



cloudera®

Modernizing Business Intelligence and Analytics

Justin Erickson
Senior Director, Product Management

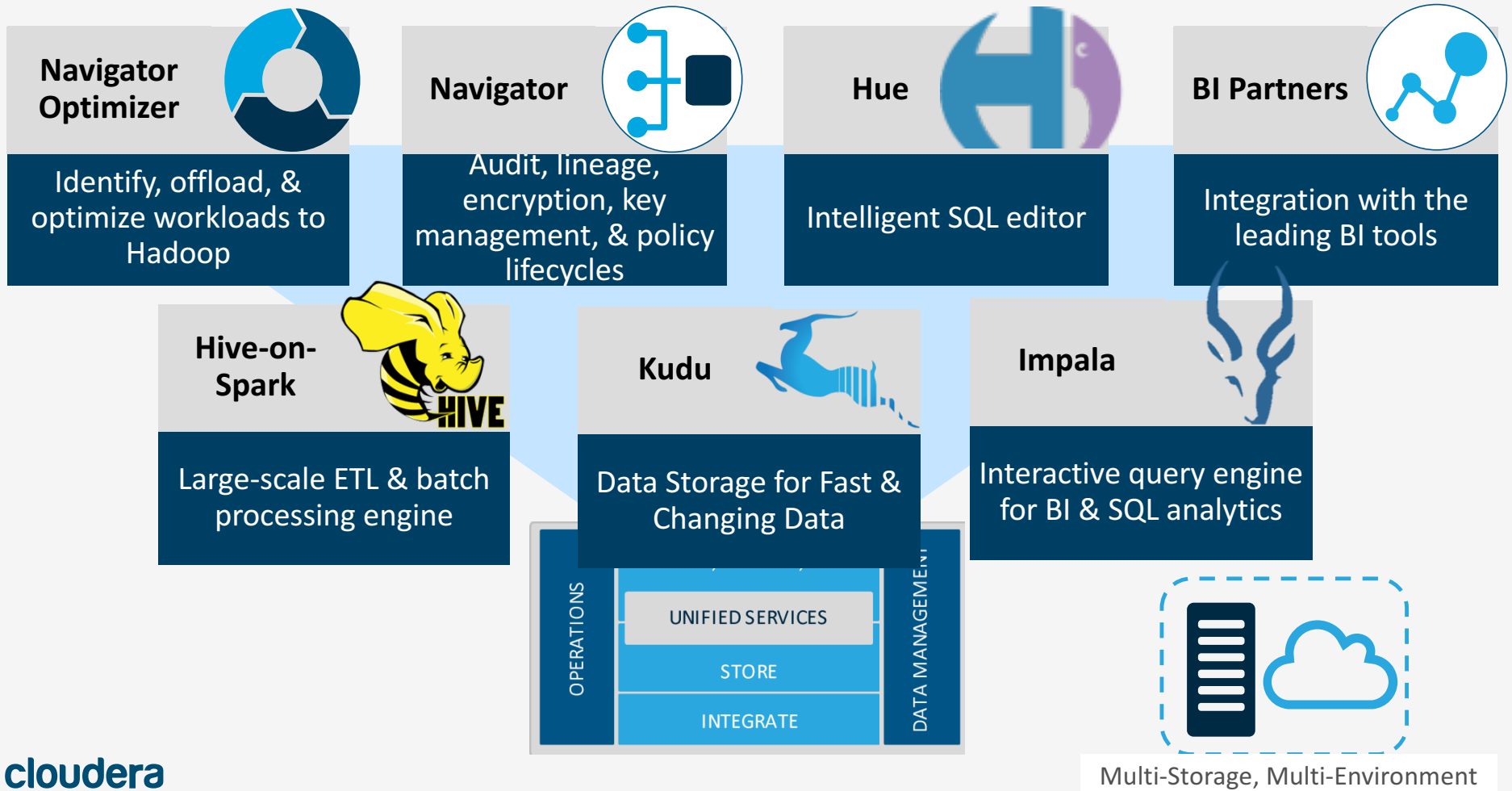
Agenda

- What benefits can I achieve from modernizing my analytic DB?
- When and how do I migrate from current systems?
- How does it work in the cloud?

