aws

COMMUNITY DAY

PUNE 2022

aws
COMMUNITY DAY
PUNE 2022

Modernize Data Pipelines

Chetan Hirapara | 10th Dec 2022

aws

COMMUNITY DAY
PUNE 2022

Chetan Hirapara

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

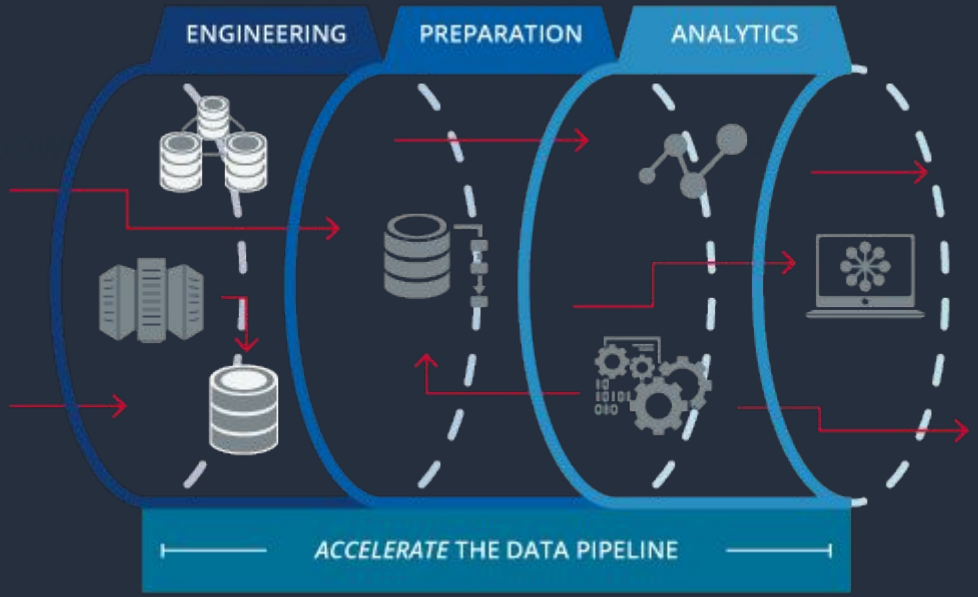
Connect: https://linktr.ee/chetan_hirapara

aws
COMMUNITY DAY
PUNE 2022

Agenda

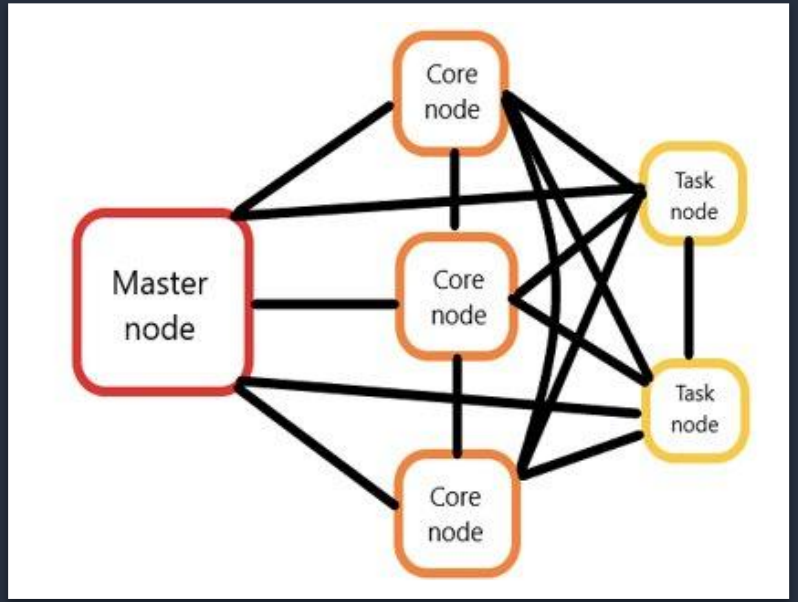
- Introduction of Data pipeline
- EMR Vs EMR Serverless
- Delta Lake on AWS
- Glue
- Amazon Managed Workflows for Apache Airflow (MWAA)
- Demo

