

Project Overview

Working backwards, from “what does my customer need”

The requirement from (our simulated) customer:

- Launching new data-driven campaign
- Main advertising channel : YouTube
- Initial questions to answer:
 - “how to categorise videos, based on their comments and statistics”
 - “What factors affect how popular a YouTube video will be”



Why YouTube?

Top three most-visited websites (monthly)

- Google: 92.5 billion
- YouTube: 34.6 billion
- Facebook: 25.5 billion

Source: visualcapitalist.com/the-50-most-visited-websites-in-the-world/



Goals and Success Criteria

How my customer will measure success?

Data Ingestion

Ingest data, one-offs and incrementally

Data Lake

Design and build a new Data Lake architecture

AWS Cloud

AWS as the cloud provider



ETL Design

Extract, transform and load data efficiently

Scalability

The data architecture should scale efficiently

Reporting

Build a Business Intelligence tier, incl. Dashboards

What you will learn in this course

- To build a data lake from scratch in Amazon S3

Joining semi-structured and structured data

- Lake House architecture design

Best practices —> cost and performance.

- Data Lake vs. Data Warehouse



What you will learn in this course

- Data lake design in layers, partitioned for cost-performance

e.g. landing, cleansed as SSOT, reporting for BI users

WORM model / Write Once Read Many

- AWS Data Catalogue



What you will learn in this course

- ETL in AWS Glue Spark jobs

Amazon SageMaker Jupyter Notebooks.

- Amazon SNS for alerting
- SQL using Amazon Athena and Spark SQL

i.e. impact of querying the optimized data layers



What you will learn in this course

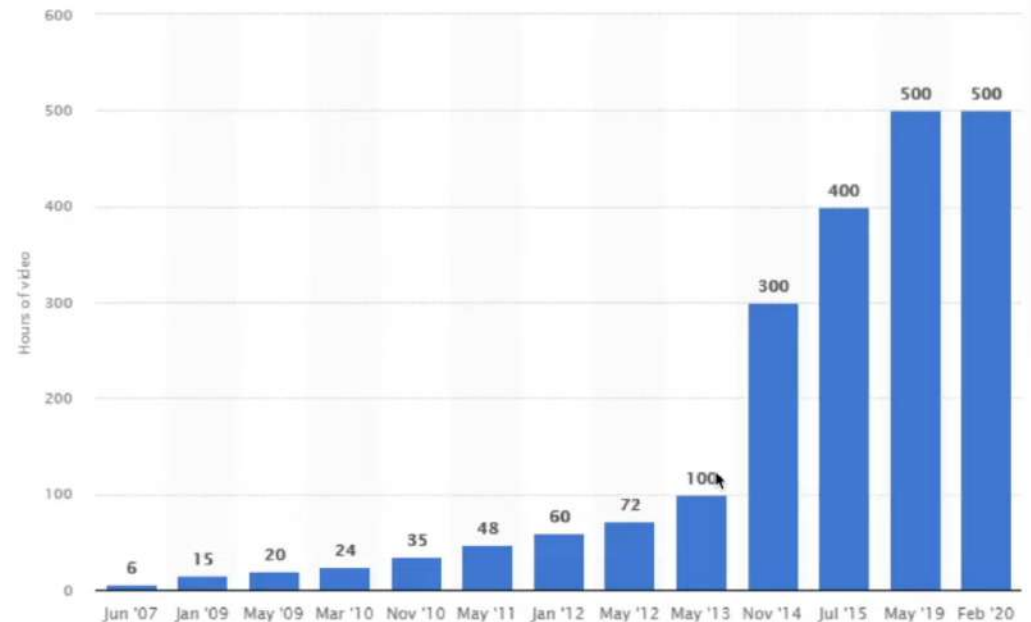
- Ingest changes incrementally and schema evolution
- BI dashboards in Amazon QuickSight



What is Big Data

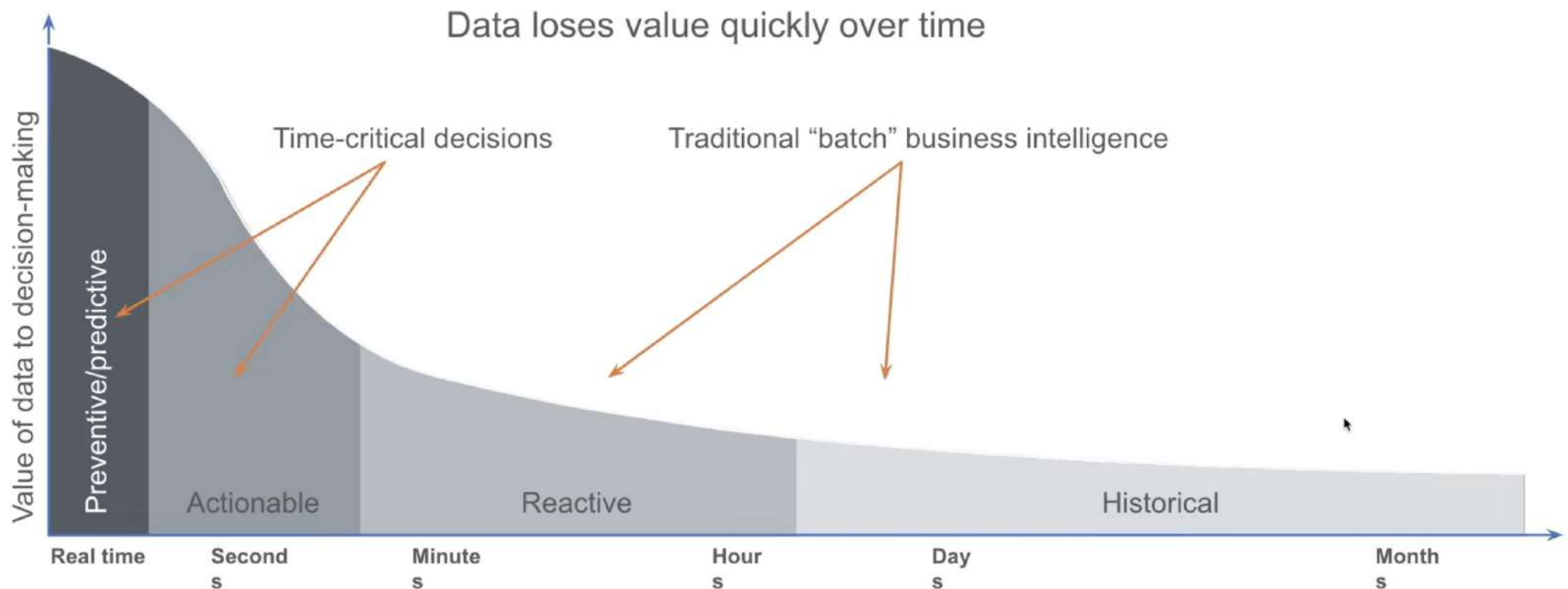
Big data is a term for:

- massive data sets, with varied and complex structure
- with the difficulties of storing and analysing
- visualizing for further processes or results.



