cloudera

DATA PROTECTION IN HYBRID DATA LAKE ENVIRONMENT

Murali Ramasami, Staff Software Engineer

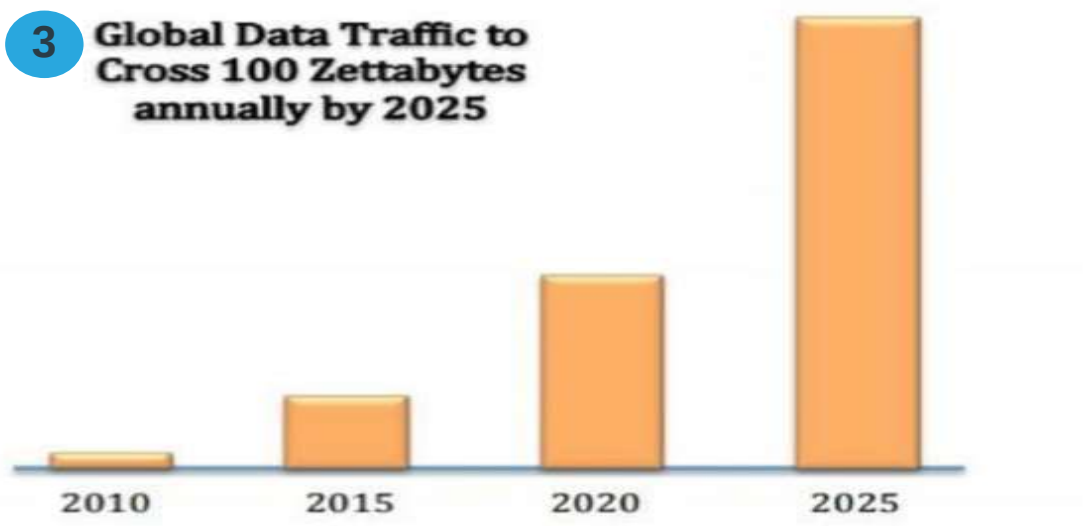
AGENDA

- Hybrid Data Lake Environment
- Data Lifecycle Manager
  - Replicate data HDFS <-> Cloud
  - Replicate data Hive <-> Cloud
- Demo

BIG DATA TREND

1 **90% of the data in the world has been created in the last two years alone**

3 **Global Data Traffic to Cross 100 Zettabytes annually by 2025**



Source: Frost & Sullivan - World's Top Global Mega Trends To 2025 and Implications to Business, Society and Cultures

