CUSTOMER DATA INGESTION

TABLE OF CONTENTS

01 Data Storage

02 Data Security

03 Data Ingestion

04 Data Processing

05 Data Scalability

06 Data Analysis

INTRODUCTION

This is a data management project that involves collecting, processing, and storing data from various sources in a centralized location for further analysis.
The Goal of this project is to make data readily available for analysis by extracting data from variety of sources and integrating it into a single repository, such as a data warehouse

01

DATA STORAGE

STORAGE USED

AWS S3

AWS S3 (Amazon Simple Storage Service) is a a highly scalable ,secure, and durable cloud storage service.

MongoDB

A popular document-oriented NoSQL database that allows for flexible and scalable data management

DATA SECURITY

02

SECURITY OF DATA

VPC

Virtual Private Cloud is a customizable and secure virtual network within AWS that allows users to launch AWS resources in a defined virtual network environment which is isolated

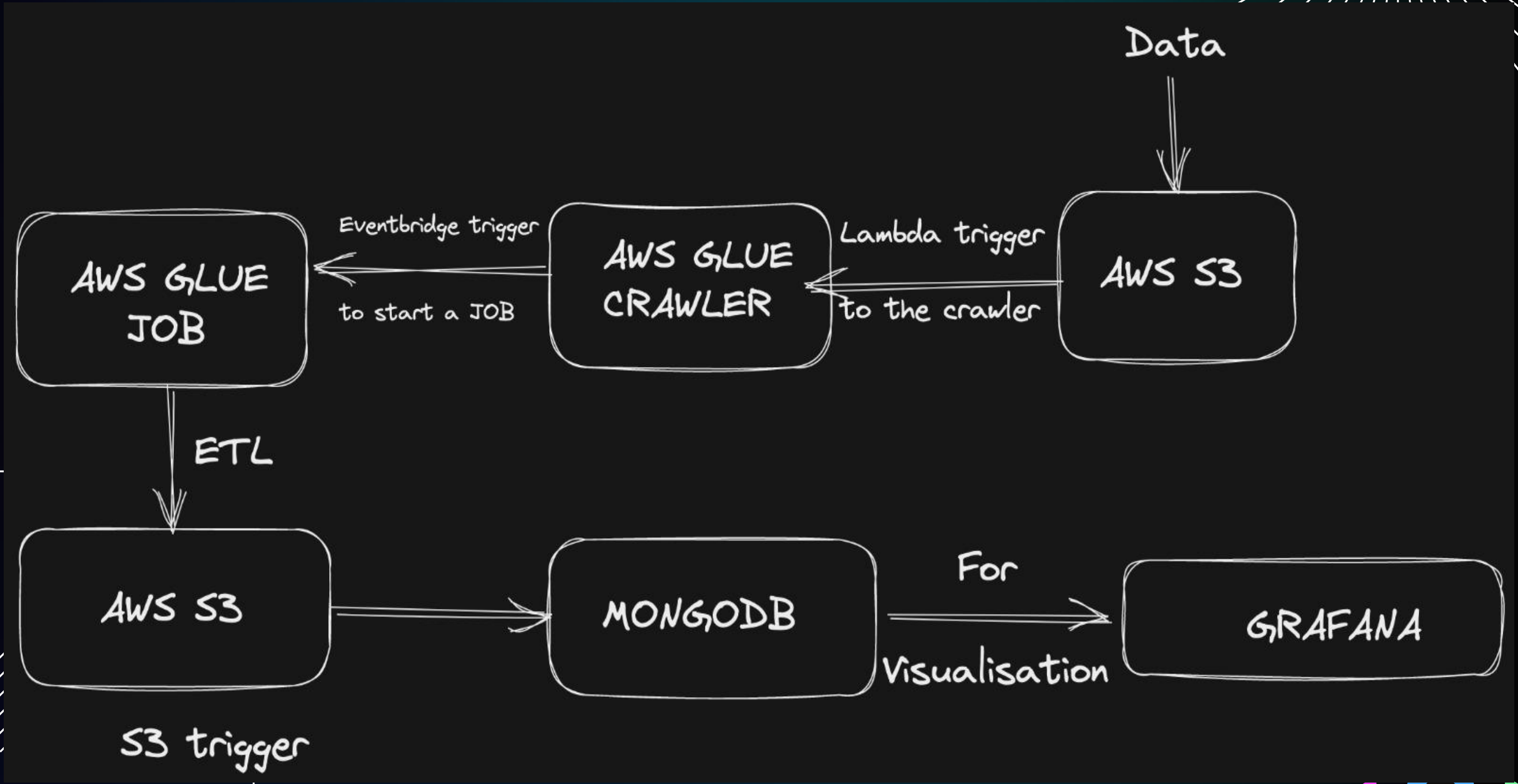
Authenticated Users

Authenticated users are individuals or entities who have successfully proven their identity and been granted access to a system or application with the appropriate permissions.