Source: Big data: A review. IEEE, 2013, DOI: 10.1109/CTS.2013.6567202

Timely decisions require new data in minutes

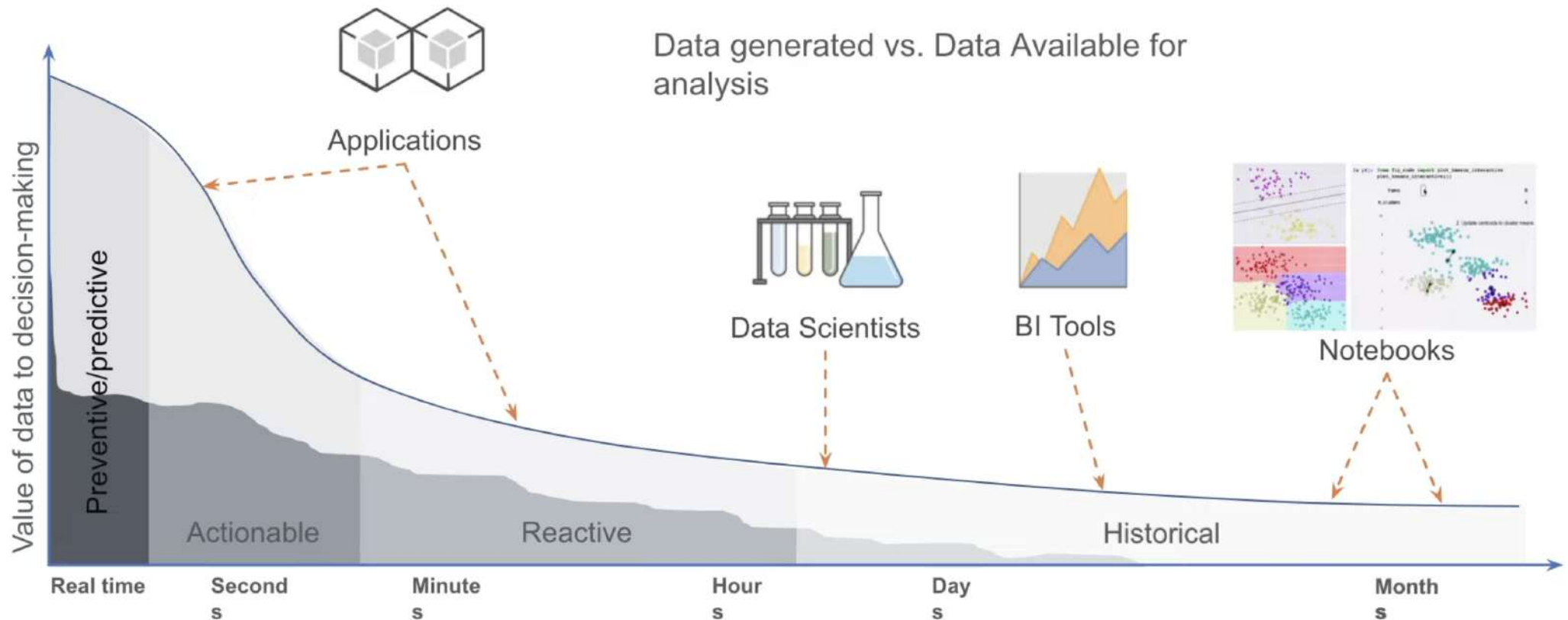


Source: Perishable insights, Mike Gualtieri, Forrester

Challenges working with data

Do you think a single Production DB can handle all these users?

Data generated vs. Data Available for analysis



Our dataset from YouTube

- Top trending videos
- What is “*Trending*”?

YouTube uses factors, including users interactions
e.g. number of views, shares, comments and likes.

Not the most-viewed videos overall for the calendar year



- Source: Kaggle. Data collected using YouTube API

Source: <https://www.kaggle.com/datasnaek/youtube-new>

< US_category_id.json (8.5 kB)

```
10 : {...} 4 items
11 : { 4 items
  "kind" : string "youtube#videoCategory"
  "etag" : string "\"m2yskBQFythfE4irbTieOgYYfBU/UVB9oxX2Bvqa_w_y3vXSLVK5E_s\""
  "id" : string "24"
  "snippet" : { 3 items
    "channelId" : string "UCBR8-60-B28hp28mDPdntcQ"
    "title" : string "Entertainment"
    "assignable" : bool true
  }
}
```

< USvideos.csv (62.76 MB)

Detail Compact Column					16 of 16 columns	
title	channel_title	category_id	publish_time	tags		
6455 unique values	2207 unique values			[none] 4% ABC "americanidol" ... 0% Other (39327) 96%		
The Trump Presidency: Last Week Tonight with John Oliver (HBO)	LastWeekTonight	24	2017-11- 13T07:30:00.000Z	last week tonight trump presidency "last week tonight donald trump "john oliver trump "donald trum...		

What is an on-premise data centre

- You / your company has its own hardware
- You will need to maintain it; e.g.
 - Purchase, fix and upgrade hardware
 - Install and maintain Operating Systems and other software



What is the cloud?

- Applications delivered as services over the Internet
- Hardware and systems software in data centres, that provide those services

Image source: techspot.com,



Image source: amazon.com,
2021

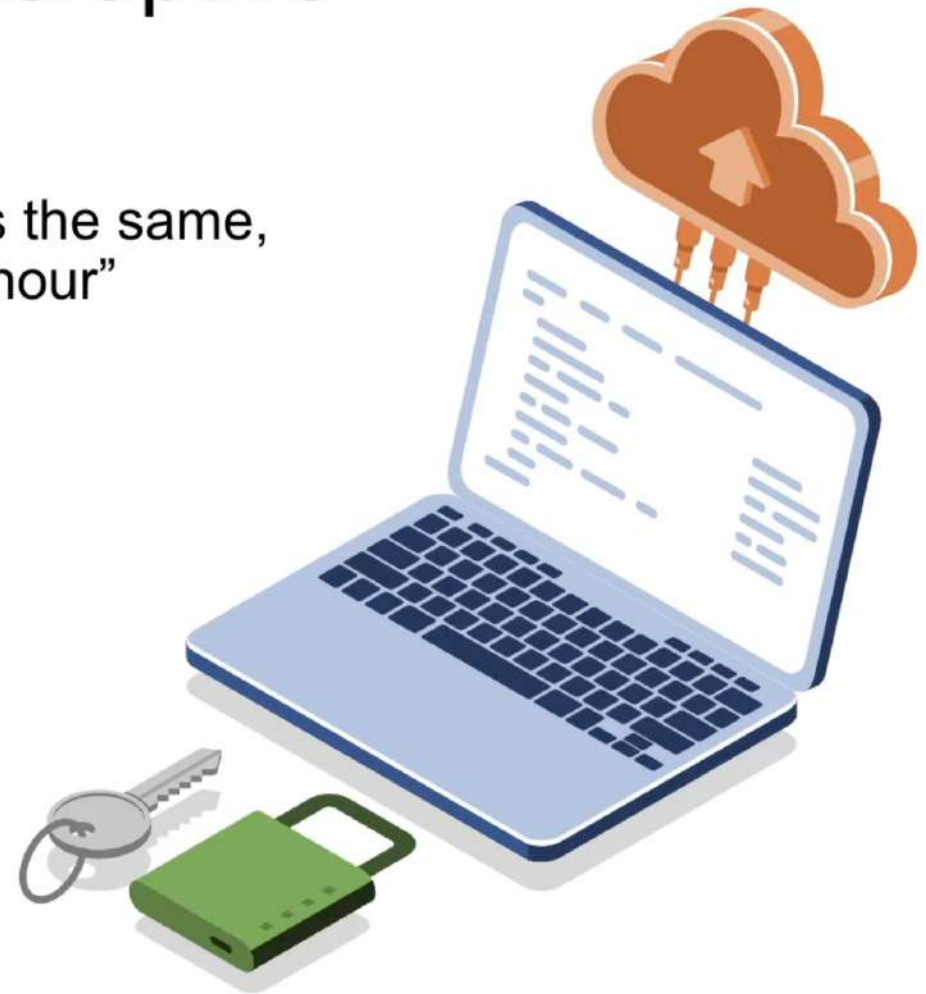
Source: "A view of cloud computing"; doi.acm.org/10.1145/1721654.1721672

Why the cloud is being so disruptive

“A 10-node cluster running for 10 hours costs the same,
as a 100-node cluster running for one hour”

