CLOUDERA

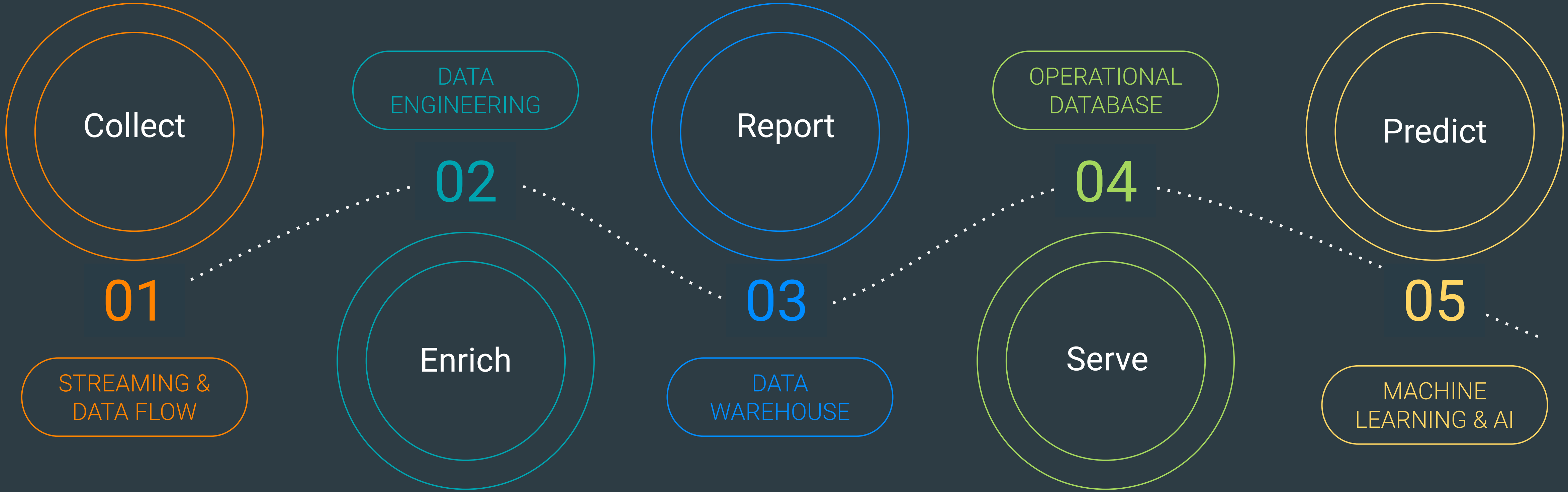
Cloudera Datawarehouse

The Enterprise Solution for Modern Analytics

Cloudera Data Platform

CLOUDERA - THE ENTERPRISE DATA CLOUD COMPANY

Manage and secure the data lifecycle in any cloud or datacenter

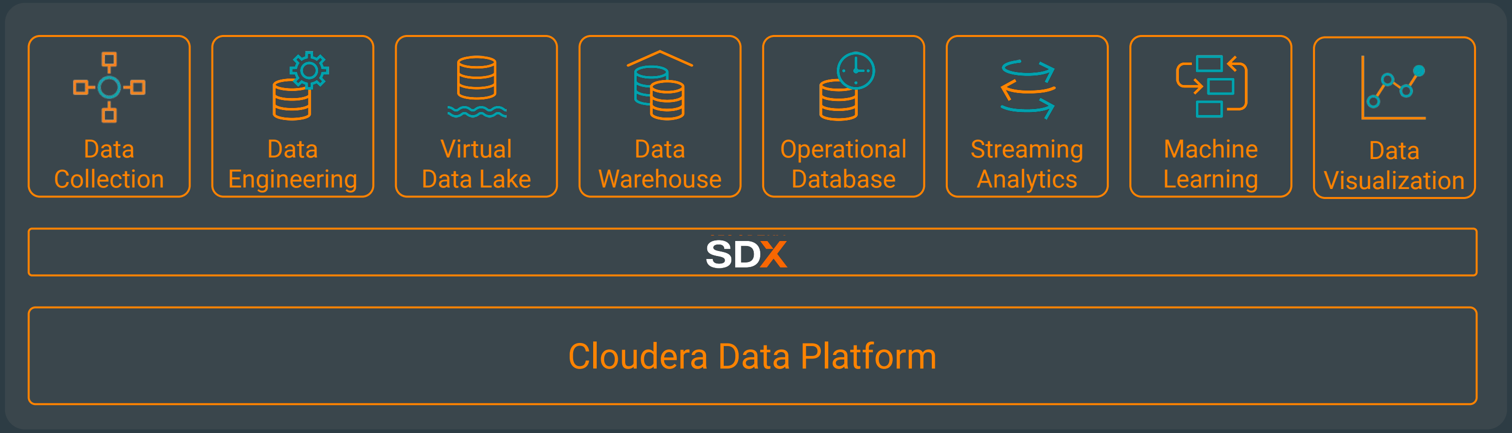


CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

A HYBRID / MULTI-CLOUD DATA PLATFORM *AND* AN INTEGRATED SUITE OF SECURE ANALYTIC APPS

Real-time
Batch
Structured
Unstructured

Data Sources



Analysts
Engineers
Scientists
Developers

Data Users

Data Lifecycle integration for better user productivity and faster time to value

Hybrid & Multi-Cloud to leverage existing investments and reduce risk

Secure & Governed to simplify data protection, sharing and compliance

Open & Extensible to support more use cases faster and at lower cost

DATA HUB CLUSTERS AND EXPERIENCES

What are the consumption options?

