aws

SUMMIT ONLINE

ANA09

Build an SQL-based data processing pipeline in minutes

Melody Yang

Data & Analytics Specialist SA
Amazon Web Services

aws SUMMIT ONLINE

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

+6.5%
Trends
More info
78%
Shares
More info
95%
Performance
More info
85%
60%
1.17
1.14
1.11
1.08
1.05
CTMX 0.45 +0.45%
FTR -0.23 -2.34%
CSCO -1.01 -1.89%
CHK 0.02
AAPL +2.
PRTO
AMZN
TSLA
AVGO
SIRI -0.65

Challenges

Delivery Speed

Security

Limited skill

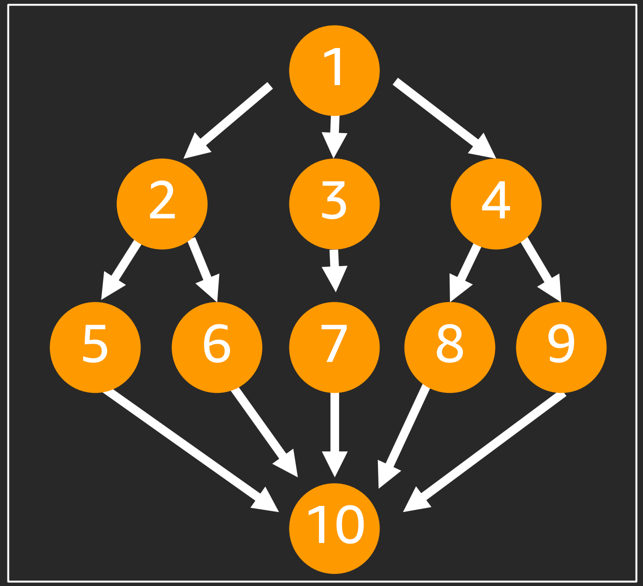
An example of the challenge

10

data processing jobs

3 **Layers**

dependency



7 **Days**

effort per job

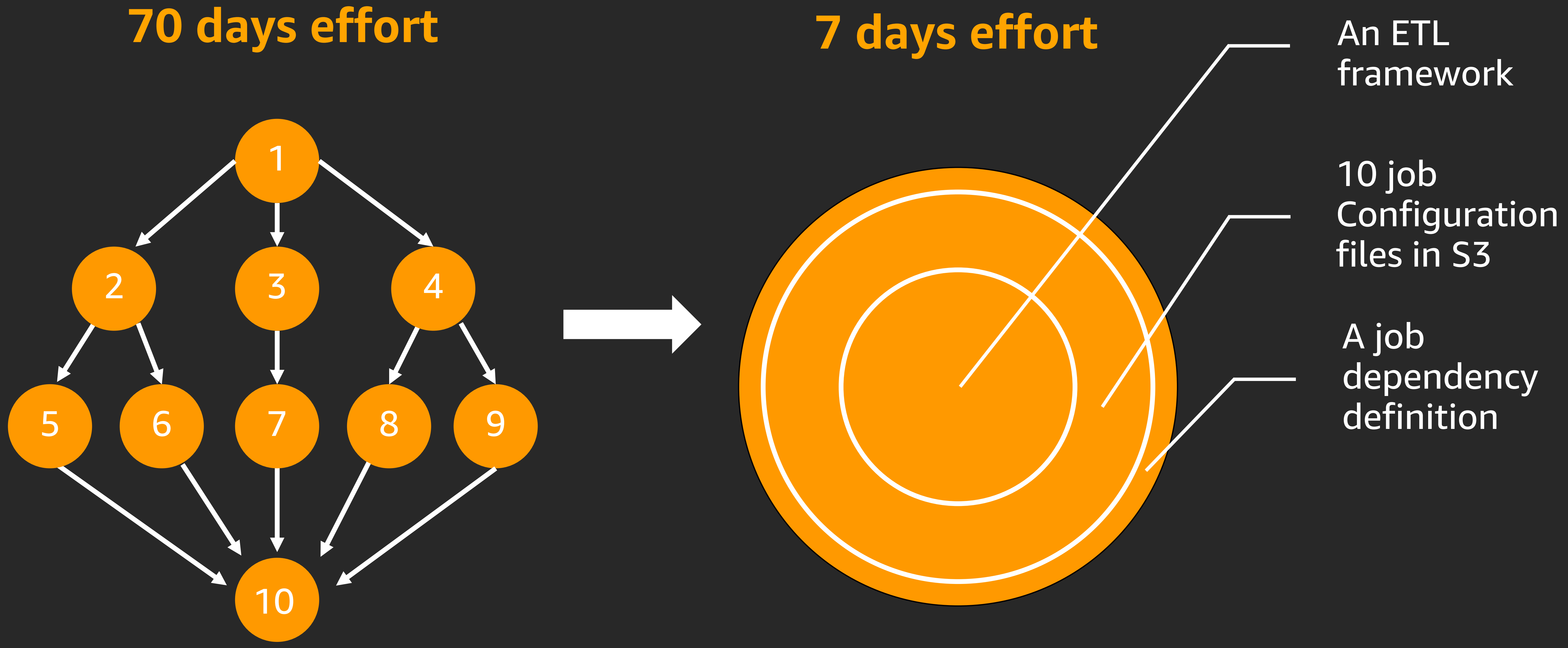
70 **Days**

total effort

Design principles

Requirements	Design Principles
Shorten analytics lifecycle	Microservices for batch and stream processes
Repeatable and scalable	Configuration driven and codeless ETL
Use existing analysis skill to improve data quality	SQL first approach

