Apache Hive 2
(대규모 고속 SQL)

최종욱, Hortonworks
tchoi@apache.org

HORTONWORKS®
POWERING THE FUTURE OF DATA™

데이터의 시대

오픈소스는 표준이고,

아파치는 그 중심에 있습니다.

THE APACHE SOFTWARE FOUNDATION
Founded: 1999

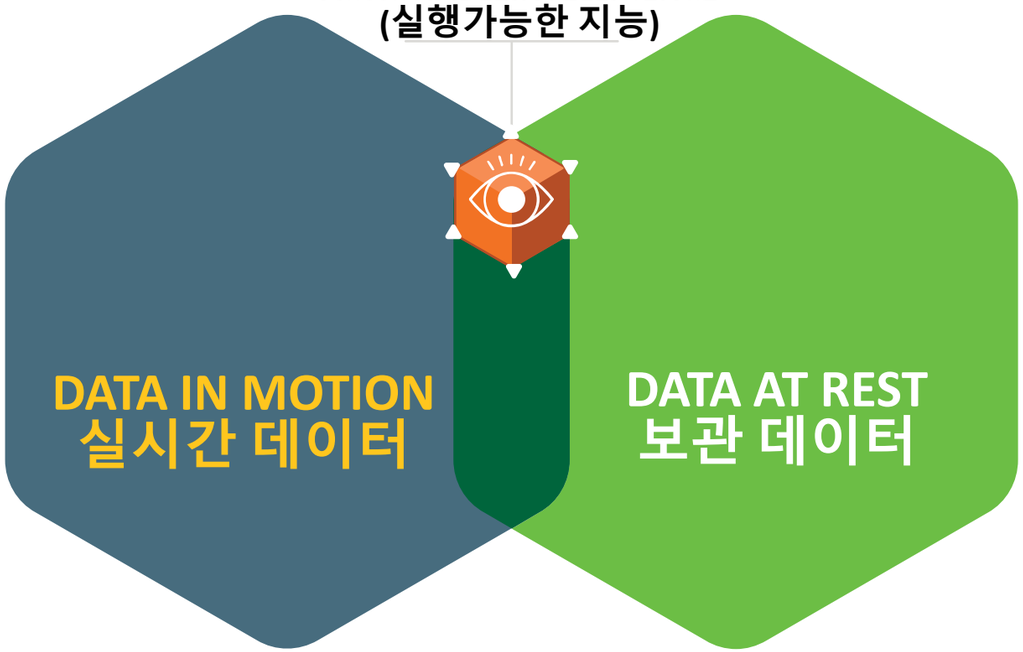
HORTONWORKS
POWERING THE FUTURE OF DATA
Founded: 2011
IPO: 2014

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

호튼웍스의 접근과 배포판(HDP, HDF)

Modern Data Applications
(선구적인 데이터 어플리케이션)

ACTIONABLEINTELLIGENCE
(실행가능한 지능)



Hortonworks DataFlow
(Powered by Apache NIFI)

Hortonworks Data Platform
(Powered by Apache Hadoop)

HORTONWORKS

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

호튼웍스 회사 소개

ONLY **100%**

오픈소스

Apache Hadoop Data Platform

2011년 Yahoo! 에서
하둡팀이 분사

1 번째 하둡 상장회사

IPO 4Q14 (NASDAQ: HDP)
유일한 흑자회사

2,100+ 파트너사

기술파트너
전략 파트너
컨설턴트
리셀러

US 포천 100대 기업의 60%, Global 포천 500대기업의 30%가 구독고객

1,100+ 구독고객

1,110+ 직원

19 개국

[1] *Operating billings is an operating measure defined as the aggregate value of all invoices sent to our customers in a given period*

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS
POWERING THE FUTURE OF DATA