Key Applications

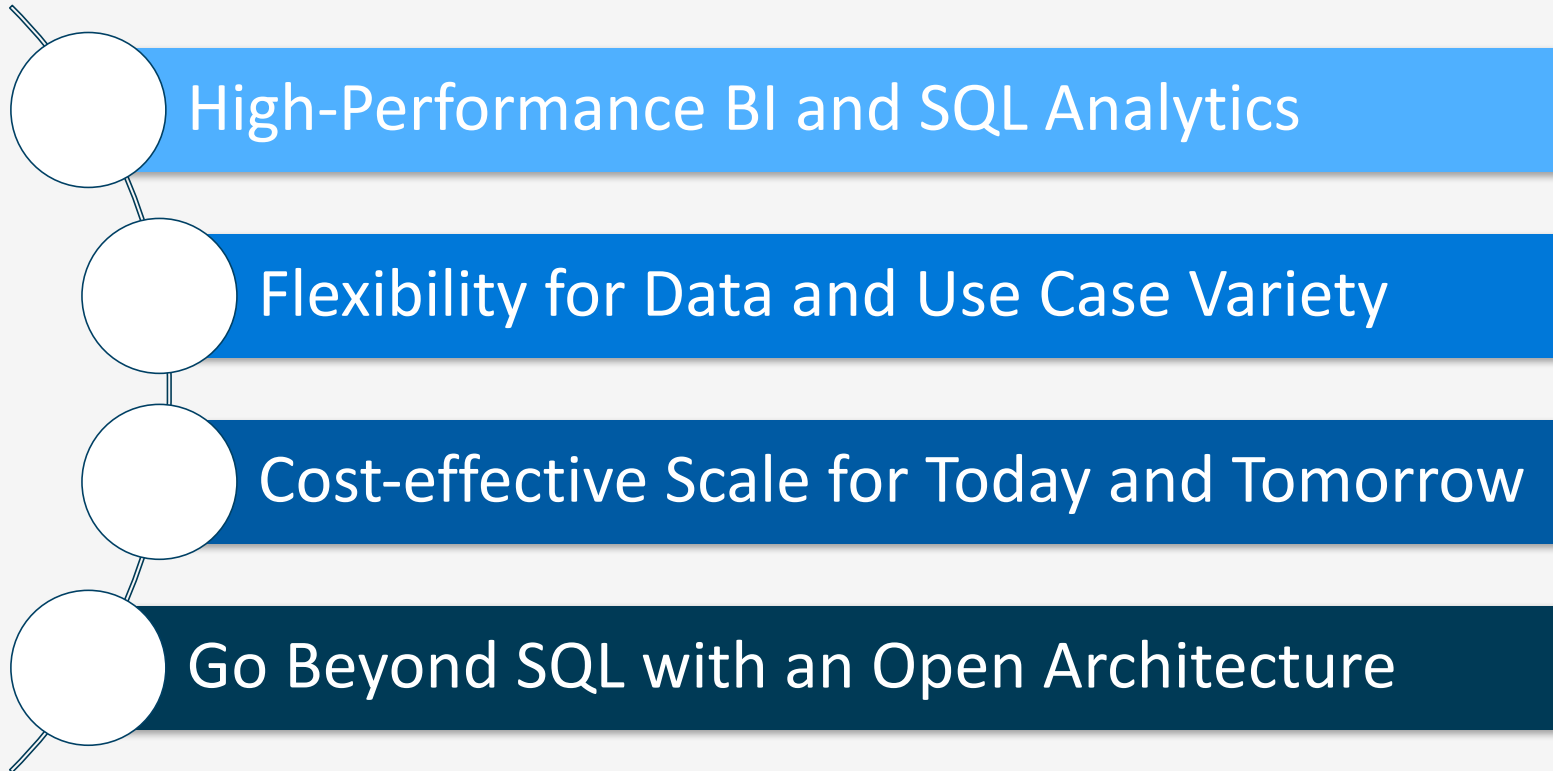


Cloudera's Analytic Database



Key Benefits

An analytic database designed for Hadoop



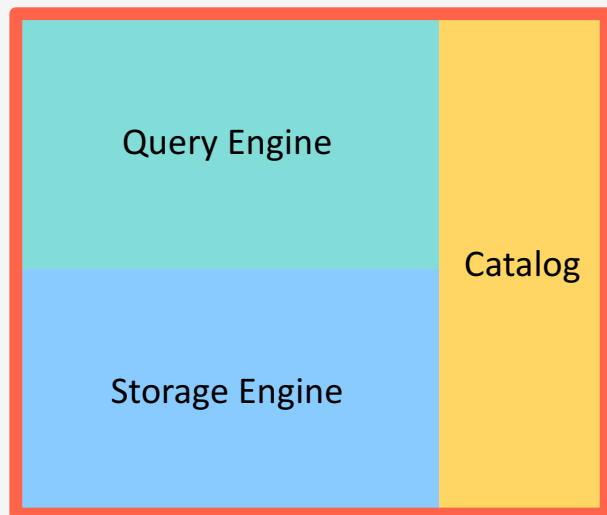
Analytic DB Anatomy

Built for self-service and hybrid cloud

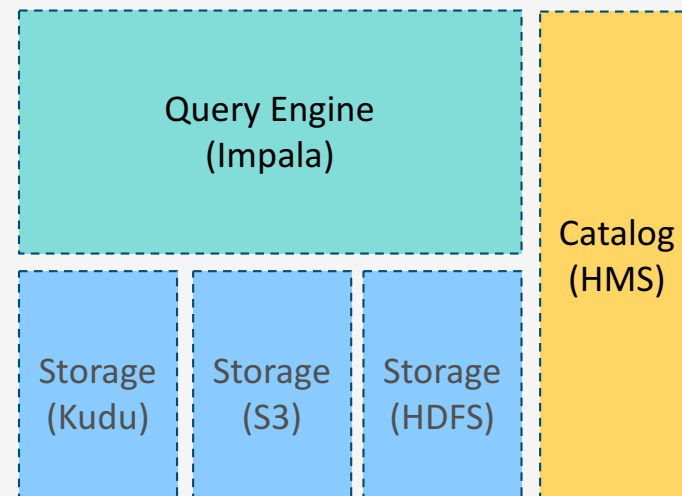
Anatomy of an Analytic Database

Cloudera Decoupled by Design

Monolithic Analytic Database



Modern Analytic Database



Pain Points

Traditional Monolithic Analytic Databases



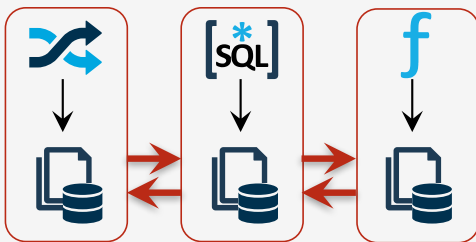
Rigid Data Model

- Tightly coupled storage and compute



Static Sizing

- Major maintenance to add capacity/nodes



Limited to SQL only

- Maintain data copies for non-SQL



Poorly Designed for Cloud

- No elasticity or integration with object storage

Benefits of Cloudera's Modern Approach

Cloud-Native & On-Premise



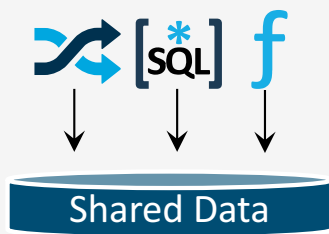
Data Flexibility

- Faster, more agile data acquisition
- Data portability: Open formats and open storage



Cost-Effective Scalability

- Elastic scale on-prem or in the cloud
- Cloud-native pay-per-use and transience
- Proven at big data scale



Go Beyond SQL

- Open Architecture: Open formats and open storage
- Shared data across SQL and non-SQL workloads



Hybrid

- Runs across multi-cloud & on-prem
- Multi-storage over S3, HDFS, Kudu, Isilon, DSSD, etc