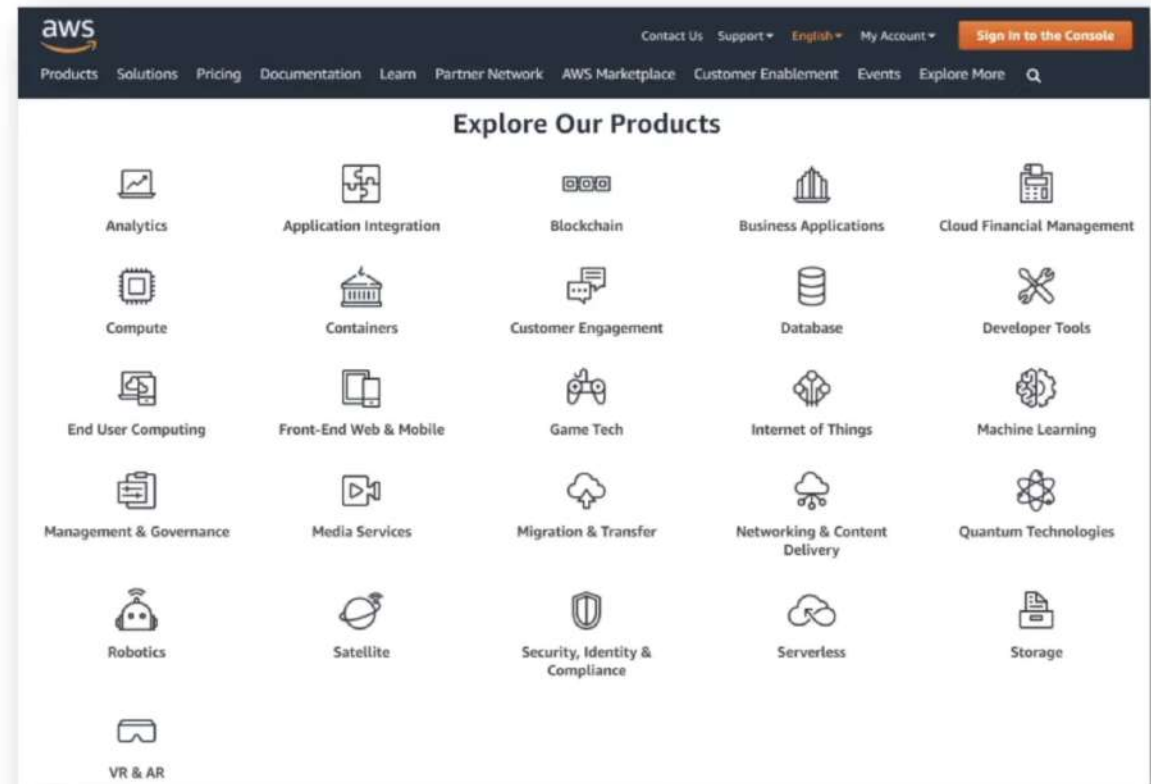
On-premise, extra-hardware for a cluster
means...

- Procurement, approval, licensing
- Shipping, installation, electricity, cooling systems,...
- Underutilized vs overutilized
- Pay even when nobody is using the hardware
- Sustainability issues



[Cont.] Why the cloud is being so disruptive

- 81% of enterprises have, at least, one application in the cloud
- 13% of +1k employees enterprises have migrated their entire IT environment to the cloud.



Source: "32% Of IT Budgets Will Be Dedicated To The Cloud By 2021". Forbes, 2020

Steps to get our data

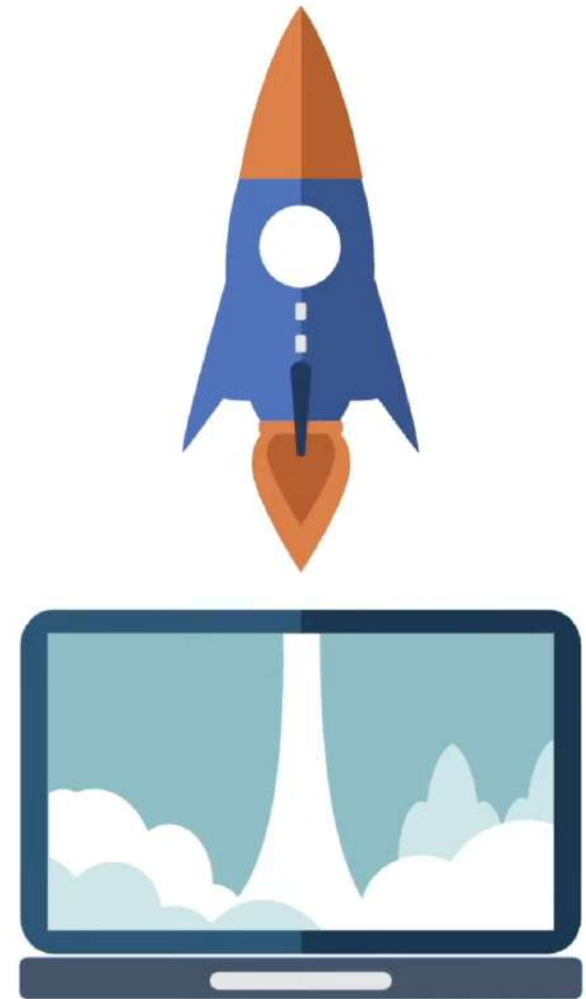
- Download from kaggle.com/datasnaek/youtube-new
- Create an Amazon S3 bucket, for our landing bucket
e.g.

```
s3://company-raw-awsregion-awsaccountID-  
env/source/source_region/tablename/year=yyyy/  
month=mm/day=dd/table_<yearmonthday>.<file_format>  
  
env = dev, test, prod  
source = name or indicator of source  
source_region = region of data source
```

- Copy the data to S3, using our AWS CLI

Agenda

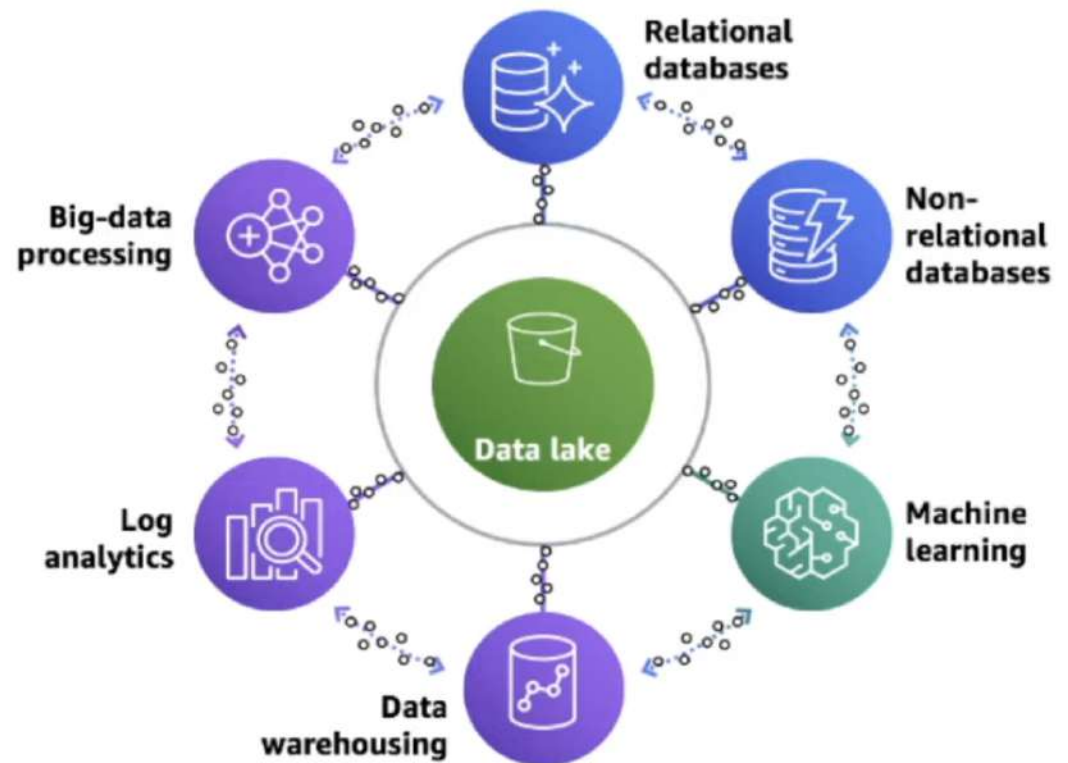
- Basics of Lake House architecture
- What is the AWS Glue Data Catalog
- Catalog our YouTube data



