cloudera ATSCALE

CONSOLIDATE YOUR DATA MARTS FOR FAST, FLEXIBLE ANALYTICS

Alex Gutow | Cloudera
Josh Klahr | AtScale

TRENDS WITHIN DATA WAREHOUSING

Expand to Do More

Nearly 40% want greater capacity for growing data, users, reports, analyses, etc[1]

Improve Responsiveness

67% of users are requesting to do more BI/analytics on their own[2]

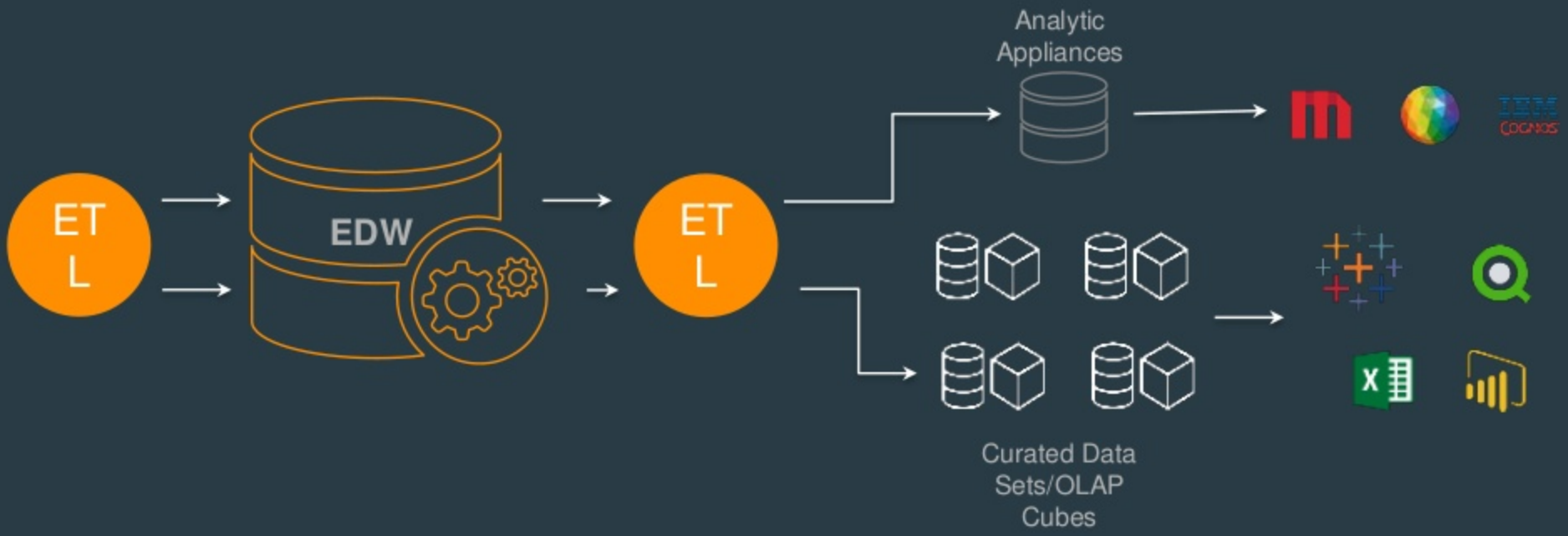
Balance Business Critical & Exploration

42% looking to augment EDW with modern platform[1]

[1]Data Warehouse Modernization In the Age of Big Data Analytics, TDWI, 2016
[2]Achieving Greater Agility with Business Intelligence, TDWI, 2014