Agenda

Scalable Data Warehousing on Hadoop

Overview

Solution Architecture

Customer Use Cases

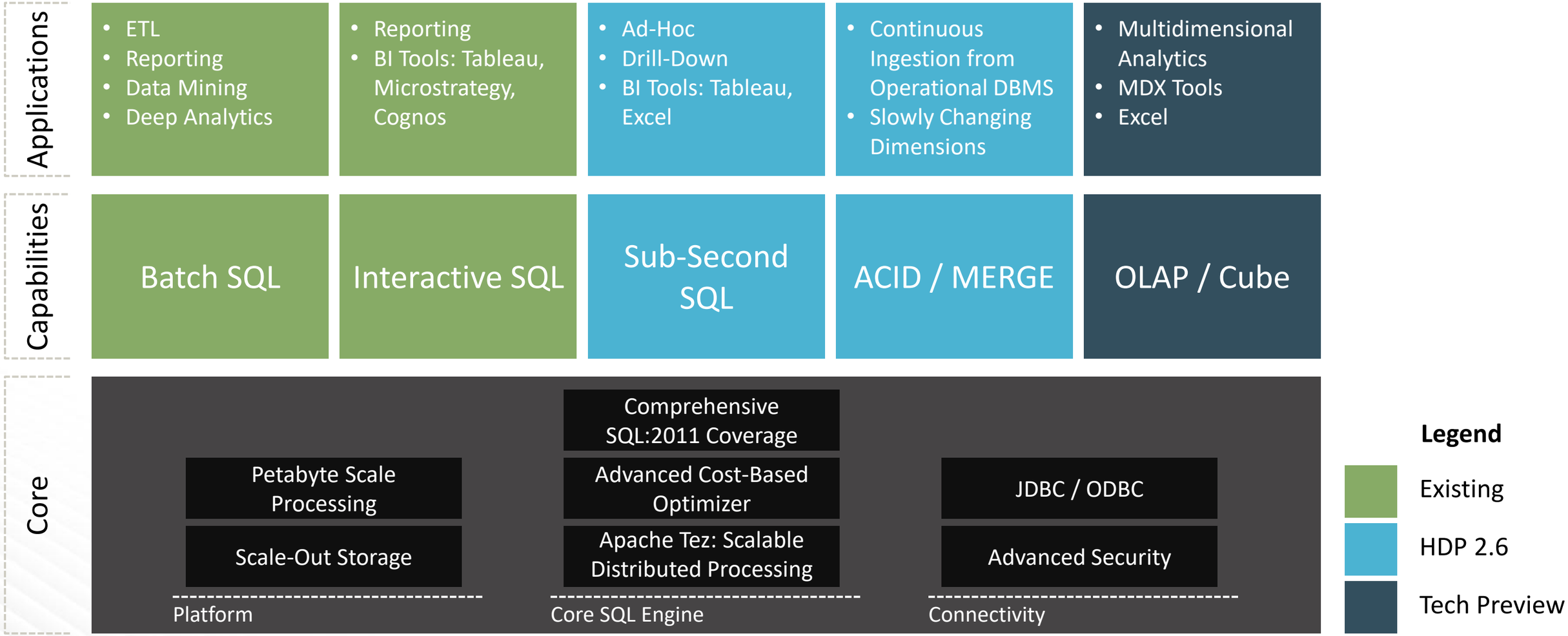
What's New

Roadmap

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Scalable Data Warehousing on Hadoop



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Hive's Unique Advantages in HDP 2.6

Why Hive:

- Comprehensive ANSI SQL including 99 TPC-DS Queries.
- The **only** Hadoop SQL with ACID MERGE for easy updates.
- In-Memory caching for MPP performance at Hadoop scale.
- Per-User dynamic row and column security.
- Replication and DR for critical workloads.
- Compatible with every major BI Tool.
- Proven at 300+ PB Scale.

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Apache Hive: Fast Facts

Most Queries Per Hour

100,000 Queries Per Hour
(Yahoo Japan)

Analytics Performance

100 Million rows/s Per Node
(with Hive LLAP)

Largest Hive Warehouse

300+ PB Raw Storage
(Facebook)

Largest Cluster

4,500+ Nodes
(Yahoo)

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Agenda

Scalable Data Warehousing on Hadoop

Overview

Solution Architecture

Customer Use Cases

What's New

Roadmap

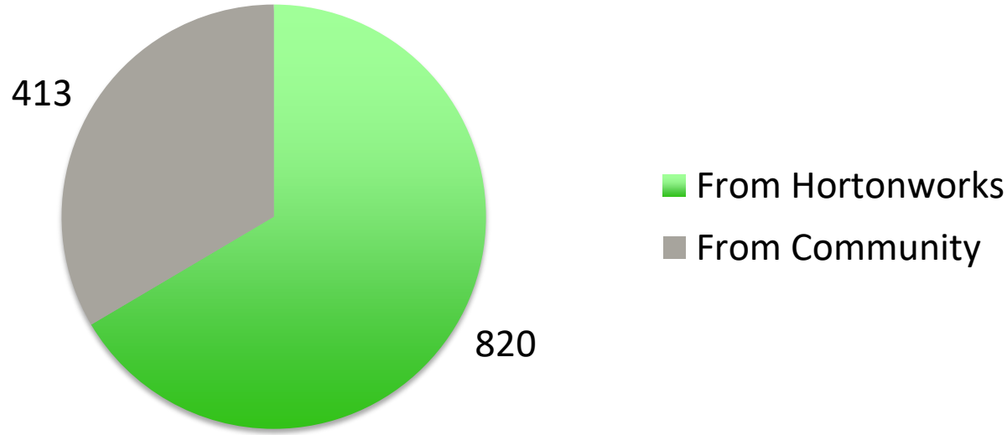
© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

HDP 2.6 Continues Strong Momentum for Hive

- At a High Level:
  - 1200+ features, improvements and bug fixes in Hive since HDP 2.5.
  - 400+ of these from outside of Hortonworks.
- Major Improvements:
  - Hive LLAP Now GA
  - ACID MERGE
  - SQL: All 99 TPC-DS out-of-the-box with only trivial rewrites
  - Hive View 2.0: Great Features for DBAs
  - Diagnostics: Tez UI Total Timeline View
  - Tech Preview: Hive OLAP Indexes powered by Druid