2 **Digital content is doubling every 18 months**

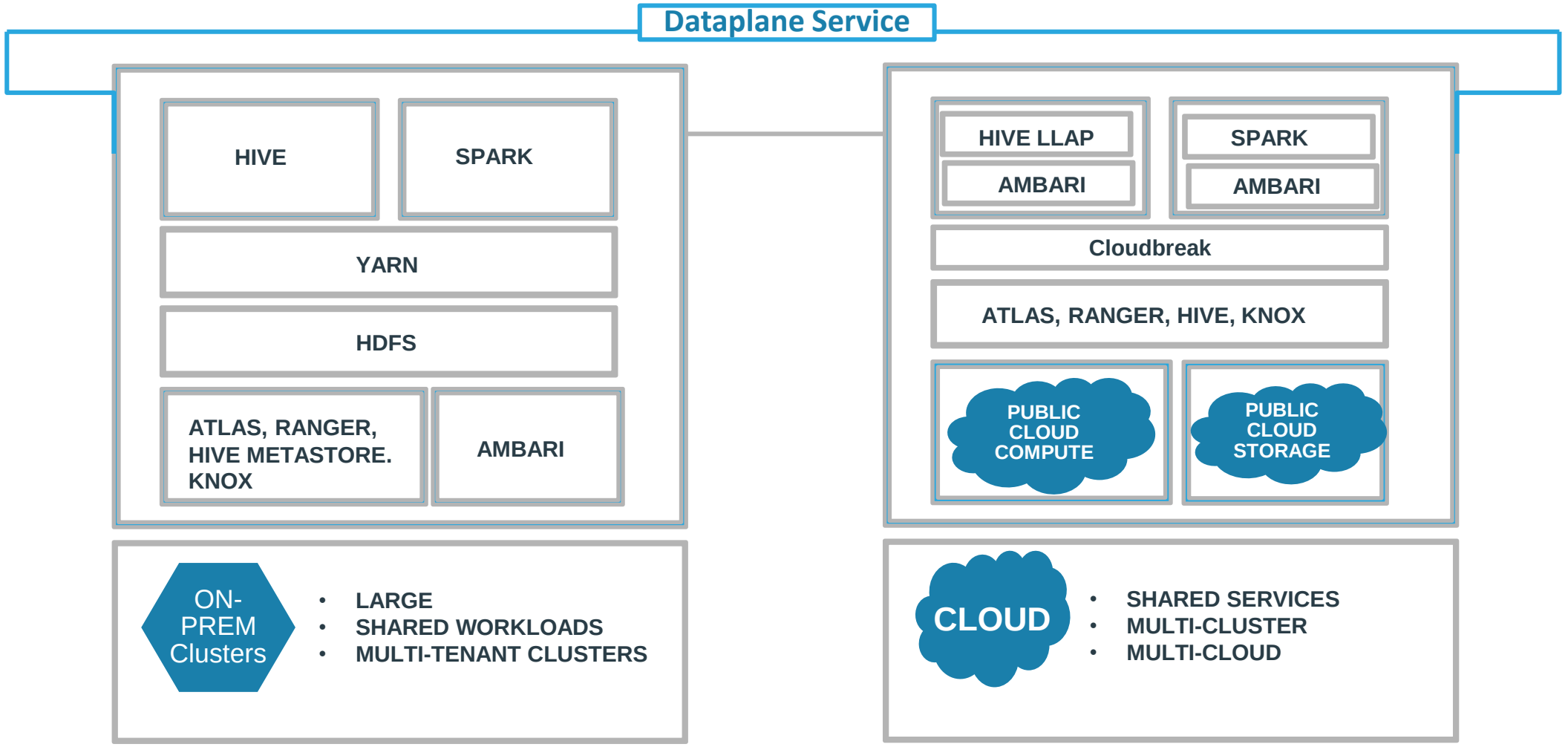
Structured Data

- Database
- Data Warehouse
- ERPs
- CRMs

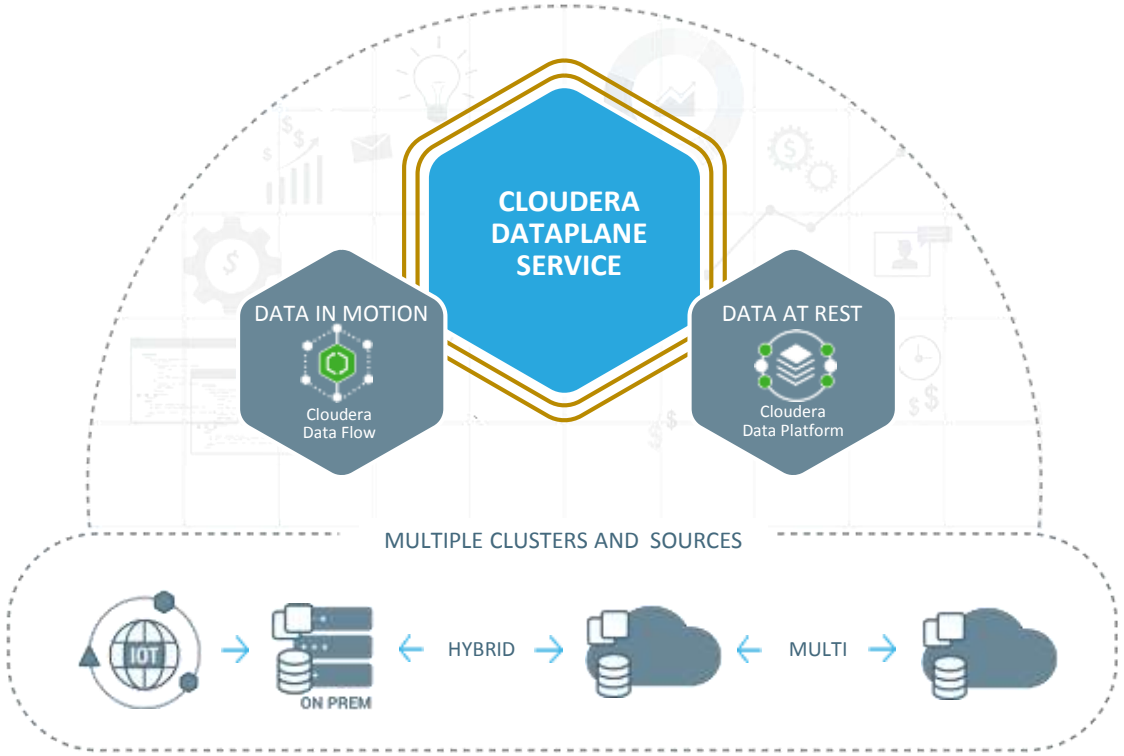
Unstructured Data

- Web blogs
- Social media
- Audio, Video
- Software file-systems

HYBRID ENTERPRISE DATA LAKE ENVIRONMENT

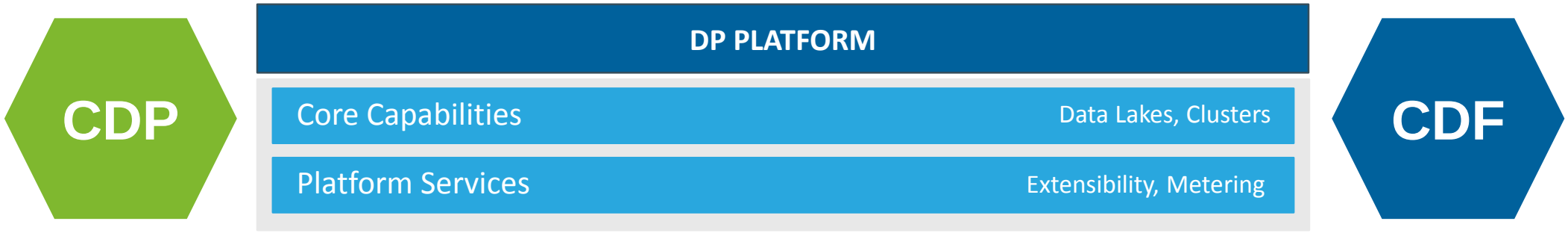


CLOUDERA DATAPLANE



Cloudera DataPlane
is a portfolio of **data solutions** that enable the enterprise to **manage & discover data** across **hybrid** environments.

DATAPLANE: AT A GLANCE



DATAPLANE APPLICATIONS

Data Lifecycle Manager

Data Lifecycle Manager (DLM) is a service that safeguards company's data by replicating it in on-premises data center(s) or in the cloud.

Data Steward Studio

Data Steward Studio (DSS) enables companies to discover, understand and govern its data across on-premises and cloud clusters.

