Serverless Data Processing (CSCI 5410)

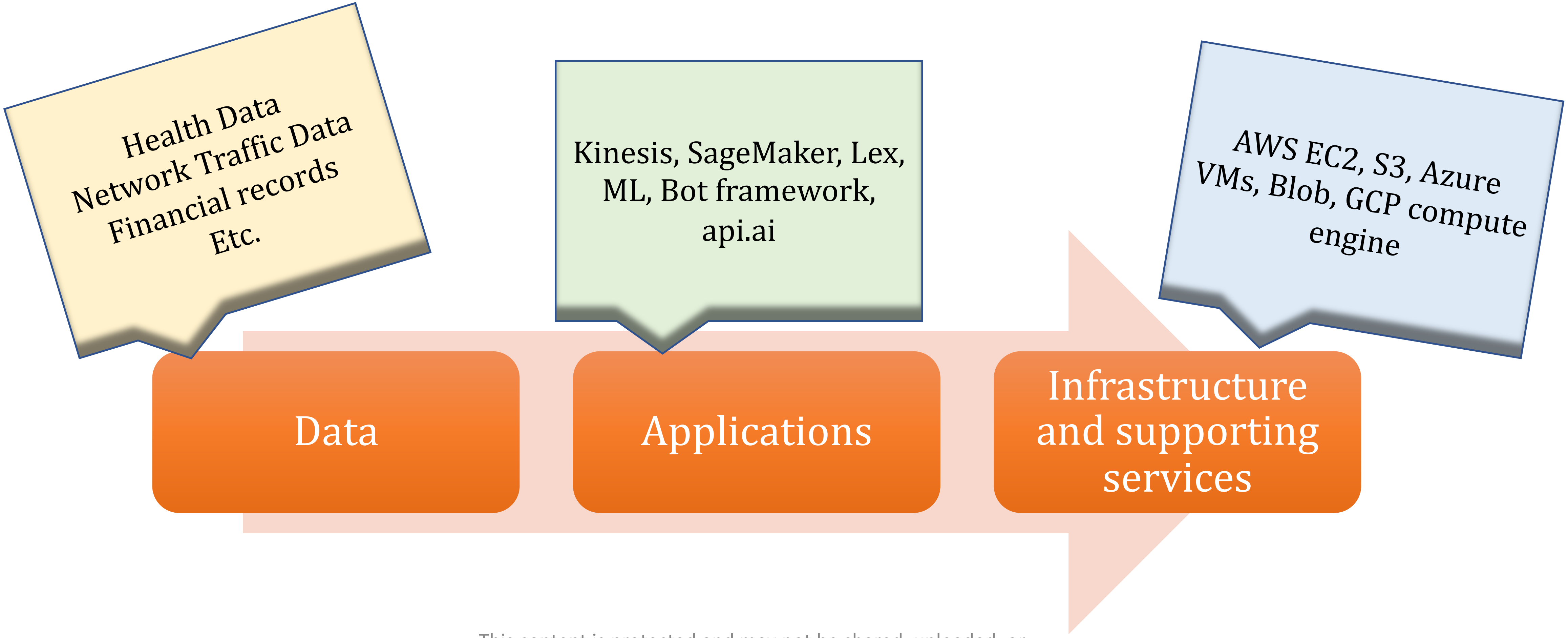
Dr. Saurabh Dey

Images and contents used in these slides are from web search and/or text books, and are used for academic purposes only

Outline

1. Data Processing in Cloud
2. ETL in Cloud

Data Processing in Cloud



Health Data
Network Traffic Data
Financial records
Etc.