The new solution



70 days effort

7 days effort

An ETL framework

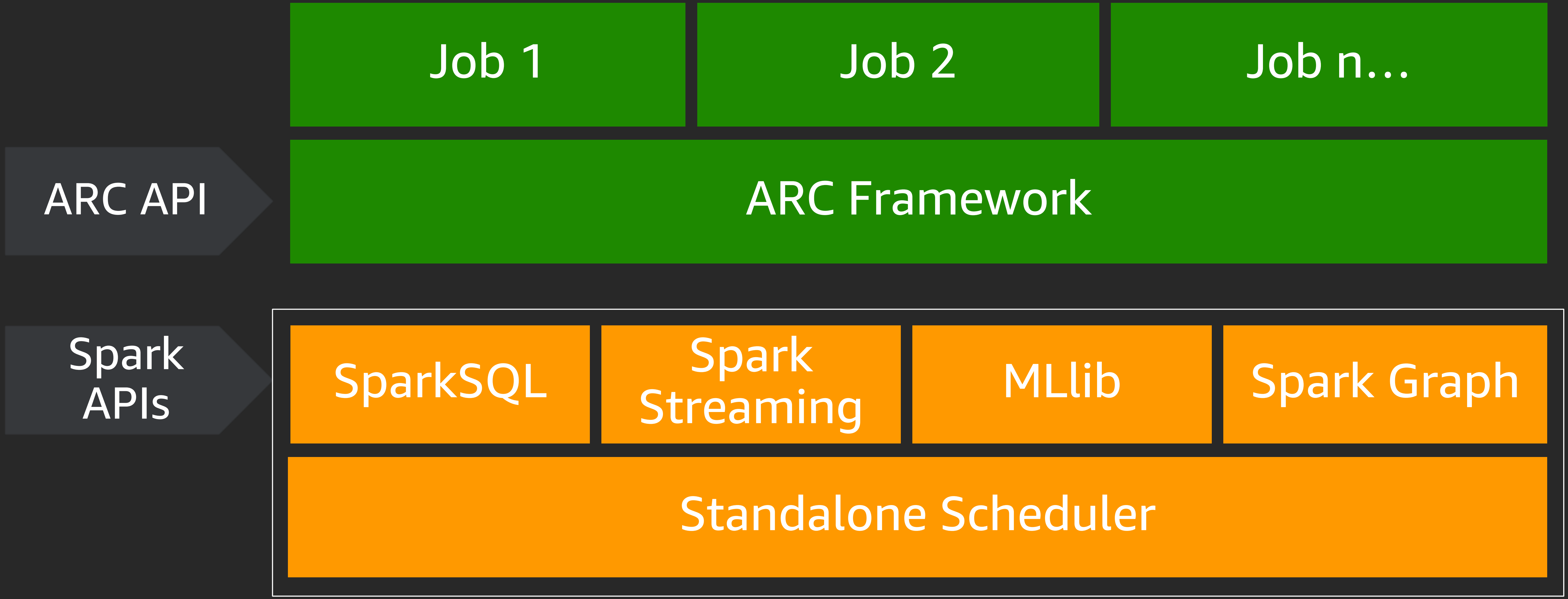
10 job Configuration files in S3