FIRST GENERATION

THE RISE OF NEW TOOLS

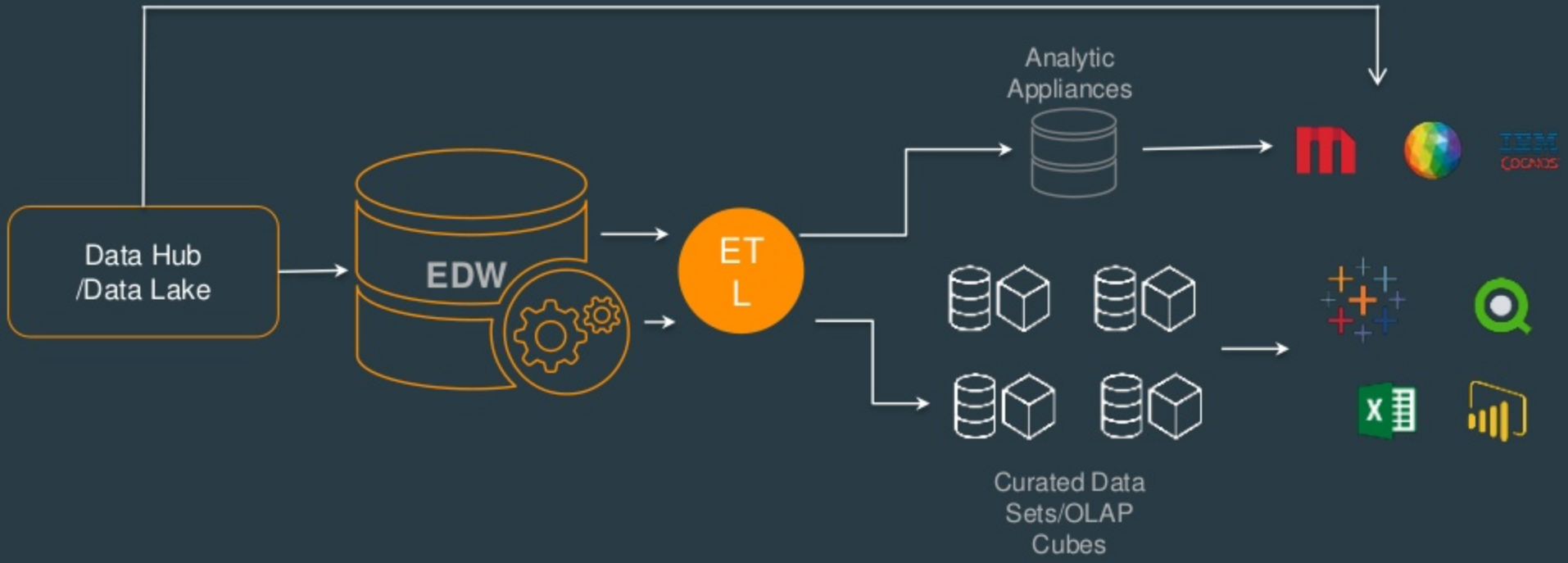


POLL QUESTION #1

What analytic appliances or data marts does your organization use? (Select all that apply)

- Netezza
- Vertica
- AWS Redshift
- Snowflake
- Exadata
- BI Tool Extracts (Tableau, Qlik, etc)
- EDW
- None/Unknown
- Other

RISE OF THE DATA LAKE



LIMITATIONS OF EXISTING INFRASTRUCTURE

- Not able to take on more reports, use cases, users, etc.
- Constrained exploration to prevent risking critical SLAs

- Need to contain costs for existing workloads
- Difficult to justify budget and maintenance for expansion
- Struggle to do more with less

- Proliferation of data silos to address additional workloads
- Maintain data copies causes inefficiencies for storage, processing, and people

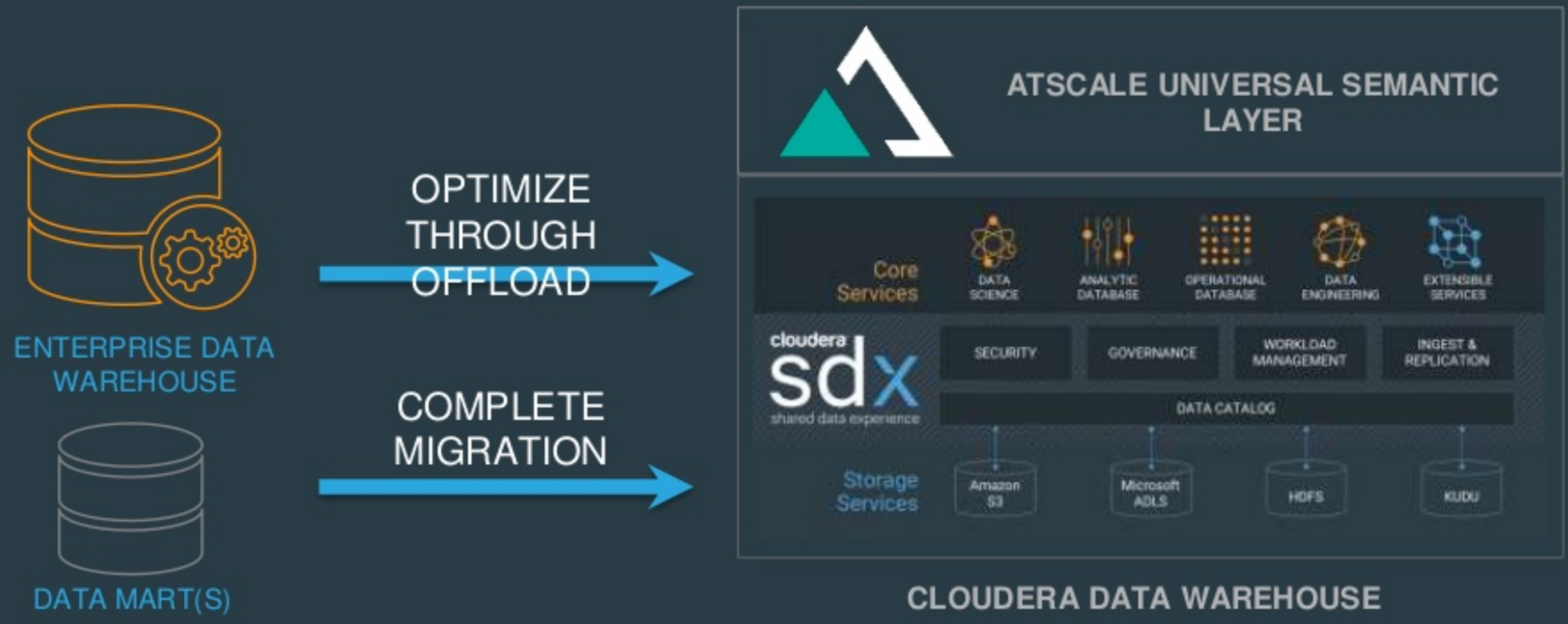
- Designed for curated reports, not iterative, self-service analytics
- Not built for elasticity or object store integration