What is a Lake House architecture

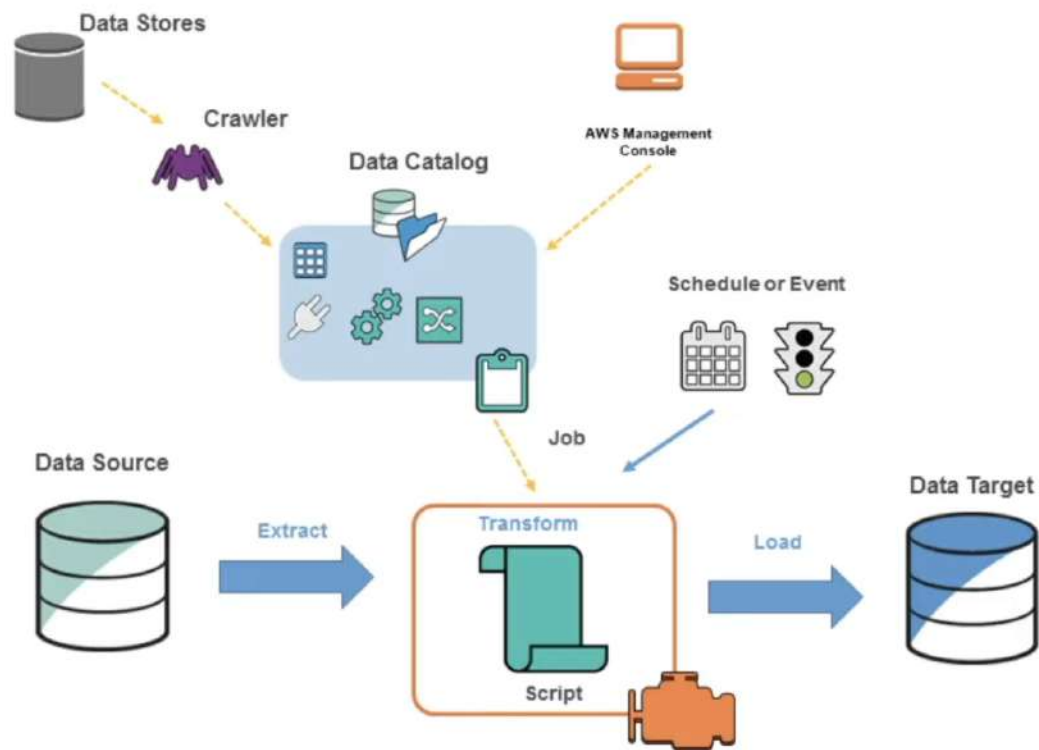
Key elements:

- Scalable Data Lakes
- Purpose-built Data Services
- Seamless Data Movement
- Unified Governance
- Performant and Cost-effective