Kinesis, SageMaker, Lex, ML, Bot framework, api.ai

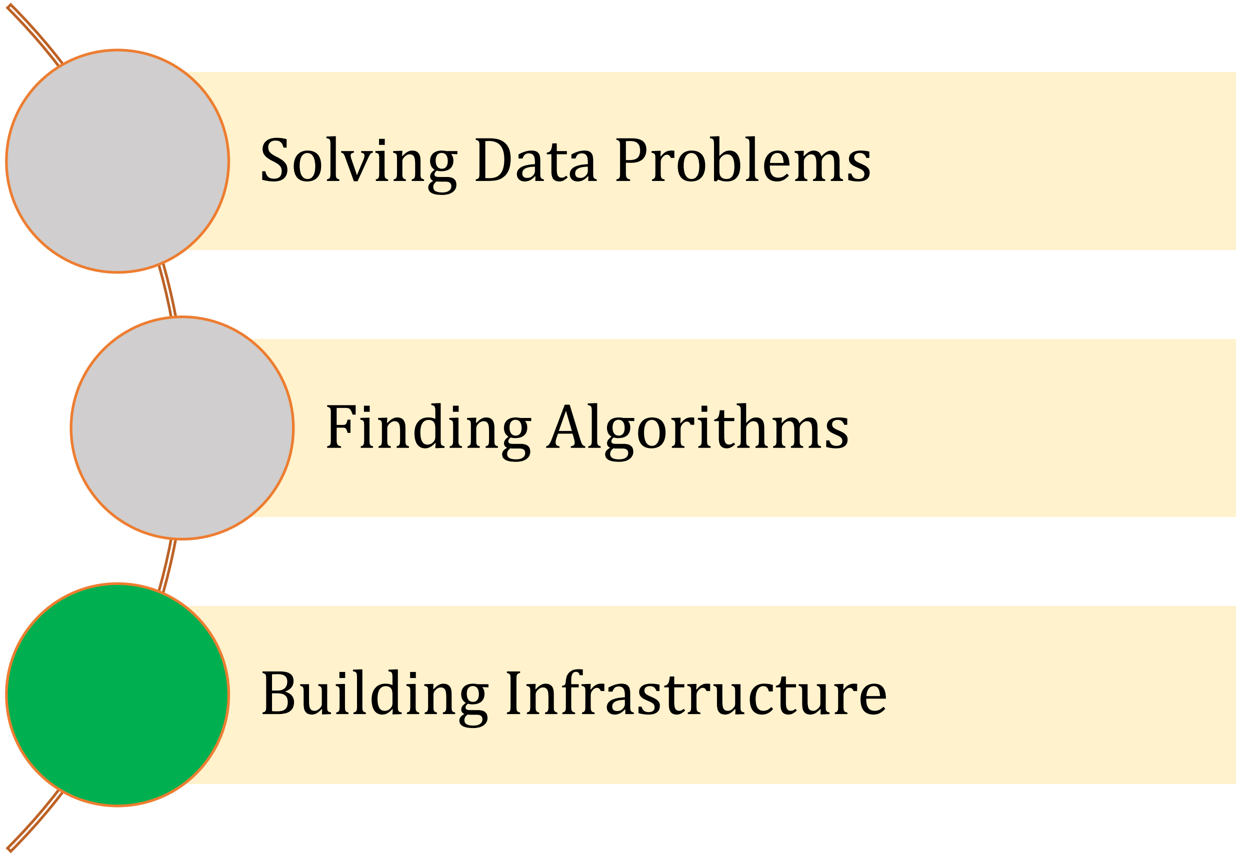
AWS EC2, S3, Azure VMs, Blob, GCP compute engine