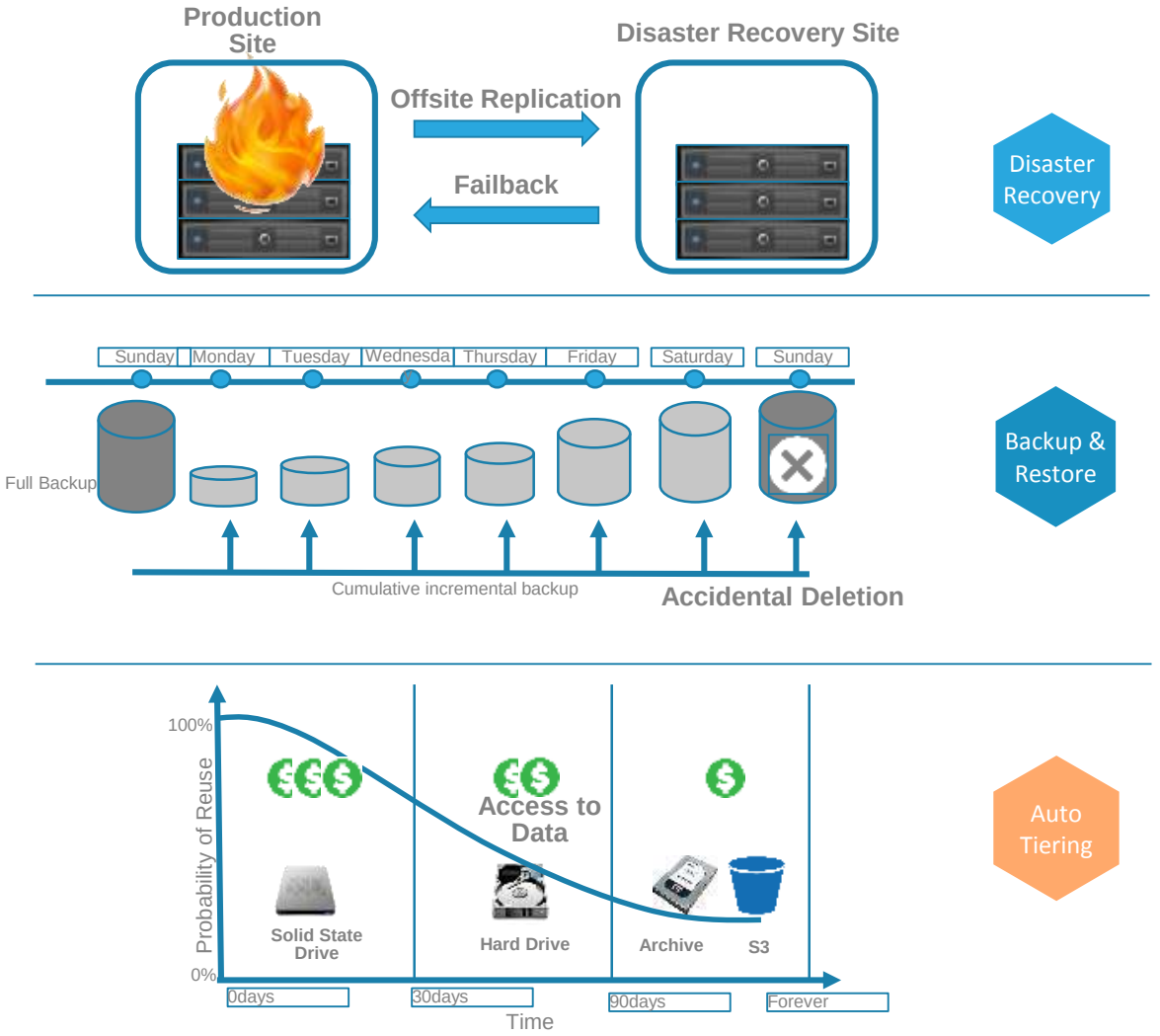
Data Analytics Studio

Data Analytics Studio (DAS) provides diagnostic tools and intelligent recommendations to make business analysts and IT teams become self-sufficient and productive.

Streams Messaging Manager

Streams Messaging Manager is a management and monitoring tool for Apache Kafka.