EDW Optimization

Expand the Value of Your Data Warehousing Landscape

Motivations for Optimizing the EDW



Cost containment for existing workloads
Limited budget for expansion



Unable to take on new workloads
Unable to keep up with changing business needs

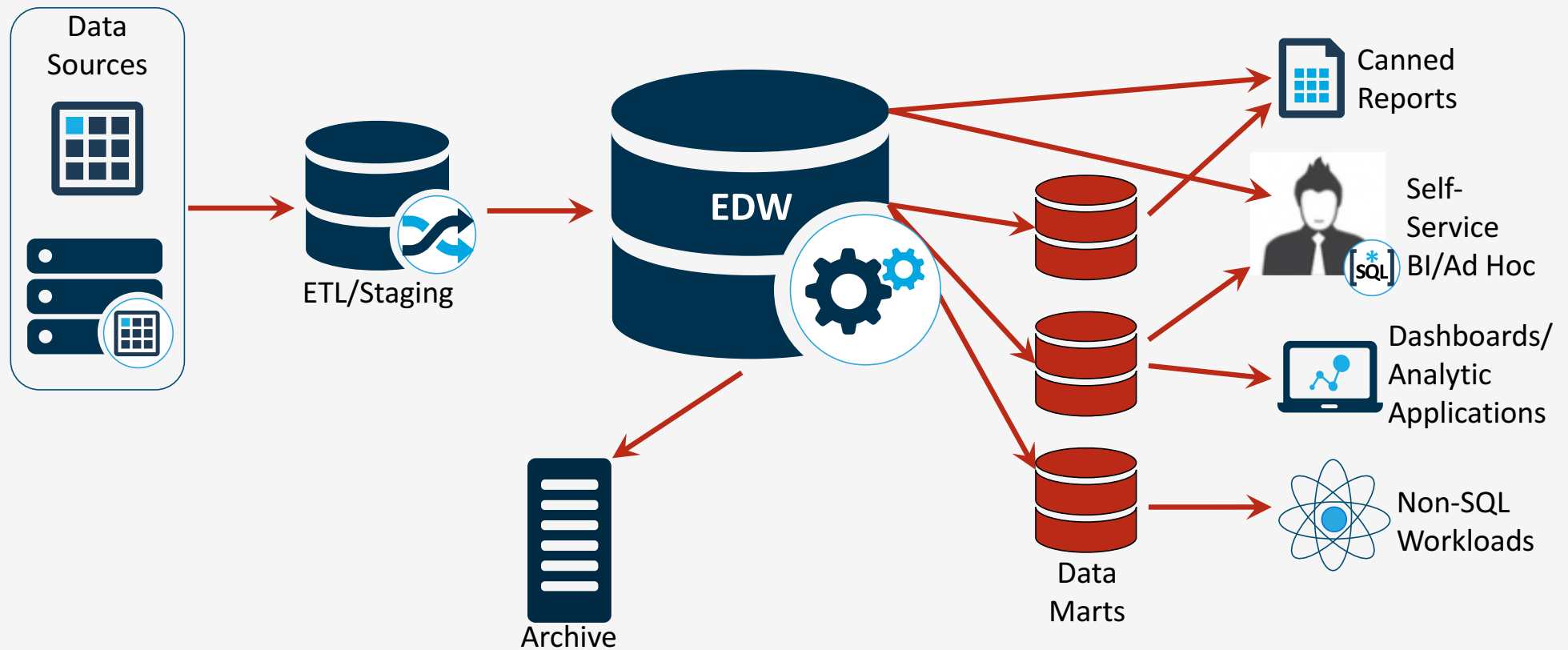


Difficulty handling both fixed-SLA reports and self-service exploration



Growing importance of self-service BI, advanced analytics, and cloud

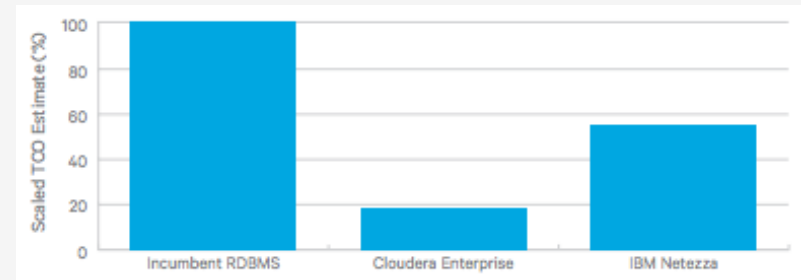
Existing EDW Landscape



Optimizing the EDW with Cloudera

- Cost-Effective Scale
 - Say yes to more without the risk
- Go Beyond SQL
 - Exploration, advanced analytics, and more all in one platform
- Modernize the Data Warehouse Landscape
 - Maximize the EDW while enabling iterative, self-service access/BI
 - Well-suited for on-prem, cloud, and hybrid deployments