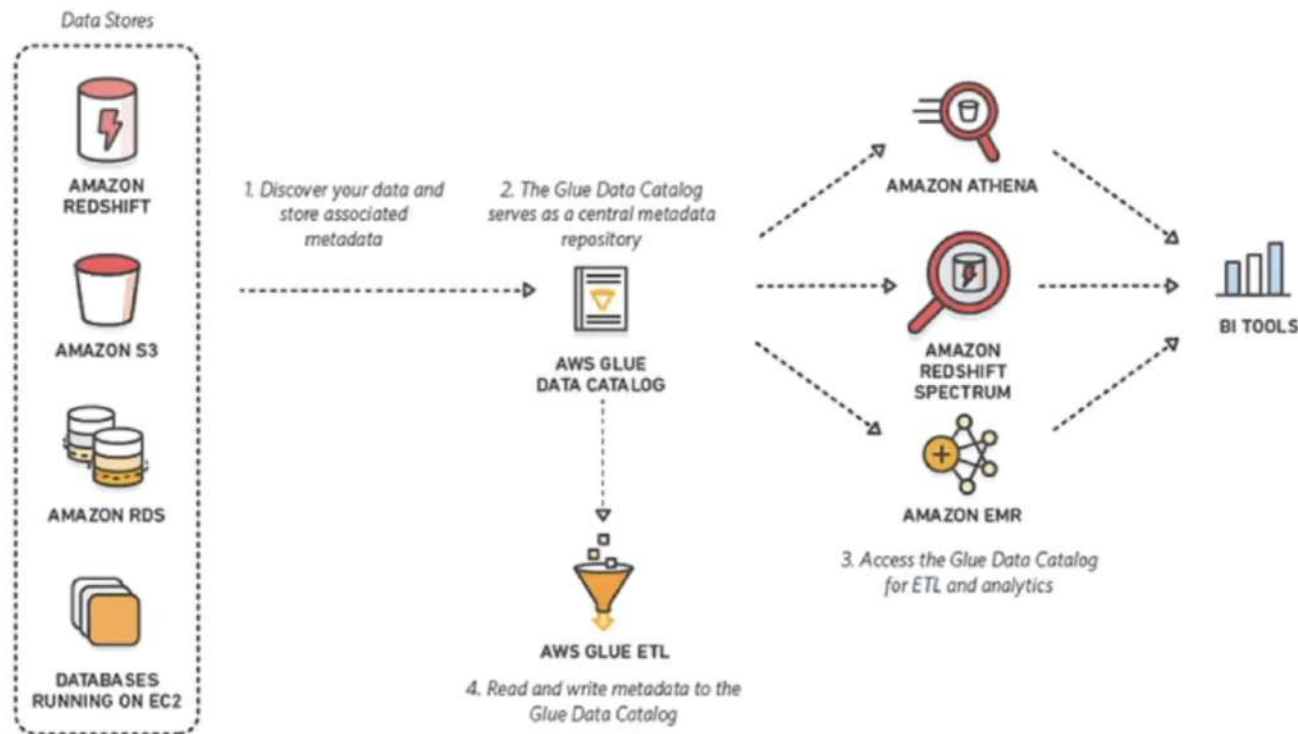


Source: aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/

What is the AWS Glue Catalog

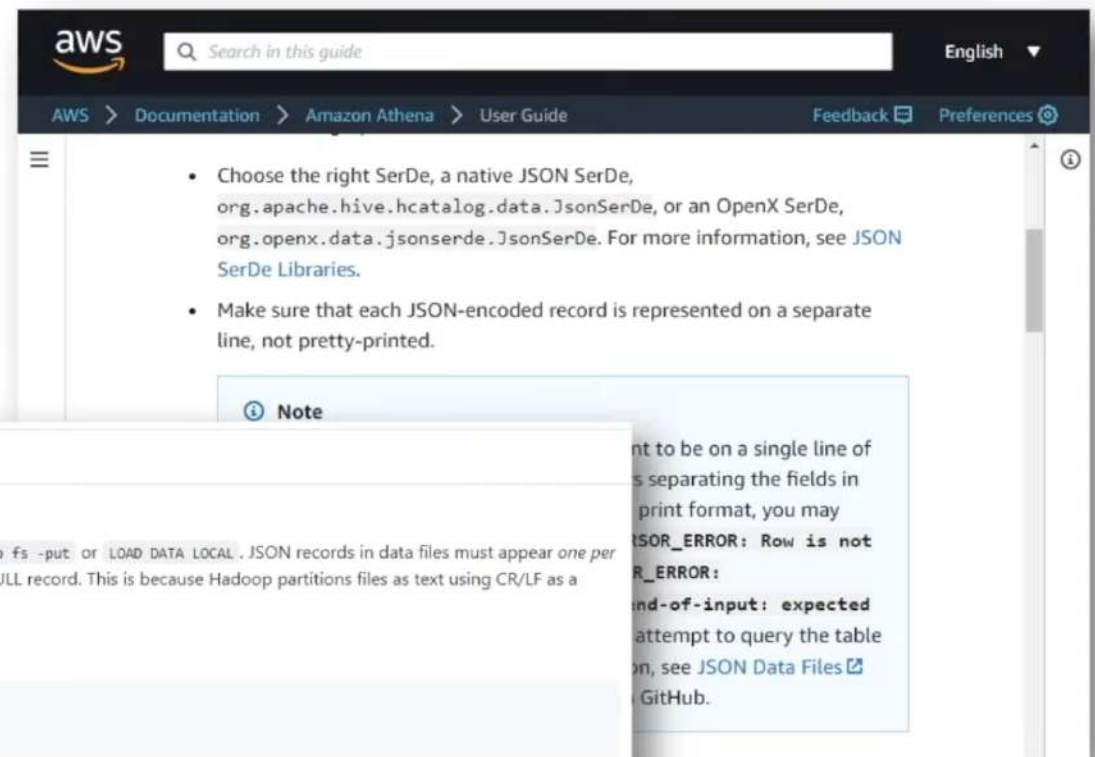


What is the AWS Glue Catalog



Source: aws.amazon.com/blogs/big-data/harmonize-query-and-visualize-data-from-various-providers-using-aws-glue-amazon-athena-and-amazon-quicksight/

JSON SerDe Libraries and Support



GitHub
Hive

README.md

JSON Data Files

Upload JSON files to HDFS with `hadoop fs -put` or `LOAD DATA LOCAL`. JSON records in data files must appear *one per line*, an empty line would produce a NULL record. This is because Hadoop partitions files as text using CR/LF as a separator to distribute work.

The following example will work.

```
{ "key" : 10 }  
{ "key" : 20 }
```

The following example will not work.

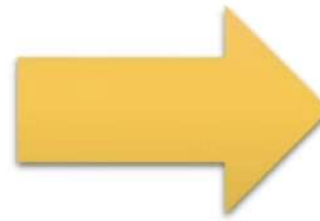
```
{  
  "key" : 10  
}  
{  
  "key" : 20  
}
```

Goal: Data Cleansing

Create our light ETL: JSON to Apache Parquet

```
1 {
2   "kind": "youtube#videoCategoryListResponse",
3   "etag": "\"ld9b1NPKjA9gJV7EZ4EKegRhao/1v2mrzYSY06onNLT2qTj13hkQZk\"",
4   "items": [
5     {
6       "kind": "youtube#videoCategory",
7       "etag": "\"ld9b1NPKjA9gJV7EZ4EKegRhao/Xy1mB4_vLrHy_BmKmPBggy2mZQ\"",
8       "id": "1",
9       "snippet": {
10         "channelId": "UCBR8-60-B28hp2BmDPdntcQ",
11         "title": "Film & Animation",
12         "assignable": true
13       }
14     },
15     {
16       "kind": "youtube#videoCategory",
17       "etag": "\"ld9b1NPKjA9gJV7EZ4EKegRhao/UZ1oLIz2dxIn045ZTFR5a3NyTA\"",
18       "id": "2",
19       "snippet": {
20         "channelId": "UCBR8-60-B28hp2BmDPdntcQ",
21         "title": "Autos & Vehicles",
22         "assignable": true
23       }
24     }
25   ]
26 }
```

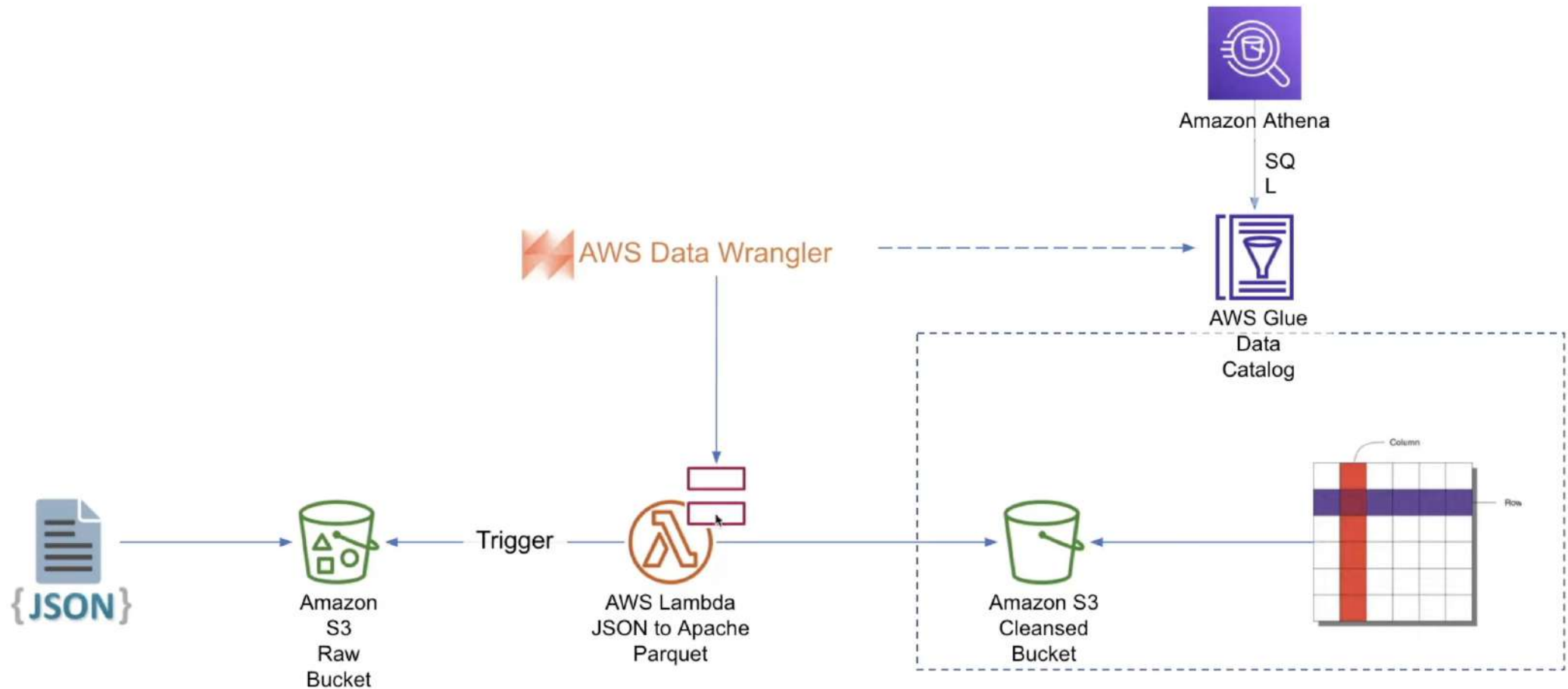
CA_category_id.json



a	b	c	...	zf
a1	b1	c1	...	zf1
a2	b2	c2	...	zf2
a3	b3	c3	...	zf3
a4	b4	c4	...	zf4
a5	b5	c5	...	zf5
a6	b6	c6	...	zf6
a7	b7	c7	...	zf7

Data Cleansing

Semi-structured data to Structured pipeline



Why Lambda to process our JSON payload?