A job dependency definition

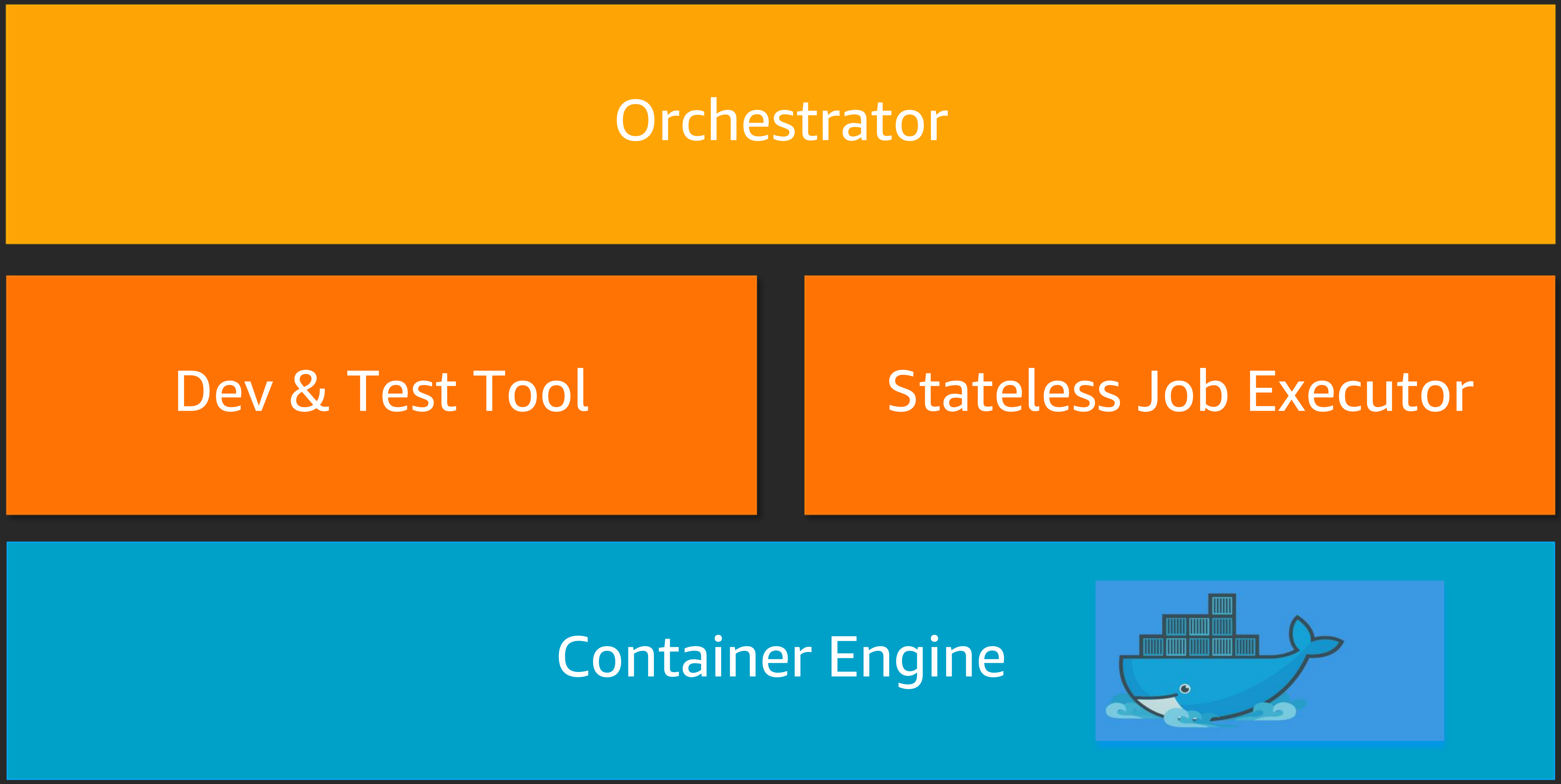
ARC framework:
Open source data processing tool