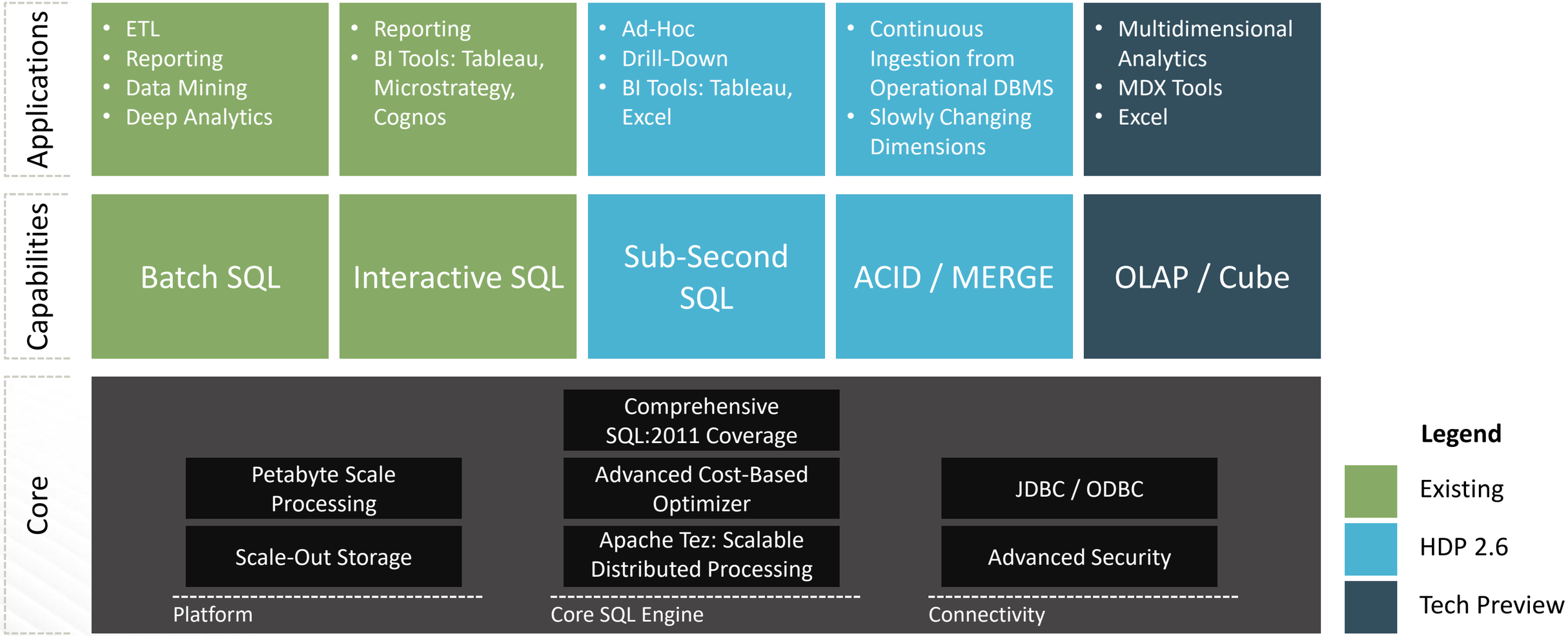
HDP 2.6 Improvements



- Hive LLAP GA
- SQL MERGE
- All TPC-DS Queries

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

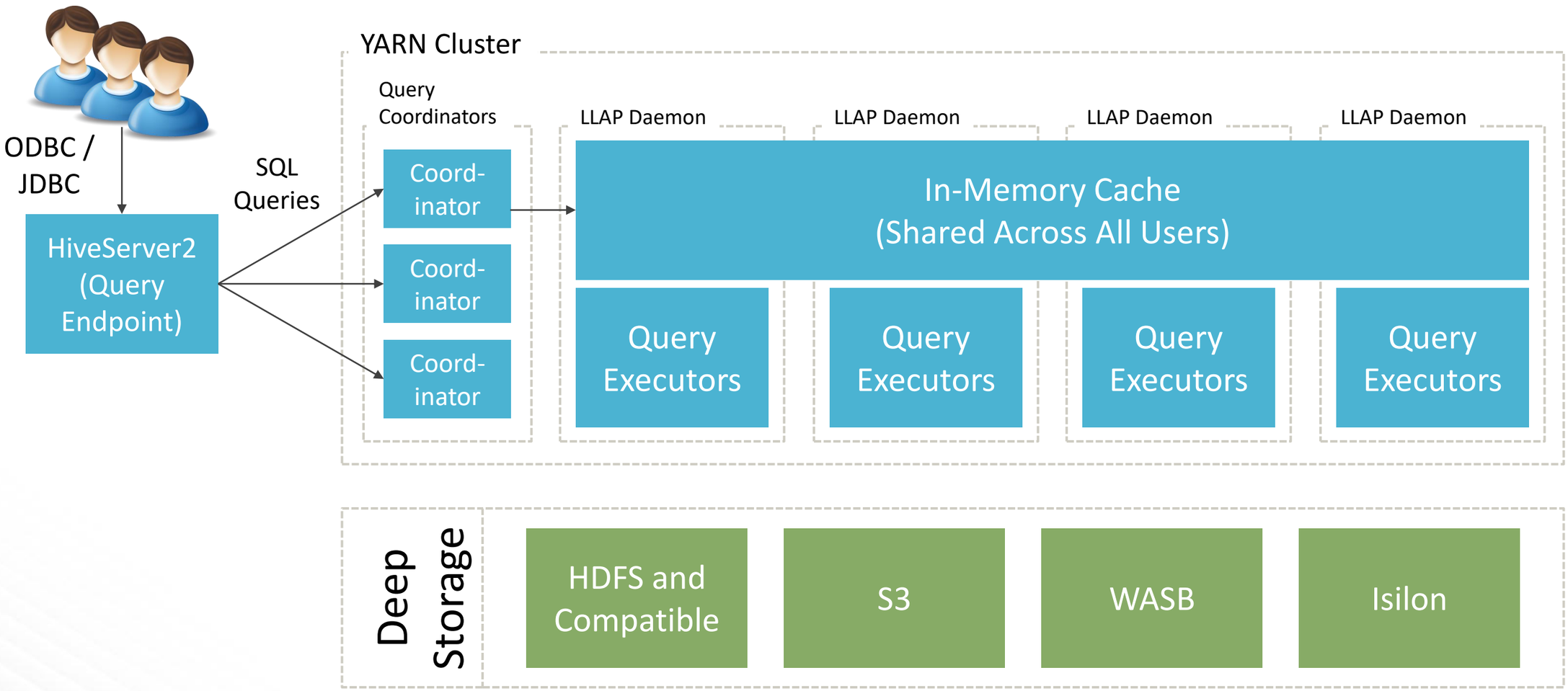
Scalable Data Warehousing on Hadoop