POLL QUESTION #2

What pain points or limitations are you facing with your data mart(s)? (Select all that apply)

- Cannot meet internal SLA for delivering data
- Struggle to contain costs
- Long wait times for new data
- Limited data for querying
- Data exists across silos
- Cannot connect BI tools
- Migrating to cloud
- Other

DATA WAREHOUSING OPTIMIZATION & CONSOLIDATION



ADVANTAGES OF MODERN DATA WAREHOUSING

High-Performance SQL +

Data Flexibility

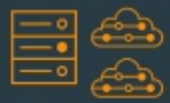
- Iterative modeling and self-service accessibility
- Portability: No proprietary formats or storage lock-in

Cost-Effective Scalability

- Elastic scale in any environment
- Cloud-native integration for optimized pay-per-use costs
- Proven at massive scale

Go Beyond SQL

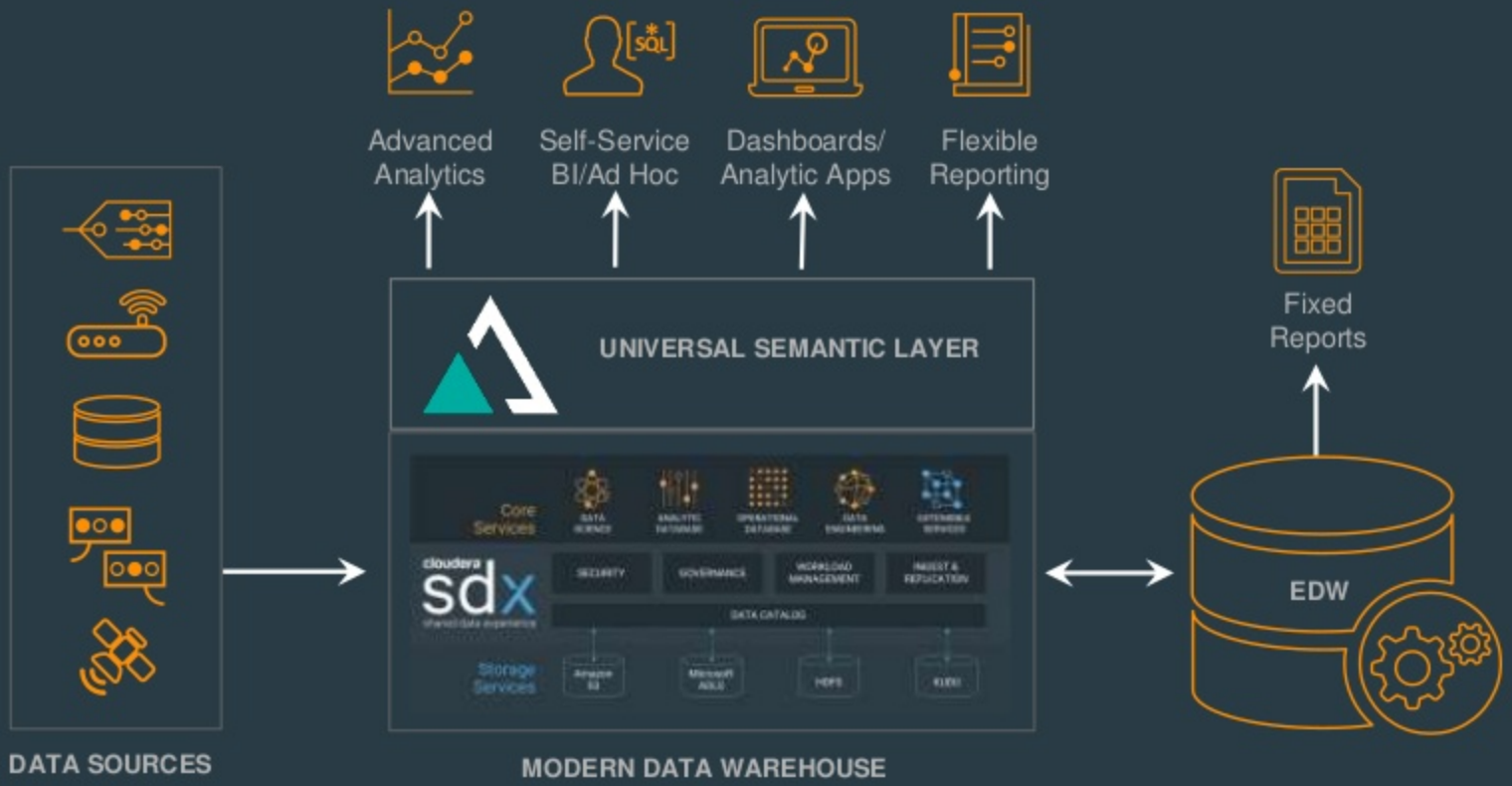
- Consolidate data silos with an open architecture
- Shared data across SQL and non-SQL workloads