SIEMENS 90% less per TB vs RDBMS and 75% less vs Netezza

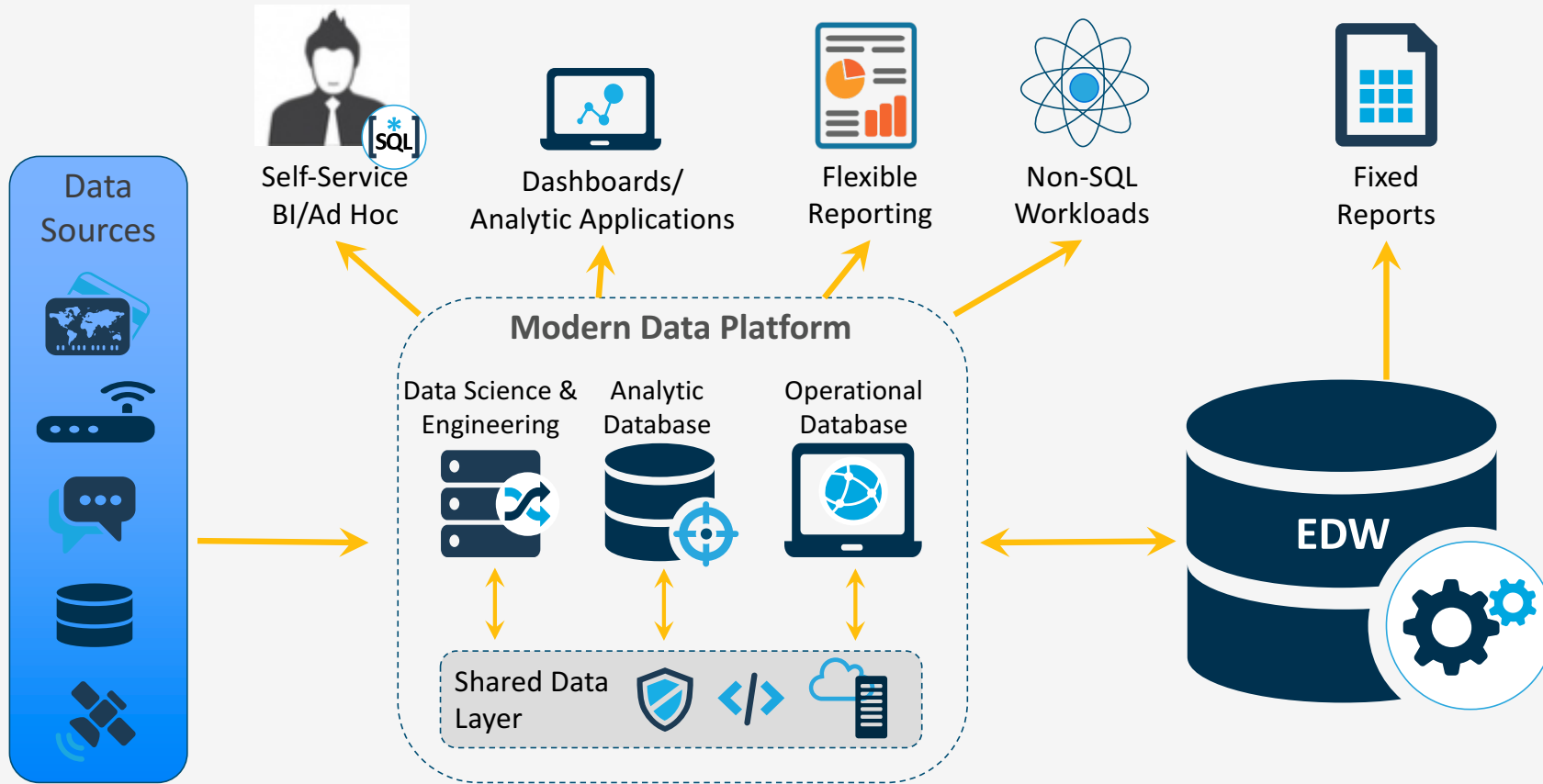


Augmented its Oracle EDW with **multi-tenant** Cloudera system with their BI tool configured to allow users to **pull reports from both**



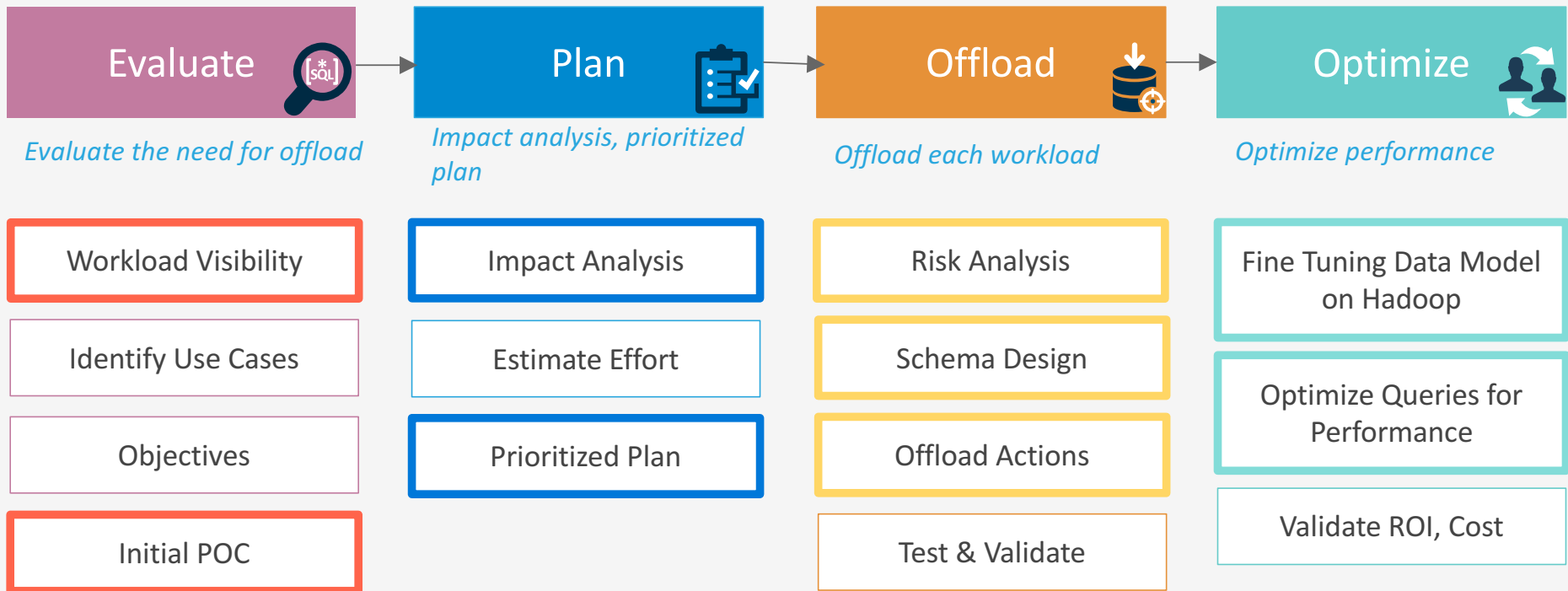
Saved **tens of millions** by offloading DBMS to Cloudera in the **cloud**