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Standardised API



Standardised practice



Orchestrator

Dev & Test Tool

Stateless Job Executor

Container Engine

Security assurance

Networking

- Private connection
- Isolated network
- Load balancer

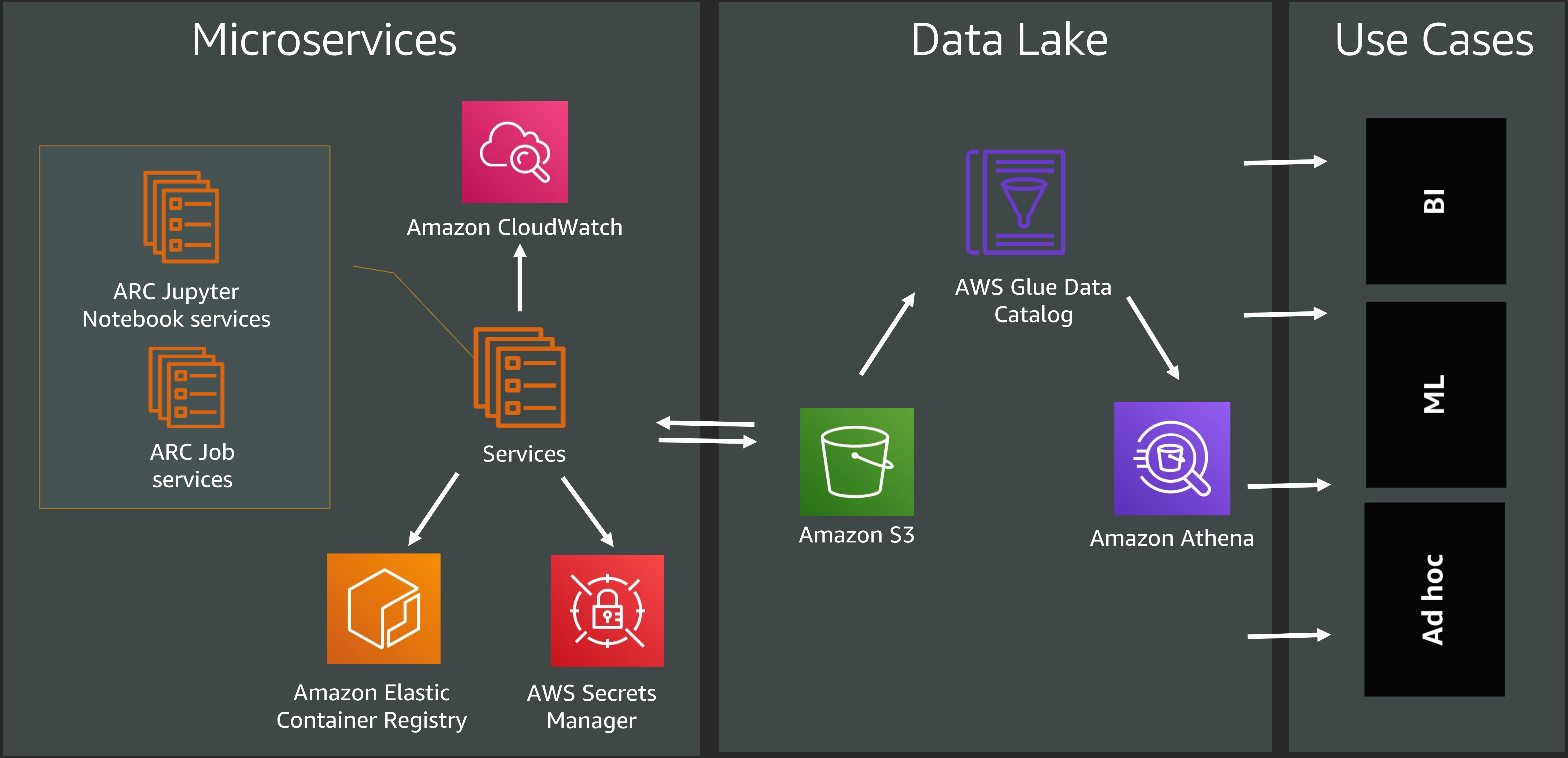
Access control

- Role based control
- Secret protection
- Encryption

Visibility

- Granular level logging
- Integrated alert service
- Readable business logic

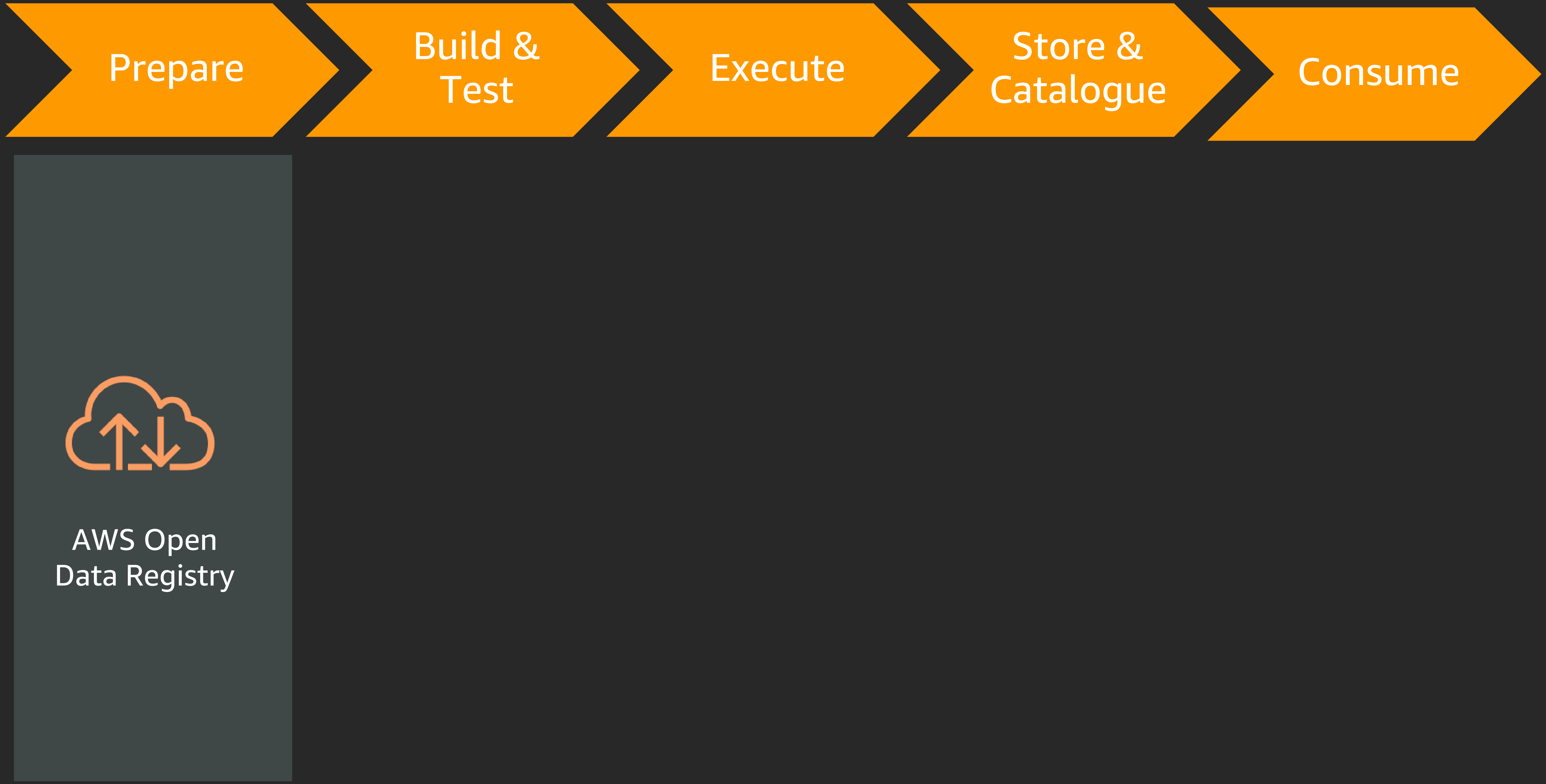