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

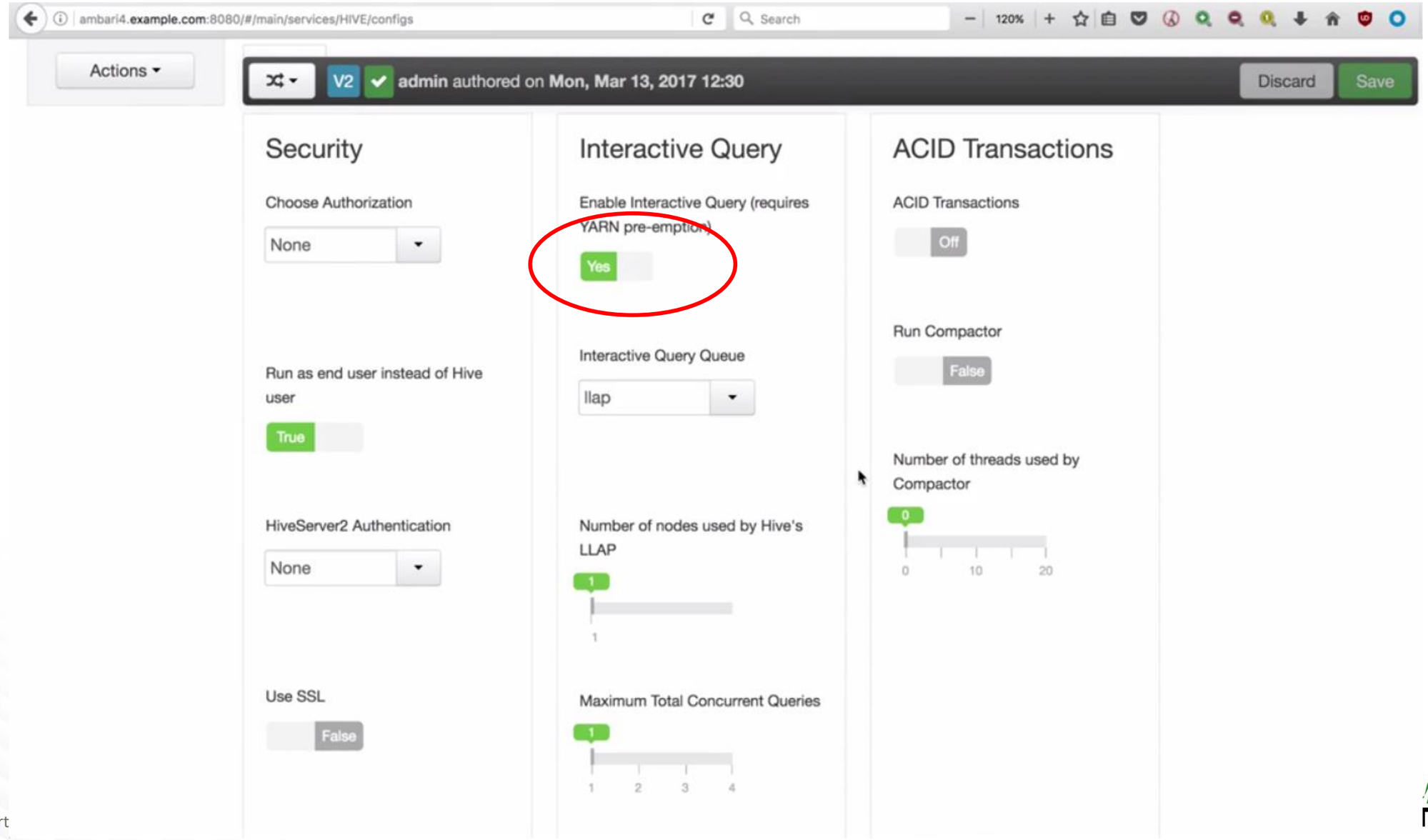
Hive LLAP -- MPP Performance at Hadoop Scale



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Hive LLAP: One-Touch Provisioning