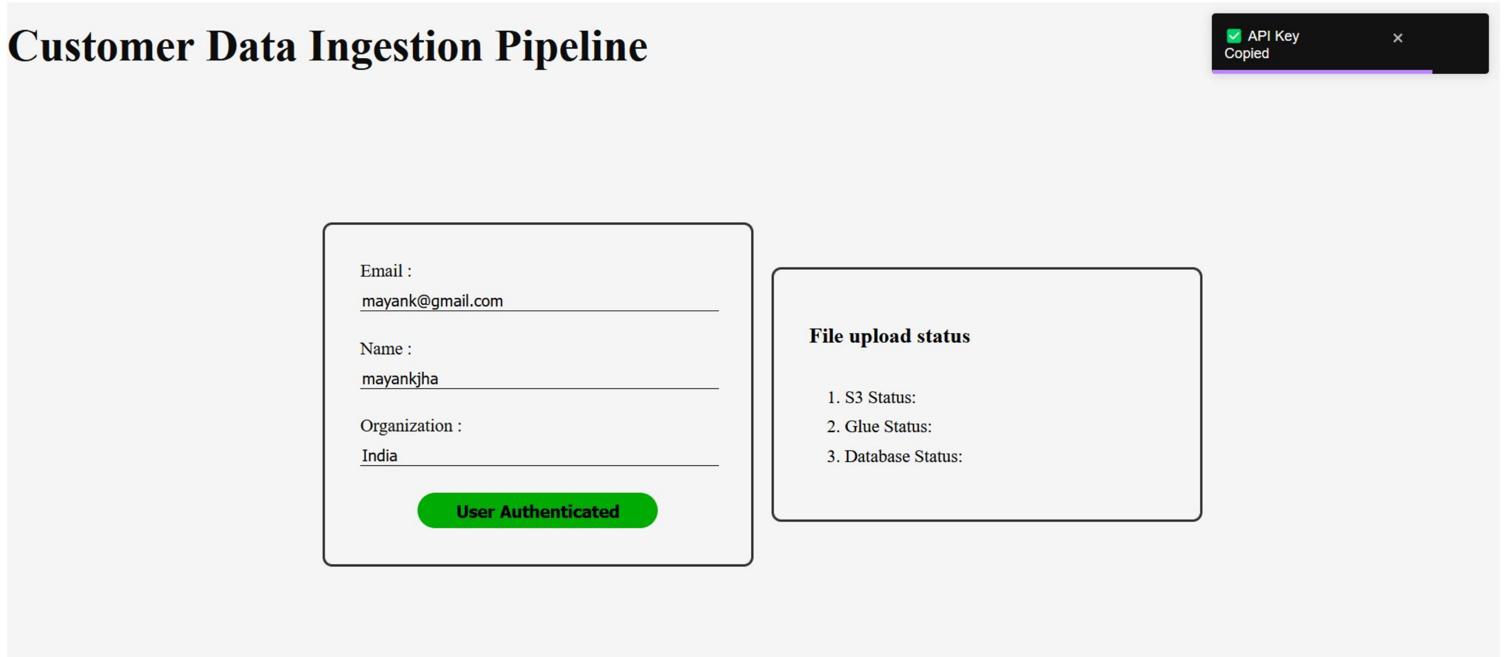
03

DATA INGESTION

INGESTION FLOW



GET THE API KEY



Customer Data Ingestion Pipeline

✅ API Key Copied ×

Email : mayank@gmail.com

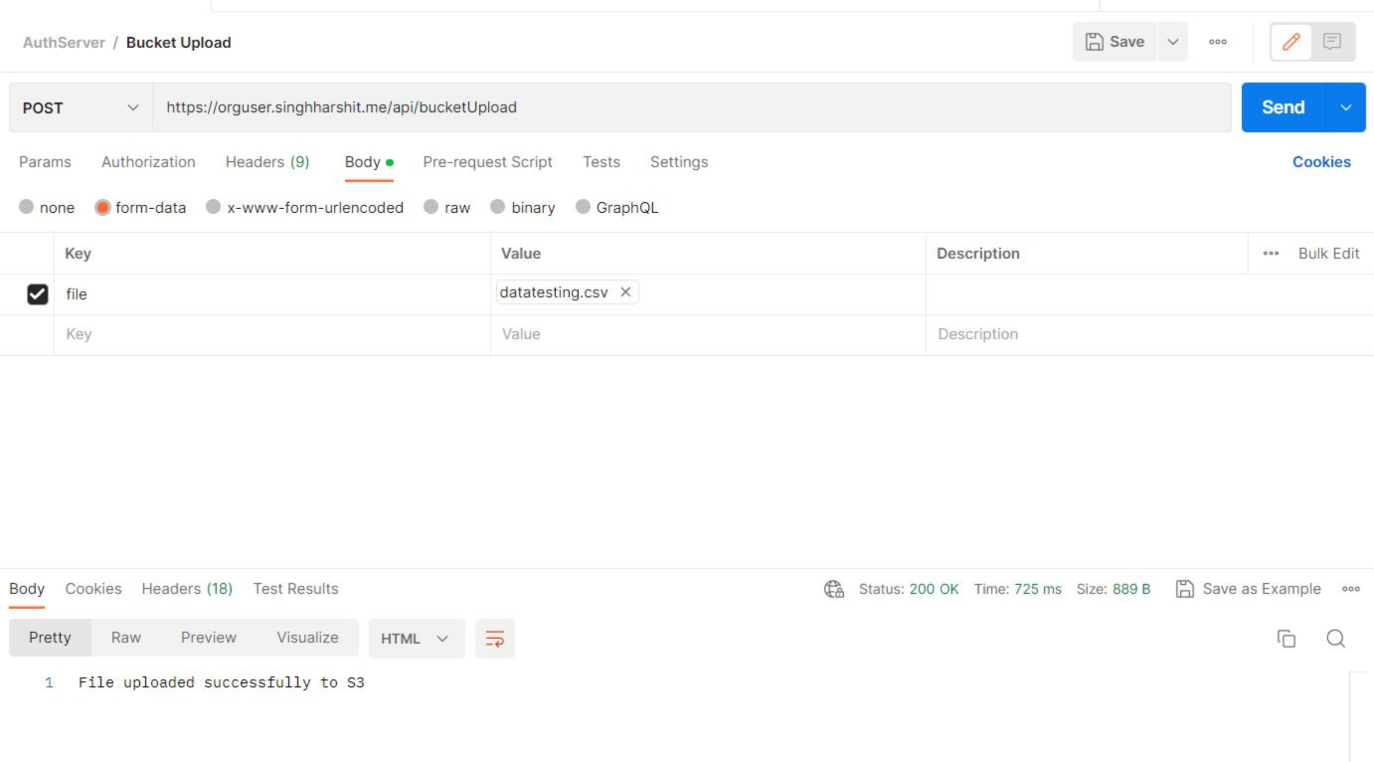
Name : mayankjha

Organization : India

File upload status

1. S3 Status:
2. Glue Status:
3. Database Status:

UPLOAD FILE FROM POSTMAN



AuthServer / Bucket Upload

POST https://orguser.singhharshit.me/api/bucketUpload

Body: ● form-data

	Key	Value	Description
☑	file	datatesting.csv	

Status: 200 OK Time: 725 ms Size: 889 B

```
File uploaded successfully to S3
```

