wizard. The first step, 'Select engine', is highlighted with an orange border. It lists several database engines: Amazon Aurora, MySQL (selected), MariaDB, PostgreSQL, Oracle, and Microsoft SQL Server. Below the selection area, there is a detailed description of MySQL and a bulleted list of its features.

**Select engine**

**Engine options**

- Amazon Aurora
- MySQL
- MariaDB
- PostgreSQL
- Oracle
- Microsoft SQL Server

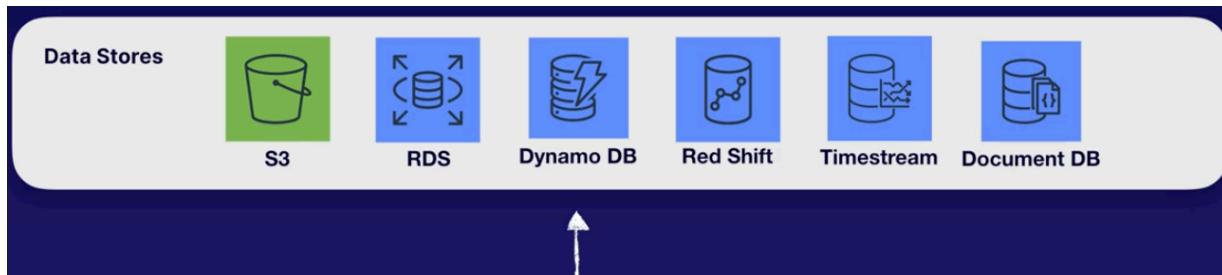
**MySQL**  
MySQL is the most popular open source database in the world. MySQL on RDS offers the rich features of the MySQL community edition with the flexibility to easily scale compute resources or storage capacity for your database.

- Supports database size up to 32 TiB.
- Supports General Purpose, Memory Optimized, and Burstable Performance instance classes.
- Supports automated backup and point-in-time recovery.
- Supports up to 5 Read Replicas per instance, within a single Region or cross-region.

Feedback English (US) © 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

## No SQL DataStore:

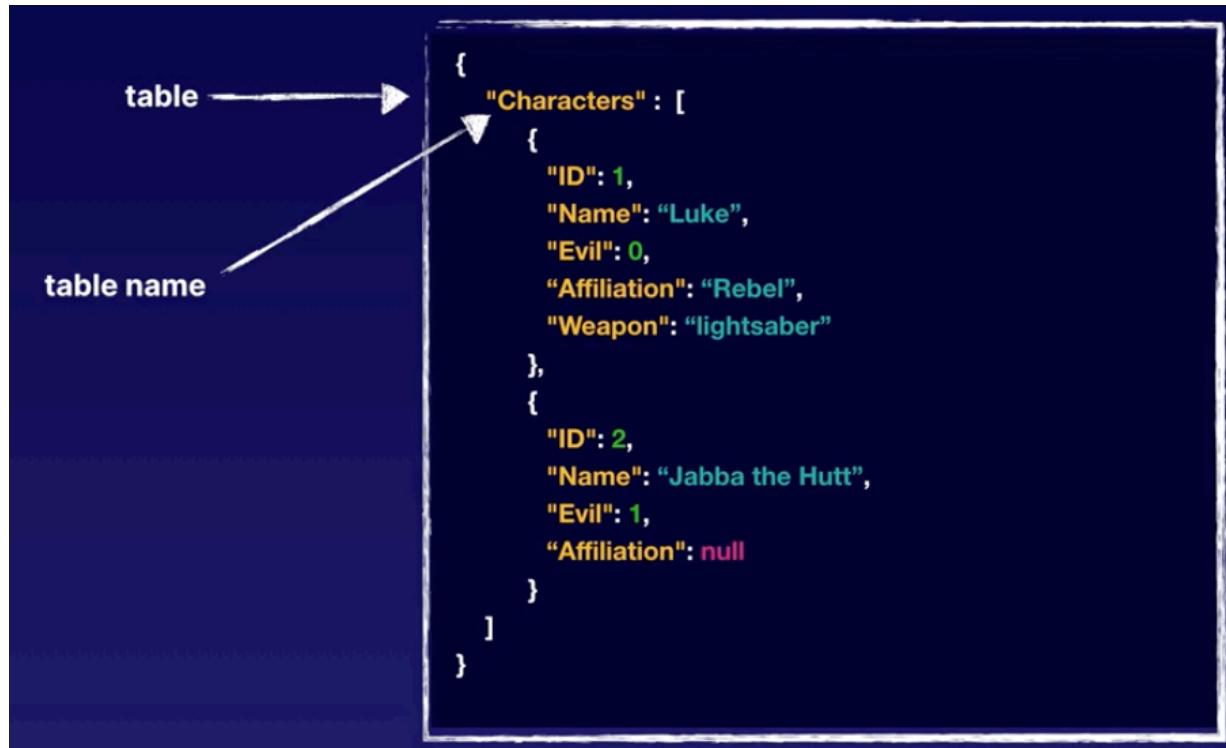
For schema less data or unstructured



Example: JSON data

```
{  
  "Characters": [  
    {  
      "ID": 1,  
      "Name": "Luke",  
      "Evil": 0,  
      "Affiliation": "Rebel",  
      "Weapon": "lightsaber"  
    },  
    {  
      "ID": 2,  
      "Name": "Jabba the Hutt",  
      "Evil": 1,  
      "Affiliation": null  
    }  
  ]  
}
```

In dynamo DB, this json dataset is the TABLE, and the first property is the TABLE Name



```
{  
  "Characters": [  
    {  
      "ID": 1,  
      "Name": "Luke",  
      "Evil": 0,  
      "Affiliation": "Rebel",  
      "Weapon": "lightsaber"  
    },  
    {  
      "ID": 2,  
      "Name": "Jabba the Hutt",  
      "Evil": 1,  
      "Affiliation": null  
    }  
  ]  
}
```

The diagram illustrates the structure of the JSON object. On the left, there is a vertical blue bar labeled "key". On the right, there is a vertical white bar labeled "value". A horizontal arrow points from the "key" bar to the "Weapon" field of the first character object, and another arrow points from the "value" bar back to the same "Weapon" field. This visualizes how a specific key maps to a specific value within the data structure.

```
{  
  "Characters": [  
    {  
      "ID": 1,  
      "Name": "Luke",  
      "Evil": 0,  
      "Affiliation": "Rebel",  
      "Weapon": "lightsaber"  
    },  
    {  
      "ID": 2,  
      "Name": "Jabba the Hutt",  
      "Evil": 1,  
      "Affiliation": null  
    }  
  ]  
}
```

```
{  
  "Characters": [  
    {  
      "ID": 1,  
      "Name": "Luke",  
      "Evil": 0,  
      "Affiliation": "Rebel",  
      "Weapon": "lightsaber" } attribute  
    },  
    {  
      "ID": 2,  
      "Name": "Jabba the Hutt",  
      "Evil": 1,  
      "Affiliation": null  
    }  
  ]  
}
```

Showing the items in the console:

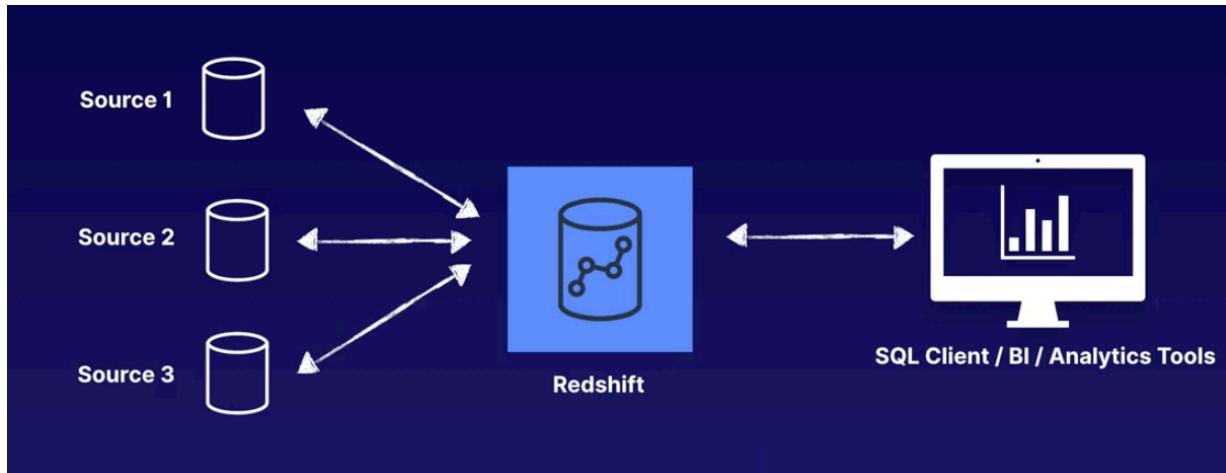
The screenshot shows the AWS DynamoDB console. On the left, the navigation pane includes 'Dashboard', 'Tables' (selected), 'Backups', 'Reserved capacity', and 'Preferences'. Under 'DAX', it lists 'Dashboard', 'Clusters', 'Subnet groups', 'Parameter groups', and 'Events'. The main area displays the 'ml-demo-table' table. A search bar at the top left shows 'ml-demo-table'. Below it, a table view shows two items:

ID	Affiliation	Evil	Name	Weapon
1	Rebel	false	Luke	lightsaber
2	null	true	Jabba the Hutt	

### Amazon Redshift:

Fully managed clustered peta bytes data warehouse solutions that congregates data from S3, DynamoDB,... and can handle any data type  
Once in Redshift, can use SQL like tools or BI tools to query it





In console, we can set # of nodes

Launch your Amazon Redshift cluster - Quick launch | [Switch to advanced settings](#)

**Amazon Redshift Pricing** offers on-demand and reserved instance pricing options. Save up to 75% through reserved instances.

Node type*	dc2.large	Storage type: SSD	Storage: 0.16 TB/node	Compute optimized	0.25 USD/node
Nodes*	2	x 0.16 TB/node = 0.32 TB storage available			

Cluster identifier\*  Cluster identifier

Database name  Database port\*

Master user name\*  Master user name

Master user password\*  Confirm password\*

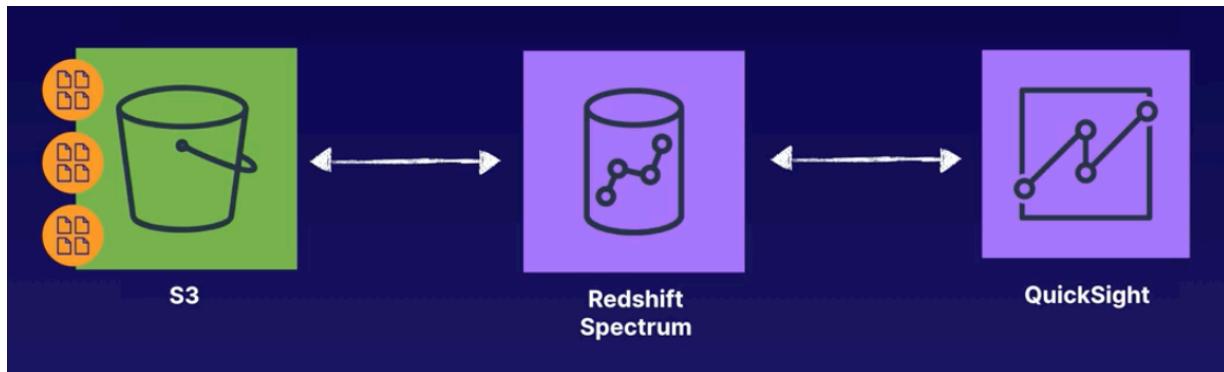
**Cluster permissions - optional**  
Your cluster needs permission to access other AWS services on your behalf. For the required permissions, add an IAM role now or after you launch the cluster. [Learn more](#)

Available IAM roles

Default settings [Switch to advanced settings](#)

We'll apply some default settings for network, security, backup, and maintenance to get you started. Switch to advance settings if you want to change the defaults.

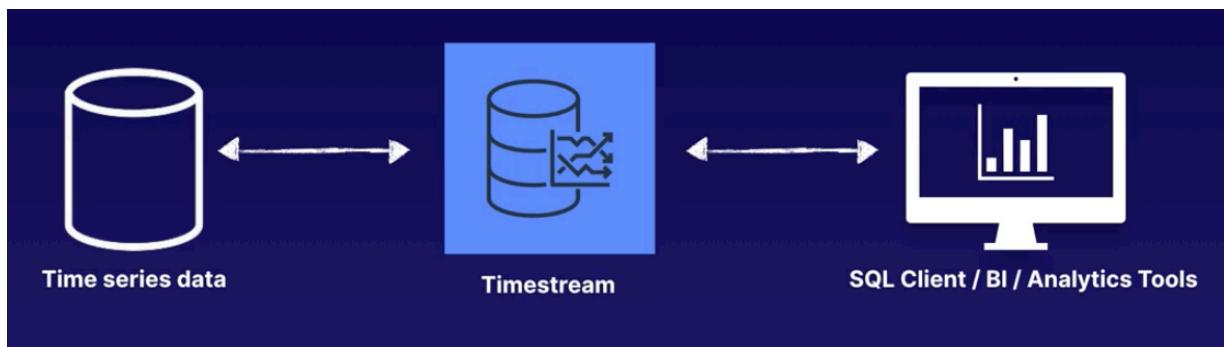
Redshift spectrum can query data from S3 and use tools like Quicksight to visualize it