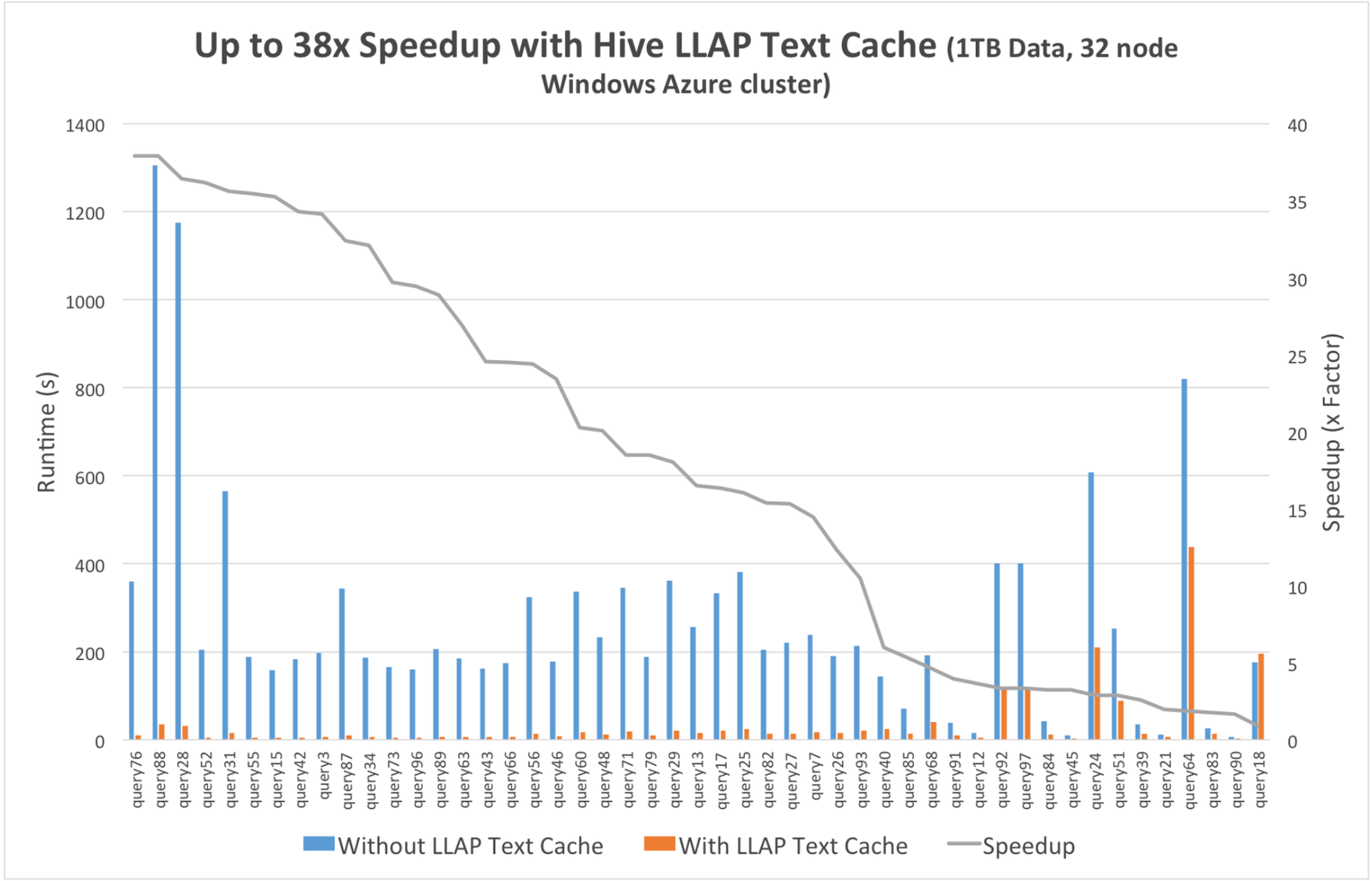


© Hort

Hive LLAP Text Cache Delivers Zero ETL Analytics

Highlights

- Fast interactive analytics on CSV and JSON data.
- No ETL / conversion to ORCFile needed, just load and go.
- Once data is cached, analytics become dramatically faster.



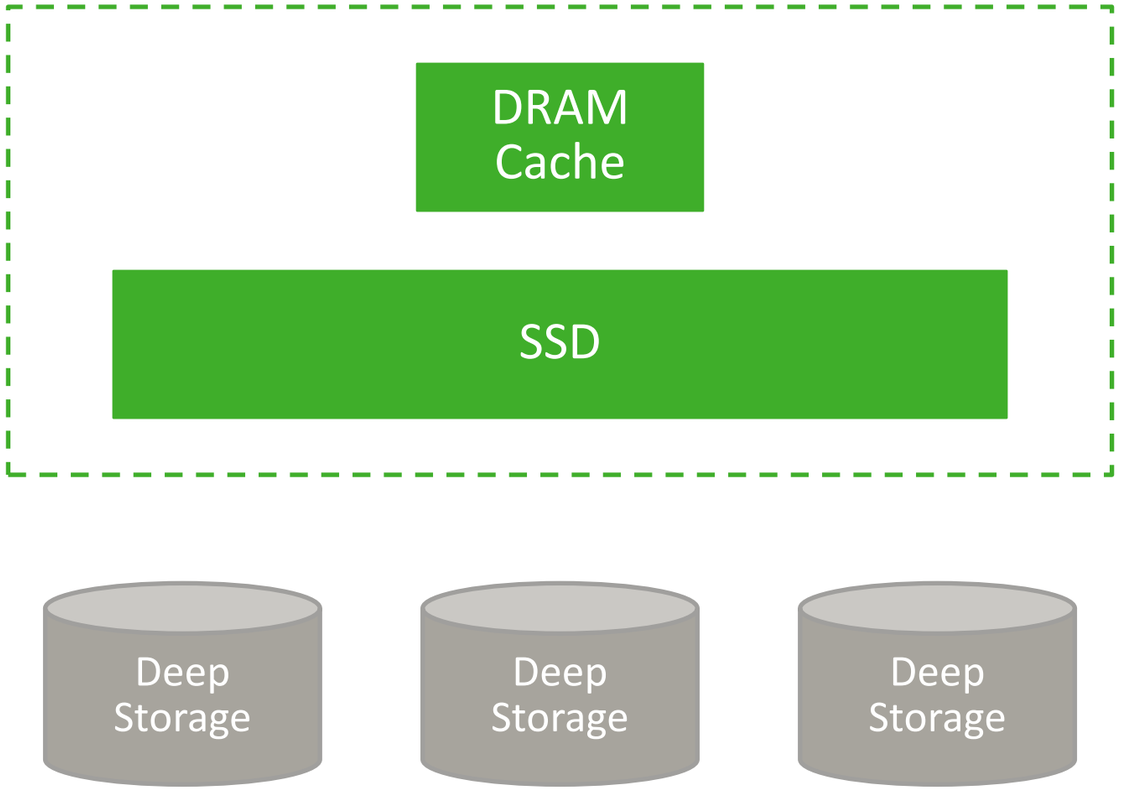
© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Cache 4x More Data with Hive LLAP SSD Cache

Highlights

- Use the combination of DRAM and SSD to dynamically cache data.
- Cache 4x more data than using DRAM alone.
- Deliver fast analytics on larger datasets with higher concurrency.