DATA LIFECYCLE MANAGER (DLM) SERVICE

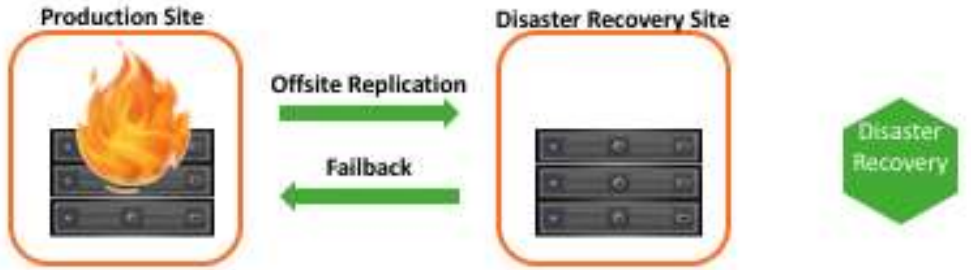


A Portfolio of Service

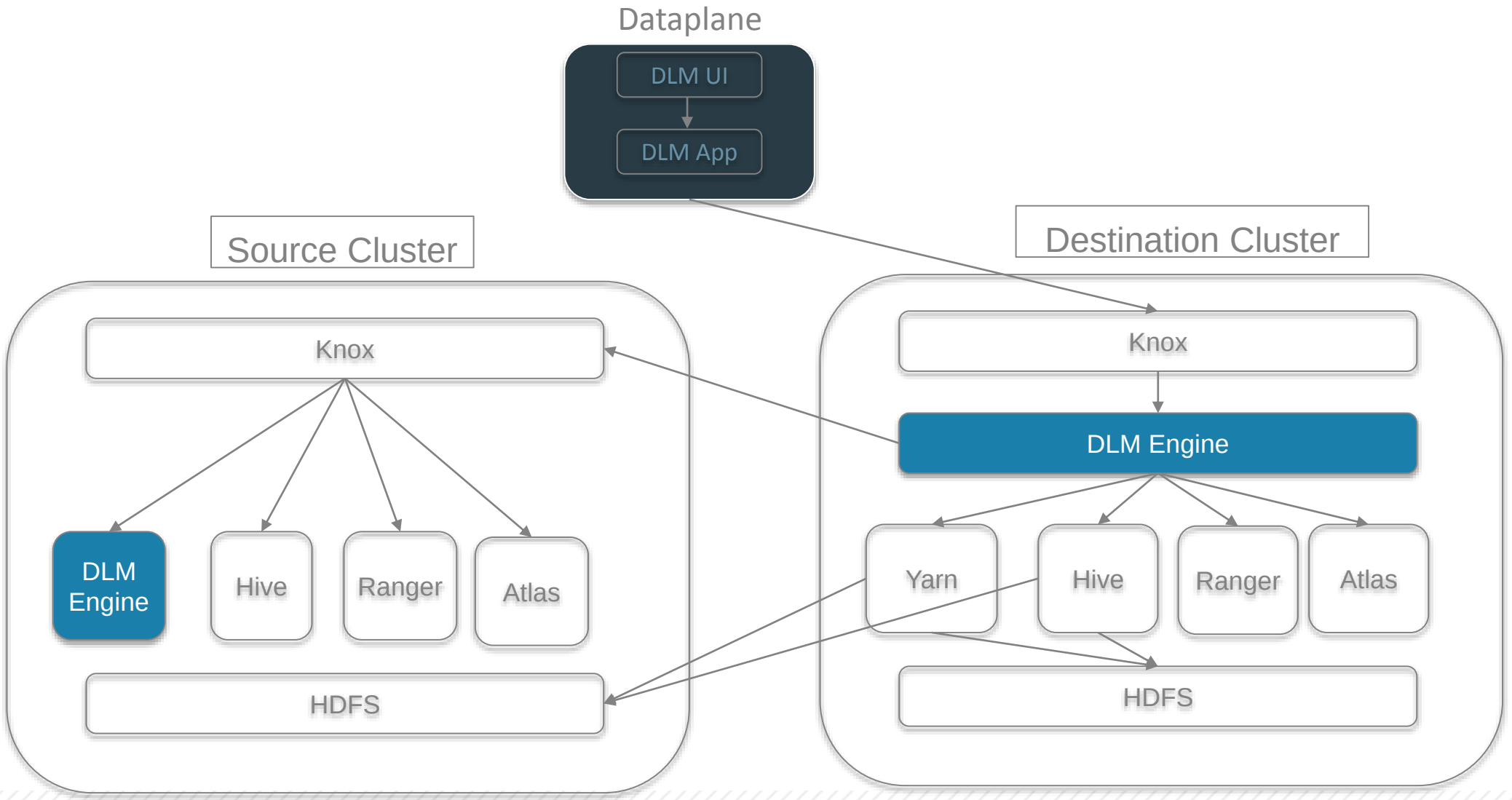
- Replication/failback to another cloud/on-prem site for Disaster Recovery
- Backup & Restore of business critical data, for protection against accidental deletion
- Auto Tiering of hot/warm/cold data to cloud object storage/on-prem for TCO reduction.

DLM FEATURES

- Incremental Hive replication & Hive metadata
- HDFS snapshot based replication between HDP clusters
- Ranger policy replication
- Atlas tag/lineage replication
- Cloud storage replication (AWS, Azure, GCP)
- Active/standby behavior on DR site using Ranger
- TDE & TLS support, Support multiple keys/KMS