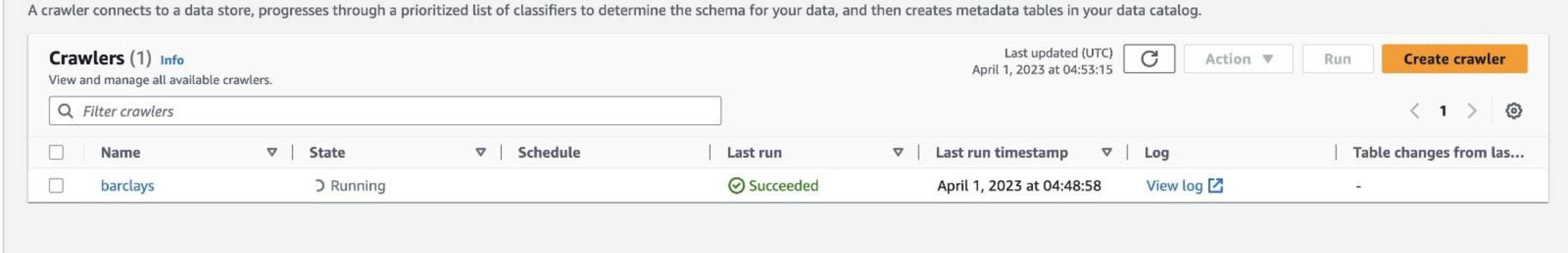
04

DATA PROCESSING

APACHE SPARK AS ETL FRAMEWORK

Apache Spark is a powerful ETL (Extract, Transform, Load) framework that can efficiently process large volumes of data from various sources and transform it into a desired format for analysis and reporting.

CRAWLER RUNS



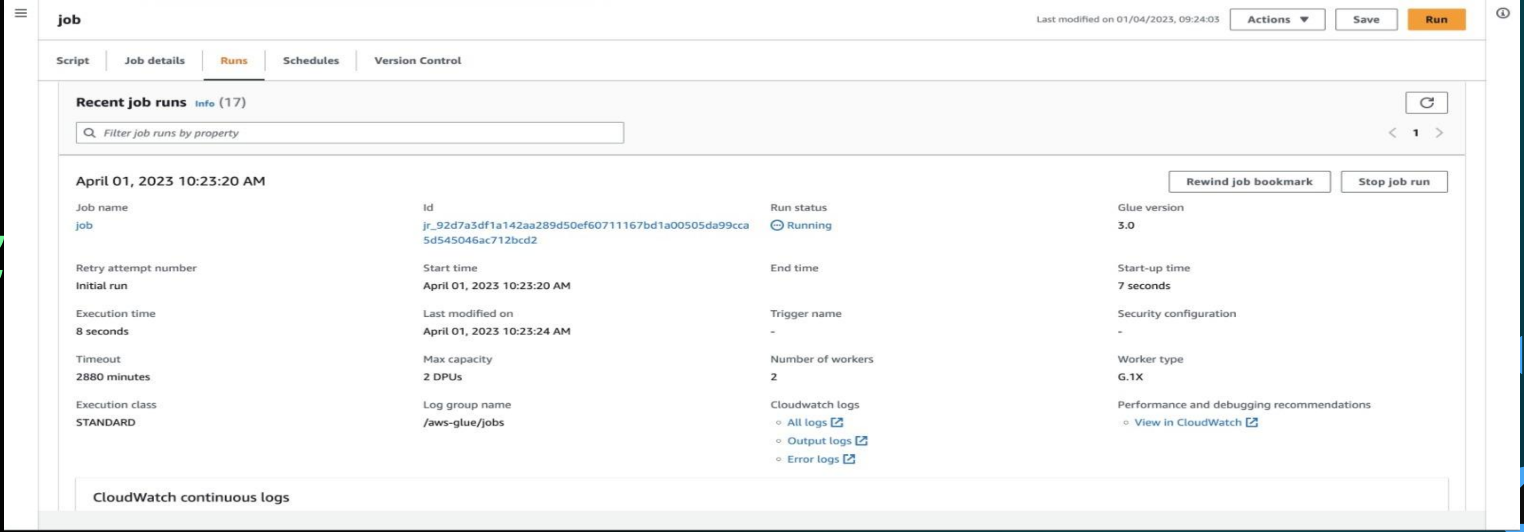
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (1) Info

View and manage all available crawlers.

