



# COMMUNITY DAY

PUNE 2022



**COMMUNITY DAY**

**PUNE 2022**

# Modernize Data Pipelines

Chetan Hirapara | 10<sup>th</sup> Dec 2022



COMMUNITY DAY

PUNE 2022

# Chetan Hirapara

Lead Data Scientist @Vedity Software  
Expertise on Data Engineering and Data Science.  
Youtuber | Blogger | tech speaker

Connect: [https://linktr.ee/chetan\\_hirapara](https://linktr.ee/chetan_hirapara)





**COMMUNITY DAY**

**PUNE 2022**



## Agenda

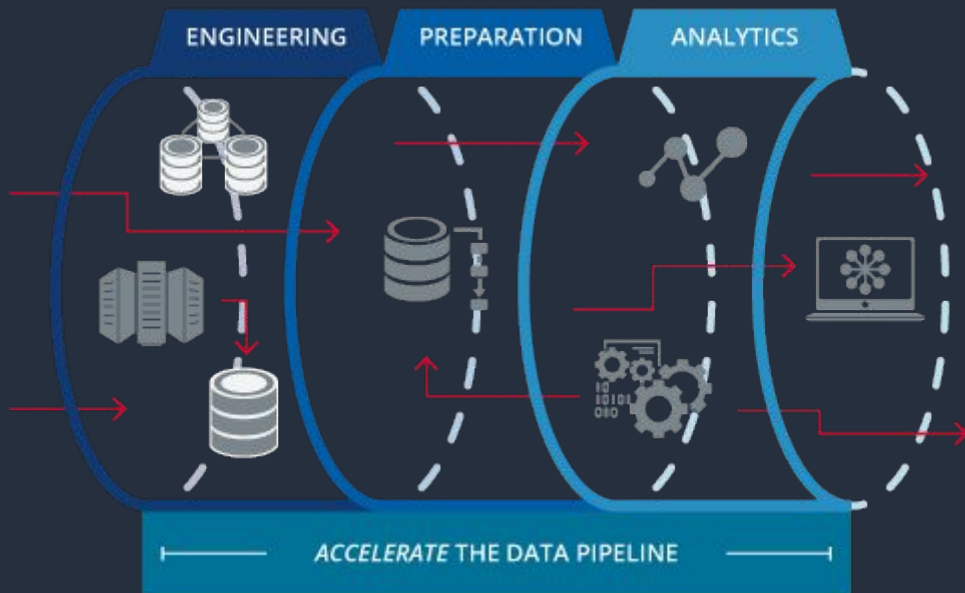
- Introduction of Data pipeline
- EMR Vs EMR Serverless
- Delta Lake on AWS
- Glue
- Amazon Managed Workflows for Apache Airflow (MWAA)
- Demo



COMMUNITY DAY

PUNE 2022

## Data pipeline in nutshell



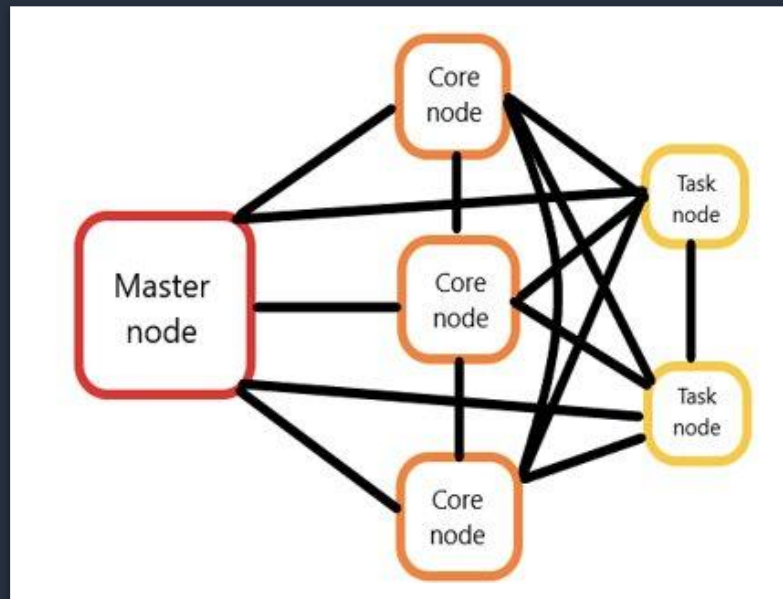


# COMMUNITY DAY

## PUNE 2022

## EMR

- Fully managed Hadoop cluster in AWS to store, process and analyze big data systems
- Map + Reduce
- HDFS / EMRFS / Local File system (Instance store/EBS)
- Supported Software's
- Spark, Hive, HBase, Hue, Pig, JupyterLab, etc.