Data

Applications

Infrastructure and supporting services

Challenges in Large Scale Data Processing



Solving Data Problems

Finding Algorithms

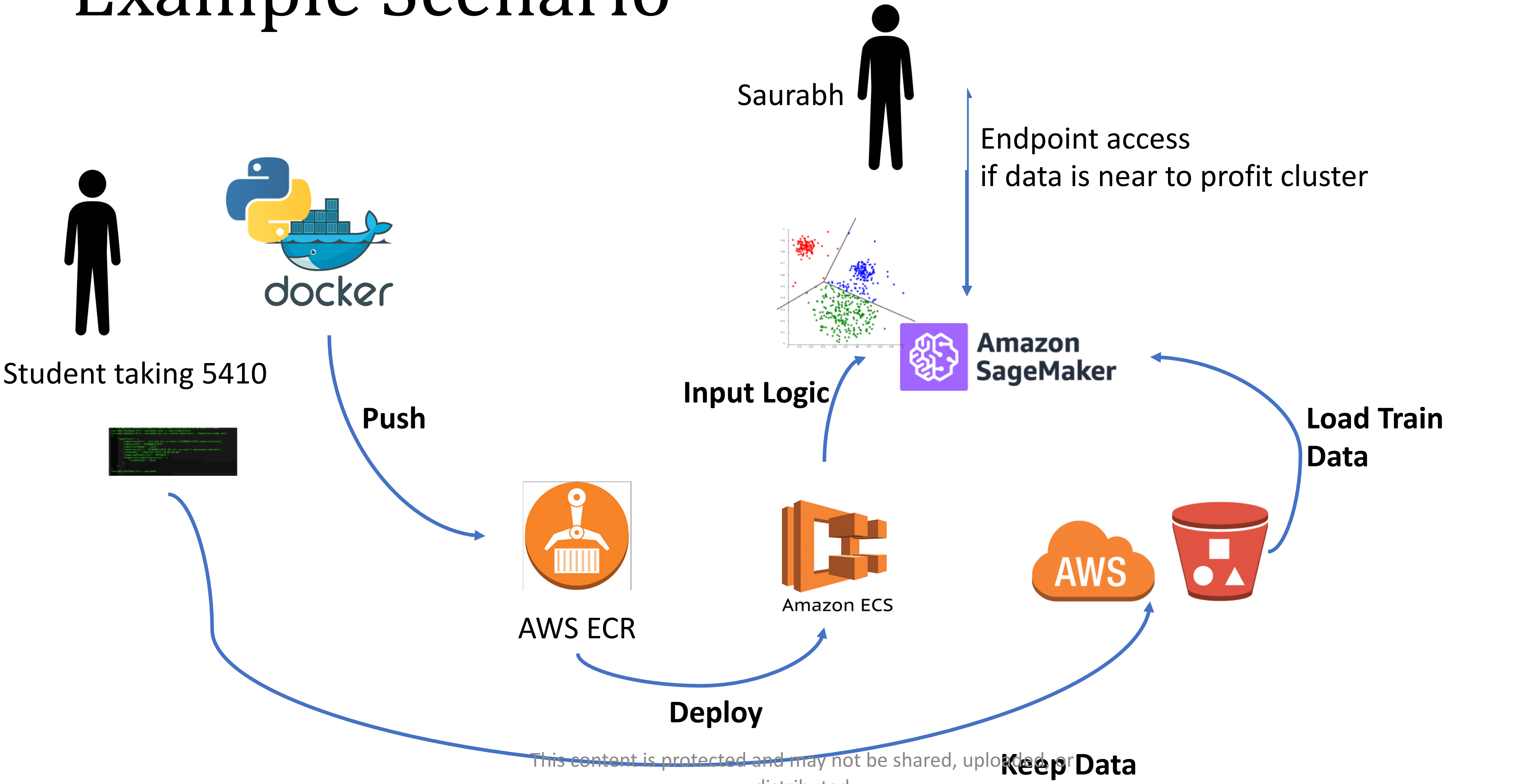
Building Infrastructure

Issues in Building Infrastructure

- **Robust and Uptime**
  - Can we guarantee that the system will not fail?
- **Connected Services**
  - Can we guarantee that we addressed correctness, freshness, and latency related problems for connected services?
- **Data Security**
  - Can we guarantee that there will be no data loss, privacy issues?
- **Cost Estimation**
  - Can we guarantee that the infrastructure cost will be less?

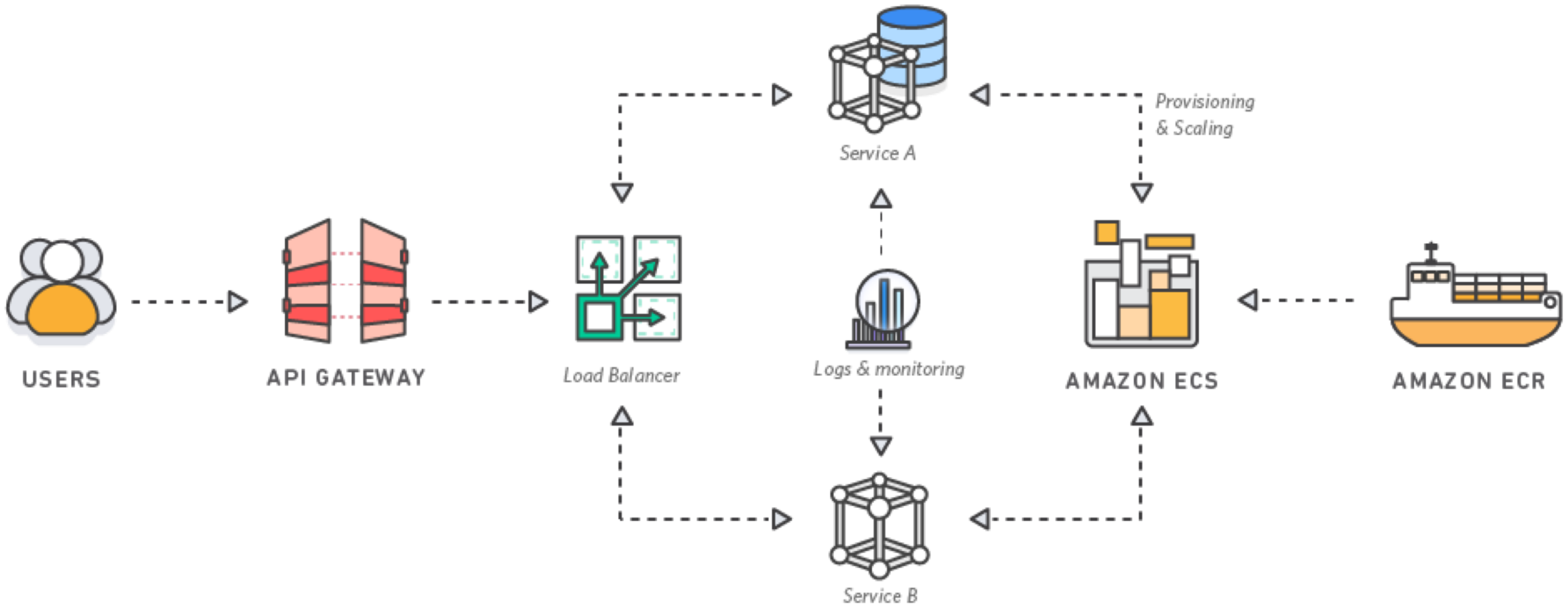
Example Scenario

If higher return
on this investment (quickmoney deposit scheme)