Modern Data Warehouse Environment



Navigator Optimizer

Built to help you through the optimization process



Workload Visibility

Get insights into what's happening today

Evaluate Queries

- Top queries
- Query duplication
- Query complexity
- Common access patterns

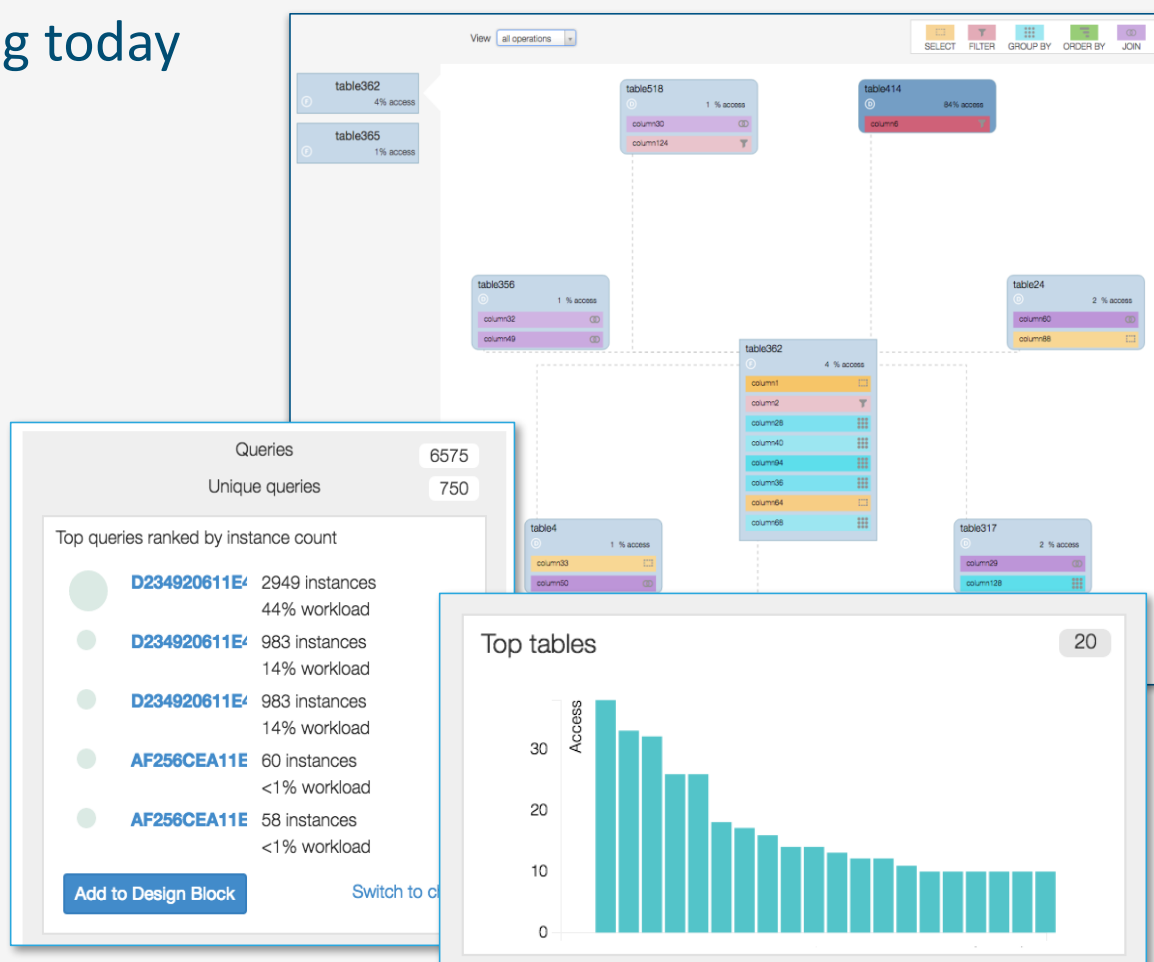
Evaluate Data Access

- Top tables, top columns
- Usage-based ER diagram
- All tables/columns in use

Evaluate POC

- Identify initial workload piece for PoC
- Get partitioning key suggestions

cloudera



Impact Analysis & Prioritized Plan

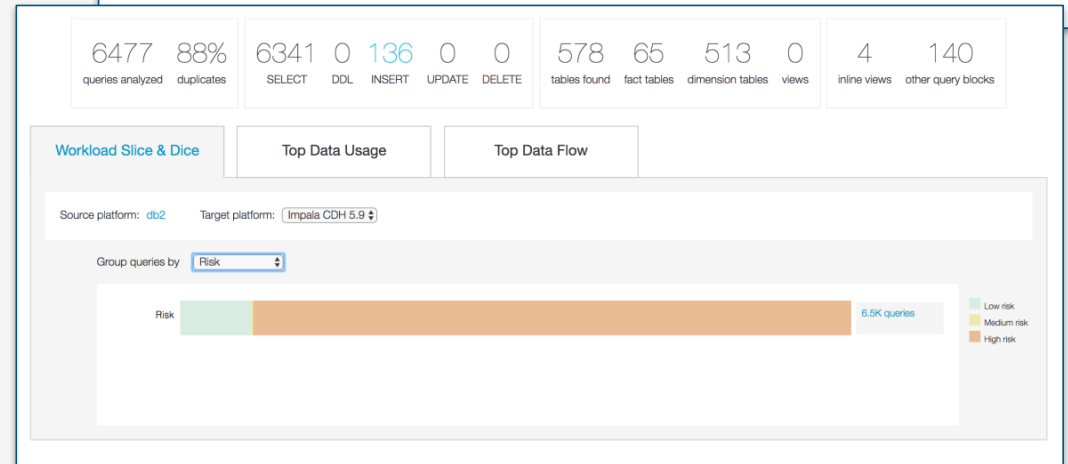
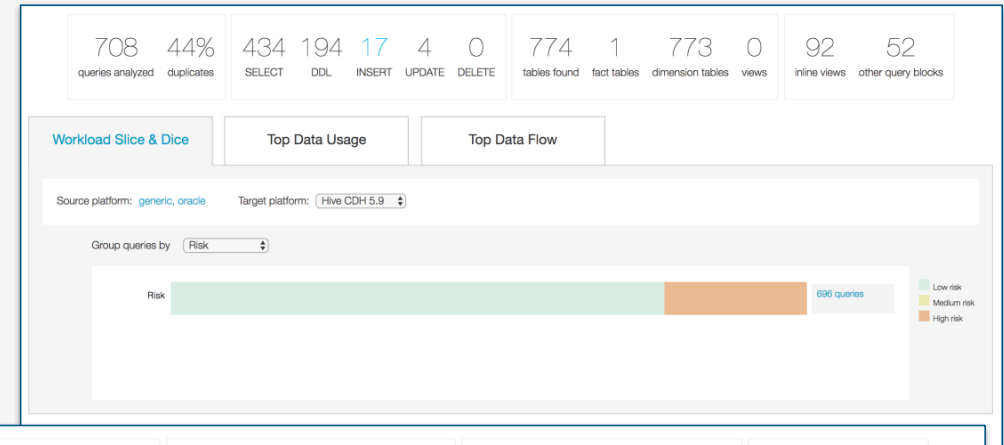
Understand what it takes to offload