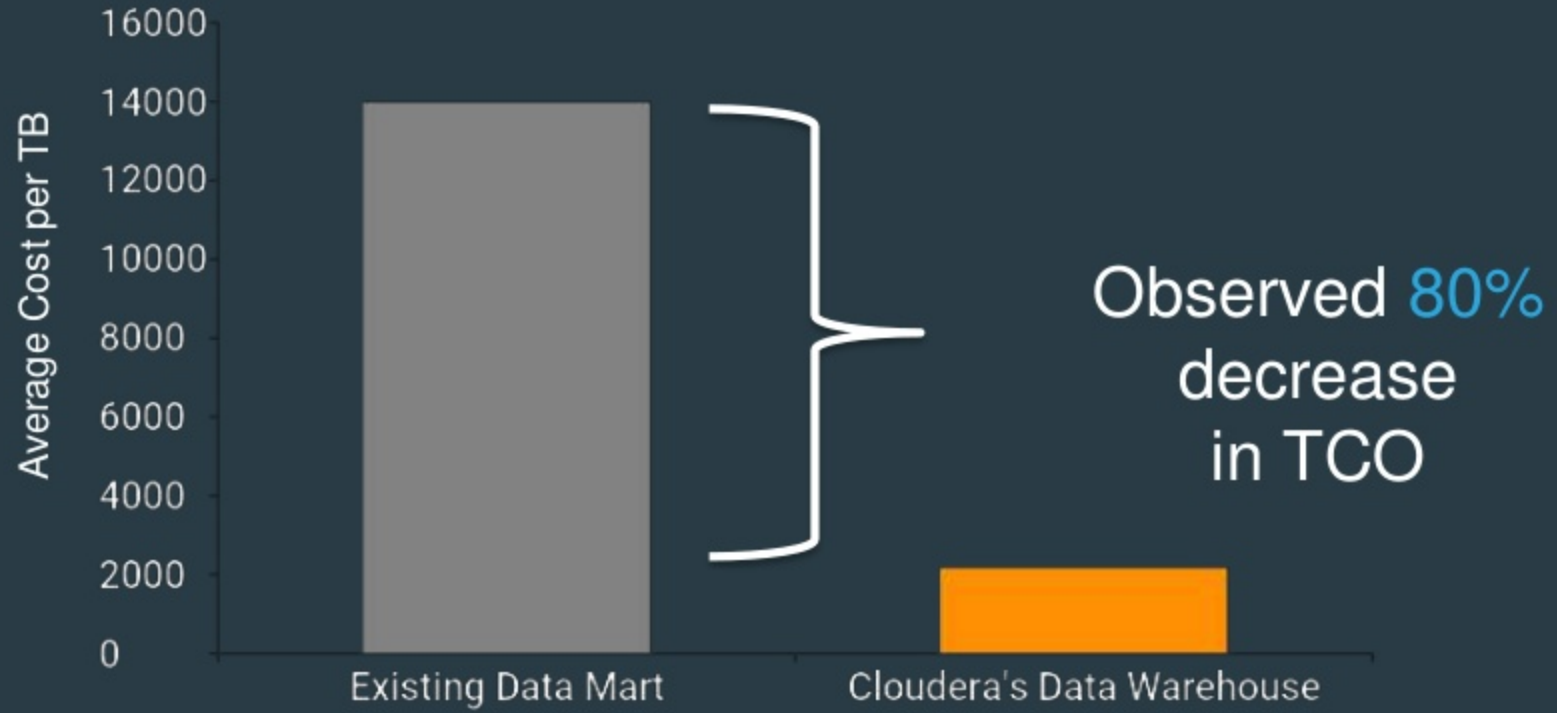
Hybrid Decoupled Architecture

- Runs across multi-cloud & on-prem for zero lock-in
- Multi- storage over S3, ADLS, HDFS, Kudu, Isilon, etc

MODERNIZING FOR THE DATA AGE



MORE VALUE AT A LOWER COST/TB



Note: TCO calculations and migration opportunity is dependent on your environment and types of workloads. Please work with Cloudera to calculate potential savings for your environment

PROOF OF PERFORMANCE & OLAP FUNCTIONALITY

IMPALA OUTPERFORMS ANALYTIC DATABASES

- Impala outperforms on both single and multi-user tests at 10TB
- Impala lead expands with concurrency
- Other SQL on Hadoop engines failed at 10TB