Container based Infrastructure

- Easy to identify problem areas
- Easy to modify and perform incremental update
- Balances the load

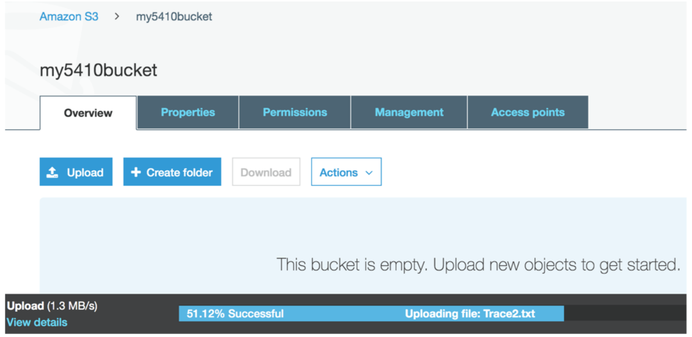


https://aws.amazon.com/ecs/getting-started/

Cloud Data Analytics

- Numerous services provide support for analytics.
- AWS Athena, GCP Data fusion are some of the popular tools.
- AWS Athena interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL
    - Connect data in S3. Define the schema, and start querying using the built-in query editor.

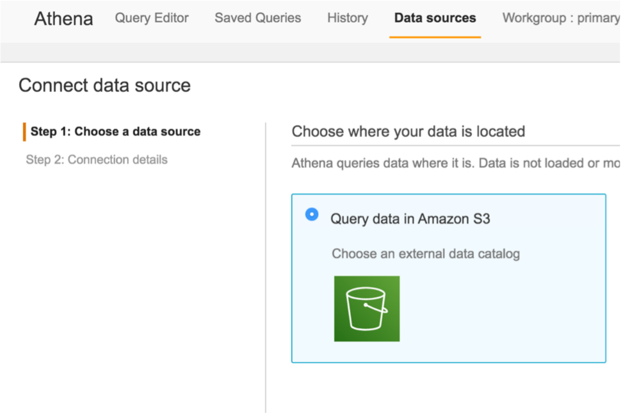
https://aws.amazon.com/athena/?c=a&sec=srv



This bucket is empty. Upload new objects to get started.

51.12% Successful

Uploading file: Trace2.txt



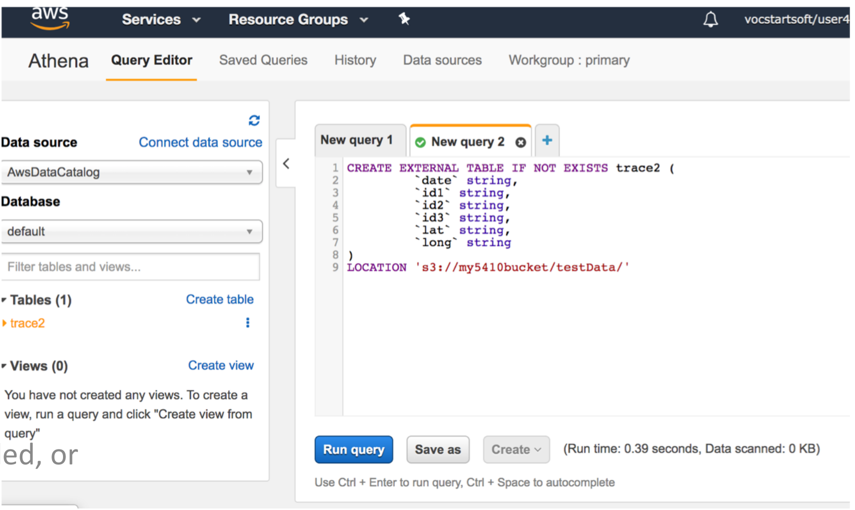
Connect data source

Choose where your data is located

Athena queries data where it is. Data is not loaded or mo

● Query data in Amazon S3

Choose an external data catalog



Data source: AwsDataCatalog

Database: default

You have not created any views. To create a view, run a query and click "Create view from query"