High level architecture



Demo

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Solution workflow



Prepare

Build & Test

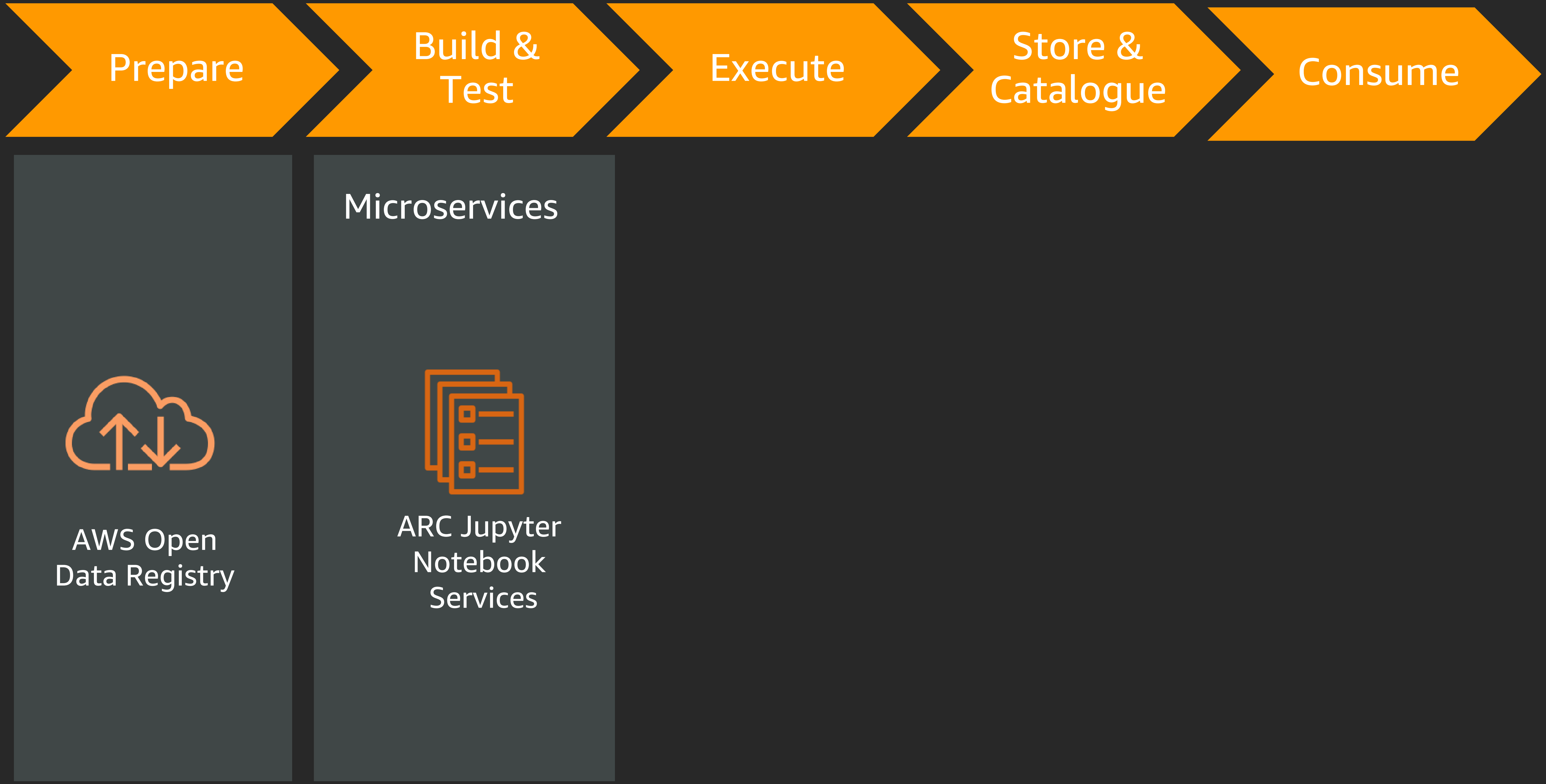
Execute

Store & Catalogue

Consume

AWS Open Data Registry

Solution workflow



Prepare

Build & Test

Execute

Store & Catalogue

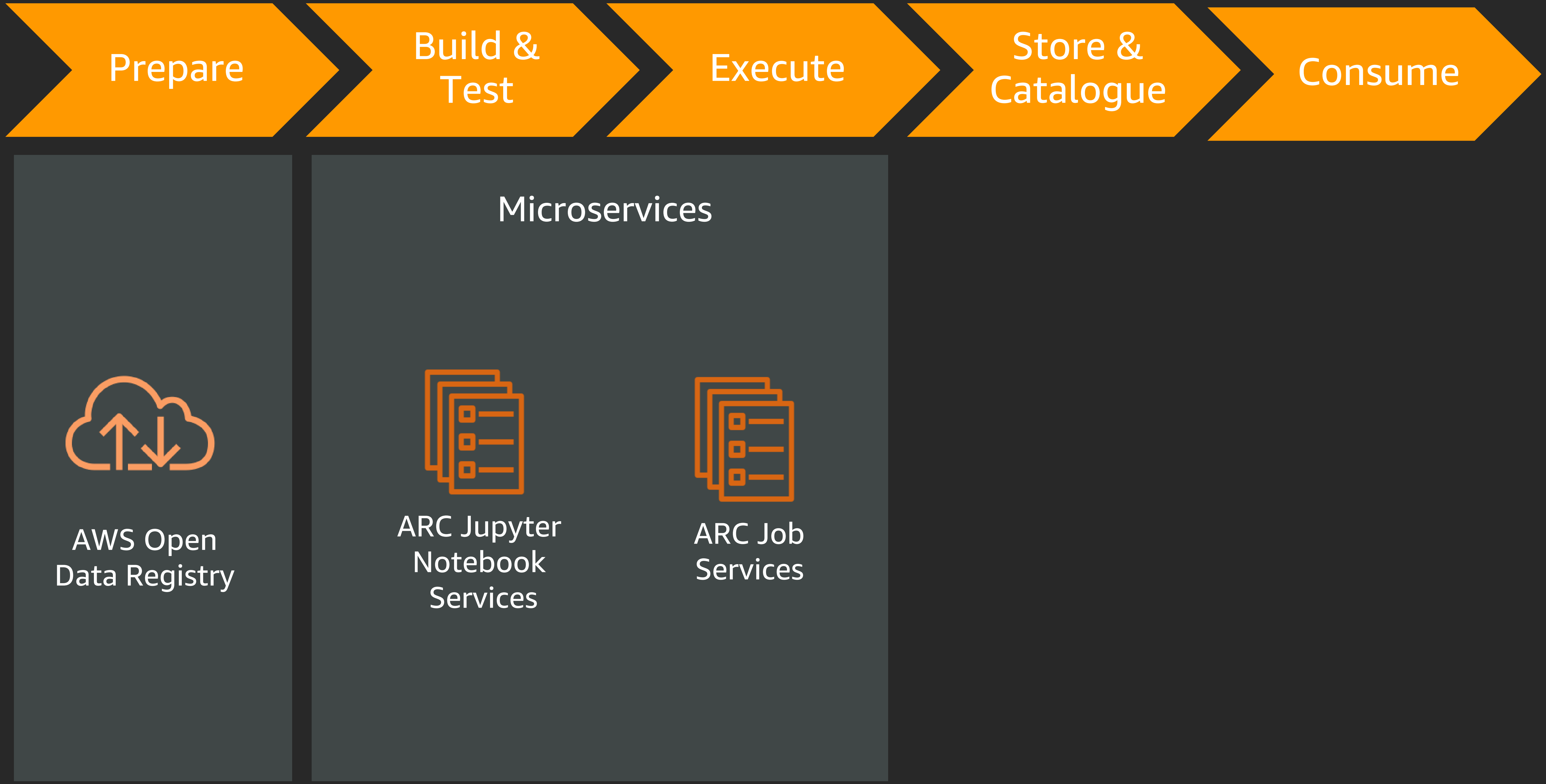
Consume

AWS Open Data Registry

Microservices

ARC Jupyter Notebook Services

Solution workflow



Prepare

Build & Test

Execute

Store & Catalogue