Single-user Test – 10TB

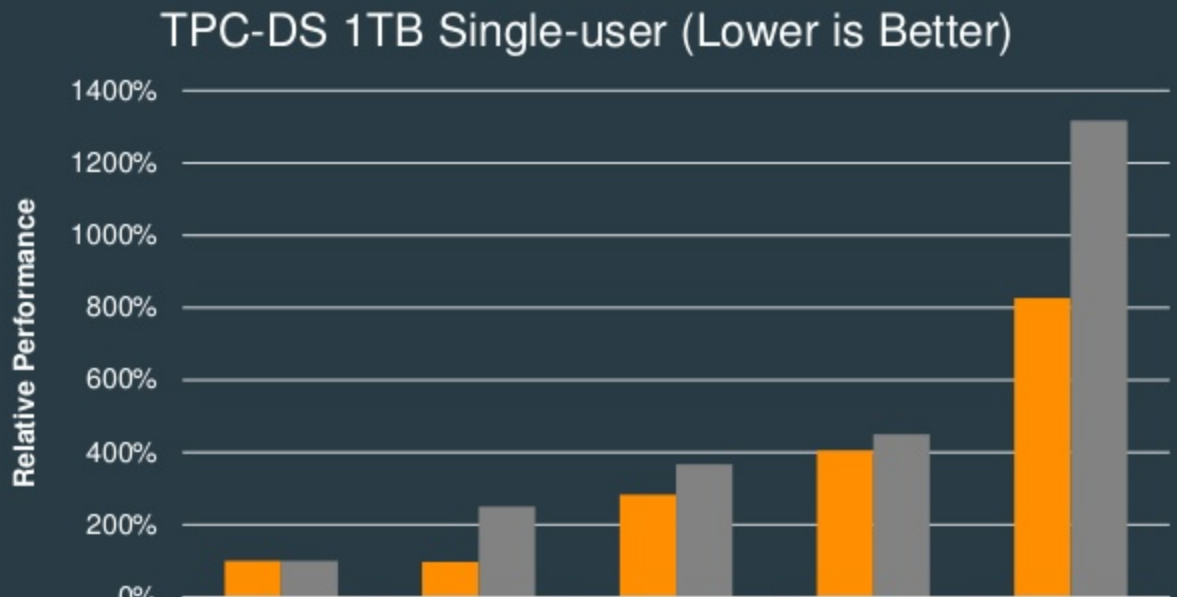
Metric	Impala	MPP DB	Impala Times Better
Total seconds	11,898	21,093	1.77x
Geometric mean	33	92	2.79x

Multi-user Test – 10TB

Streams	Impala QpH	MPP DB QpH	Impala Times Better
2	41	20	2.05x
4	75	20	3.75x
8	133	16	8.31x

TPC-DS 1TB: SINGLE-USER

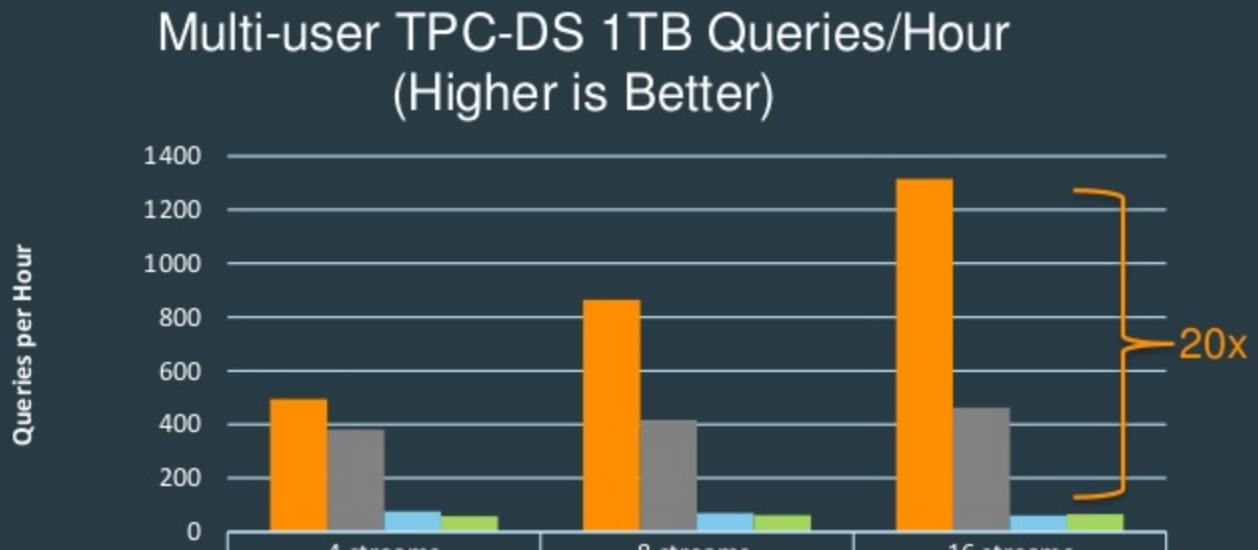
- The analytical db cohort (Impala / MPP) leads SQL on Hadoop
- Impala outperforms
  - Presto by 8.3x
  - Hive w/ LLAP by 4x
  - Spark SQL by 2.8x



	Impala	Greenplum	Spark SQL	Hive with LLAP	Presto
Total	100%	97%	283%	405%	826%
Geomean	100%	250%	367%	450%	1317%

TPC-DS 1TB: MULTI-USER

- At 1TB the analytic db cohort (Impala / MPP) expands lead with concurrency
- With 16 streams Impala outperforms
  - Spark by ~22x
  - Hive by ~20x
  - Greenplum by ~3x