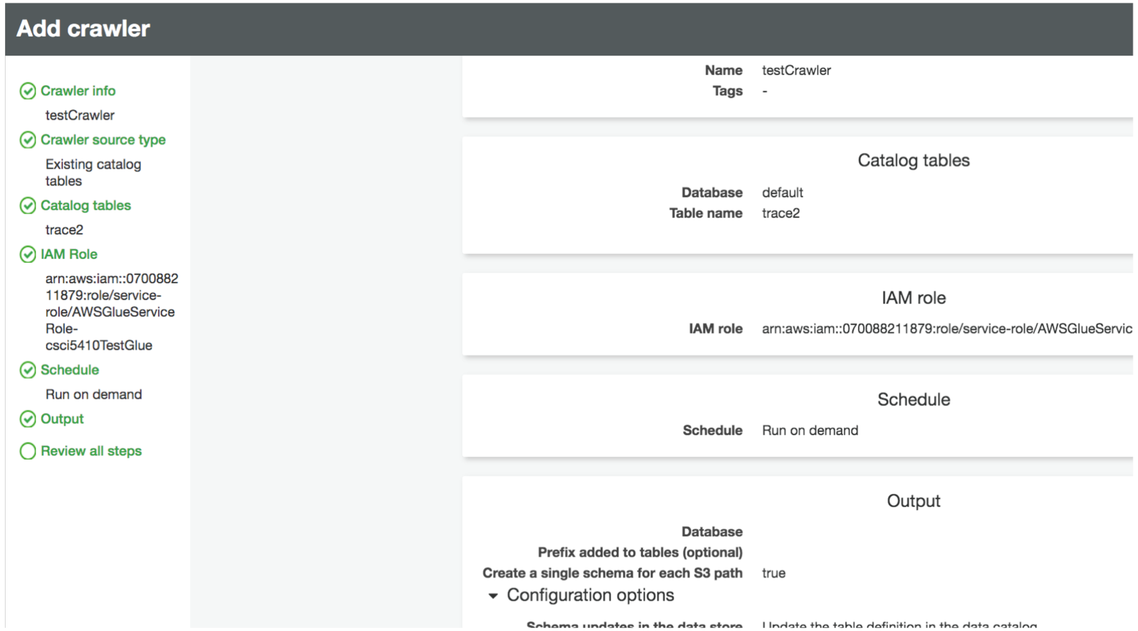
```
CREATE EXTERNAL TABLE IF NOT EXISTS trace2 (
        `date` string,
        `id1` string,
        `id2` string,
        `id3` string,
        `lat` string,
        `long` string
)
LOCATION 's3://my5410bucket/testData/'
```

(Run time: 0.39 seconds, Data scanned: 0 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Cloud ETL

E.g. AWS Glue



Add crawler

- ✓ Crawler info
  testCrawler
- ✓ Crawler source type
  Existing catalog tables
- ✓ Catalog tables
  trace2
- ✓ IAM Role
  arn:aws:iam::070088211879:role/service-role/AWSGlueServiceRole-csci5410TestGlue
- ✓ Schedule
  Run on demand
- ✓ Output
- ○ Review all steps