Data Hub Clusters

DataFlow

Data Engineering

Data Warehouse

Operational Database

Machine Learning

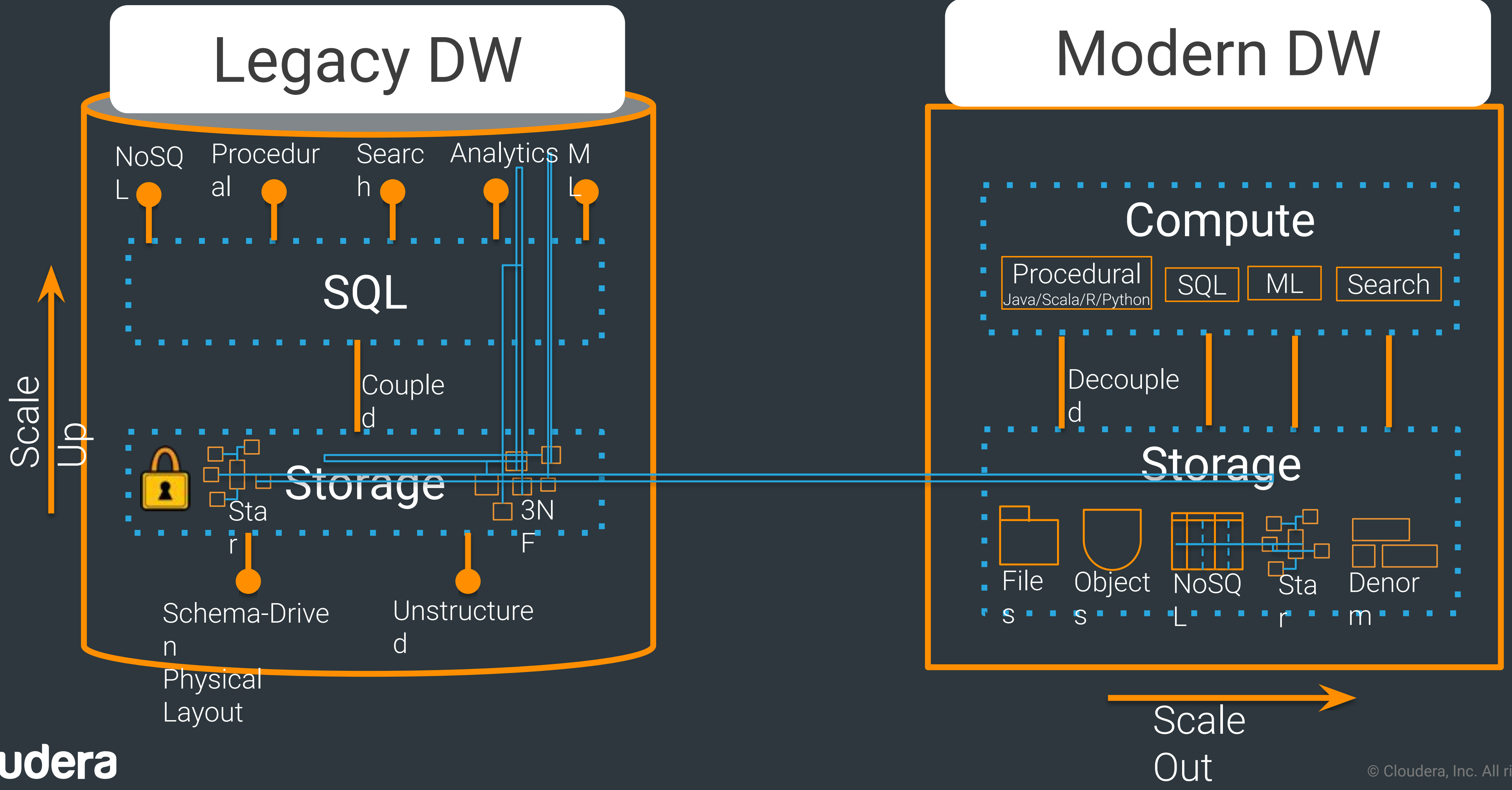
A **Data Hub Cluster** is a customizable environment that runs like a traditional Hadoop cluster, but is designed to leverage Cloud Storage.

An **Experience** is a container-based compute environment for specific purposes: ML, DW, DE, OD, DF