- One of the AWS compute services
- Serverless
- High available and scalable
- Limits at this moment:
 - Deployment package is 50MB
 - You can use /tmp or mount EFS volumes: more storage and share it across executions
 - 10 GB for memory
 - 6 vCPUs
 - 15 min timeout

Comparing the different data storage options

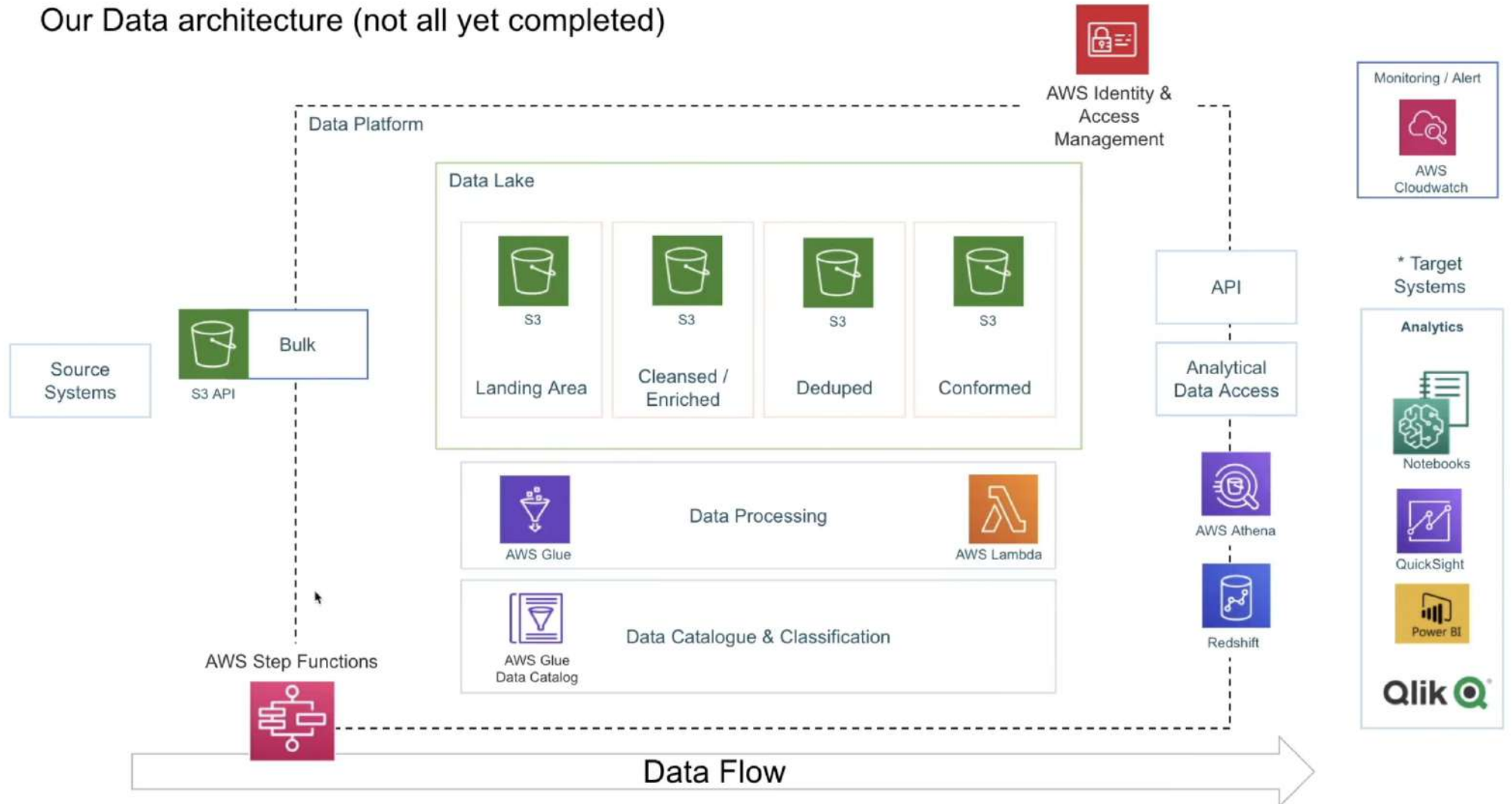
This table compares the characteristics of these four different data storage options for Lambda:



	Amazon S3	/tmp	Lambda Layers	Amazon EFS
Maximum size	Elastic	512 MB	50 MB (direct upload; larger if from S3).	Elastic
Persistence	Durable	Ephemeral	Durable	Durable
Content	Dynamic	Dynamic	Static	Dynamic
Storage type	Object	File system	Archive	File system
Lambda event source integration	Native	N/A	N/A	N/A
Operations supported	Atomic with versioning	Any file system operation	Immutable	Any file system operation
Object tagging	Y	N	N	N
Object metadata	Y	N	N	N
Pricing model	Storage + requests + data transfer	Included in Lambda	Included in Lambda	Storage + data transfer + throughput
Sharing/permissions model	IAM	Function-only	IAM	IAM + NFS
Source for AWS Glue	Y	N	N	N
Source for Amazon QuickSight	Y	N	N	N
Relative data access speed from Lambda	Fast	Fastest	Fastest	Very fast

Source: aws.amazon.com/blogs/compute/choosing-between-aws-lambda-data-storage-options-in-web-apps/

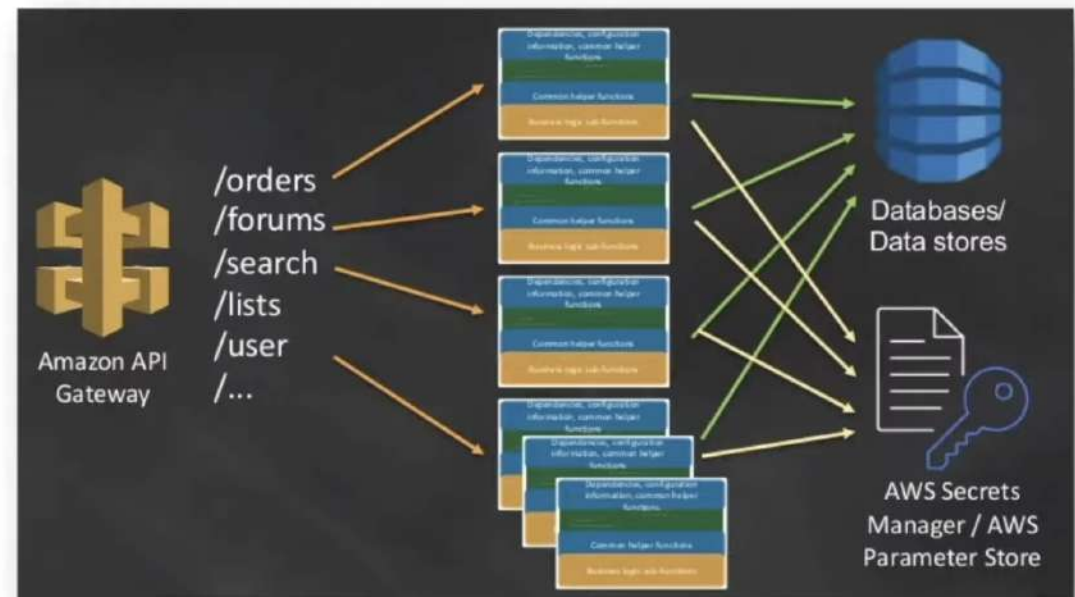
Our Data architecture (not all yet completed)



* Not all target services will be used

Lambda layers

- convenient way to package libraries and other dependencies
- reduces the size of uploaded deployment, so it's faster to deploy your code
- promote code sharing and separation of responsibilities
 - Simpler iterations, so faster on writing business logic.