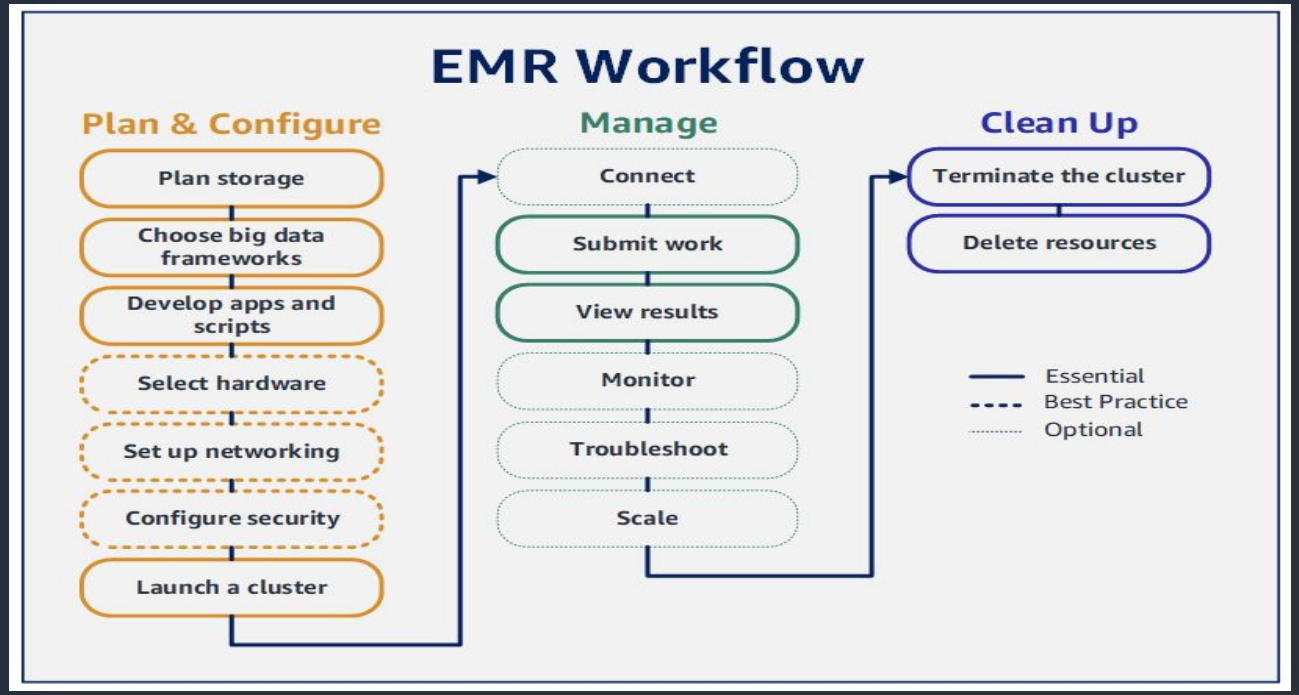
Data pipeline in nutshell



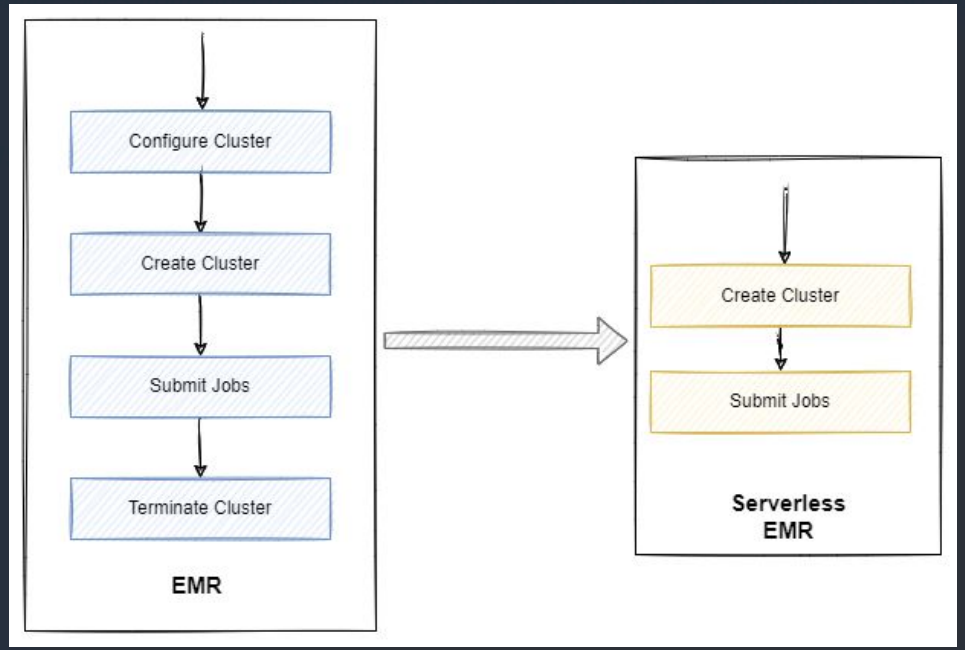
EMR

- 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.

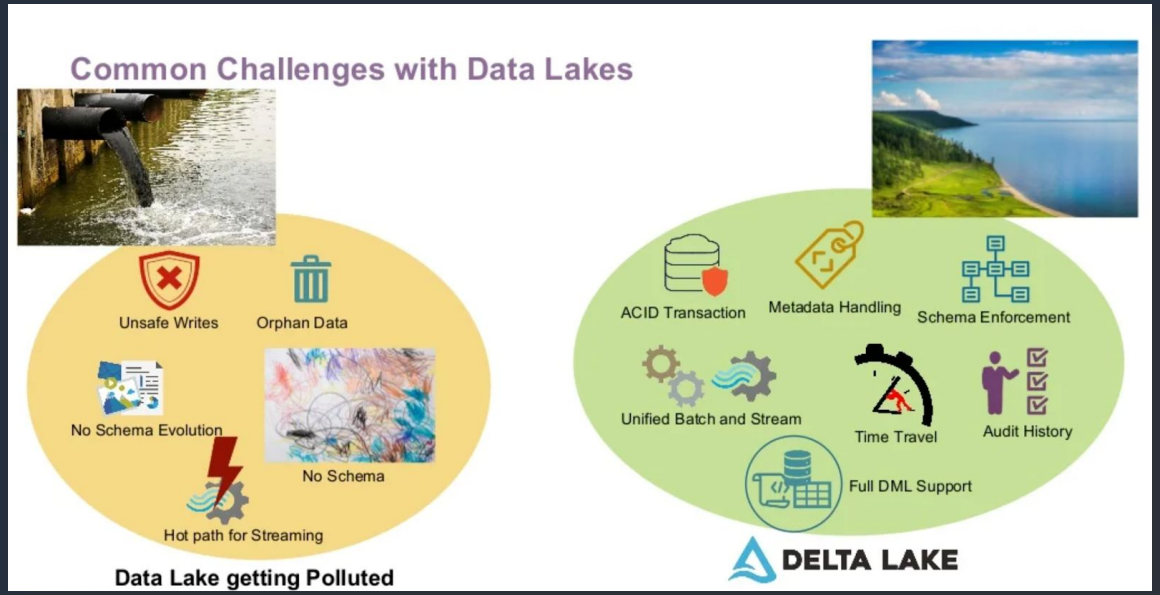



EMR Workflow
Plan & Configure
Plan storage
Choose big data frameworks
Develop apps and scripts
Select hardware
Set up networking
Configure security
Launch a cluster
Manage
Connect
Submit work
View results
Monitor
Troubleshoot
Scale
Clean Up
Terminate the cluster
Delete resources
Essential
Best Practice
Optional