	4 streams	8 streams	16 streams
Impala	495	865	1,315
Greenplum	381	417	462
Spark SQL	76	70	61
Hive with LLAP	58	62	66

COST-EFFECTIVE PERFORMANCE

With Cloudera + AtScale

Cloudera + AtScale can meet or exceed the performance needs of traditional database users

- Impala provides leading analytic database performance at high user concurrency
- AtScale's Intelligence Platform improves query performance
- Dimensional Calc Engine allows complex OLAP-style calculations to run directly on cluster

Fortune 50 Insurance Company			
Use Case	Netezza	AtScale + Cloudera	Speedup Factor
Metric Trends	4.0	0.3	15.6 times faster
Customer Growth	6.8	0.3	20.6 times faster
Segment Benchmark	2.4	0.9	2.7 times faster
Time Series	5.5	1.1	4.9 times faster
Transaction Dashboard	5.5	0.4	12.7 times faster
Top Customers	5.4	0.9	6.1 times faster
Product Heatmap	5.5	0.2	34.8 times faster
Top Products	5.5	1.9	2.9 times faster
Top Suppliers	5.5	0.8	6.9 times faster

BI ON BIG DATA IS ABOUT MORE THAN JUST PERFORMANCE

AtScale adds OLAP to Cloudera's Analytic DB

Benefits of AtScale + Impala

- OLAP is the language of BI
  - Provides business users access to functions like: time-intelligence, drill-down, and slicing and dicing
- AtScale is optimized for Impala
  - Provides intelligent queries, optimized data structures

DATA SCIENTISTS AND BUSINESS USE THE SAME DATA

Curate the data once - deliver self-service BI and Data Science

Benefits the AtScale Universal Semantic Layer

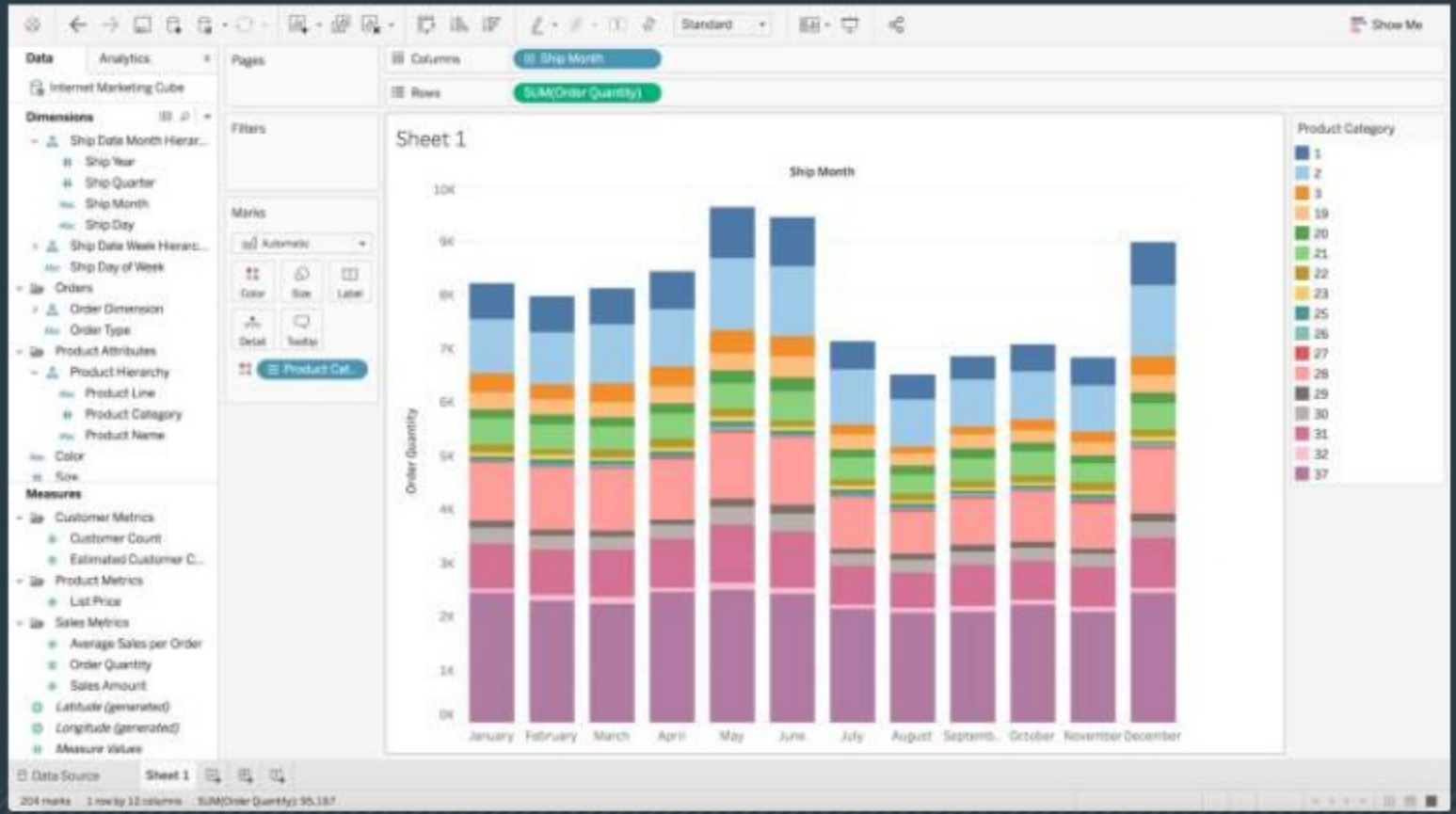
- Keeps the data in the Cloudera cluster
- Avoids extractions or data marts
- Simplifies and modernizes your data architecture
- Allows any BI tool to access the data, including Excel, PowerBI, Tableau, etc.