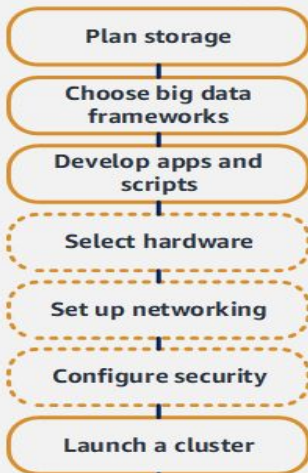


# COMMUNITY DAY

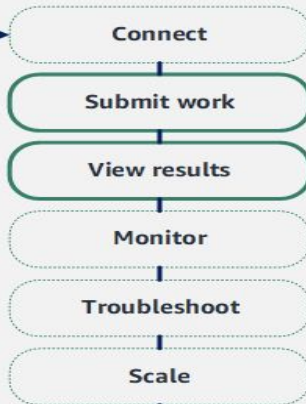
## PUNE 2022

## EMR Workflow

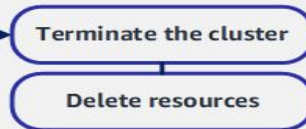
### Plan & Configure



### Manage



### Clean Up



- Essential
- - - Best Practice
- ..... Optional

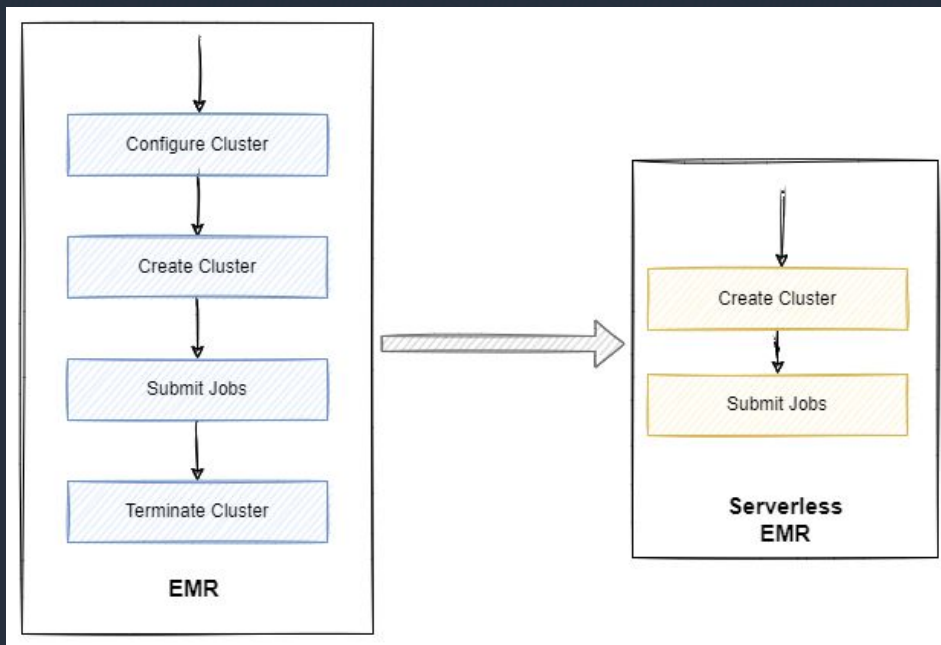




COMMUNITY DAY

PUNE 2022

## EMR Vs Serverless EMR







# COMMUNITY DAY

## PUNE 2022

## Delta Lake features

### Common Challenges with Data Lakes



Unsafe Writes



Orphan Data



No Schema Evolution



No Schema



Hot path for Streaming

**Data Lake getting Polluted**



ACID Transaction



Metadata Handling



Schema Enforcement



Unified Batch and Stream



Time Travel



Audit History



Full DML Support





# COMMUNITY DAY

## PUNE 2022

# Setup Delta Lake on EMR

### Additional Options

☐ EMRFS consistent view ⓘ

Custom AMI ID  ⓘ

#### ▼ Bootstrap Actions

Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. [Learn more](#) ⓘ

Bootstrap action type	Name	JAR location	Optional arguments
Custom action	Install Delta Lake Core	s3://[REDACTED]/emr-bootstrap-scripts/bootstrap-delta-core.sh	

```
#!/bin/bash
```