Impact Analysis

- Focus efforts by identifying duplication
- Workload risk assessment based on complexity and best practices
- Understand query compatibility

Prioritized Plan

- Estimate effort
- Identify easiest pieces to start for fast success
- Prioritize workloads for offload



Predictable Offload

Remove the guesswork

Understand offload requirements

- Determine most common workload patterns
- Develop data-/usage-driven offload strategy

Actionable recommendations

- Complexity assessment for riskier areas
- Focus efforts by identifying duplication
- Design recommendations for best results

The screenshot shows the 'Evaluate' tool interface. On the left, a table lists 13 unique queries with their IDs and risk levels. On the right, a detailed view shows 'Tables accessed' (55) and 'Evaluate risk' information, including a high risk alert with 7 items.

Select	Query ID	Risk
<input checked="" type="checkbox"/>	128049	Low
<input checked="" type="checkbox"/>	126883	Medium
<input checked="" type="checkbox"/>	126737	Medium
<input checked="" type="checkbox"/>	126758	Medium
<input checked="" type="checkbox"/>	126781	Medium
<input checked="" type="checkbox"/>	126813	Medium
<input checked="" type="checkbox"/>	126682	High
<input checked="" type="checkbox"/>	127030	High
<input checked="" type="checkbox"/>	127768	High
<input checked="" type="checkbox"/>	127769	High
<input checked="" type="checkbox"/>	127931	Low
<input checked="" type="checkbox"/>	126681	High
<input checked="" type="checkbox"/>	128132	High

High risk alert (7 items):

- >5 table joins or >10 join conditions found. (3 unique queries)
- High cardinality GROUP BY column found. (1 unique queries)
- >50 query blocks present in large query. (1 unique queries)
- >10 Inline Views present in query. (1 unique queries)
- Query has no filters. (2 unique queries)
- Unsupported commands: UPDATE or DELETE. (1 unique queries)

The 'Optimization Recommendation' panel provides the following suggestions:

- 7.8K queries can benefit from a new partitioning strategy on 1.3K tables. [Design Partition Key](#)
- 261 queries can be benefit from denormalization. [Denormalize Tables](#)
- 416 queries can benefit by inline view materialization. [Materialize Inline Views](#)
- 7.4K queries can benefit from table aggregation. [Aggregate Tables](#) [Download](#)

Optimizing within Hadoop

Maintain peak performance

Understand usage and keep up with data needs

- Understand most common usage patterns
- Identify optimization opportunities
- Proactively adjust data models

Performance optimizations

- Best practice guidance for Hive and Impala
- Query performance optimization
- Increase platform adoption

Tables accessed **774**

Evaluate risk

Table volume and column statistics are not detected. To get accurate risk alerts, upload table volume and column statistics.

High risk alert **1**

>=5 table joins or >=10 join conditions found. [111 unique queries](#)

Medium risk alert **6**

Query has no filters. [686 unique queries](#)

>=50 query blocks present in large query. [4 unique queries](#)

>=10 Inline Views present in query. [19 unique queries](#)

Query with inline views found. [26 unique queries](#)

Cartesian or CROSS join found. [9 unique queries](#)

>=10 columns present in GROUP BY list. [11 unique queries](#)

Optimization Recommendation

7 queries can benefit from table aggregation. [Aggregate Tables](#)

93 queries can benefit from denormalization. [Denormalize Tables](#)

Built for hybrid cloud

What's Driving Analytics to the Cloud?

Big data deployments in cloud are accelerating:

- Executive Mandate: Minimize on-prem datacenter footprint
- Increased Agility: End-user self-service
- Elasticity: Optimize infrastructure usage
- Lower Overall TCO



Most Organizations Are or Will be Hybrid Cloud

- 76% will embrace hybrid cloud (Gartner¹)
- 82% will have a multi-cloud strategy (RightScale²)
- 50% will “repatriate” at least one public cloud workload back to private cloud or on-prem for cost reasons (451³)
- 50% of Cloudera’s cloud customers run a hybrid environment

¹Gartner, *Market Trends: Cloud Adoption Trends Favor Public Cloud With a Hybrid Twist 2015*

²RightScale *2016 State of the Cloud Report*

³451 Research: *AWS Lambda: new and exciting, old and rehashed, more vendor lock-in (or all the above)?, November 22, 2016*

Why is this a critical strategy?