POLL QUESTION #3

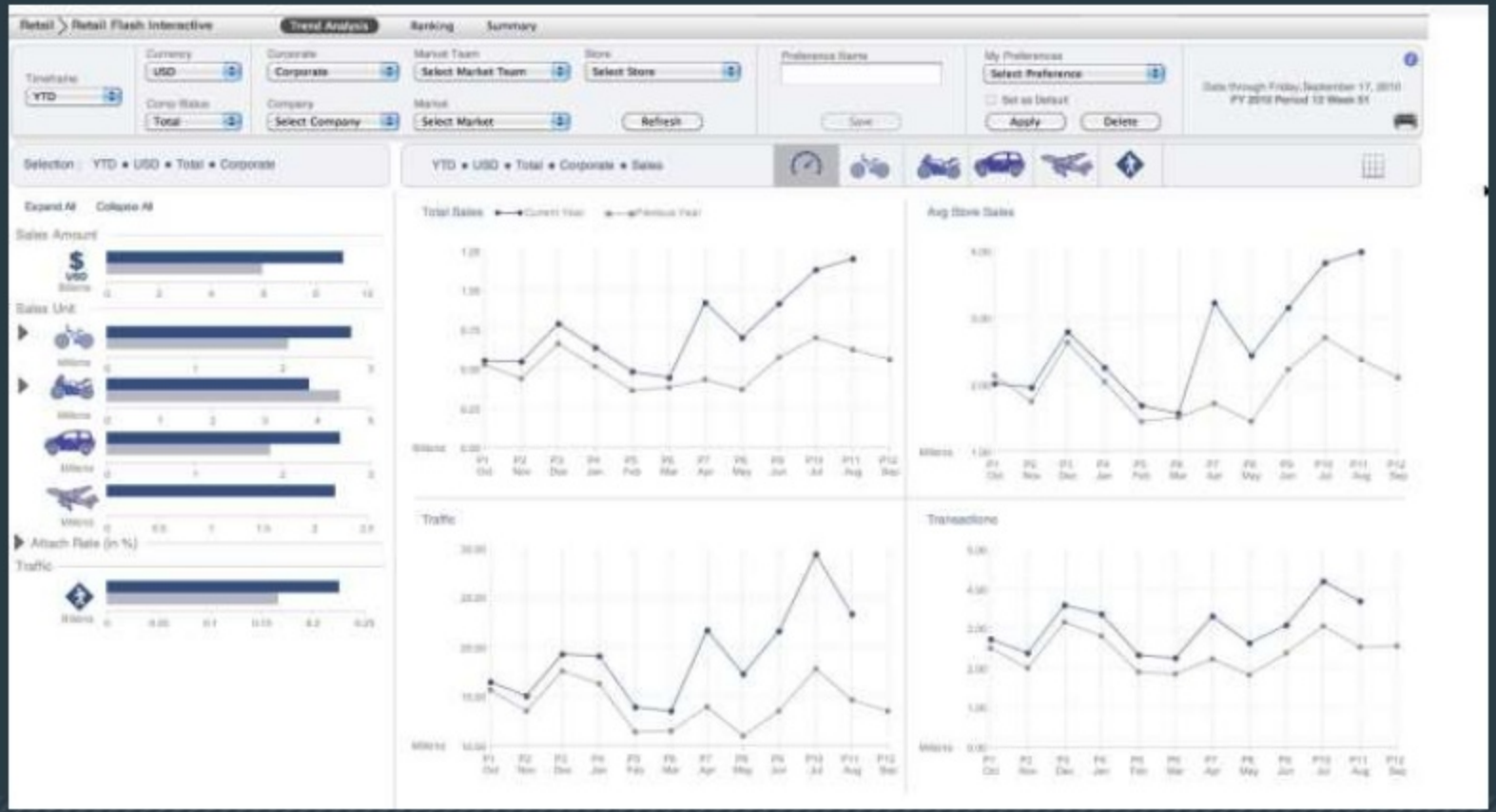
Which BI tools are you using? (Select all that apply)

- Tableau
- SAP Business Objects
- Excel
- PowerBI
- Qlik
- Spotfire
- Microstrategy
- Zoomdata
- Cognos
- Other