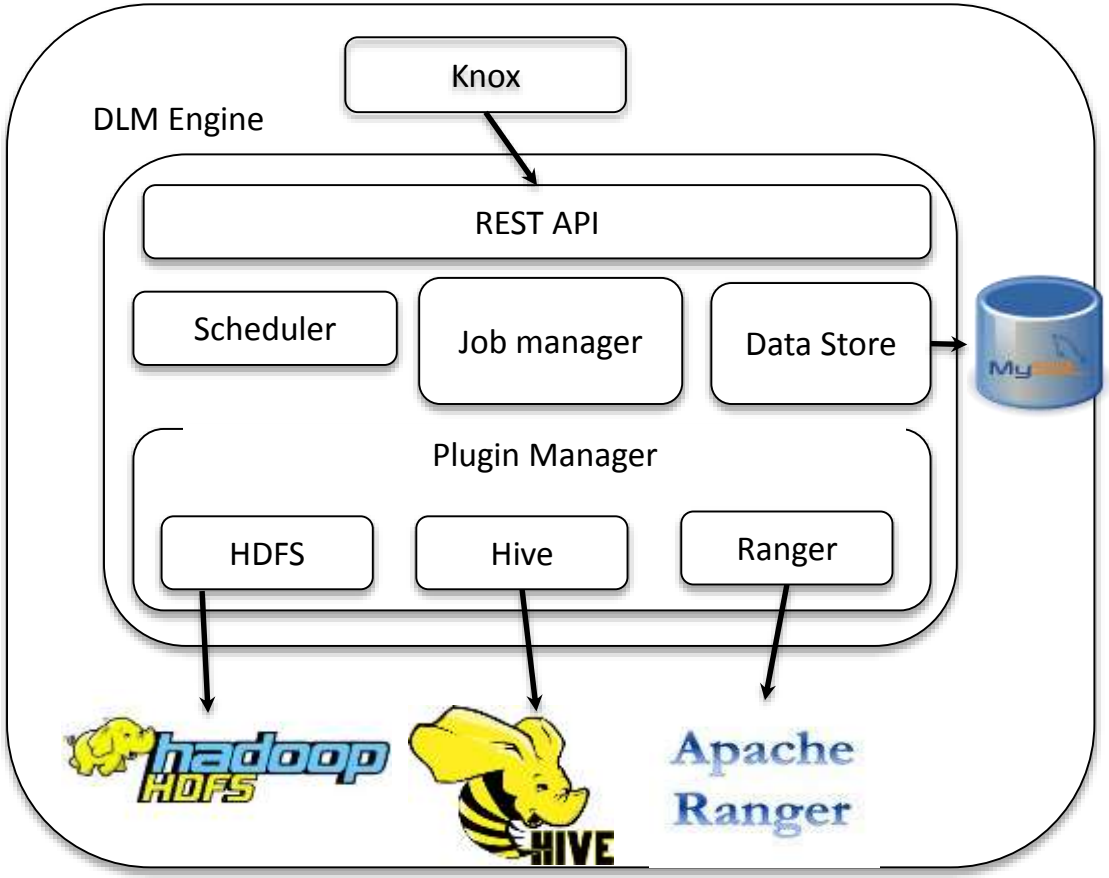


DLM ARCHITECTURE



DLM ENGINE INTERNALS

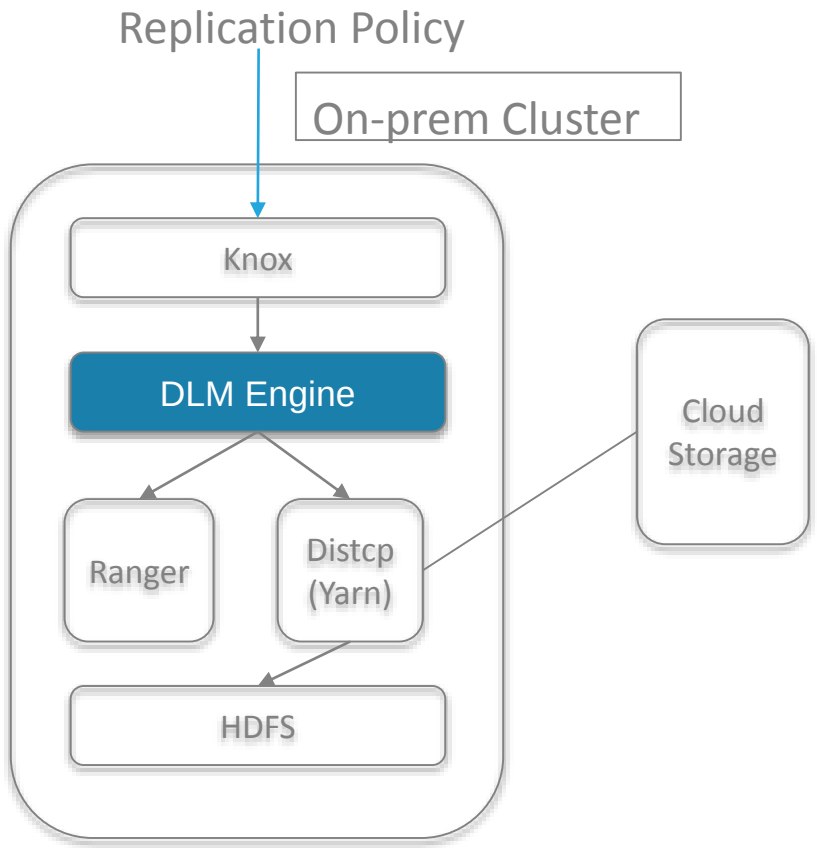
- DLM engine is stateless server, state in external DB.
- Schedules replication jobs using Quartz.
- Job manager handles job concurrency, failures, retries and recovery.
- Metric collection for job and data transfer
- Specific plugins handles replication of the specific data from source service to target service.



REPLICATING DATA ON-PREM <-> CLOUD