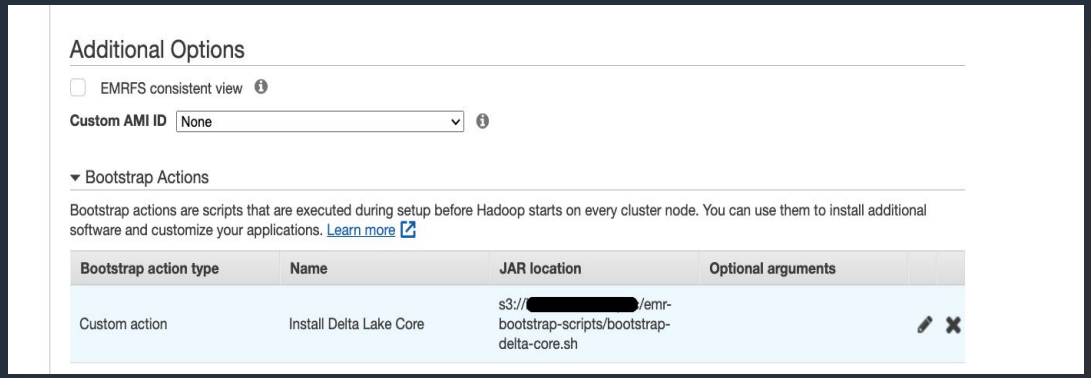
EMR Vs Serverless EMR



Delta Lake features



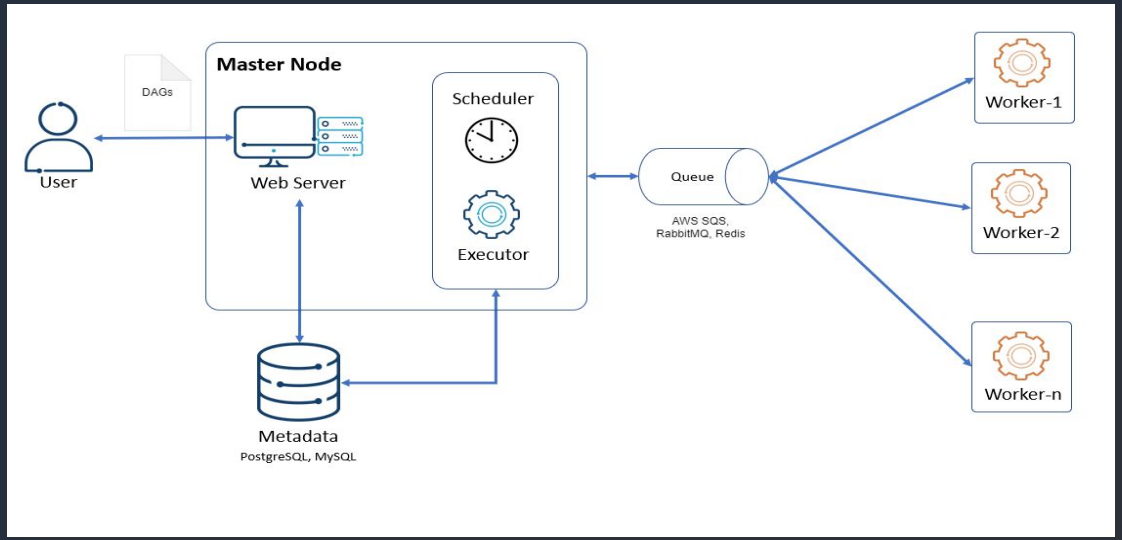
Setup Delta Lake on EMR



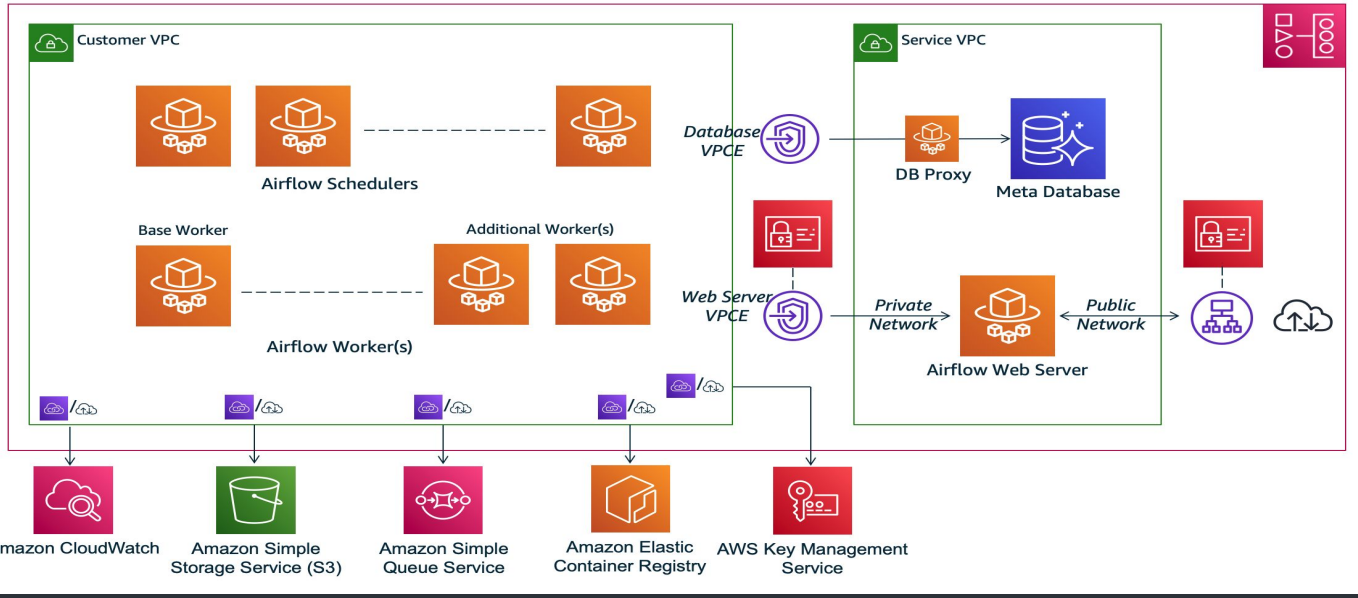

```
#!/bin/bash

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/
```