Press **esc** to exit full screen

Steps

- Try changing the data type in the data catalog!
- Error? See next slide

Query 1 × Query 2 × Query 3 × Query 4 × Query 5 × Query 6 × Query 7 × **Query 8 ×**

```
1 SELECT ref.snippet_title, stats.*
2 FROM raw_statistics stats
3 INNER JOIN cleansed_statistics_reference_data ref on (stats.category_id = CAST(ref.id as INT))
4 limit 10;
```

SQL Ln 4, Col 10

Run again Cancel Save as Clear Create ▼

✔ Completed Time in queue: 0.143

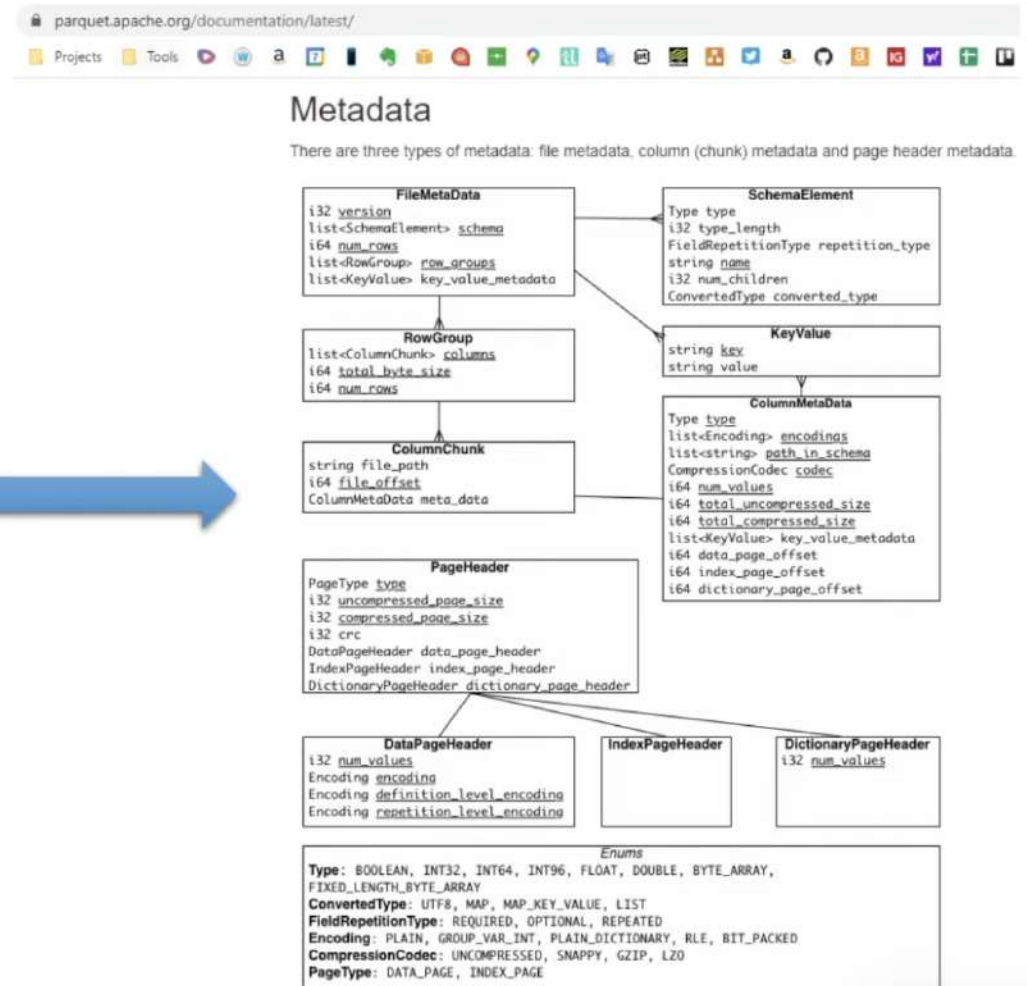
Results (10)

Search rows

snippet_title ▼	video_id ▼	trending_date ▼	title ▼
People & Blogs	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT O...
News & Politics	ilxy3JN3-jc	17.14.11	LeBron James admits he was ri...
Sports	NZFhMSgbKKM	17.14.11	Dennis Smith Jr. and LeBron Ja...

Apache Parquet

Apache Parquet files have headers!



Steps

1. Keep the data type change in the data catalogue
2. Delete our testing JSON file
3. Confirm APPEND in Lambda
4. Run Test event in Lambda
5. Copy again our data, from our laptops (AWSCLI)
6. Add the S3 Trigger to Lambda

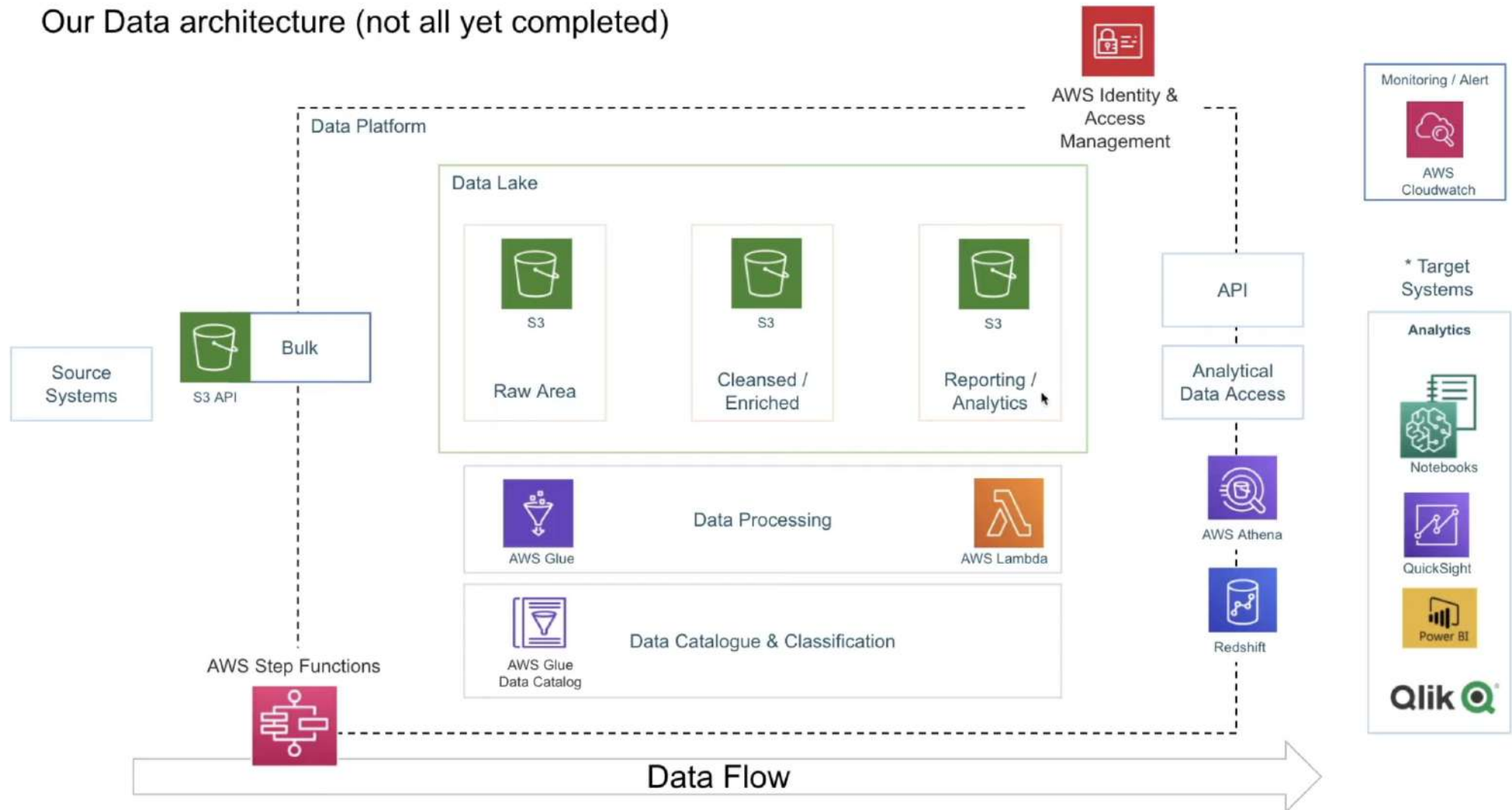
The screenshot shows a SQL query interface with a tab for 'Query 8'. The query is as follows:

```
1 SELECT ref.snippet_title, stats.*
2 FROM raw_statistics stats
3 INNER JOIN cleansed_statistics_reference_data ref on (stats.category_id = CAST(ref.id as INT))
4 limit 10;
```