New **TimeStream**, announced in 2018

To handle time series data

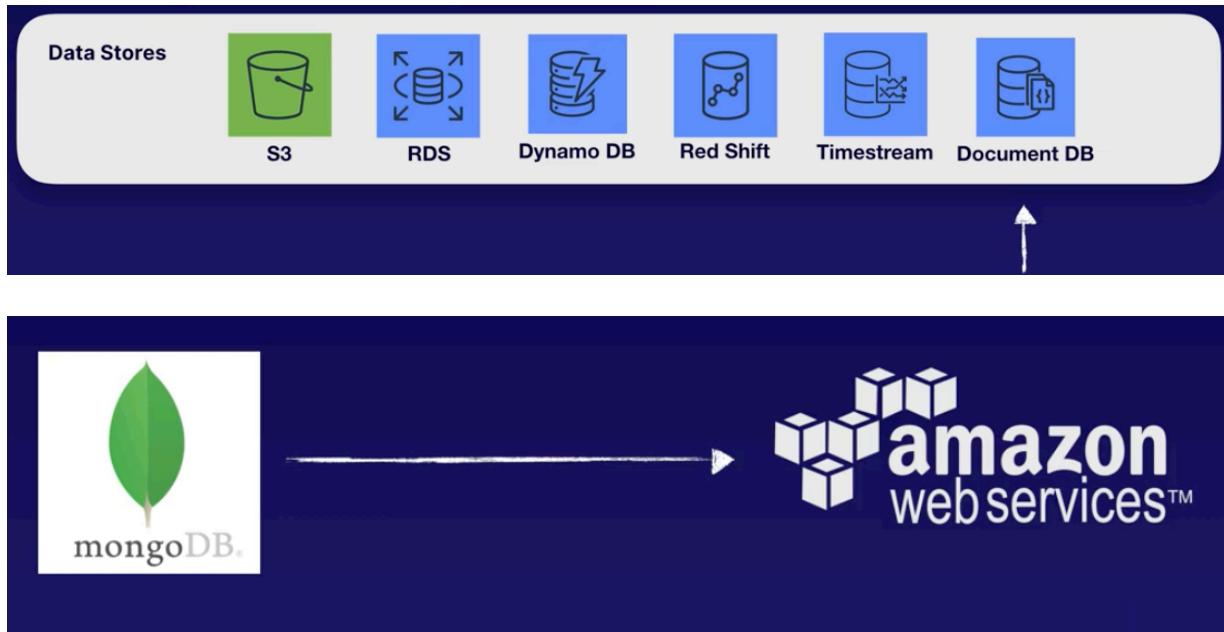
Will probably Not be in the exam



### Document DB

Place to migrate MongoDB data

Probably not be in the exam



These services are used at a high level in the exam  
 Can still go more in depth - solutions architect developer course will provide more details.

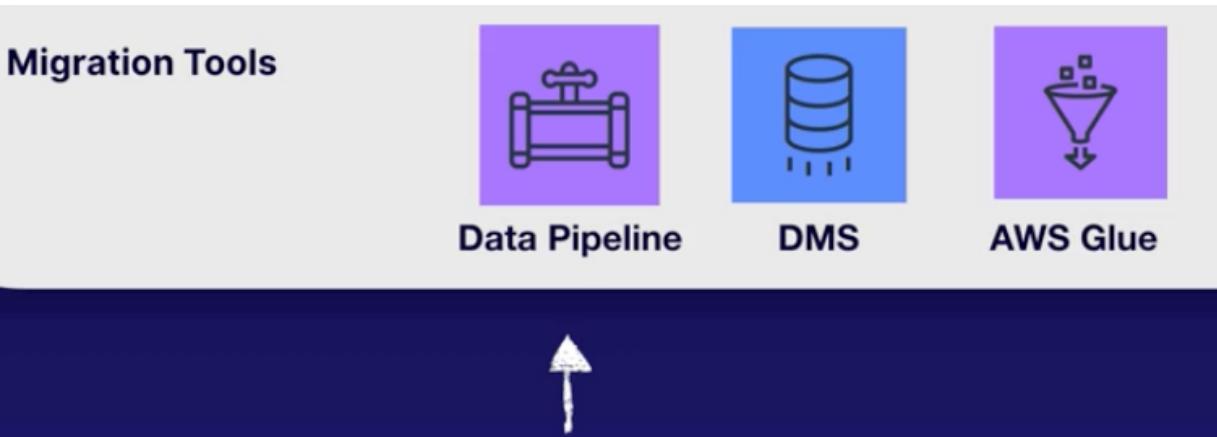
### AWS Migration Tools

To help get our data into S3 so it's ready for ML process

# AWS Migration Tools

Can be a question: If data is in one location, what is the best tool or strategy to get it into another location

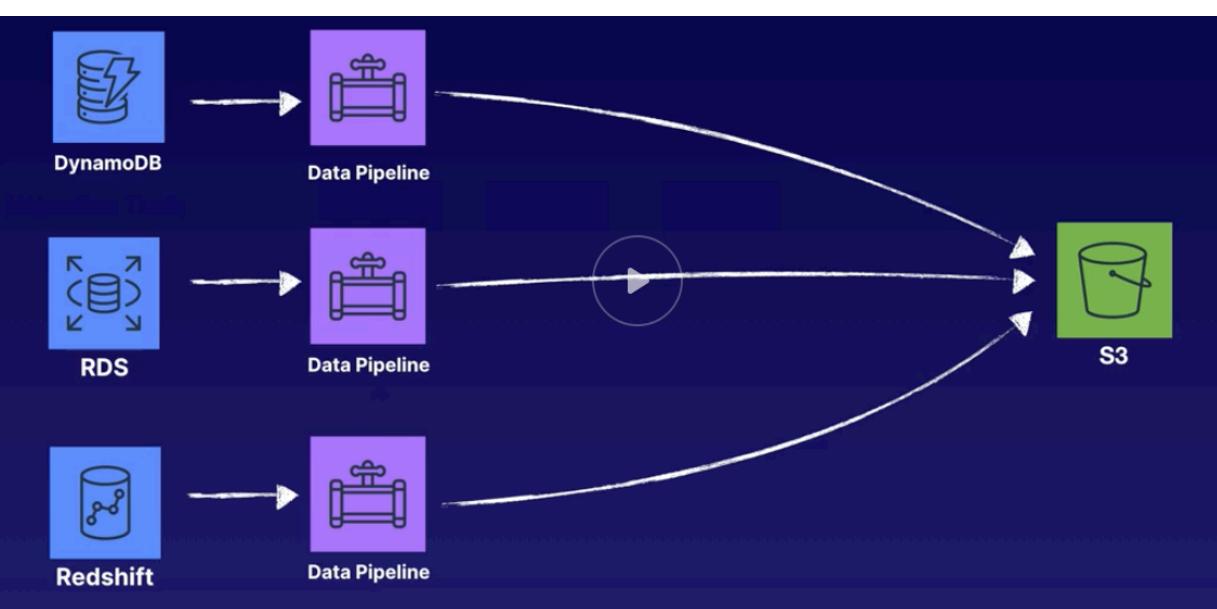
### Data Pipeline



## AWS Data Pipeline

AWS Data Pipeline helps you move, integrate, and process data across AWS compute and storage resources, as well as your on-premises resources. AWS Data Pipeline supports integration of data and activities across multiple AWS regions.