Cloudera Data Warehouse

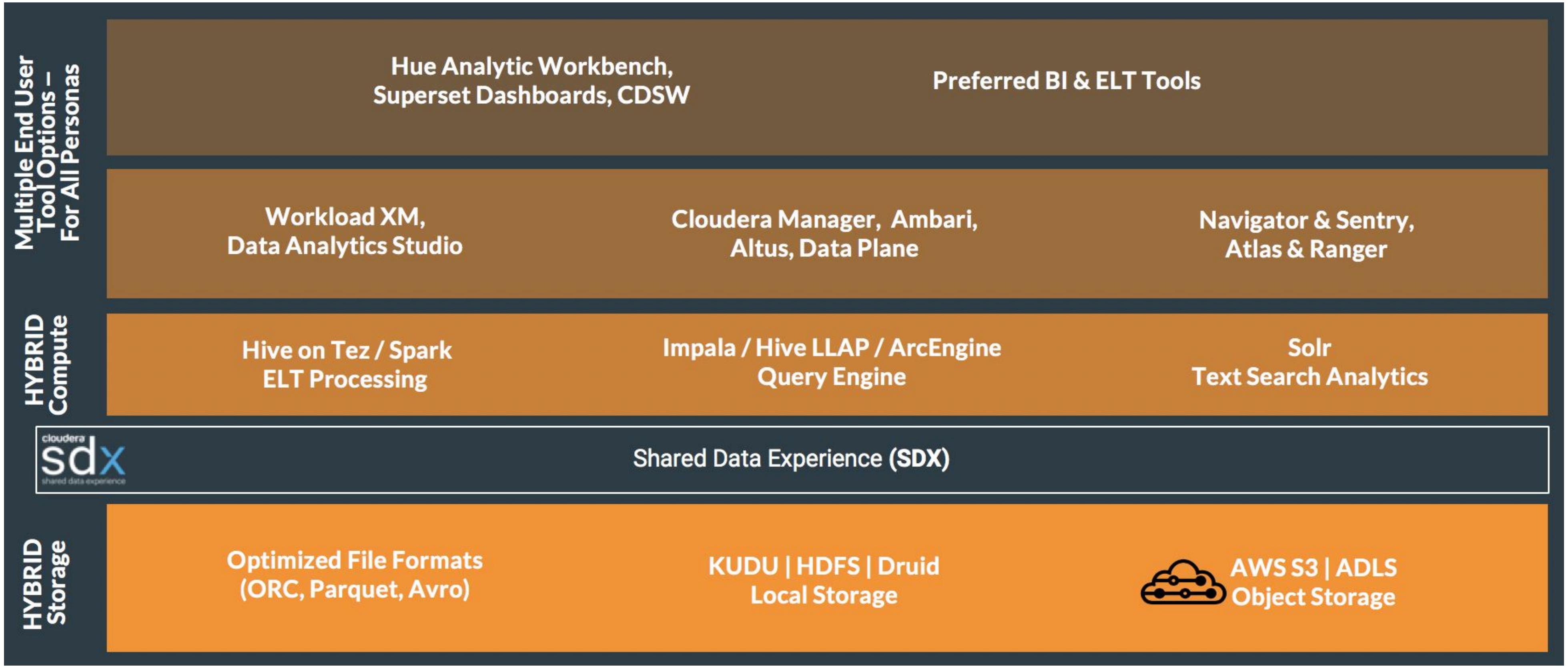
Cloudera and Modern Data Warehousing

Platform Comparison



CLOUDERA'S MODERN DATA WAREHOUSE

A well-established product that runs business critical use cases for the largest global enterprises



Multiple End User Tool Options – For All Personas

Hue Analytic Workbench, Superset Dashboards, CDSW

Preferred BI & ELT Tools

Workload XM, Data Analytics Studio

Cloudera Manager, Ambari, Altus, Data Plane

Navigator & Sentry, Atlas & Ranger

HYBRID Compute

Hive on Tez / Spark ELT Processing

Impala / Hive LLAP / ArcEngine Query Engine

Solr Text Search Analytics

cloudera sdx shared data experience

Shared Data Experience **(SDX)**

HYBRID Storage

Optimized File Formats (ORC, Parquet, Avro)