Name testCrawler
Tags -

Catalog tables

Database default
Table name trace2

IAM role

IAM role arn:aws:iam::070088211879:role/service-role/AWSGlueServic

Schedule

Schedule Run on demand

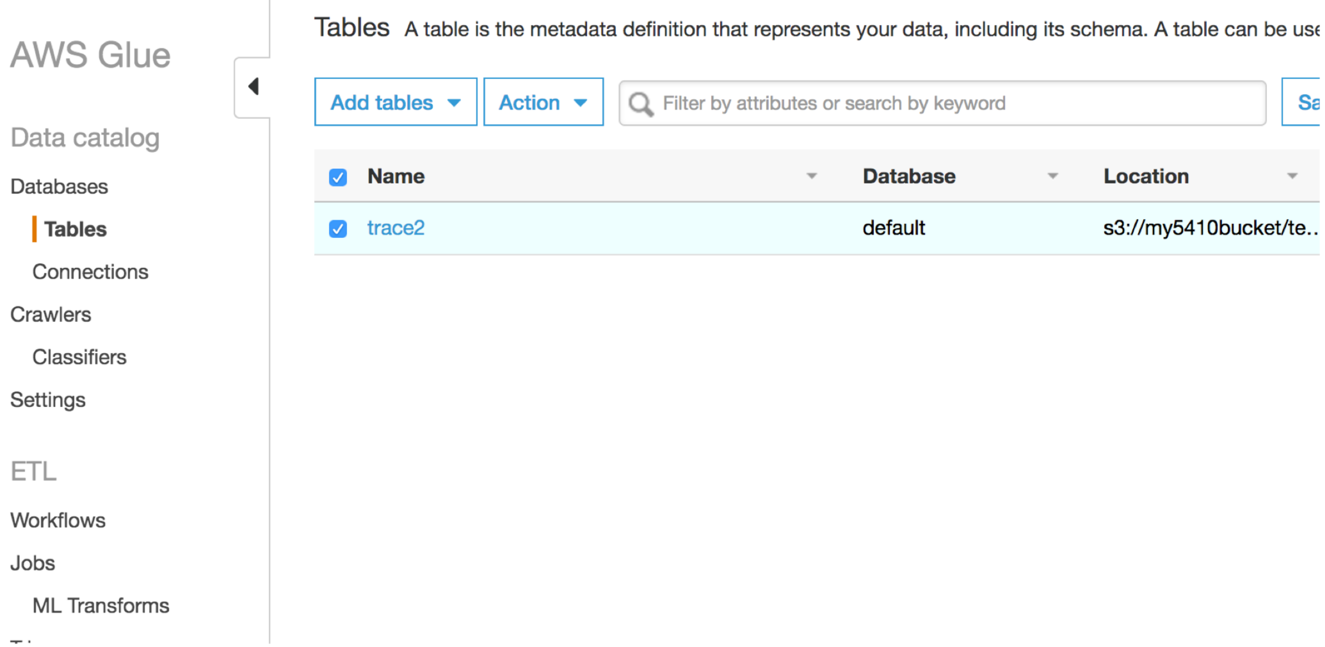
Output

Database
Prefix added to tables (optional)
Create a single schema for each S3 path true

▾ Configuration options

Schema updates in the data store Update the table definition in the data catalog

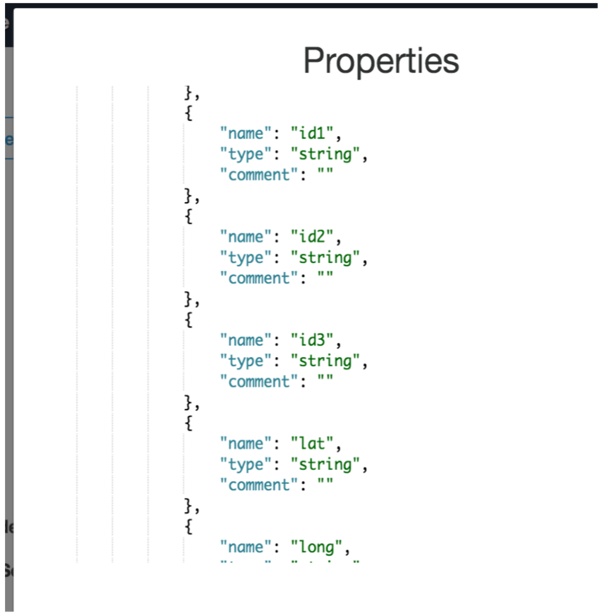
- AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics.
- After connecting stored data, discovering the data - Glue builds a catalogue with metadata
- AWS Glue automates much of the effort in building, maintaining, and running ETL jobs. AWS Glue crawls the data sources, identifies data formats, and suggests schemas and transformations.



Tables A table is the metadata definition that represents your data, including its schema. A table can be use

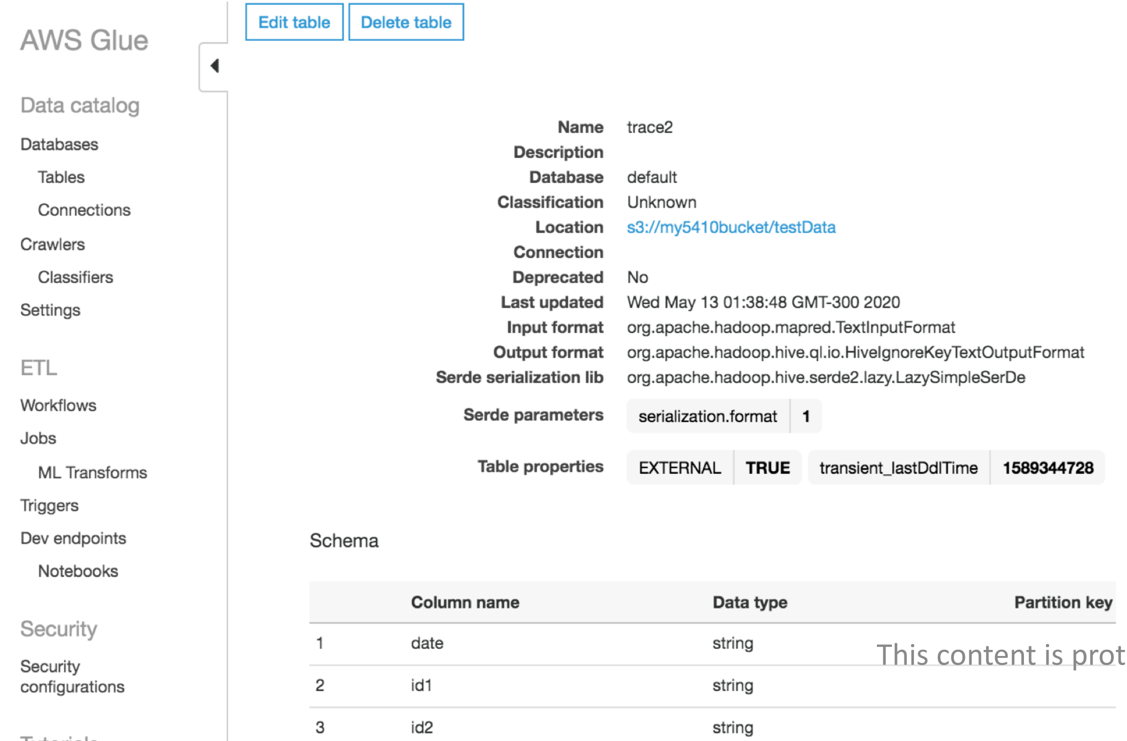
Filter by attributes or search by keyword