Last updated (UTC)
April 1, 2023 at 04:53:15

□	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from las...
□	barclays	Running		✓ Succeeded	April 1, 2023 at 04:48:58	View log	-



Recent job runs Info (17)

April 01, 2023 10:23:20 AM

Job name
job

Id
jr_92d7a3df1a142aa289d50ef60711167bd1a00505da99cca5d545046ac712bcd2

Run status
Running

Glue version
3.0

Retry attempt number
Initial run

Start time
April 01, 2023 10:23:20 AM

End time

Start-up time
7 seconds

Execution time
8 seconds

Last modified on
April 01, 2023 10:23:24 AM

Trigger name
-

Security configuration
-

Timeout
2880 minutes

Max capacity
2 DPUs

Number of workers
2

Worker type
G.1X

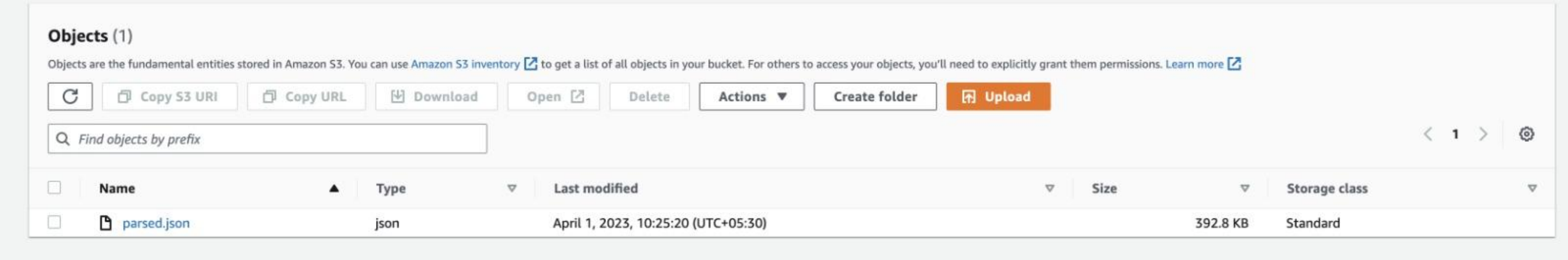
Execution class
STANDARD

Log group name
/aws-glue/jobs

Cloudwatch logs
- All logs
- Output logs
- Error logs

Performance and debugging recommendations
- View in CloudWatch

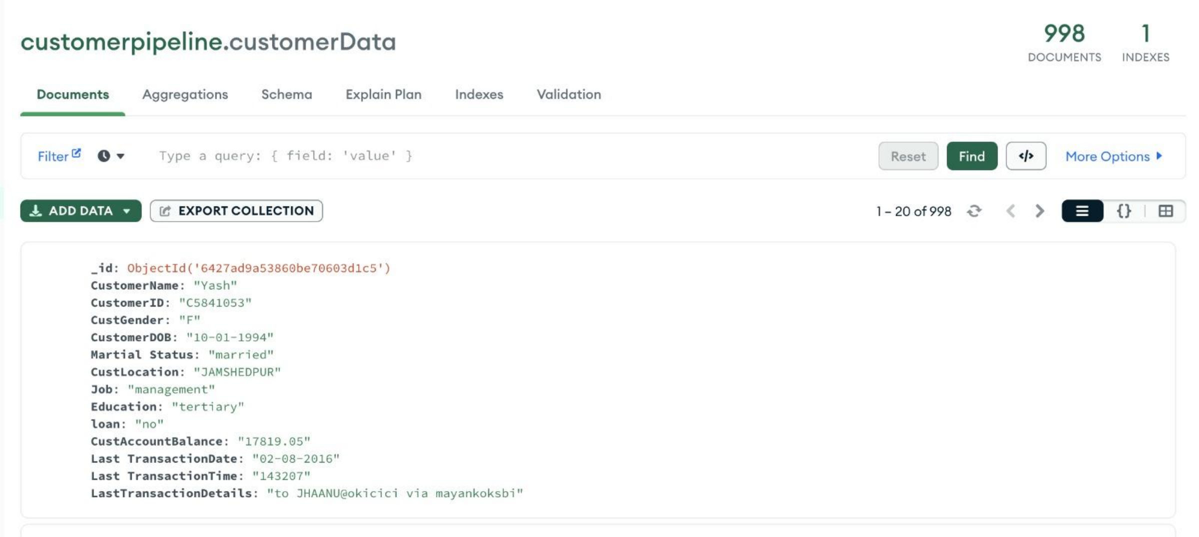
CloudWatch continuous logs



Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	parsed.json	json	April 1, 2023, 10:25:20 (UTC+05:30)	392.8 KB	Standard



customerpipeline.customerData

998 DOCUMENTS 1 INDEXES

1 – 20 of 998

```
_id: ObjectId('6427ad9a53860be70603d1c5')
CustomerName: "Yash"
CustomerID: "C5841053"
CustGender: "F"
CustomerDOB: "10-01-1994"
Martial Status: "married"
CustLocation: "JAMSHEDPUR"
Job: "management"
Education: "tertiary"
loan: "no"
CustAccountBalance: "17819.05"
Last TransactionDate: "02-08-2016"
Last TransactionTime: "143207"
LastTransactionDetails: "to JHAANU@okicici via mayankoksbi"
```