KUDU | HDFS | Druid Local Storage

AWS S3 | ADLS Object Storage

1200+
Enterprise customers

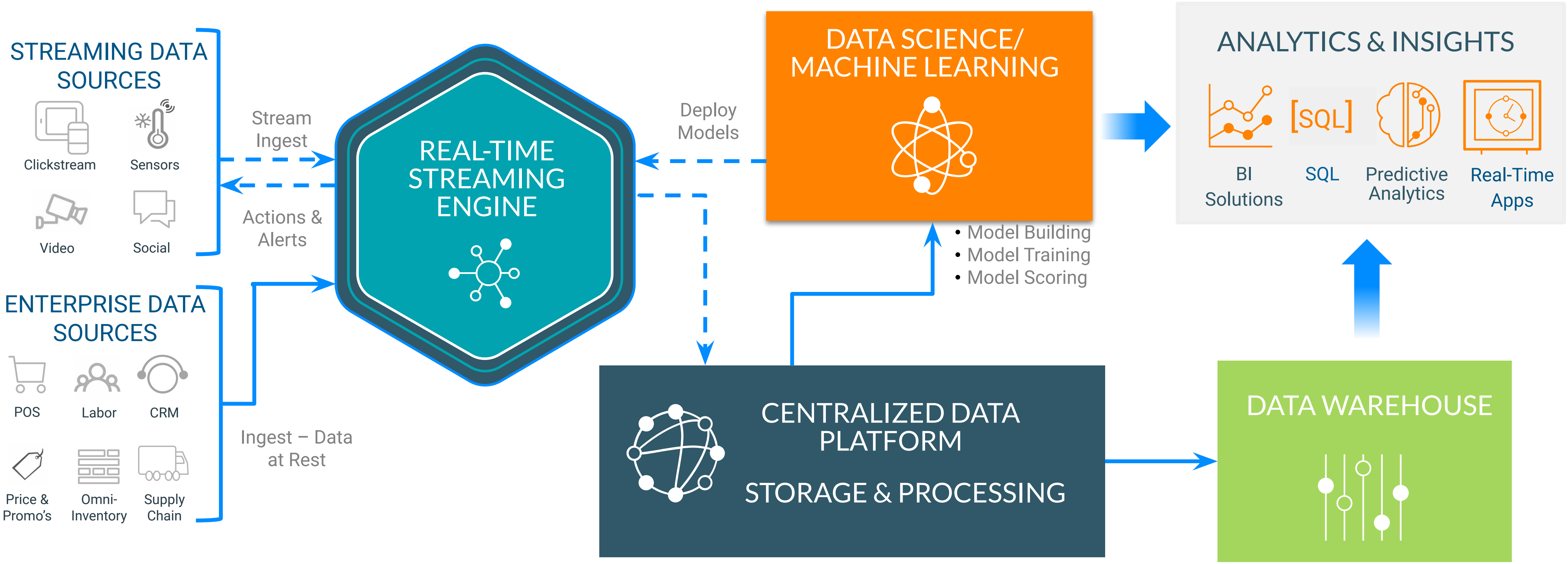
Millions
Subsecond queries daily

50K+
Nodes deployed

Gartner peerinsights customers' choice 2018 ™

ENABLING ANALYTICS & INSIGHTS ANYWHERE

Driving Enterprise Business Value



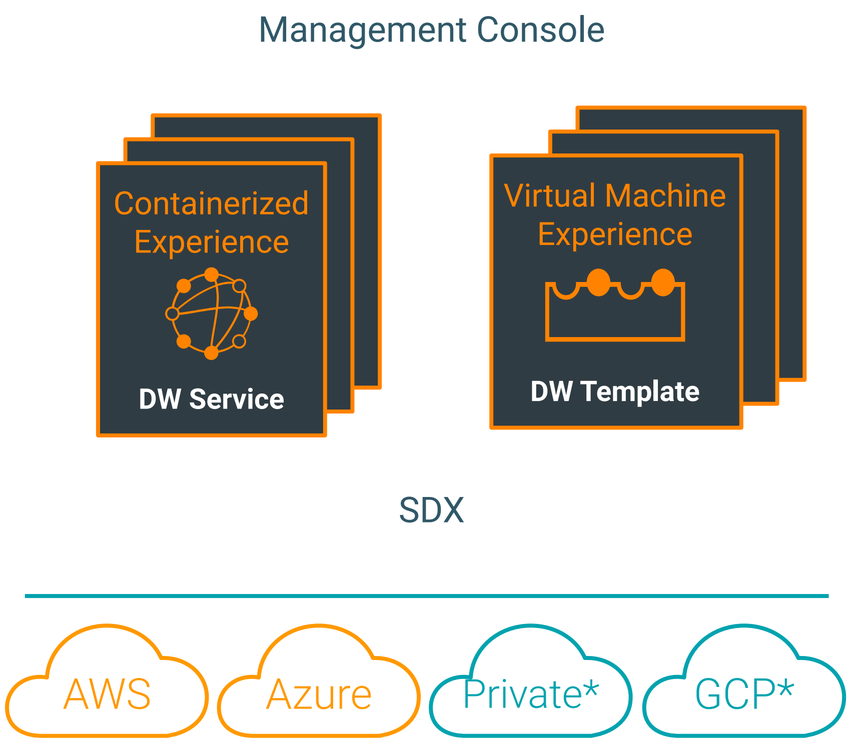
CLOUDERA

CDW is a managed data warehouse service that runs Cloudera’s **powerful engines** on a **containerized architecture** to let you **meet SLAs**, **onboard new use cases** with zero friction, and **minimize cost**

Two Cloud-Native Solutions for CDW

DW Service

- Kubernetes orchestration of container-based compute for agile clusters
- Opinionated and packaged provisioning / scaling
- Commonly administered by Line of Business
- Simplicity and ease of use over customization and control



Management Console

Containerized Experience

DW Service

Virtual Machine Experience

DW Template

SDX

AWS

Azure

Private*

GCP*

DW Template

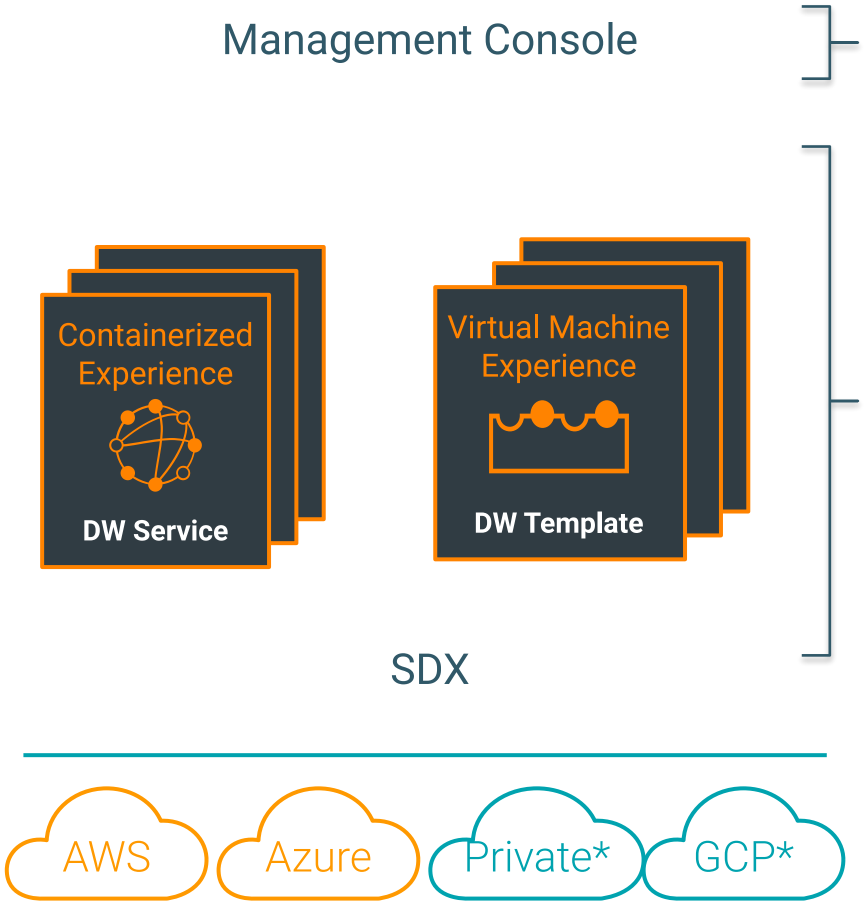
- Native VM clusters for complex long running workloads (BI, ETL)
- Bespoke and flexible provisioning / scaling
- Typically administered by Central IT
- Customization and control over simplicity and ease of use

* Future Release

COMPARISON OF SERVICE VS TEMPLATE DEPLOYMENT

	DW Service	**DW Template**
Installation	None	None
Provisioning / Deprovisioning	*Automatic*	*Manual, Automate using CDP CLI*
Sizing (in nodes)	Small (10), Medium (20), Large (40)	> 3
Node Type	Memory-Optimized, Preselected and tuned	Memory-Optimized, Customer Choice, Manual Tuning
Scaling	*Automatic (cluster size increment)*	*Manual, Automate using CDP CLI*
Caching	Results, Data, File Handle, *Warm* Compute Nodes	Results, Data, File Handle
Suspend/Resume	*Automatic*	*Manual, Automate using CDP CLI*
Multiple Database Catalogs	*Allowed*	*Not Allowed*
Shared Data Experience	Inherited from CDP - Ranger, Knox, Atlas, HMS	Inherited from CDP - Ranger, Knox, Atlas, HMS