HDFS - CLOUD STORAGE REPLICATION

- Replication to/from cloud storage directly, no cluster required on target
- Supported on AWS S3, Azure ADLS, GCP GCS
- Data is pushed/pulled from on-prem clusters using Hadoop Connectors for the filesystem
- Supports cloud native encryptions
- File ACLs, Ranger Policies and Atlas Metadata are copied in meta files for restore



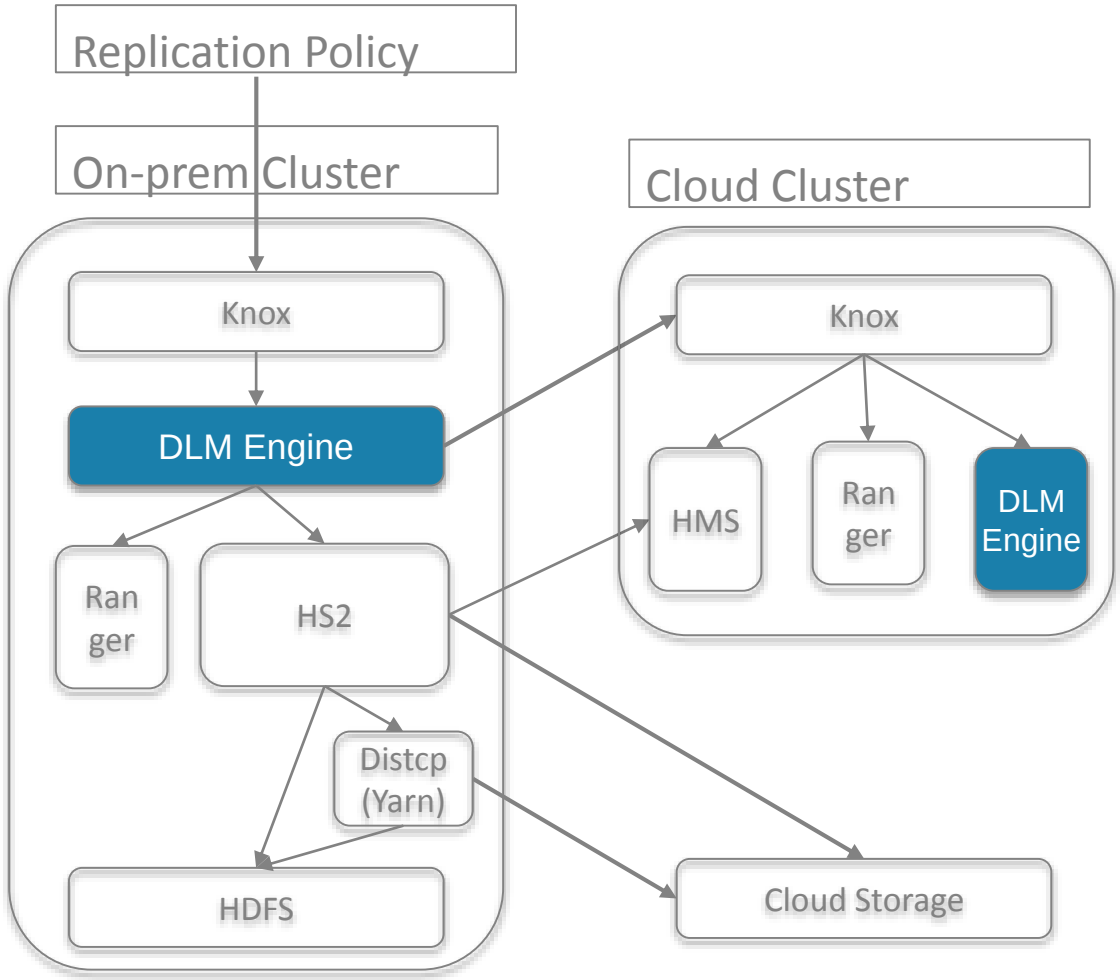
REPLICATION—SECURITY POLICIES & GOVERNANCE METADATA