aws
COMMUNITY DAY
PUNE 2022

Introduction of Airflow

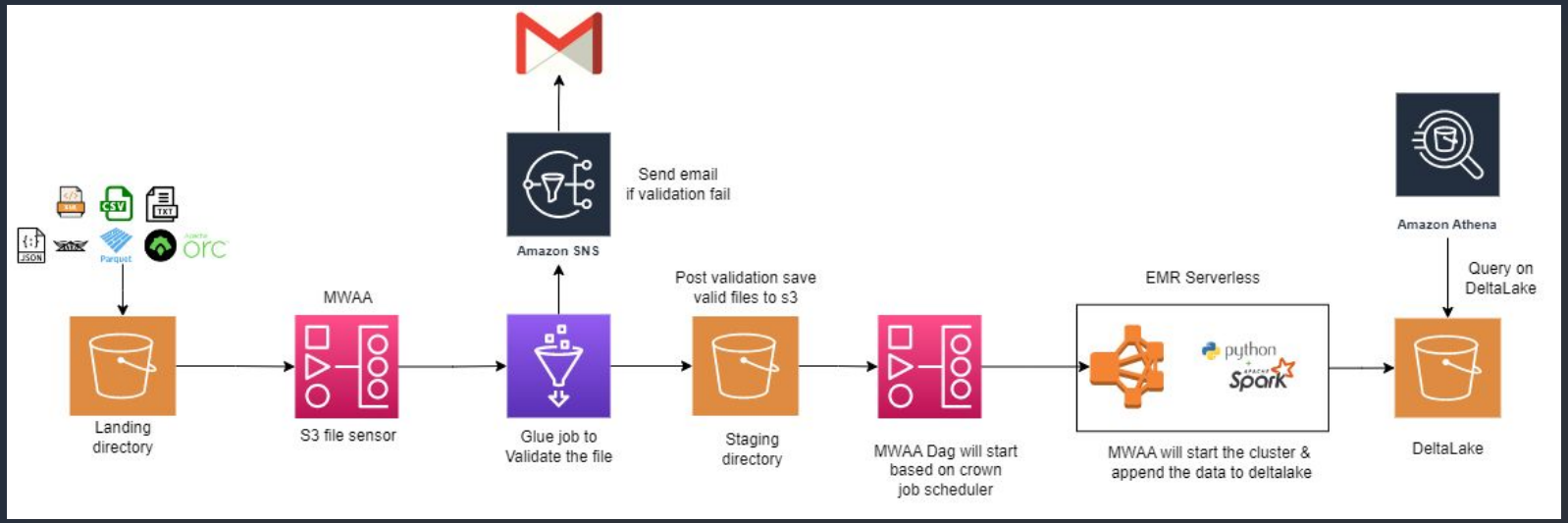


Amazon MWAA Architecture



aws
COMMUNITY DAY
PUNE 2022

Demo Use case




aws
COMMUNITY DAY
PUNE 2022
AWSUGPUN
AWS USER GROUP PUNE
Live Demo

aws
COMMUNITY DAY
PUNE 2022

QUESTIONS ?

Connect: https://linktr.ee/chetan_hirapar

/Chetan_Hirapara

aws

COMMUNITY DAY

PUNE 2022

Thank you!

See you at the AWS Community Day Pune 2023

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