COMPONENTS THAT ENABLE END-TO-END DW WORKFLOWS



Management Console: Single pane of glass for platform-wide admin

DW Template: Pre-built Virtual Machine service

DW Service: Pre-built Containerized service

SQL Workbench: A SQL tool for exploring, querying, editing

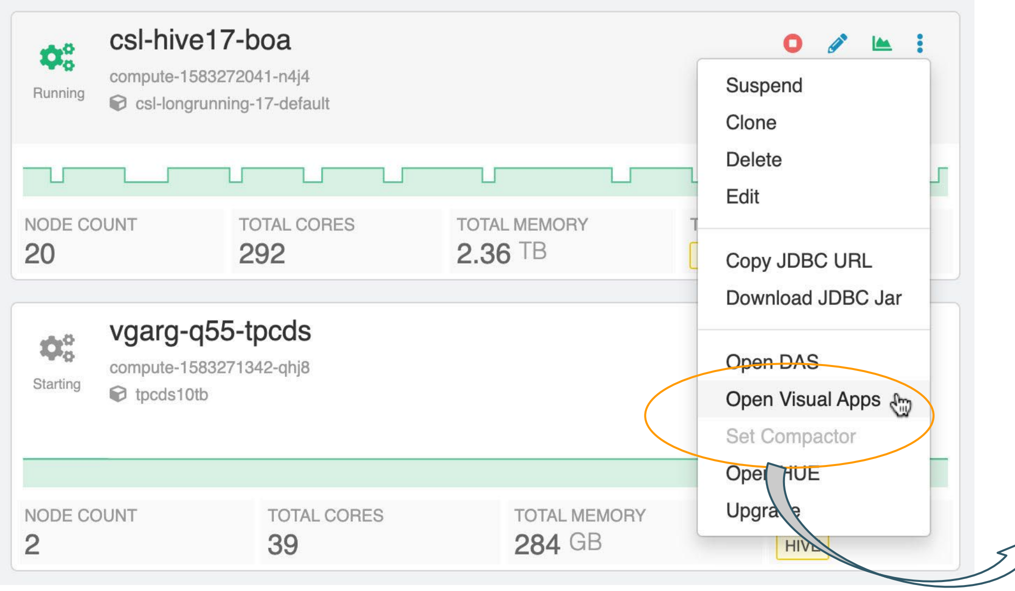
Workload Manager: Analyze and optimize SQL workloads

SQL Engines: Parallel processing engines for many analytics patterns

- **Hive** - Ideal for Dashboards, Complex Reports, Long-Running Queries
- **Impala** - Ideal for Ad Hoc Queries, High Concurrency Self Service
- **Druid+Hive** - Ideal for Time Series with Pre-Defined Aggregations
- **Kudu+Impala** - Ideal for Event Analytics, Time Series with Ad Hoc Queries