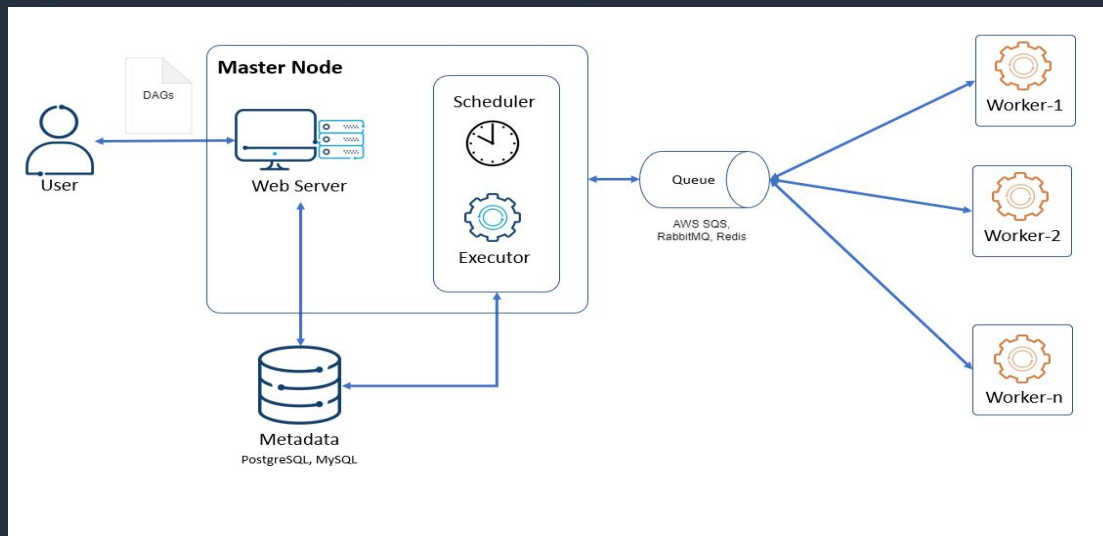
```
sudo aws s3 cp s3://mys3bucket-delta-lake-poc/emr-bootstrap-  
scripts/jars/delta-core_2.12-0.8.0.jar /usr/lib/spark/jars/
```