[Get started now](#)



# Activities

The following are the AWS Data Pipeline activity objects:

## Objects

- [CopyActivity](#)
- [EmrActivity](#)
- [HadoopActivity](#)
- [HiveActivity](#)
- [HiveCopyActivity](#)
- [PigActivity](#)
- [RedshiftCopyActivity](#)
- [ShellCommandActivity](#)
- [SqlActivity](#)

Can use these activities, specify datasource and how we want to migrate the data.

Can run on demand or on schedule

Can also be used as a transformation tool

We can manage the pipeline execution, the resources (EC2) and retry

AWS has some templates already available:

## Create Pipeline

*(i)* You can create pipeline using a template or build one using the Architect page.

Name

Description (optional)

Source  Build using a template  Choose... Choose...

- Getting Started
- AWS Command Line Interface (CLI) Templates
- Run AWS CLI command
- DynamoDB Templates
- Export DynamoDB table to S3
- Import DynamoDB backup data from S3
- Elastic MapReduce (EMR) Templates
- Run job on an Elastic MapReduce cluster
- RDS Templates
- Full copy of RDS MySQL table to S3
- Incremental copy of RDS MySQL table to S3
- Load S3 data into RDS MySQL table
- Redshift Templates
- Full copy of RDS MySQL table to Redshift
- Incremental copy of RDS MySQL table to Redshift
- Load data from S3 into Redshift

Schedule

*(i)* You can run your pipeline once or specify a schedule.

Run  never  after  occurrence(s)

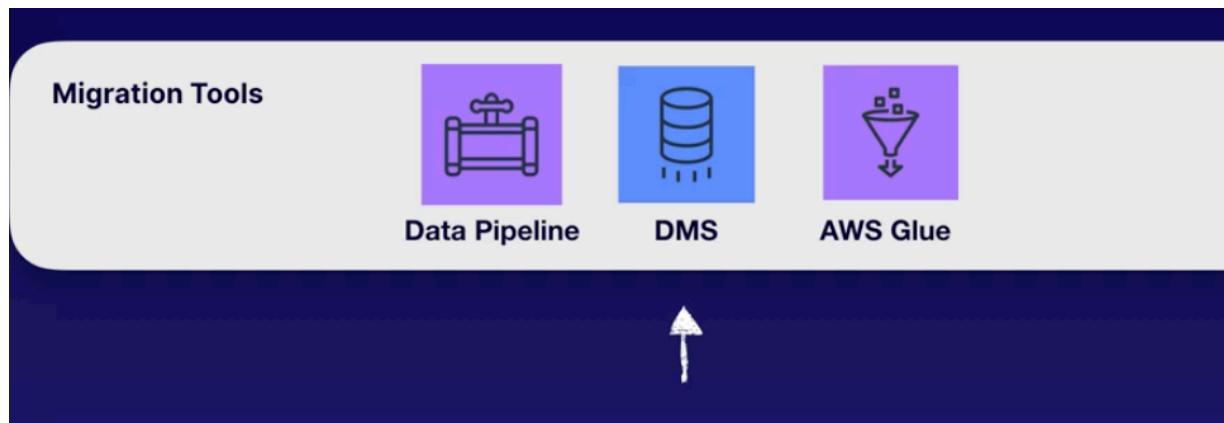
Run every  1 hour  1 day  1 week  1 month  1 year

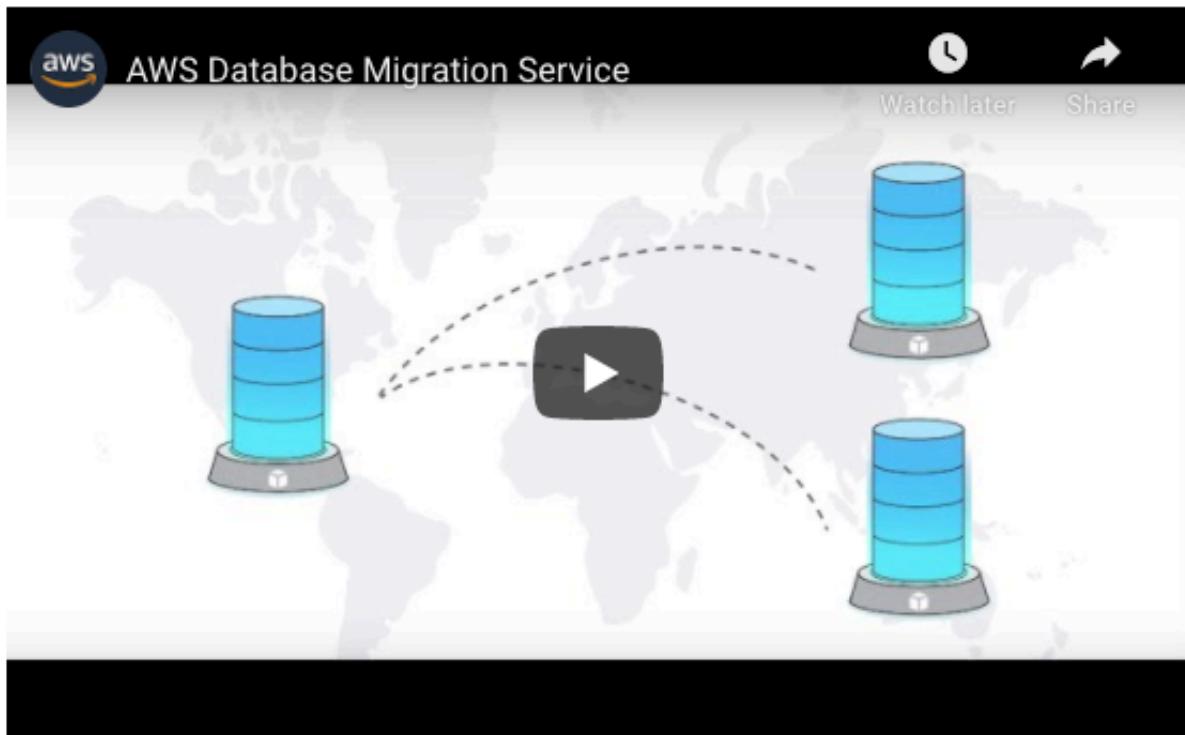
Starting  2019-03-04  UTC Offset from current time is 20:30 UTC

Ending  never  after  occurrence(s)

UTC Offset from current time is 20:30 UTC

## DMS - Database Migration Service





AWS Database Migration Service helps you migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from most widely used commercial and open-source databases.

## Use cases [Learn more](#)

### Homogeneous database migrations

You create a migration task with connections to the source and target databases, then start the migration with the click of a button. AWS Database Migration Service takes care of the rest.

### Database consolidation

You can use AWS Database Migration Service to consolidate multiple source databases into a single target database. This can be done for homogeneous and heterogeneous migrations, and you can use this feature with all supported database engines.