* Future Release

DW Viz in CDW



csl-hive17-boa

Running

compute-1583272041-n4j4

csl-longrunning-17-default

NODE COUNT	TOTAL CORES	TOTAL MEMORY
20	292	2.36 TB

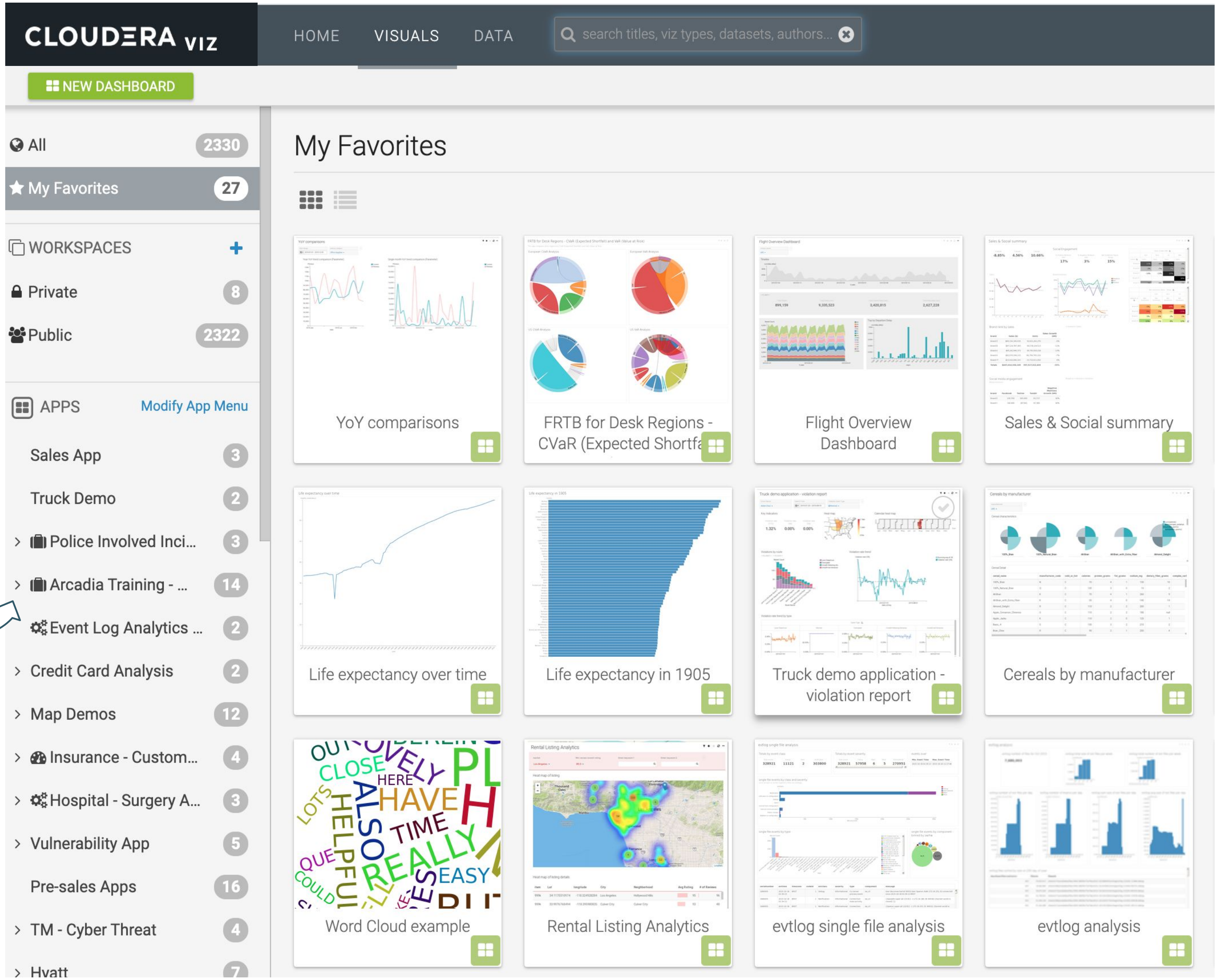
vgarg-q55-tpcds

Starting

compute-1583271342-qhj8

tpcds10tb

NODE COUNT	TOTAL CORES	TOTAL MEMORY
2	39	284 GB

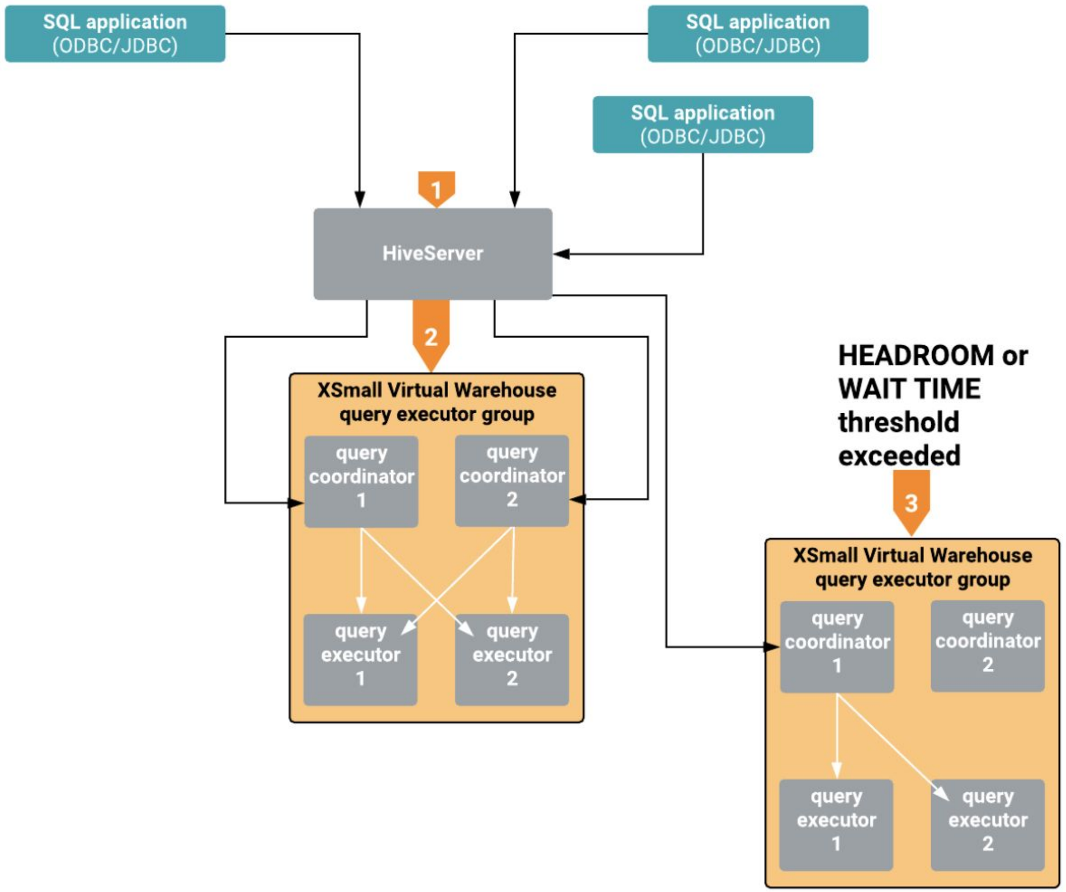


My Favorites

All	2330
My Favorites	27
Private	8
Public	2322
Sales App	3
Truck Demo	2
Police Involved Inci...	3
Arcadia Training - ...	14
Event Log Analytics ...	2
Credit Card Analysis	2
Map Demos	12
Insurance - Custom...	4
Hospital - Surgery A...	3
Vulnerability App	5
Pre-sales Apps	16
TM - Cyber Threat	4
Hyatt	7

- YoY comparisons
- FRTB for Desk Regions - CVaR (Expected Shortfa
- Flight Overview Dashboard
- Sales & Social summary
- Life expectancy over time
- Life expectancy in 1905
- Truck demo application - violation report
- Cereals by manufacturer
- Word Cloud example
- Rental Listing Analytics
- evtlog single file analysis
- evtlog analysis

Query Placement Logic within a Virtual Warehouse - Hive


SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
SQL application (ODBC/JDBC)
1
HiveServer
2
XSmall Virtual Warehouse query executor group
query coordinator 1
query coordinator 2
query executor 1
query executor 2
HEADROOM or WAIT TIME threshold exceeded
3
XSmall Virtual Warehouse query executor group
query coordinator 1
query coordinator 2
query executor 1
query executor 2