COMMUNITY DAY

PUNE 2022

# Introduction of Airflow

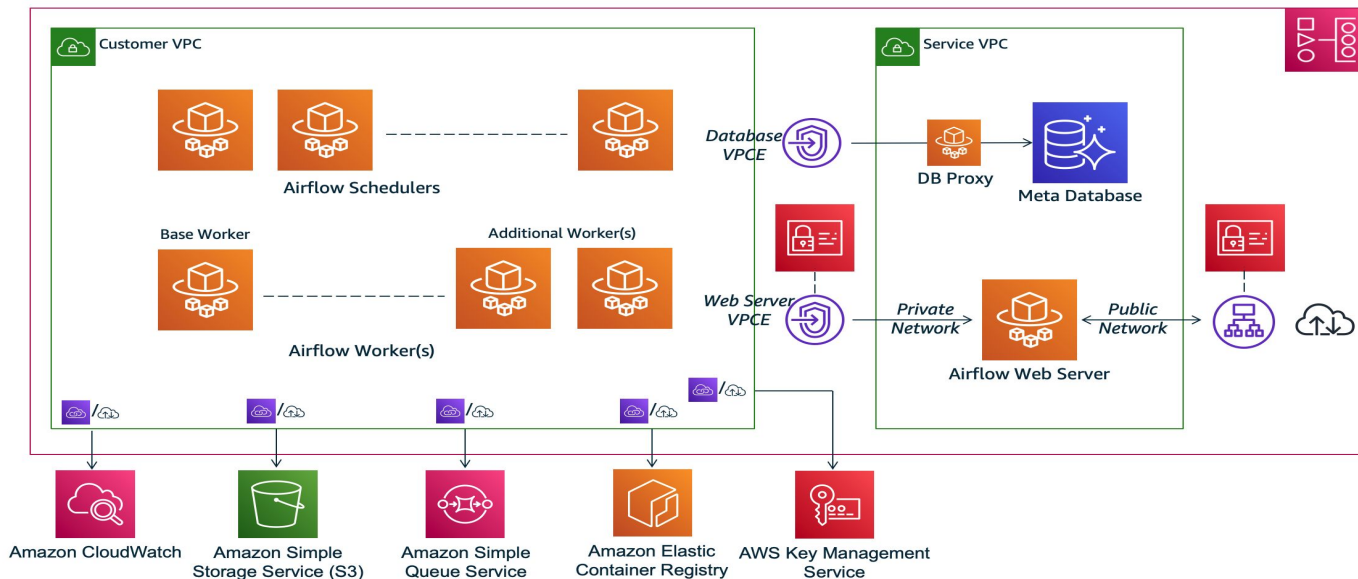




# COMMUNITY DAY

## PUNE 2022

### Amazon MWAA Architecture

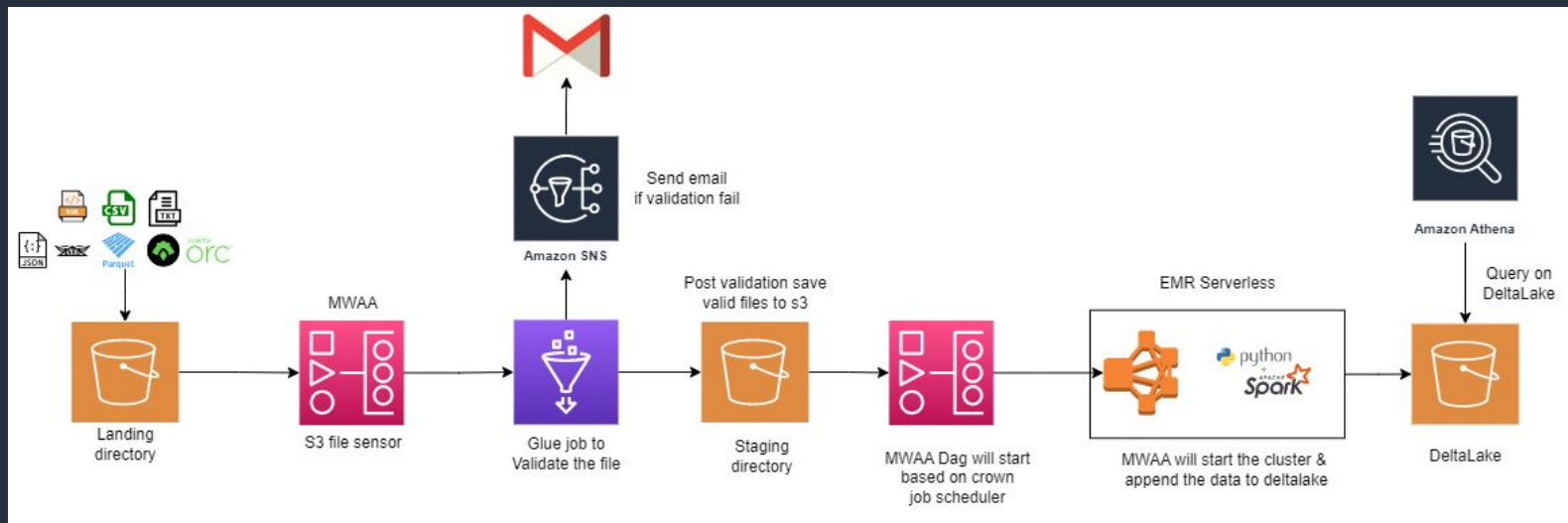




# COMMUNITY DAY

## PUNE 2022

## Demo Use case







COMMUNITY DAY

PUNE 2022

# Live Demo





**COMMUNITY DAY**

**PUNE 2022**



Connect: [https://linktr.ee/chetan\\_hirapar](https://linktr.ee/chetan_hirapar)



/Chetan\_Hirapara

# QUESTIONS ?





**COMMUNITY DAY**

**PUNE 2022**



# Thank you!

See you at the AWS Community Day Pune 2023





**COMMUNITY DAY**

**PUNE 2022**



## References

- [Data Lake on AWS EMR](#)
- [Building AWS Glue job using pyspark](#)
- [Github - Delta tables with Amazon EMR](#)
- <https://docs.delta.io/latest/quick-start.html>
- [Incremental Data processing using Delta lake with EMR](#)