The interface indicates the query is 'Completed' with a 'Time in queue: 0.143'. Below the query, there are buttons for 'Run again', 'Cancel', 'Save as', 'Clear', and 'Create'. The results are displayed in a table with 10 rows, showing columns: snippet_title, video_id, trending_date, and title.

snippet_title	video_id	trending_date	title
People & Blogs	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT O...
News & Politics	ilxy3JN3-jc	17.14.11	LeBron James admits he was ri...
Sports	NZFhM5gbKKM	17.14.11	Dennis Smith Jr. and LeBron Ja...

Our Data architecture (not all yet completed)



* Not all target services will be used

Cleansed vs Analytics layer

Using Cleansed Layer



```
1 SELECT ref.snippet_title, stats.title, stats.title
2 FROM raw_statistics stats
3     INNER JOIN cleansed_statistics_reference_data ref on (stats.category_id = ref.id)
4 WHERE ref.id=2
```

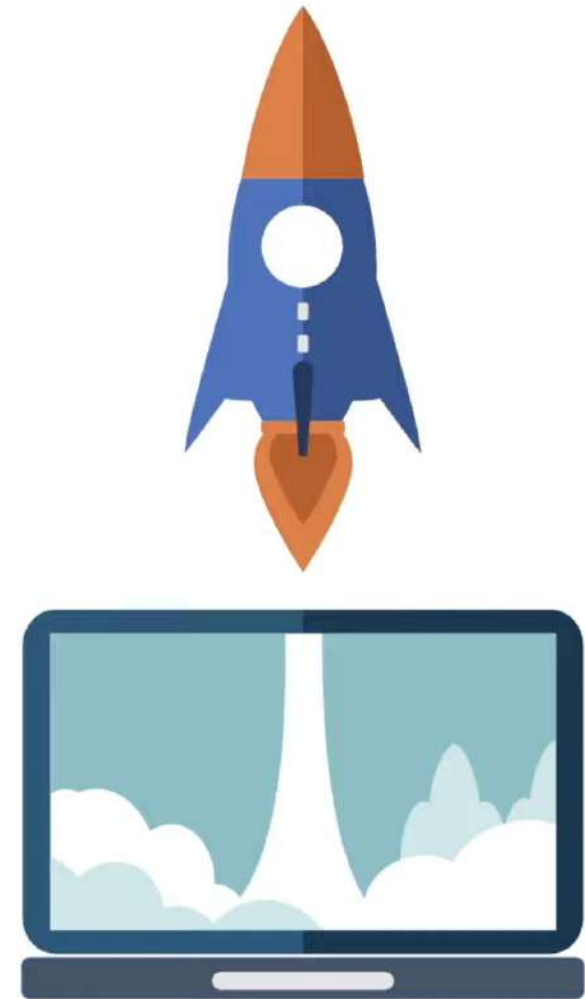
Using Reporting Layer



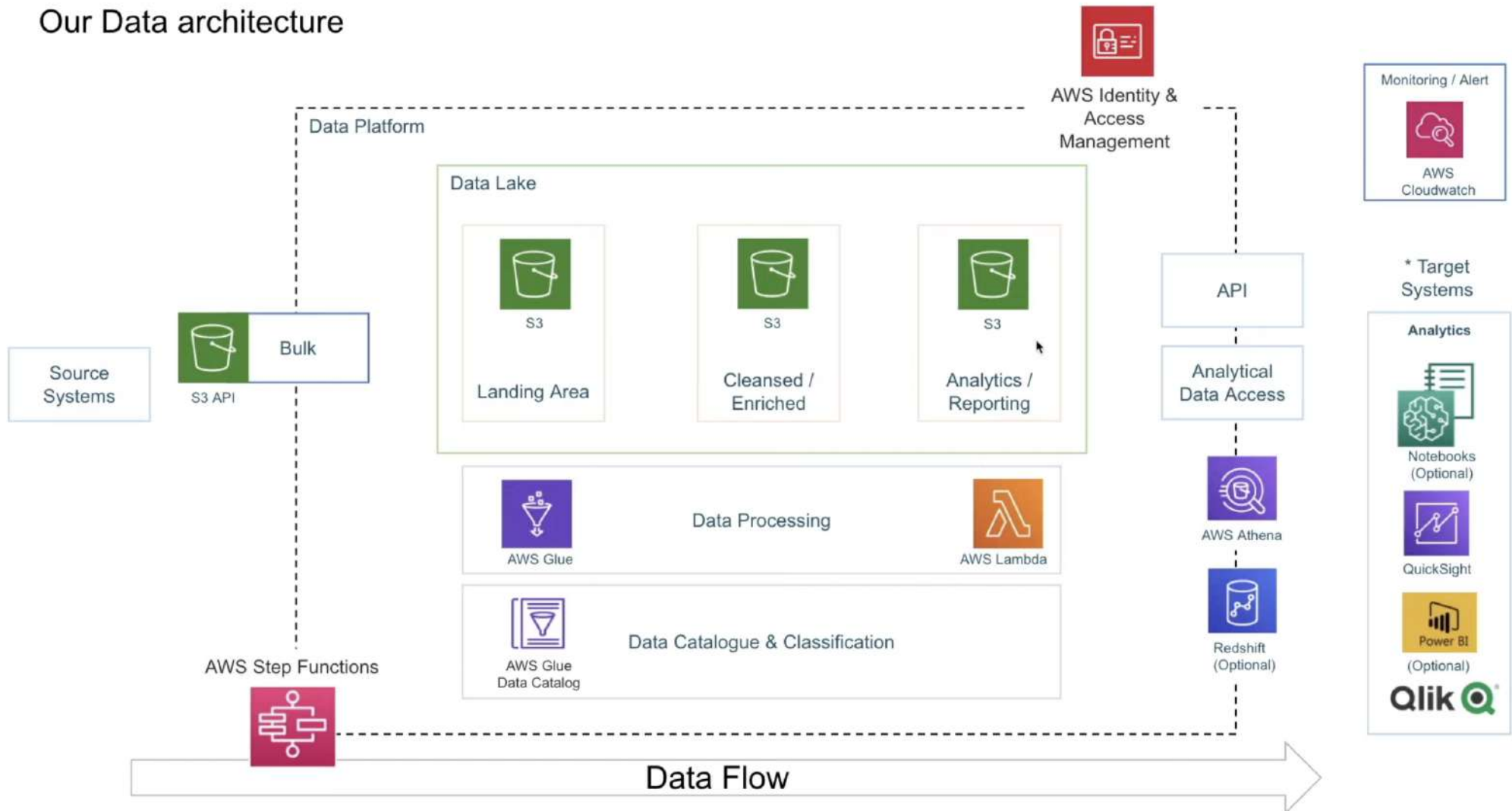
```
1 SELECT snippet_title, title, title
2 FROM rpt_youtube_statistics_categories
3 WHERE id=2
```

Agenda

- Business Intelligence, using Amazon QuickSight



Our Data architecture



* Not all target services will be used