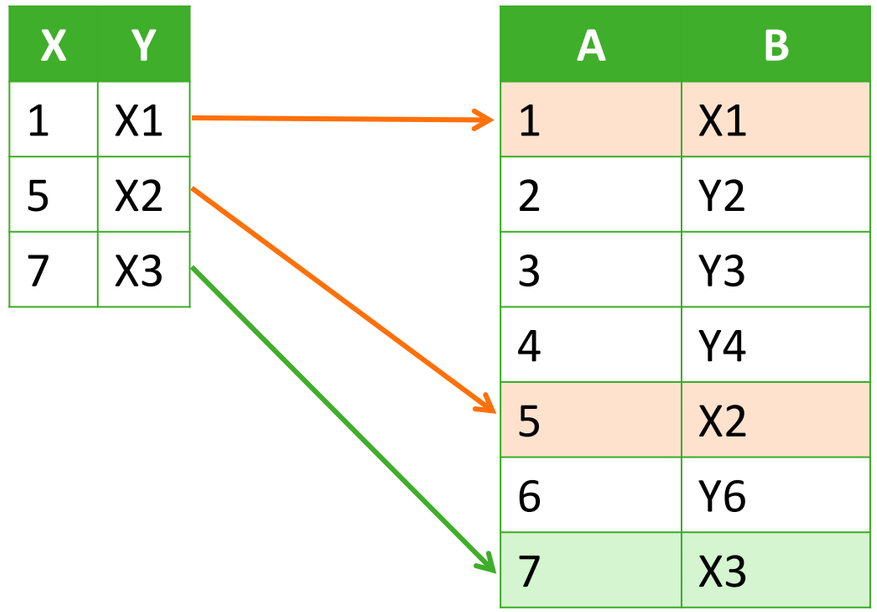
© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Hive ACID MERGE Makes Data Maintenance Simple

Highlights

- SQL Standard ACID MERGE now available in Hive.
- Efficiently perform record-level inserts, updates and deletes.
- Delivers real Data Management in Hadoop, massively simplifying updates, data restatements and change data capture.



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Comprehensive SQL in Hive Including All 99 TPC-DS Queries

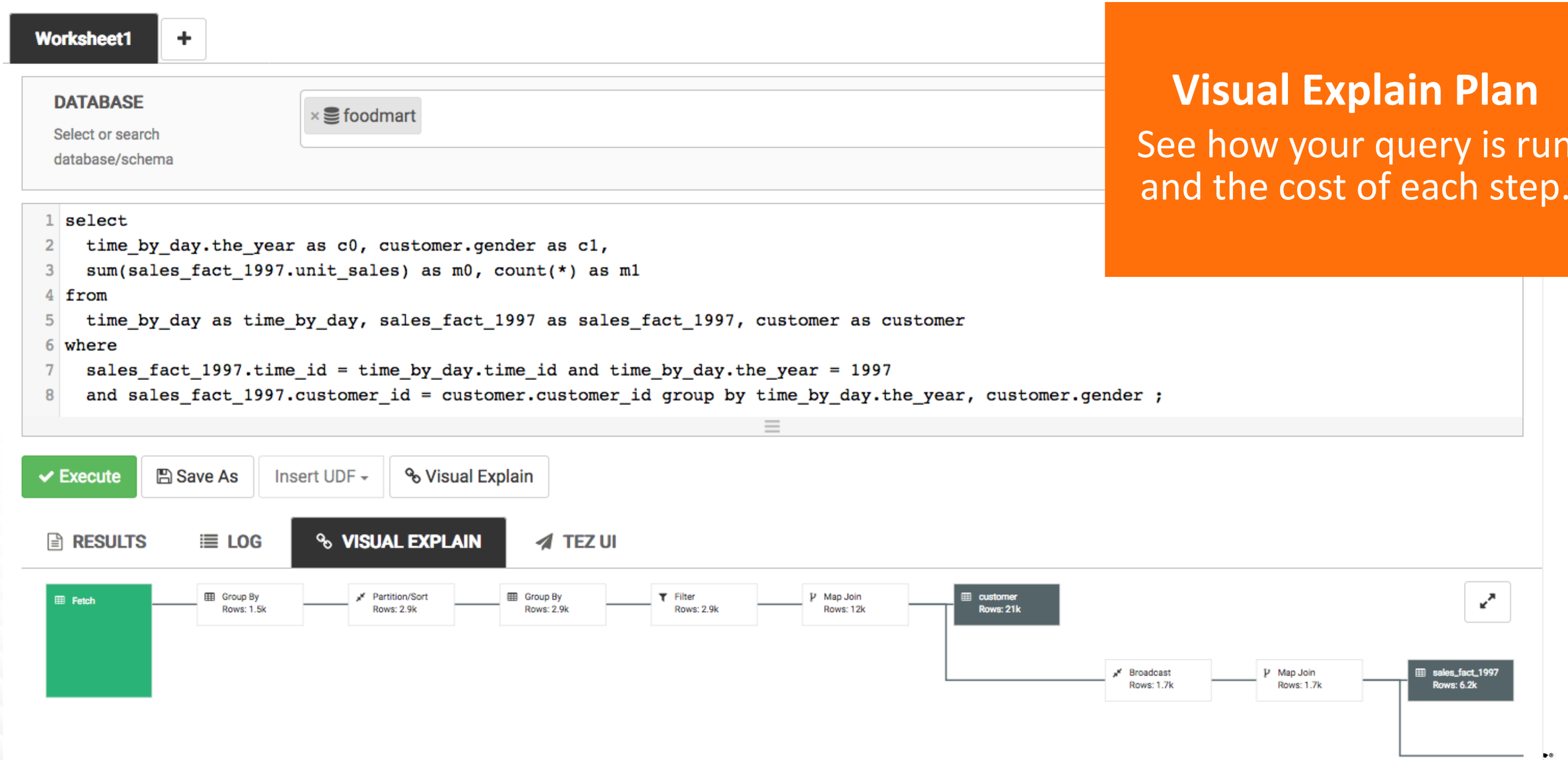
Highlights

- Multiple and Scalar Subqueries
- INTERSECT and EXCEPT
- Standard syntax for ROLLUP / GROUPING
- Syntax improvements for GROUP BY and ORDER BY
- In HDP 2.6+ Hive runs all 99 TPC-DS with only trivial re-writes.