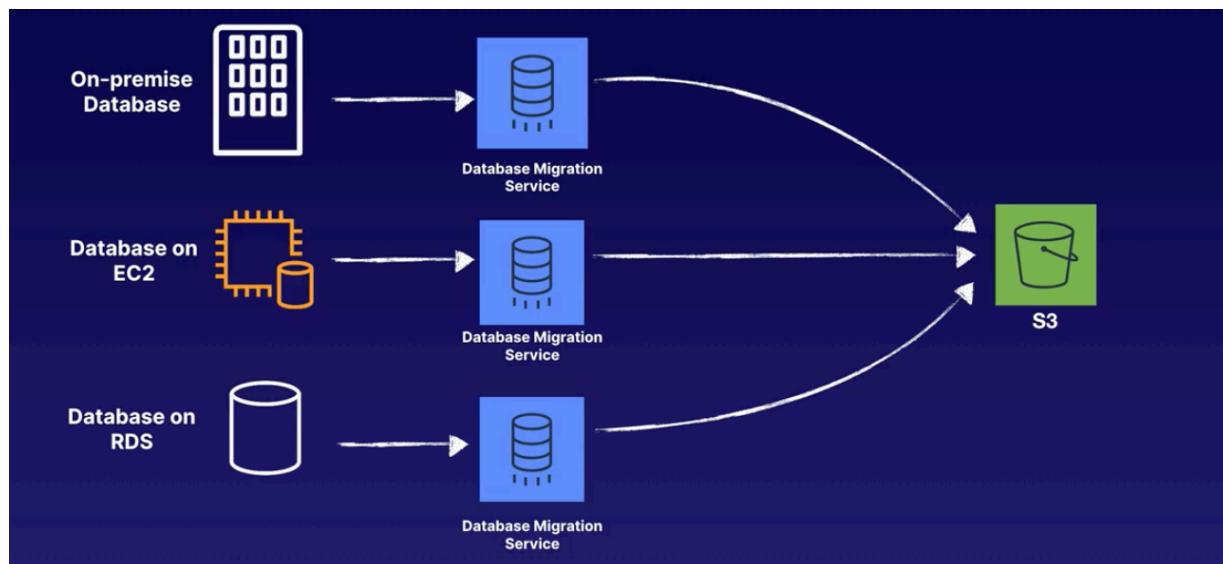
### Heterogenous database migrations

Use the AWS Schema Conversion Tool to convert the source schema and code to match that of the target database, and then use the AWS Database Migration Service to migrate data from the source database to the target database.

### Continuous data replication

Continuous data replication has a multitude of use cases including Disaster Recovery instance synchronization, geographic database distribution and Dev/Test environment synchronization. You can use DMS for both homogeneous and heterogeneous data replications for all supported database engines.

To transfer data between 2 relational DBs, but we can also output data to S3  
Can do SQL Server to S3, or SQL Server to MySQL -> heterogeneous migration



DMS does NOT transfer any transformation  
Other than just column name change

Can set up dat source, and pick target

The screenshot shows two side-by-side screenshots of the AWS DMS console. On the left, under 'Source endpoint', the 'Create endpoint' screen is displayed with the 'Source endpoint' tab selected. A dropdown menu lists various database engines: aurora, sql, db2, mysql, oracle, postgres, and sybase. The 'aurora' option is highlighted with an orange box. On the right, under 'Target endpoint', the same 'Create endpoint' screen is shown, but with the 'Target endpoint' tab selected. This dropdown also lists the same database engines, with 'aurora' again highlighted by an orange box. A large orange arrow points from the source endpoint configuration towards the target endpoint configuration.

DMS supports heterogeneous and homogenous migrations.

## Last Migration Service: AWS Clue

### AWS Glue

AWS Glue is a fully managed ETL (extract, transform, and load) service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores.





## Build your AWS Glue Data Catalog

AWS Glue automatically stores metadata in a central data catalog. It can create table definitions for many common data stores, including, S3 buckets, web logs, and AWS databases. [AWS Glue recognizes, infers, organizes, and classifies your data.](#)



## Generate and edit transformations

PySpark [transformation](#) scripts are auto generated using source and target metadata. You can store customized versions to transform your data to meet your business needs. AWS Glue provides an environment to modify your jobs.



## Schedule and run your jobs

AWS Glue [runs your ETL jobs in a serverless environment](#). You don't need to set up the infrastructure, you just use Amazon's infrastructure and pay for the resources you use. You can define triggers to run jobs based on a schedule or event. AWS Glue enables you to monitor your jobs.

When using Glue, it creates Tables that are related to metadata



Services

## AWS Glue



Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Settings

ETL

Jobs

Triggers

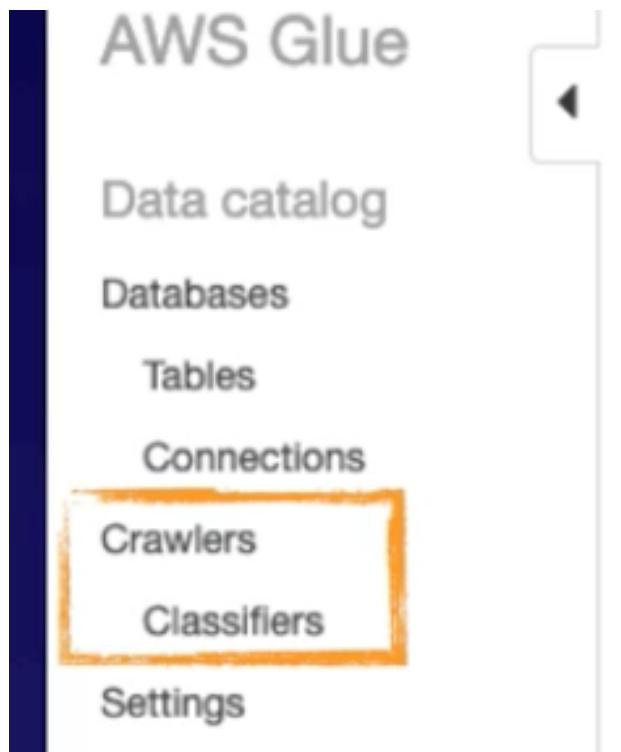
Dev endpoints

Notebooks

Security

Security  
configurations

To create Tables, we set up a Crawler, that goes out, looks at the data and determine the schema associated with that data



Within crawlers, several classifiers that work at hierarchy level and try to infer what type of data it is.

So works for schema less or JSON data

The screenshot shows the "Built-In Classifiers in AWS Glue" documentation page. It includes a table of built-in classifiers and a detailed table of log classifiers.

**Built-In Classifiers in AWS Glue**

AWS Glue provides built-in classifiers for various formats, including JSON, CSV, web logs, and many database systems.

If AWS Glue doesn't find a custom classifier that fits the input data format with 100 percent certainty, it invokes the built-in classifiers in the order shown in the following table. The built-in classifiers return a result to indicate whether the format matches (`certainty=1.0`) or does not match (`certainty=0.0`). The first classifier that has `certainty=1.0` provides the classification string and schema for a metadata table in your Data Catalog.