Portability & Cost

Functionality

Data Gravity

Cost-Efficiencies & Flexibility in the Cloud

Primary Analytic Database Patterns

Reduce Operating Costs

ETL




Only pay for what you need,
when you need it

- Transient clusters
- Object storage centric
- Cloud-native deployment




New Insights, New Revenue

BI/Analytics



Explore and analyze all data,
wherever it lives

- Long-running clusters
- Object storage or local storage
- Lift-and-shift deployment



Additive Benefits in the Cloud

Extending core performance, flexibility, scalability, and open architecture benefits



Predictable Results Whenever You Want

- Consistent query performance, even during peak times
- Multi-tenancy via isolated clusters on shared data



Add Use Cases, Analytics, and Data On-Demand

- Avoid the IT backlog with instant access to all data
- On-demand clusters query directly on shared object storage



Contention-Free ETL

- ETL anytime without impacting other workloads or risking SLAs
- Separate ETL clusters as-needed on shared data



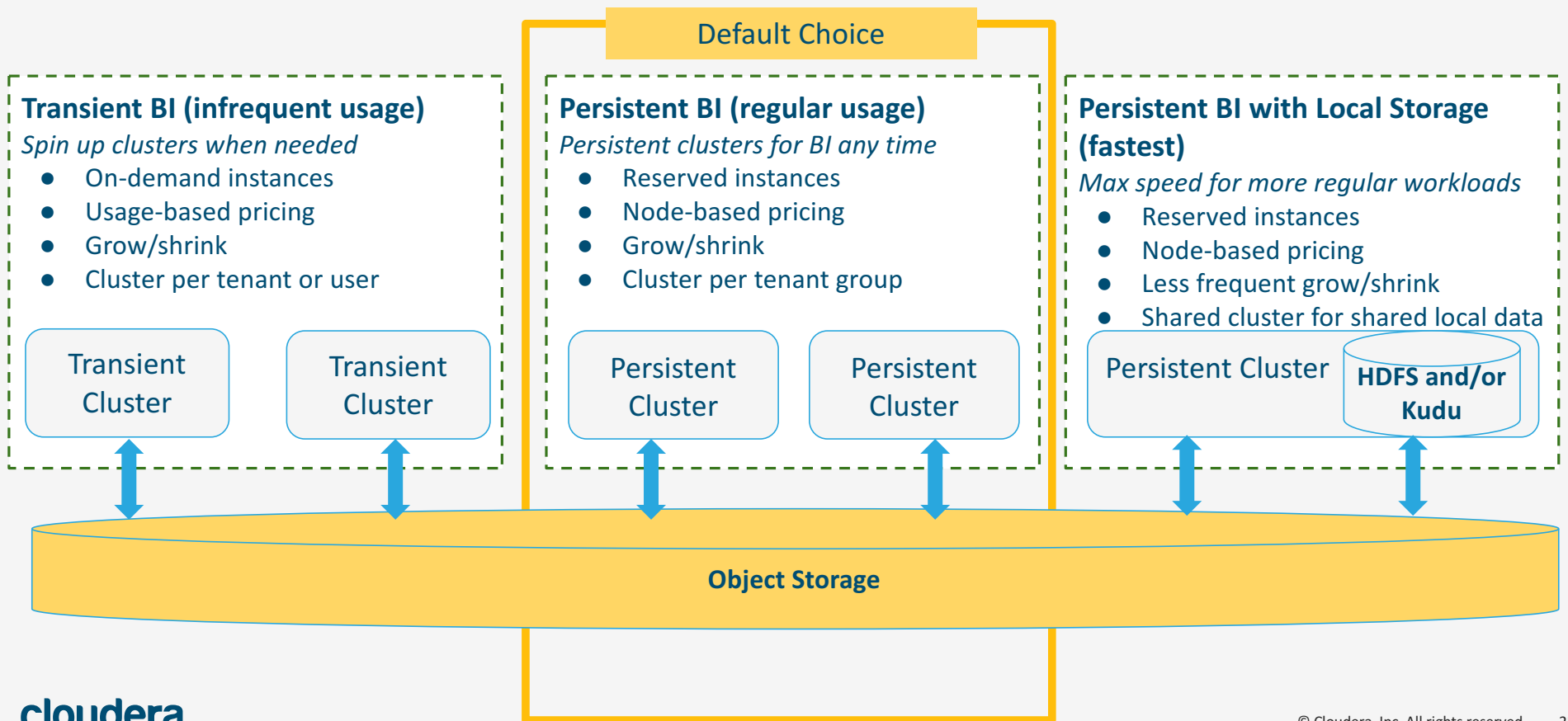
Just-in-Time Resources

- Real-time capacity for your needs, as they change
- Elastically grow/shrink your cluster via decoupled architecture