☑	Name	Database	Location
☑	trace2	default	s3://my5410bucket/te...



Properties

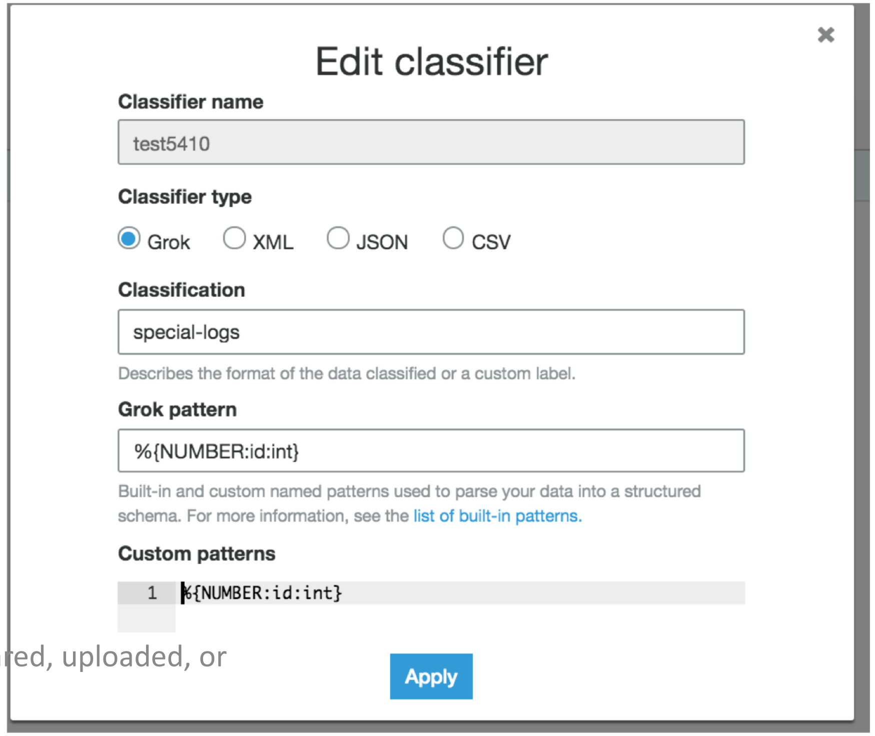
```
                    },
                    {
                        "name": "id1",
                        "type": "string",
                        "comment": ""
                    },
                    {
                        "name": "id2",
                        "type": "string",
                        "comment": ""
                    },
                    {
                        "name": "id3",
                        "type": "string",
                        "comment": ""
                    },
                    {
                        "name": "lat",
                        "type": "string",
                        "comment": ""
                    },
                    {
                        "name": "long",
```



Name trace2
Description
Database default
Classification Unknown
Location s3://my5410bucket/testData
Connection
Deprecated No
Last updated Wed May 13 01:38:48 GMT-300 2020
Input format org.apache.hadoop.mapred.TextInputFormat
Output format org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters serialization.format 1
Table properties EXTERNAL TRUE transient_lastDdlTime 1589344728