DATA SCALABILITY

05

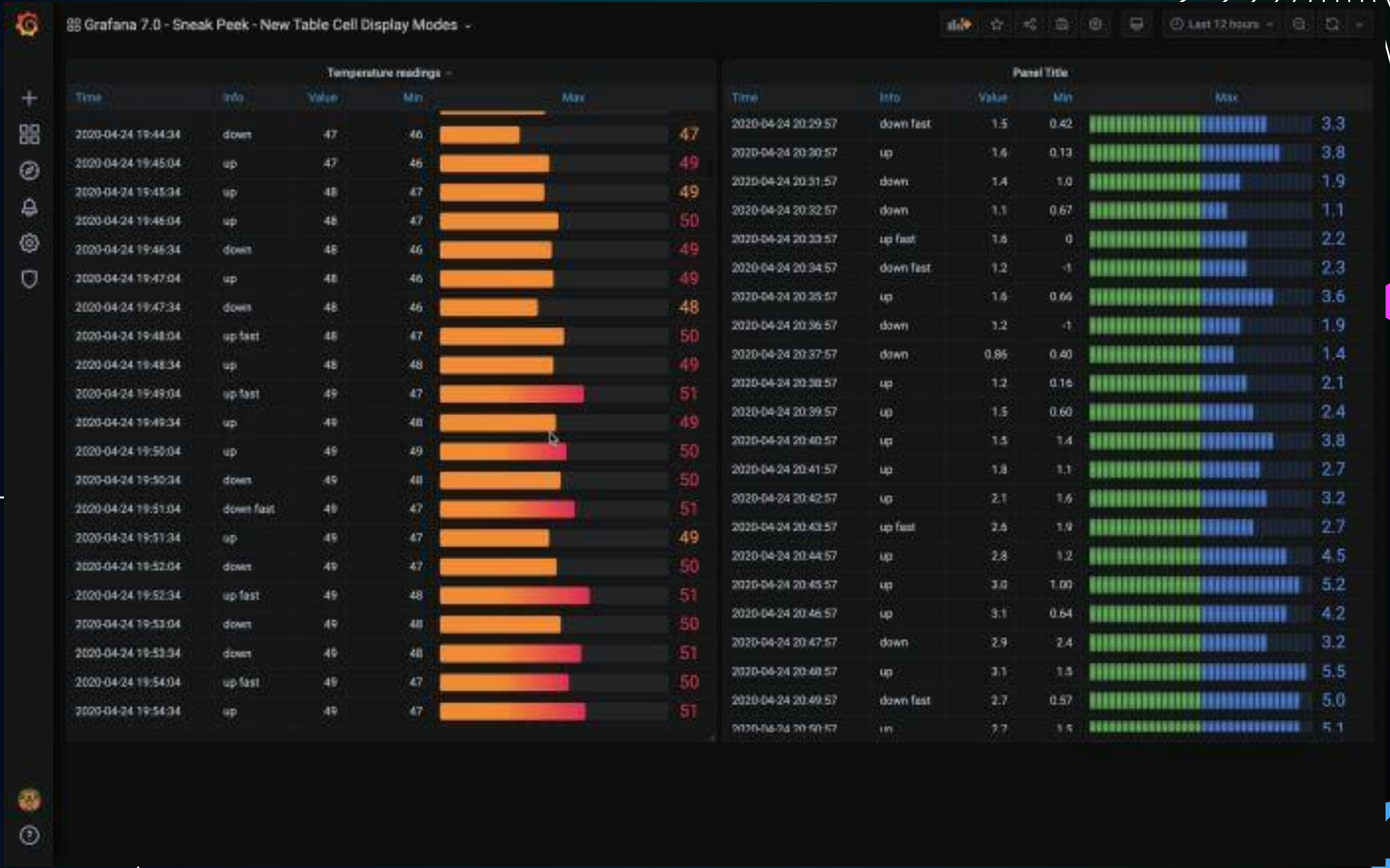
PROPERTIES FOR SCALABILITY

- Serverless
- Event Driven
- Big Data Processing
- Resilient

06

DATA ANALYSIS

GRAFANA TO BE IMPLEMENTED



Temperature readings

Time	Info	Value	Min	Max
2020-04-24 19:44:34	down	47	46	47
2020-04-24 19:45:04	up	47	46	49
2020-04-24 19:45:34	up	48	47	49
2020-04-24 19:46:04	up	48	47	50
2020-04-24 19:46:34	down	48	46	49
2020-04-24 19:47:04	up	48	46	49
2020-04-24 19:47:34	down	48	46	48
2020-04-24 19:48:04	up fast	48	47	50
2020-04-24 19:48:34	up	48	48	49
2020-04-24 19:49:04	up fast	49	47	51
2020-04-24 19:49:34	up	49	48	49
2020-04-24 19:50:04	up	49	49	50
2020-04-24 19:50:34	down	49	48	50
2020-04-24 19:51:04	down fast	49	47	51
2020-04-24 19:51:34	up	49	47	49
2020-04-24 19:52:04	down	49	47	50
2020-04-24 19:52:34	up fast	49	48	51
2020-04-24 19:53:04	down	49	48	50
2020-04-24 19:53:34	down	49	48	51
2020-04-24 19:54:04	up fast	49	47	50