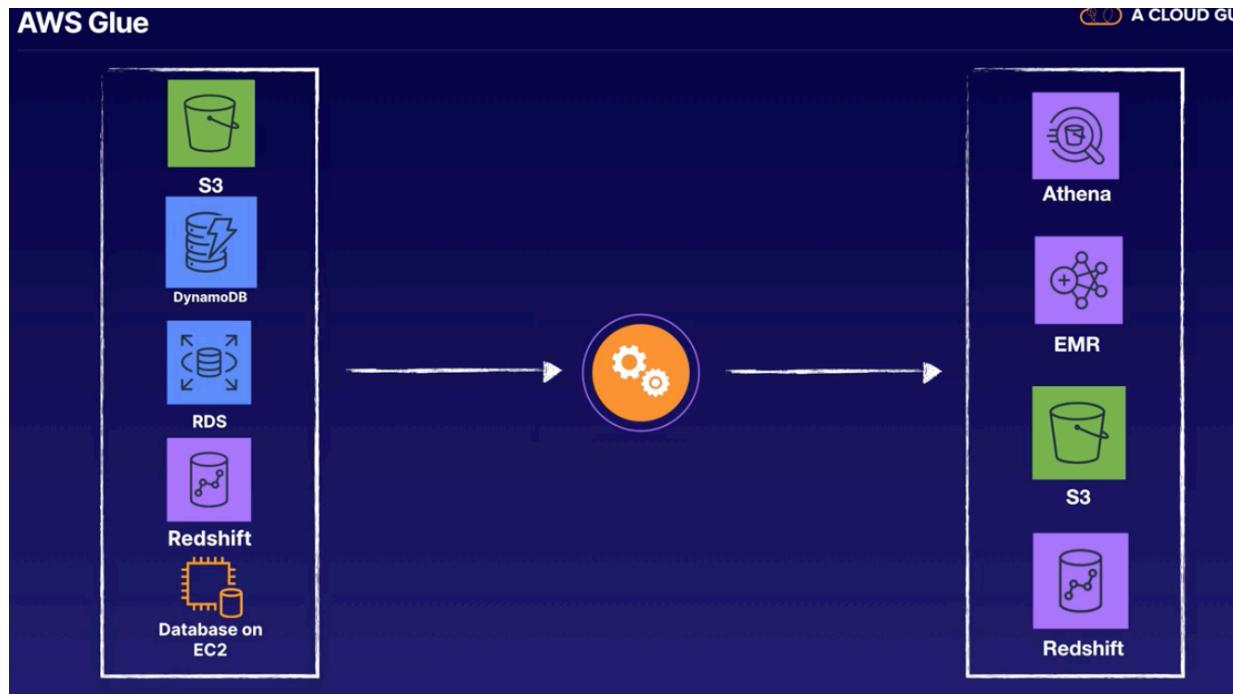
Classifier type	Classification string	Notes
Apache Avro	avro	Reads the schema at the beginning of the file to determine format.
Apache ORC	orc	Reads the file metadata to determine format.
Apache Parquet	parquet	Reads the schema at the end of the file to determine format.
JSON	json	Reads the beginning of the file to determine format.
Binary JSON	bson	Reads the beginning of the file to determine format.
XML	xml	Reads the beginning of the file to determine format. AWS Glue determines the table schema based on XML tags in the document. For information about creating a custom XML classifier to specify rows in the document, see <a href="#">Writing XML Custom Classifiers</a> .
Amazon Ion	ion	Reads the beginning of the file to determine format.
Combined Apache log	combined_apache	Determines log formats through a grok pattern.
Apache log	apache	Determines log formats through a grok pattern.
Linux kernel log	linux_kernel	Determines log formats through a grok pattern.

**Log Classifiers**

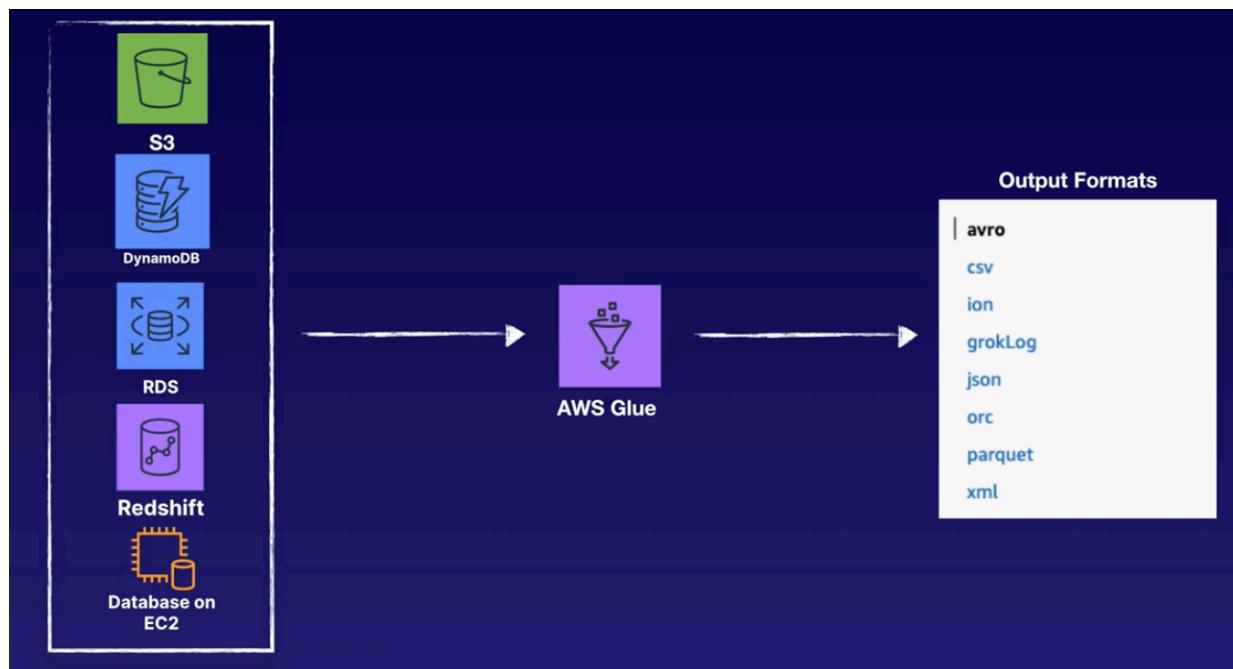
Microsoft log	microsoft_log	Determines log formats through a grok pattern.
Ruby log	ruby_logger	Reads the beginning of the file to determine format.
Squid 3.x log	squid	Reads the beginning of the file to determine format.
Redis monitor log	redismonlog	Reads the beginning of the file to determine format.
Redis log	redislog	Reads the beginning of the file to determine format.
CSV	csv	Checks for the following delimiters: comma (,), pipe (), tab (\t), semicolon (;), and Ctrl-A (\u00001). Ctrl-A is the Unicode control character for Start Of Heading.
Amazon Redshift	redshift	Uses JDBC connection to import metadata.
MySQL	mysql	Uses JDBC connection to import metadata.
PostgreSQL	postgresql	Uses JDBC connection to import metadata.
Oracle database	oracle	Uses JDBC connection to import metadata.
Microsoft SQL Server	sqlserver	Uses JDBC connection to import metadata.
Amazon DynamoDB	dynamodb	Reads data from the DynamoDB table.