HiveServer

- Receives all queries
- Always on
- Builds initial plan

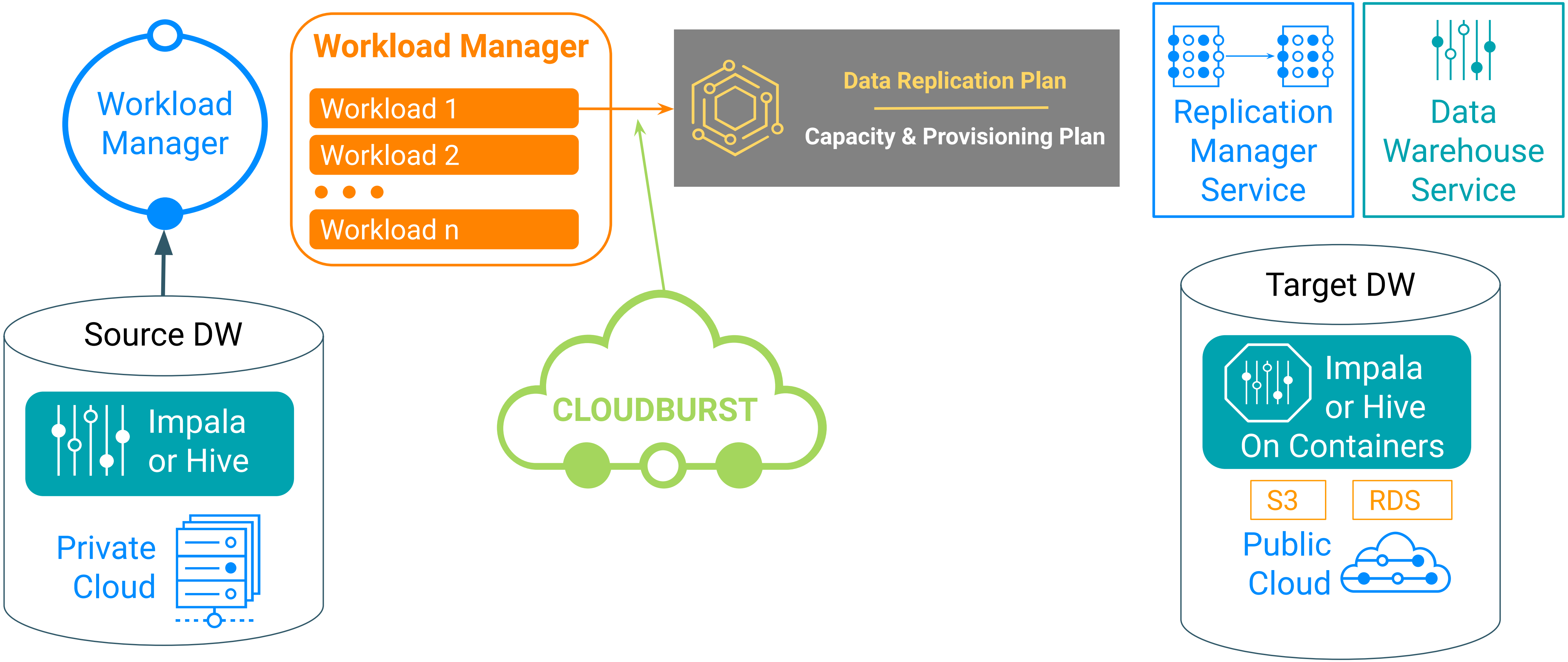
query coordinator

- Builds final plan
- Coordinates with executors
- One query at a time
- Determines parallelism

query executor

- Executes query
- Participate in up to 12 queries at a time
- Determines throughput

DATA WAREHOUSE SETUP



CDP on AWS - Security Overview

Feature Differentiators

Ranger:

- Consistent & Comprehensive Authorization Model Across Ecosystem
- Scalable & Authoritative Audits
- Dynamic Security Policies
- Robust Data Protection (TDE, Masking, 3rd Party….)

Knox:

- Flexible & Scalable Authentication Patterns (REST Proxying, AuthN Federation, SSO…)

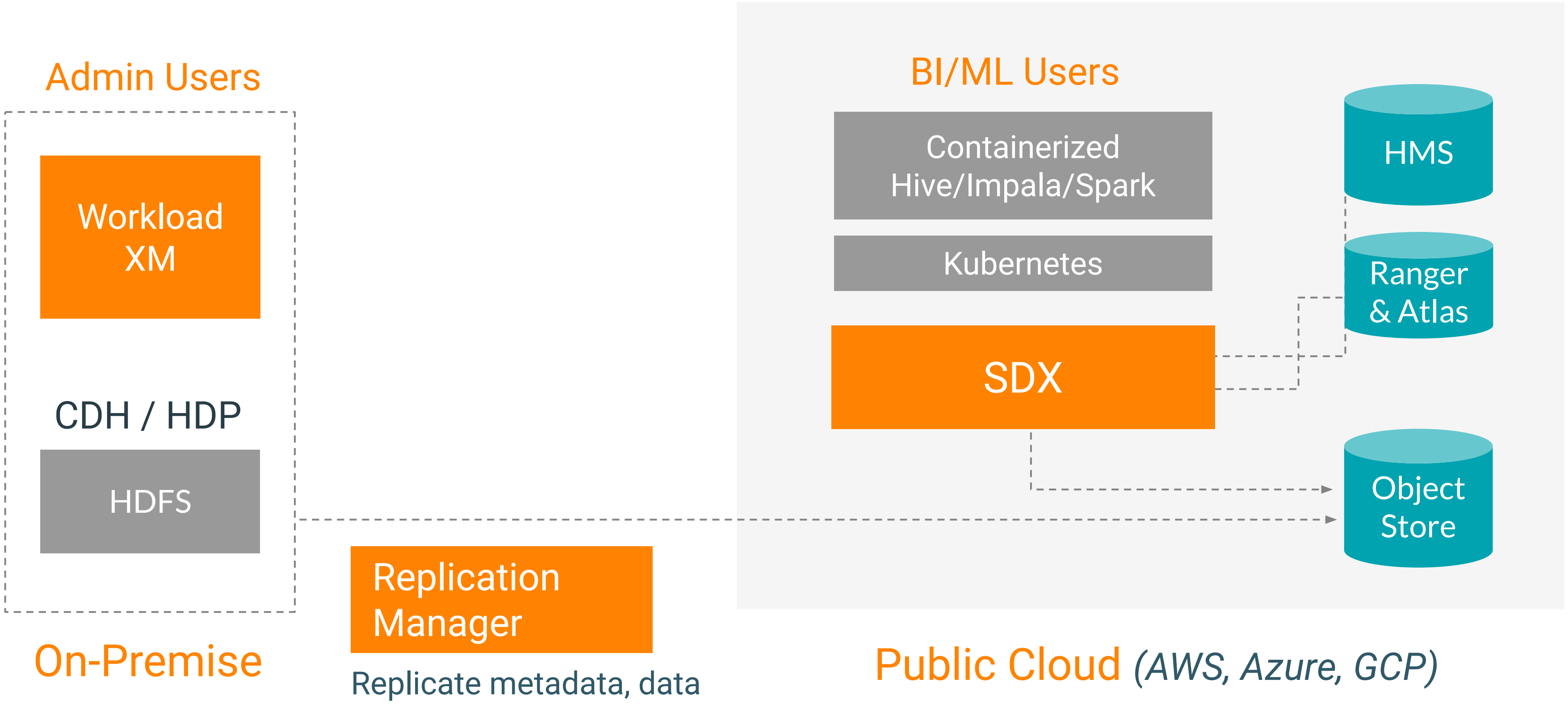
NavAtlas & Ranger:

- Fine-Grained Security via ABAC Model
- Classification Based Security

Hive & Ranger:

- Dynamic Row Filtering and Dynamic Column Masking @ Scale

WORKLOAD ANALYSIS, REPLICATION & SDX MAKE IT EASY TO TAKE NEW WORKLOADS TO THE CLOUD



CLOUD DATA WAREHOUSE PERFORMANCE TESTING

Cloudera Delivers Better Price Performance

Industry standard TPC Benchmark

20% lower costs than Amazon Redshift

40% lower costs than Microsoft Synapse

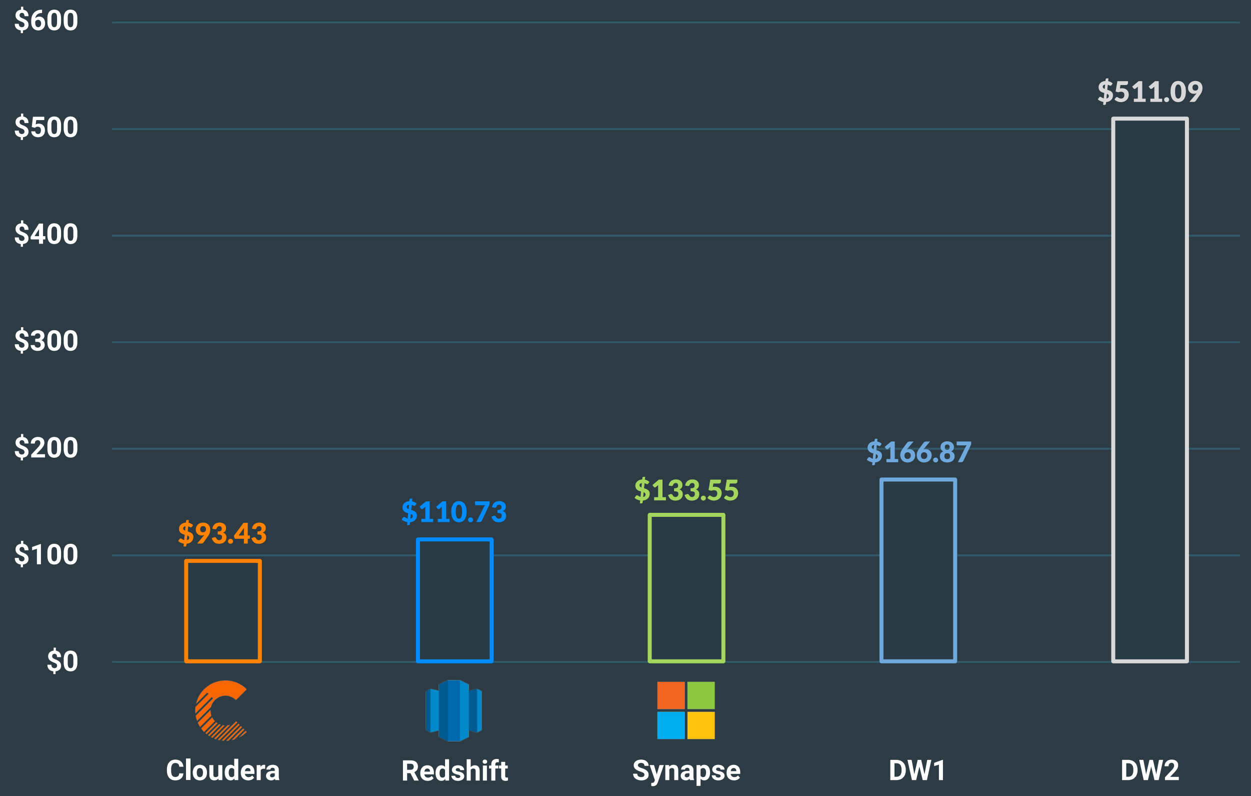
80% lower costs than “DW1”

550% lower costs than “DW2”

More can be learned about the TPC-DS benchmark at http://www.tpc.org/tpcds/.

Cloud Data Warehouse Performance Testing - January 2021

Price-Performance Comparison (Lower is Better)



Prepared by: McKnight Consulting Group, www.mcknightcg.com January 2021

ADVANTAGES OF A MODERN DATA WAREHOUSE

Data Flexibility

- Storage of **ALL data** types
- **Iterative modeling** and **self-service** accessibility
- Portability: **No proprietary** formats or storage **lock-in**

Cost-Effective Scalability

- **Elastic scale** in any environment
- Cloud-native integration for pay-per-use cost
- Proven at **massive scale**
- Workload introspection to reduce cost and improve performance

Go Beyond SQL

- Consolidate data silos with open architecture
- **Shared data** across SQL & Non-SQL workloads
- Choice for the right processing & analytics tool for the right job

Hybrid Decoupled Architecture

- Runs across multiple public clouds & on-premise for **no lock-in**
- Multiple storage options: HDFS, Kudu, Druid; S3, ADLS

Workshop

- CDW Setup
- Data Exploration – Using DAS
  - Create External & Managed Tables
  - Materialized Views
- Auto-scaling, Auto-suspend & Caching in CDW
- Security & Governance with Ranger & Atlas
- Cloudera Data Visualization with CDW

THANK YOU

CLOUDERA