- Copies resource and descendent policies for the directories/DBs replicated
- Copies row filters and column masking rules
- Adds deny policy on target
- Atlas metadata, tags associated for the directories/DBs created will also get replicate on to the target

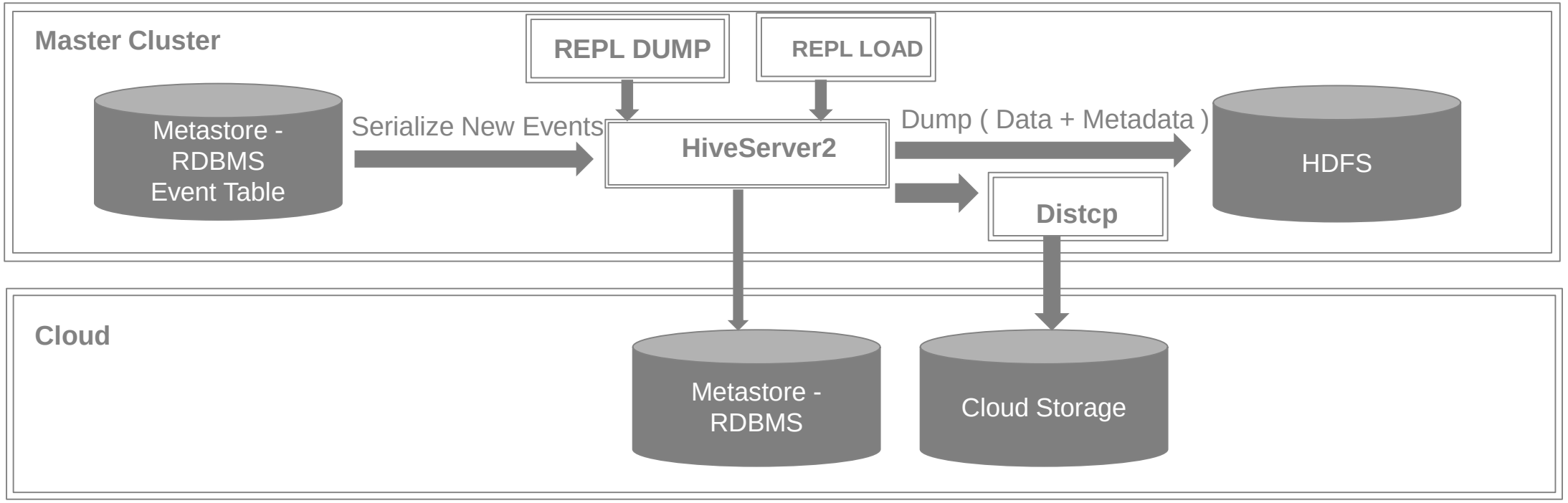
DEMO

HIVE ON-PREM TO CLOUD REPLICATION

- Minimal cluster on the cloud - shared services cluster
- Push based replication to avoid opening data ports
- Uses cloud SDK for secure data transfer
- Supports cloud storage encryption
- Uses Hadoop connector to transfer data



EVENT BASED HIVE REPLICATION (CLOUD)



DEMO

THANK YOU