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Hive View 2.0: Visual Explain Plan Makes Debugging Easier

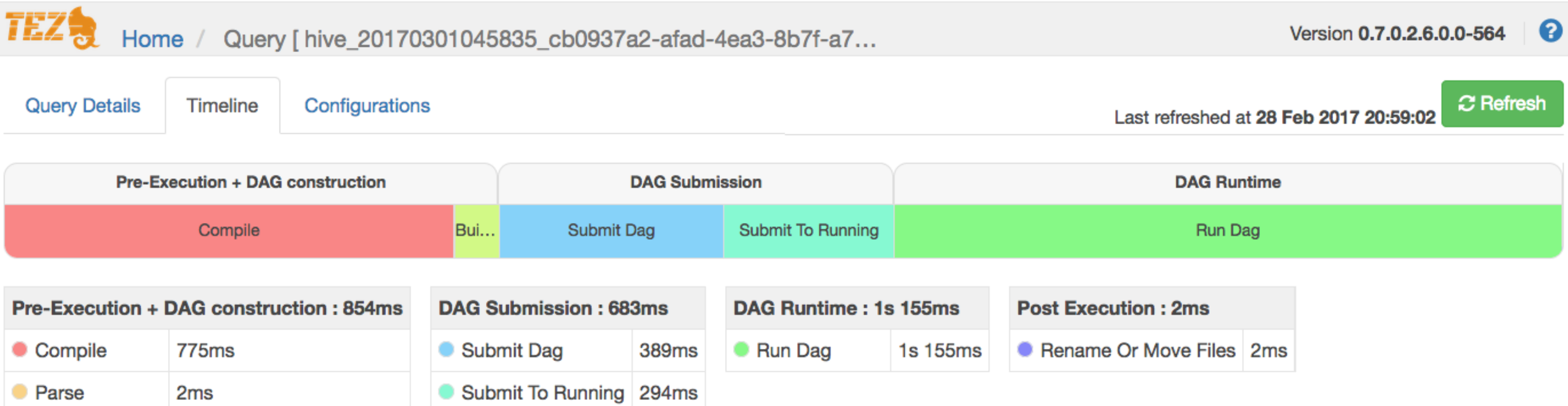


Visual Explain Plan

See how your query is run and the cost of each step.

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

Tez Total Timeline View Show Exactly Where Time Goes



Total Timeline View

See exactly where query time is spent, from compilation to execution.

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

Dynamic Tag-based Access Policies with Apache Atlas

- **Basic Tag policy** – PII example. Access and entitlements must be tag based ABAC and scalable in implementation.
- **Geo-based policy** – Policy based on IP address, proxy IP substitution maybe required. The rule enforcement but be geo aware.
- **Time-based policy** – Timer for data access, de-coupled from deletion of data.
- **Prohibitions** – Prevention of combination of Hive tables that may pose a risk together.

Key Benefits:

New scalable metadata based security paradigm

Dynamic, real-time policy

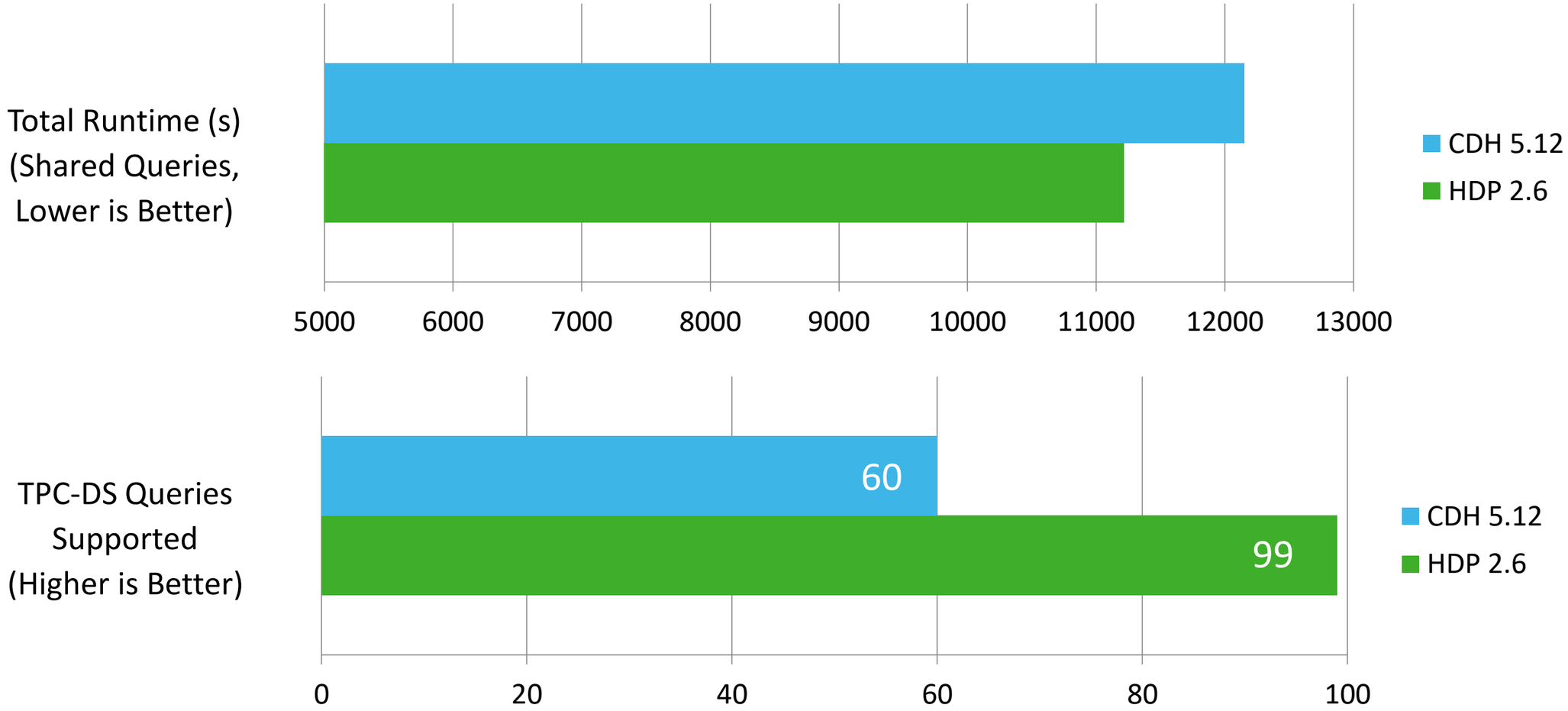
Active protection – fast updates to changes