2020-04-24 19:54:34	up	49	47	51

Panel Title

Time	Info	Value	Min	Max
2020-04-24 20:29:57	down fast	1.5	0.42	3.3
2020-04-24 20:30:57	up	1.6	0.13	3.8
2020-04-24 20:31:57	down	1.4	1.0	1.9
2020-04-24 20:32:57	down	1.1	0.67	1.1
2020-04-24 20:33:57	up fast	1.6	0	2.2
2020-04-24 20:34:57	down fast	1.2	-1	2.3
2020-04-24 20:35:57	up	1.6	0.66	3.6
2020-04-24 20:36:57	down	1.2	-1	1.9
2020-04-24 20:37:57	down	0.86	0.40	1.4
2020-04-24 20:38:57	up	1.2	0.16	2.1
2020-04-24 20:39:57	up	1.5	0.60	2.4
2020-04-24 20:40:57	up	1.5	1.4	3.8
2020-04-24 20:41:57	up	1.8	1.1	2.7
2020-04-24 20:42:57	up	2.1	1.6	3.2
2020-04-24 20:43:57	up fast	2.6	1.9	2.7
2020-04-24 20:44:57	up	2.8	1.2	4.5
2020-04-24 20:45:57	up	3.0	1.00	5.2
2020-04-24 20:46:57	up	3.1	0.64	4.2
2020-04-24 20:47:57	down	2.9	2.4	3.2
2020-04-24 20:48:57	up	3.1	1.5	5.5
2020-04-24 20:49:57	down fast	2.7	0.57	5.0
2020-04-24 20:50:57	up	2.7	1.5	5.1

OUR TEAM

Anupama Jha

Harshit Singh

Mayank Jha

Shashank Kumar