To note we can create specific classifiers that specify our data

So with a crawler and classifier, we can pull the data, map it to an appropriate schema and export to various sources like S3



We can also change the format to any ML format



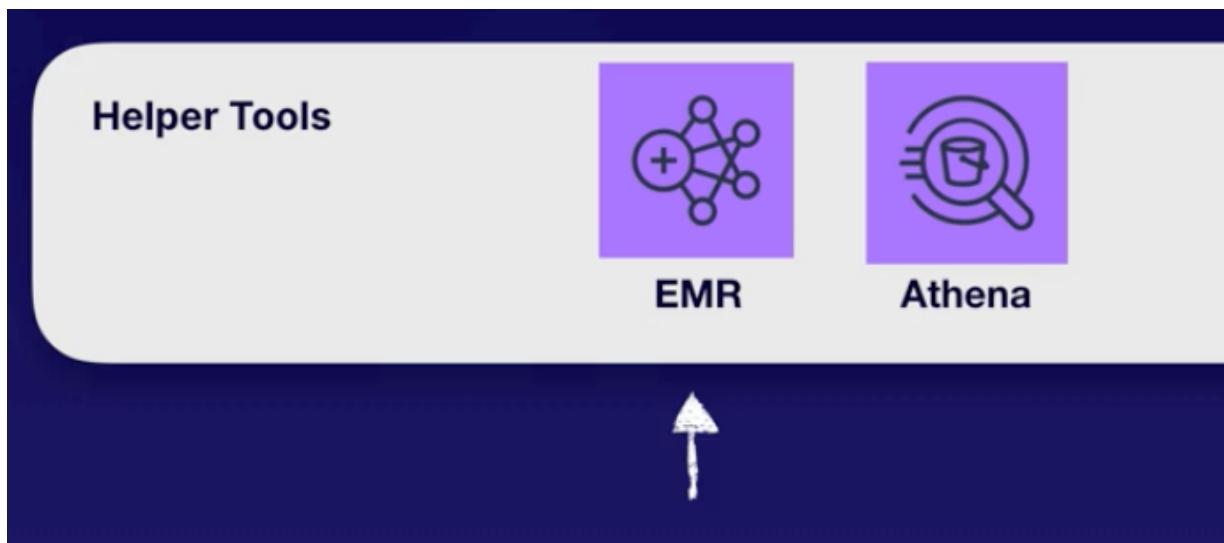
Choose the right approach for migration.  
Pull from a DB and export to S3

Datasource	Migration Tool	Why
PostgreSQL RDS instance with training data.	AWS Data Pipeline	Specify SqlActivity query and places the output into S3.
Unstructured log files in S3.	AWS Glue	Create custom classifier and output results into S3.
Clustered Redshift data.	AWS Data Pipeline AWS Glue	Use the <b>unload</b> command to return results of a query to CSV file in S3. Create Data catalog describing data and load it into S3.
On-premise MySQL instance with training data.	AWS Database Migration Service	DMS can load data in CSV format onto S3.

## AWS Helper Tools

# AWS Helper Tools

## Amazon EMR



Fully managed Hadoop Cluster ecosystem that runs on multiple EC2 instances

Pick and choose different frameworks we want to use in a cluster  
To run different workload across multiple EC2 for petabytes worth of data

**Elastic Map Reduce (EMR)**

A CLOUD GURU

The collage includes the following logos:

- APACHE PIG
- presto
- mxnet
- MAHOUT
- hadoop HDFS
- OOZIE
- TensorFlow
- HIVE
- APACHE HBASE
- Jupyter
- Zookeeper
- APACHE Flume
- Apache Spark

Spark ML and MLib ETL and Machine Learning Library	Presto SQL Query Engine	Mahout Machine Learning Framework
Hive ETL Service	Jupyter Notebooks Code Sharing	TensorFlow Machine Learning Framework
Hadoop Distributed File System Persistant Datastore	mxnet	MXNet Machine Learning Framework

We can use EMR to store vast amount of files in a different file system, to use as an Input or Training data  
If data is already in EMR, we can use frameworks/services to migrate the data into S3, or use AWS Data Pipeline to load the data into S3

When creating a cluster, we can use different frameworks:

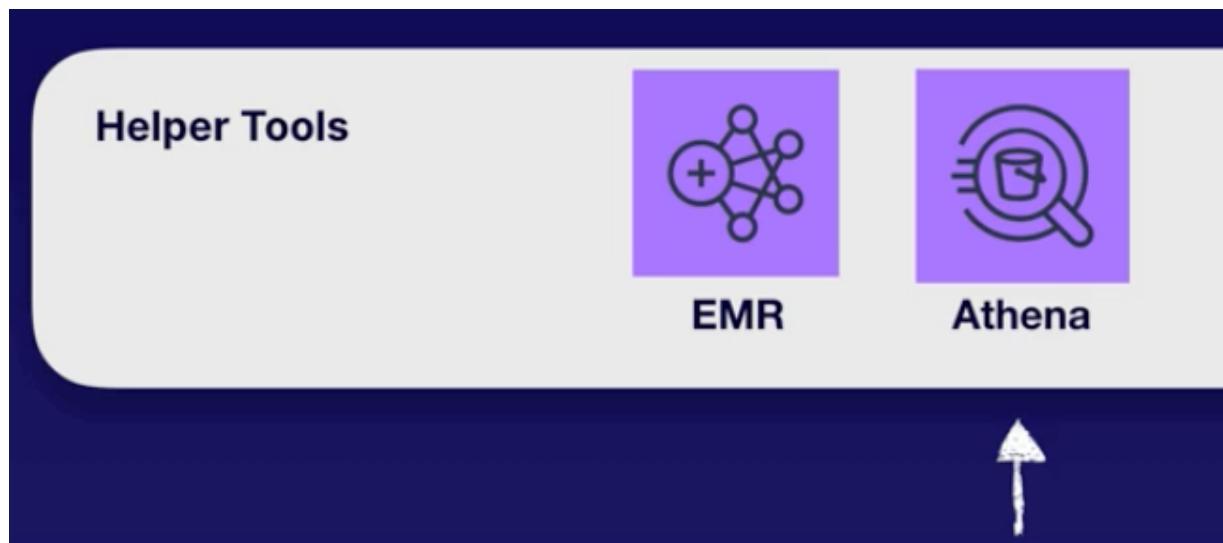
## Create Cluster - Advanced Options

### Software Configuration

Release emr-5.20.0

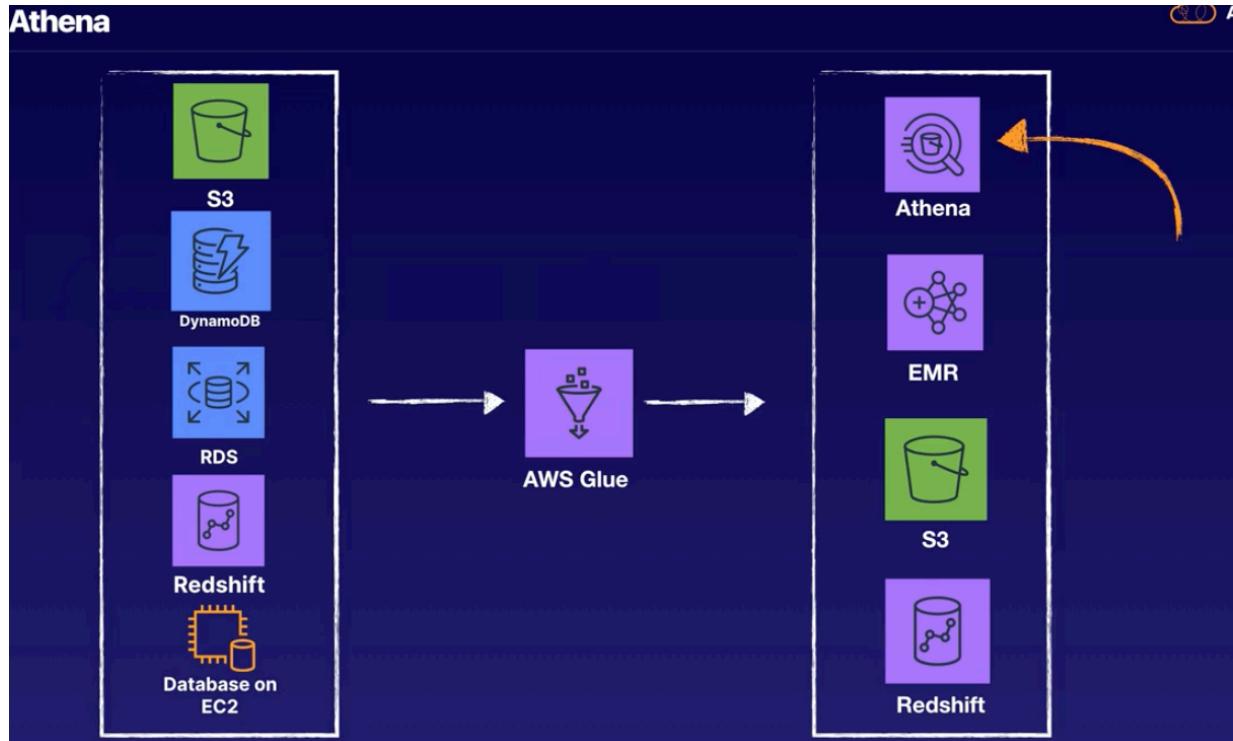
- |  |  |   |
|--|--|---|
| <input checked="" type="checkbox"/> Hadoop 2.8.5     | <input type="checkbox"/> Zeppelin 0.8.0          | <input type="checkbox"/> Livy 0.5.0                   |
| <input checked="" type="checkbox"/> JupyterHub 0.9.4 | <input type="checkbox"/> Tez 0.9.1               | <input type="checkbox"/> Flink 1.6.2                  |
| <input type="checkbox"/> Ganglia 3.7.2               | <input type="checkbox"/> HBase 1.4.8             | <input type="checkbox"/> Pig 0.17.0                   |
| <input checked="" type="checkbox"/> Hive 2.3.4       | <input checked="" type="checkbox"/> Presto 0.214 | <input type="checkbox"/> ZooKeeper 3.4.13             |
| <input checked="" type="checkbox"/> MXNet 1.3.1      | <input type="checkbox"/> Sqoop 1.4.7             | <input checked="" type="checkbox"/> Mahout 0.13.0     |
| <input type="checkbox"/> Hue 4.3.0                   | <input type="checkbox"/> Phoenix 4.14.0          | <input type="checkbox"/> Oozie 5.0.0                  |
| <input checked="" type="checkbox"/> Spark 2.4.0      | <input type="checkbox"/> HCatalog 2.3.4          | <input checked="" type="checkbox"/> TensorFlow 1.12.0 |

## Amazon Athena



Serveless platform to run SQL queries on S3

We can set up a table within data catalog within AWS Glue, and use Athena to query S3



Screenshot of the AWS Athena Query Editor interface:

- Top Bar:** Shows the AWS logo, Services dropdown, Resource Groups dropdown, Brock Tubre (User), N. Virginia (Region), and Support.
- Navigation:** Athena (selected), Query Editor, Saved Queries, History, AWS Glue Data Catalog (highlighted), Workgroup: primary, Settings, Tutorial, Help, What's New.
- Left Panel (Database):**
  - Database dropdown: adult-data-database (selected).
  - Tables (1): ml\_sandbox\_demo
    - age (bigint)
    - workclass (string)
    - fnlwgt (bigint)
    - education (string)
    - education\_num (bigint)
    - marital\_status (string)
    - occupation (string)
    - relationship (string)
    - race (string)
    - sex (string)
    - capital\_gain (bigint)
    - capital\_loss (bigint)
    - hours\_per\_week (bigint)
    - native\_country (string)
    - income (string)
  - Views (0):
  - Create view: Create view
- Right Panel (Query Editor):**
  - New query 1: A text input field containing the number "1".
  - Run query, Save as, Create buttons.
  - Format query, Clear buttons.
  - Results: An empty table.

Message at the bottom: You have not created any views. To create a view, run a query and click "Create view from query".

The screenshot shows the AWS Athena Query Editor interface. On the left, the Database dropdown is set to "adult-data-database" and the Tables section lists "ml\_sandbox\_demo" with 10 columns: age, workclass, fnlwgt, education, education\_num, marital\_status, occupation, relationship, race, sex, and capital. The Views section is empty. In the center, a query editor window titled "New query 1" contains the following SQL code:

```

1 SELECT * FROM ml_sandbox_demo
2 WHERE sex = 'Male'
3 AND age > 39 limit 10;

```

Below the query are buttons for "Run query", "Save as", and "Create". A status message indicates "(Run time: 1.68 seconds, Data scanned: 1.12 MB)". At the bottom, there are "Format query" and "Clear" buttons. On the right, the "Results" section displays a table with 9 rows of data corresponding to the query results.

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital
1	50	Self-emp-not-inc		Bachelors		Married-civ-spouse	Exec-managerial	Husband	White	Male	
2	53	Private		11th		Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
3	52	Self-emp-not-inc		HS-grad		Married-civ-spouse	Exec-managerial	Husband	White	Male	
4	42	Private		Bachelors		Married-civ-spouse	Exec-managerial	Husband	White	Male	
5	40	Private		Assoc-voc		Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	
6	40	Private		Doctorate		Married-civ-spouse	Prof-specialty	Husband	White	Male	
7	43	Private		11th		Married-civ-spouse	Transport-moving	Husband	White	Male	
8	56	Local-gov		Bachelors		Married-civ-spouse	Tech-support	Husband	White	Male	
9	54	?		Some-college		Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	

Difference between Athena and Redshift spectrum

<h1>Redshift Spectrum</h1> <h2>Athena</h2>	<ul style="list-style-type: none"> <li>• Query S3 data</li> <li>• Must have Redshift Cluster</li> <li>• Made for existing Redshift customers</li> </ul>
	<ul style="list-style-type: none"> <li>• Query S3 data</li> <li>• No need for Redshift cluster</li> <li>• New customers quickly want to query S3 data</li> </ul>

Overall, S3 is our one stop shop for Data store for SageMaker