Consume

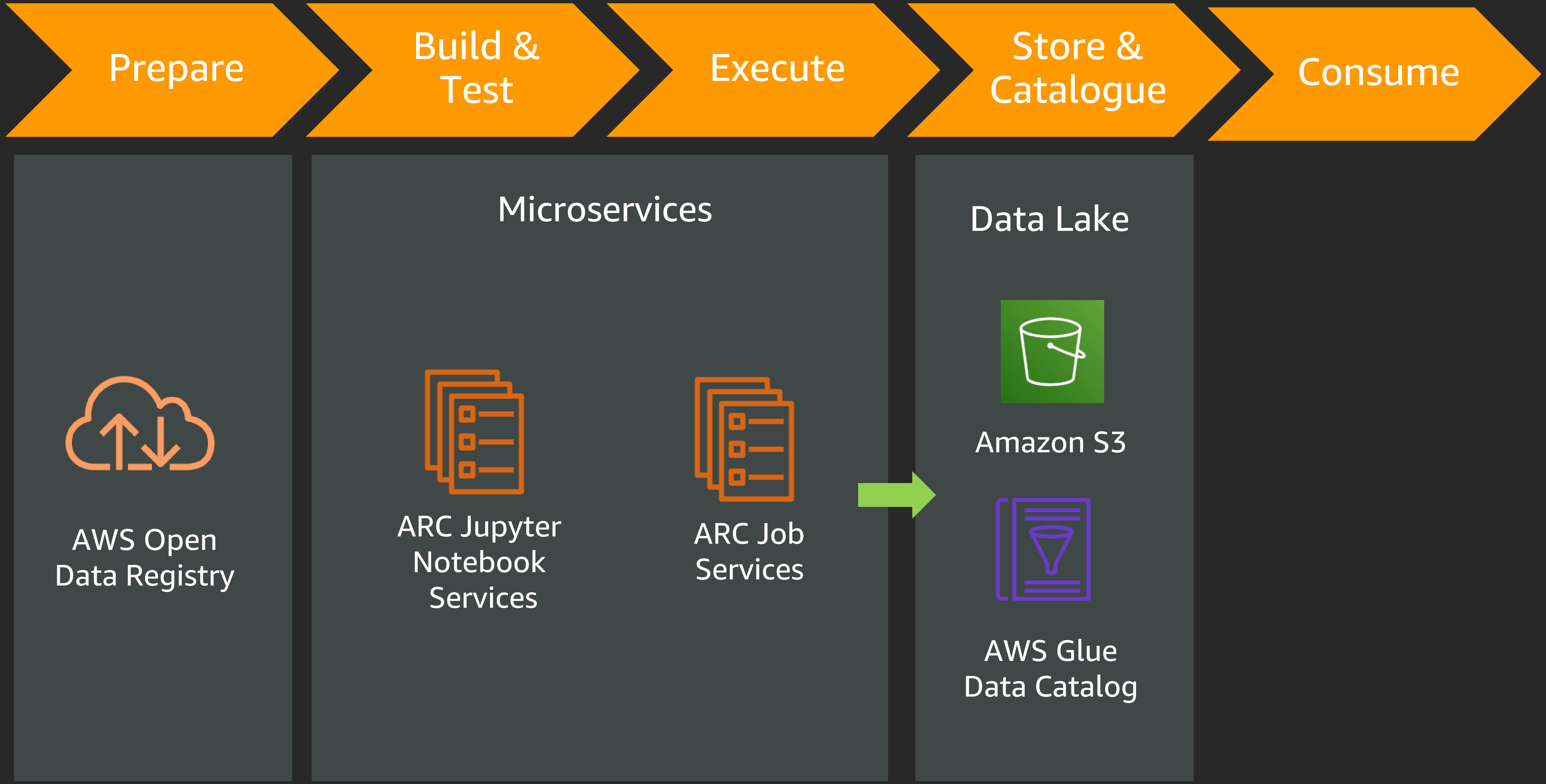
AWS Open Data Registry

Microservices

ARC Jupyter Notebook Services

ARC Job Services

Solution workflow



Prepare

Build & Test

Execute

Store & Catalogue

Consume

AWS Open Data Registry

Microservices

ARC Jupyter Notebook Services

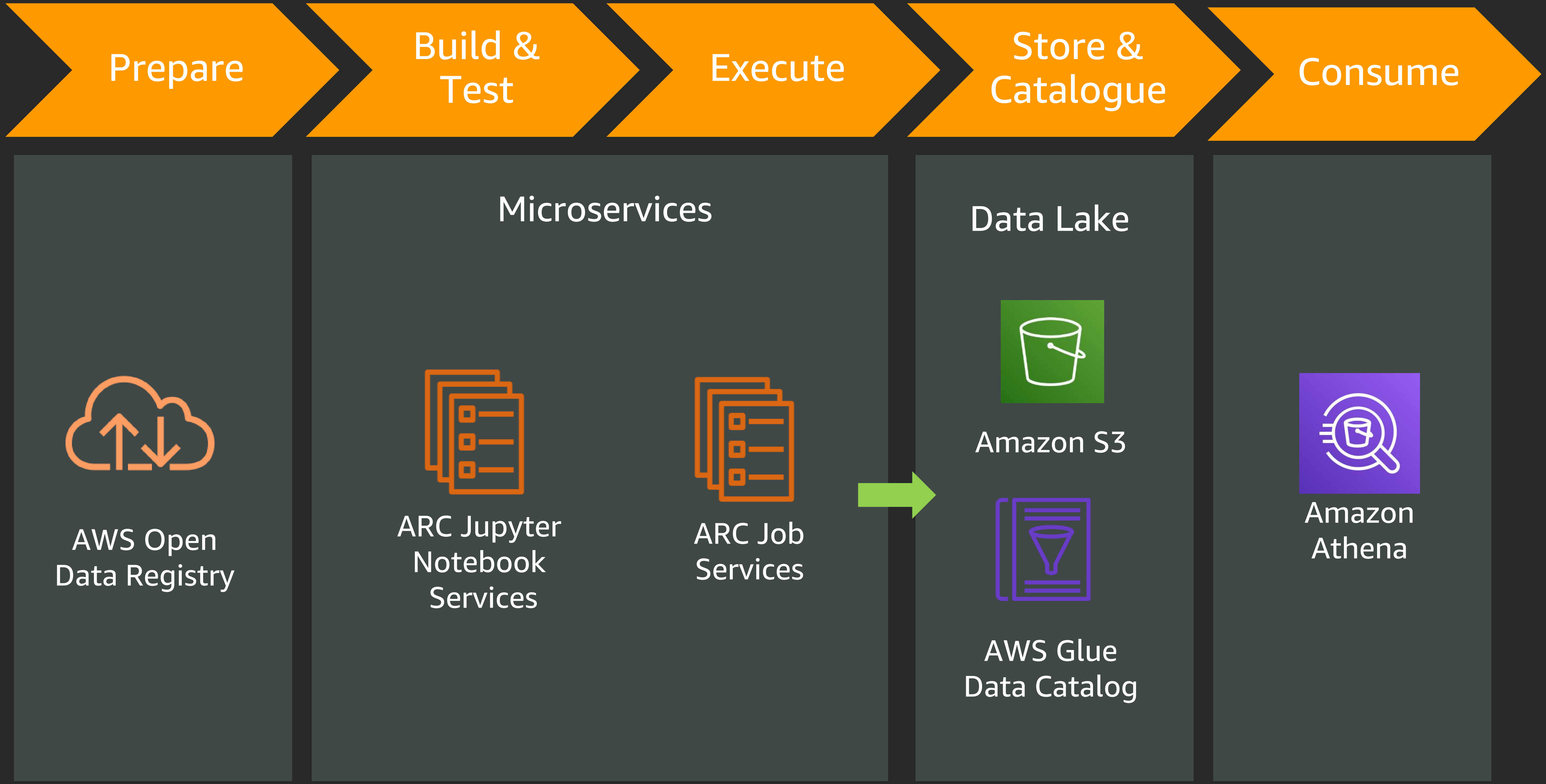
ARC Job Services

Data Lake

Amazon S3

AWS Glue Data Catalog

Solution workflow



Prepare

Build & Test

Execute

Store & Catalogue

Consume

AWS Open Data Registry

Microservices

ARC Jupyter Notebook Services

ARC Job Services

Data Lake

Amazon S3

AWS Glue Data Catalog

Amazon Athena

Summary

Build a data pipeline in minutes, not weeks

- SQL first approach, no more custom-code
- Empower business users with self- service ETL

Leverage microservices

- Move away from monolithic architecture with improved productivity and speed
- Fault tolerance and easy to scale

Highly secure and transparent

- Multiple layers of security controls
- Granular level logging, validation and alert

Resources

AWS Reference Architecture

- SQL Based Data Processing in Amazon ECS

ARC reference

- Documentation and tutorial - https://arc.tripl.ai/tutorial/
- CI/CD Example - https://github.com/tripl-ai/deploy
- Forum - https://github.com/tripl-ai/question
- Source Code - https://github.com/tripl-ai/arc
- Docker Hub - https://hub.docker.com/u/triplai

Thank you!

Melody Yang

meloyang@amazon.com

aws SUMMIT ONLINE

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.