cloudera

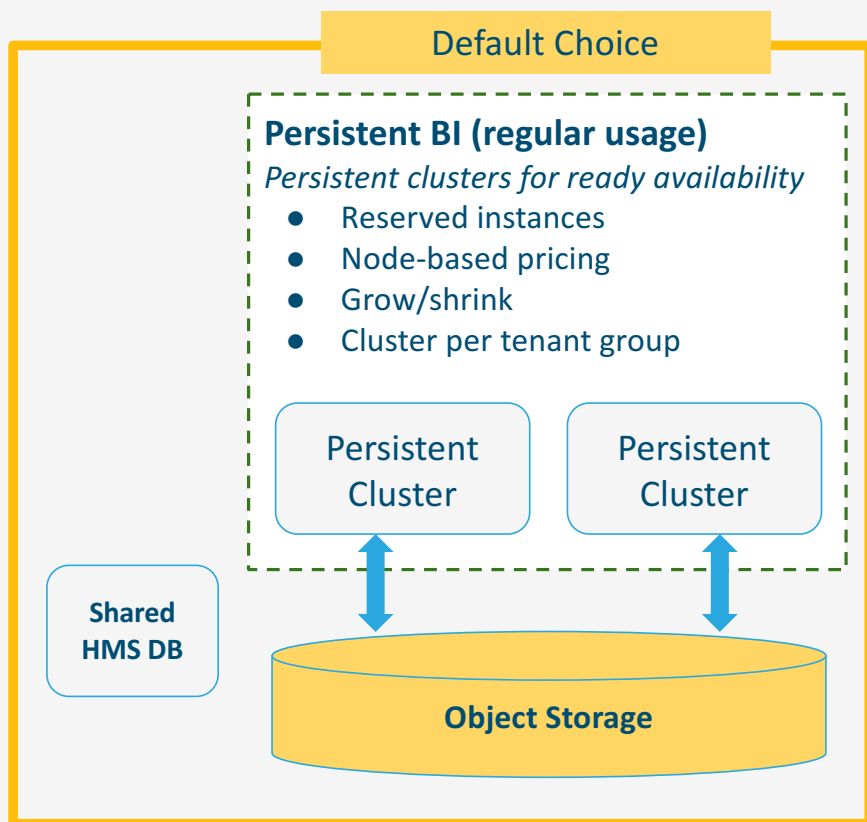
BI/Analytics in the Cloud

Three Architectures Options to Optimize Price/Performance



Persistent BI on Object Storage

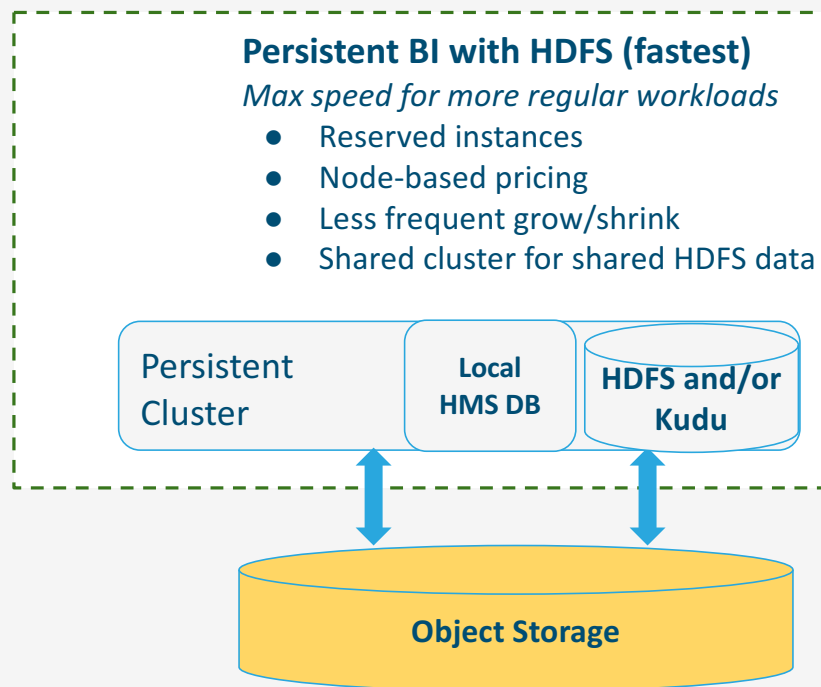
Best for elasticity (and speed vs transient)



- **This is usually the best choice**
- Best when workloads are:
 - Flexible and changing
 - Frequent during most working days
 - Not scheduled for fixed hours
- Benefits include:
 - Predictable results readily available
 - Full multi-tenant isolation
 - Common data in shared object storage
 - Grow/shrink for TCO efficiency
- Tradeoffs:
 - Per node perf of object storage (use more, cheaper nodes)

Persistent BI with Locally Attached Storage

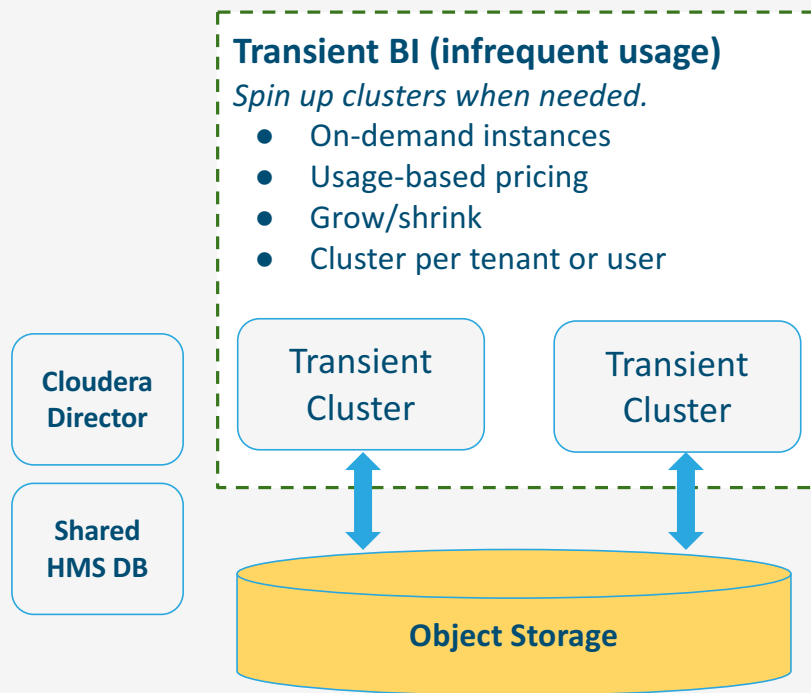
Best performance for consistent workloads



- Best when workloads are:
 - Regular and consistent
 - Consistently querying common data
 - Tight SLAs for performance
 - Fast changing data (that needs Kudu)
 - Running without object storage (eg. Azure, GCE)
- Benefits include:
 - Faster performance per node on local data
 - Ability to query object storage for rest of data
- Tradeoffs:
 - Less elastic than object stored based clusters
 - Less isolation for multi-tenant workloads using same HDFS data
 - Cost if there are off-peak hours

Transient BI on Object Storage

Best TCO for infrequent usage



- Best when workloads are:
 - Infrequent or scheduled
- Benefits include:
 - Lowest TCO with clusters only when needed
 - Full multi-tenant isolation
 - Common data in shared object storage
- Tradeoffs:
 - Delay to spin-up clusters when needed
 - Capability of BI users to spin up clusters
 - Per node perf of object storage (use more, cheaper nodes)

cloudera

Thank You

Justin Erickson