Schema

	Column name	Data type	Partition key
1	date	string	
2	id1	string	
3	id2	string	



Edit classifier

Classifier name
test5410

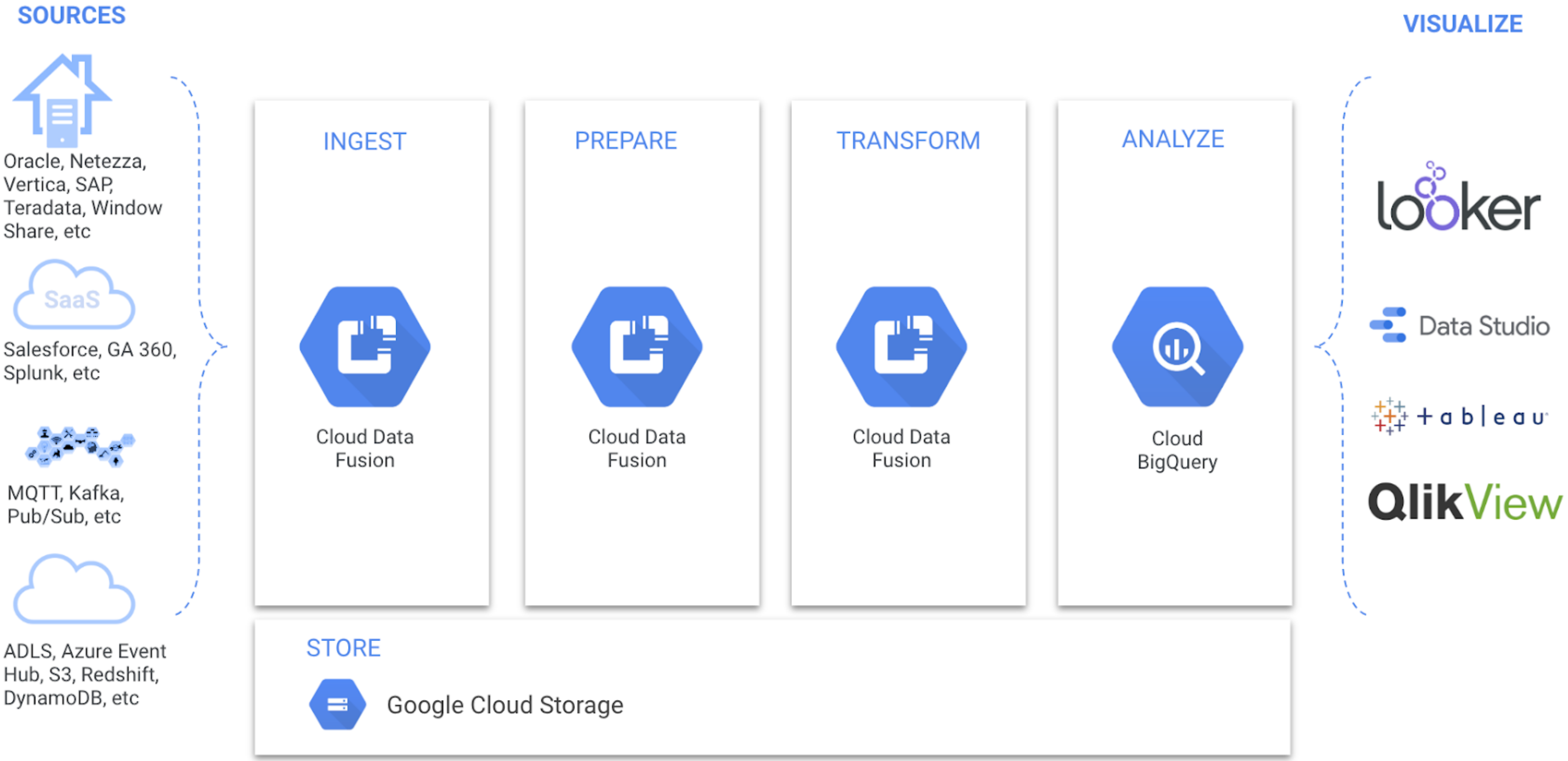
Classifier type
● Grok ○ XML ○ JSON ○ CSV

Classification
special-logs
Describes the format of the data classified or a custom label.

Grok pattern
%{NUMBER:id:int}
Built-in and custom named patterns used to parse your data into a structured schema. For more information, see the list of built-in patterns.

Custom patterns
```
%{NUMBER:id:int}
```

E.g. GCP Data Fusion



SOURCES

Oracle, Netezza, Vertica, SAP, Teradata, Window Share, etc

Salesforce, GA 360, Splunk, etc

MQTT, Kafka, Pub/Sub, etc

ADLS, Azure Event Hub, S3, Redshift, DynamoDB, etc

INGEST

Cloud Data Fusion

PREPARE

Cloud Data Fusion

TRANSFORM

Cloud Data Fusion

ANALYZE

Cloud BigQuery

STORE

Google Cloud Storage

VISUALIZE

looker

Data Studio

tableau

QlikView

Questions to Consider

- How to use AWS Glue to migrate data from Google BigQuery to AWS S3?
- What is AWS Glue equivalent in Azure, and how does it work?