ATSCALE PRESENTS DATA FOR BI USERS, NOT DATA ENGINEERS



DASHBOARDS DEMAND CONCURRENCY, THROUGHPUT



JOINT CUSTOMER SUCCESS

TOYOTA

Self-Service BI for Finance & IoT

Cloudera + AtScale

- Deliver sub-second queries
- Supports ad hoc Tableau and PowerBI
- Data science embedded with business units and goals
- Enable and protect enterprise data consumption

8k
Self-service users

2,200
Self-delivered dashboards

95%
Faster time-to-insights

1,000s
Databases migrated

GLOBAL PHARMACEUTICAL

R&D Information Platform

Cloudera + AtScale

- Consistent, shared data access through BI tools (Tableau, Tibco)
- Interactive query speeds
- OLAP capabilities
- HIPAA compliant

Time-to-insights in minutes, not years:

- Reduced cost and time to identify clinical trial groups
- Accelerate new drug development
- 1st time metrics and monitoring on compliance data

8,700

Analytic Users

70% Execs/Managers
25% Analysts
5% Data Scientists

250+

Use cases

8PB

structured + unstructured

100% unstructured data captured

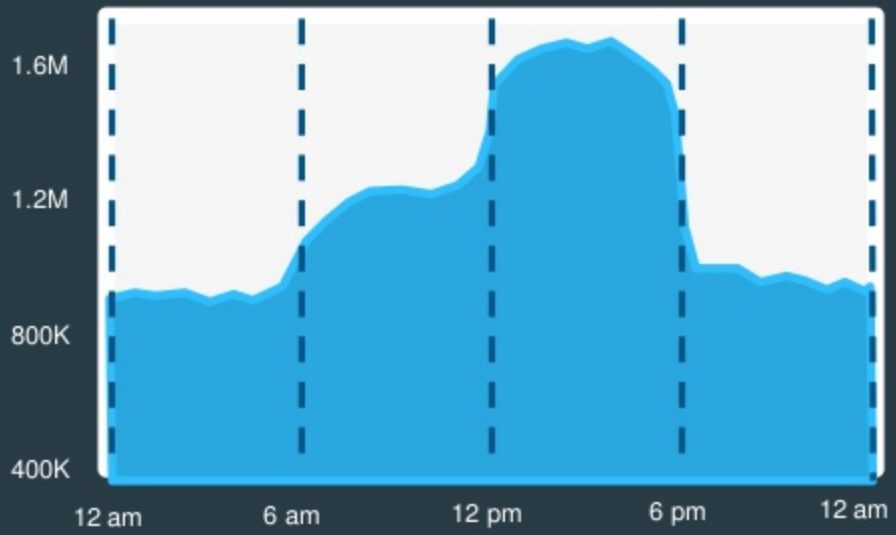
200x

Reduction in silos

THE OFFLOAD PROCESS

WHERE DO YOU START?

Query volume can be huge



Numerous databases, thousands of tables, many users and applications

Queries can be very complex

- How do you determine what workloads to run on Cloudera's platform?
- Will the queries run efficiently?
- What does it take to migrate?
- How do you prioritize?

DATA WAREHOUSE OPTIMIZATION

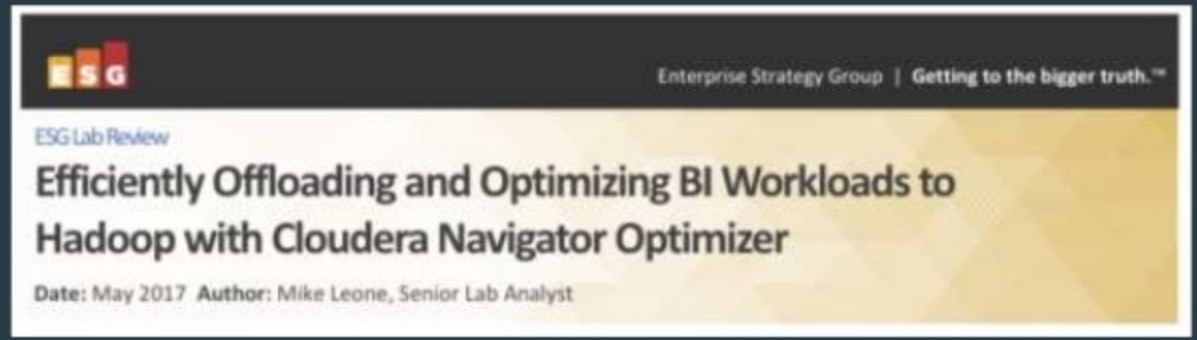
Tools & Framework to help you through the process

WHERE TO LEARN MORE

Contact us or check out:

Learn more how GSK and Toyota brought BI to the Data Lake:
http://info.atscale.com/webinar_do_dont_bi_on_data_lake_2018-0

Get the report for how to offload BI workloads:
https://www.cloudera.com/content/dam/www/marketing/resources/analyst-reports/efficiently-offloading-and-optimizing-bi-workloads-to-hadoop-with-cloudera-navigator-optimizer.pdf.landing.html



THANK YOU

cloudera

ATSCALE