Centralized and simple to manage policy

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Head to Head: TPC-DS Hive versus Impala

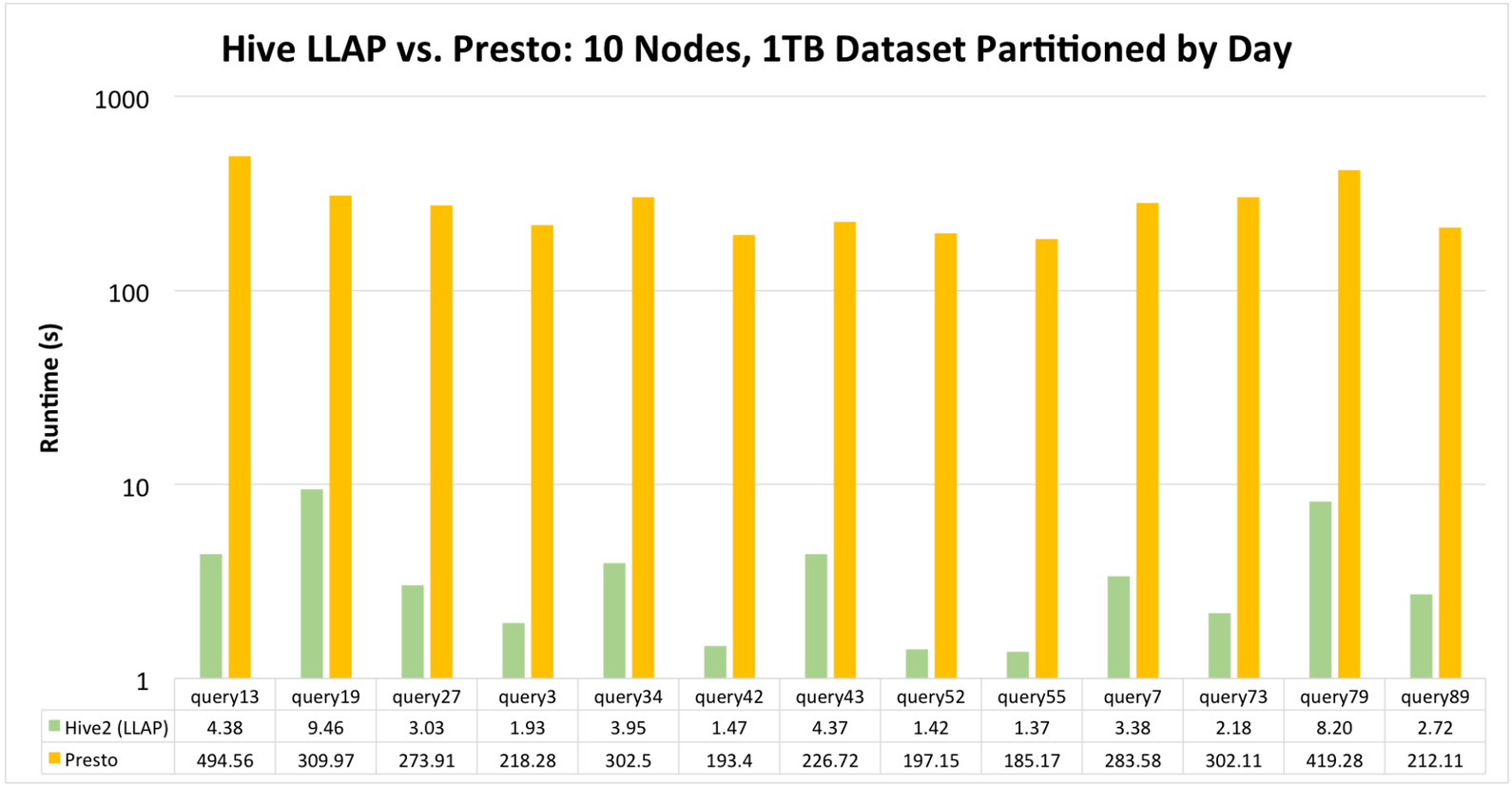


TPC-DS; Scale 10,000; 9 Nodes; Identical, Trivially Modified SQL Queries

Performance: Apache Hive vs. Presto on a partitioned 1TB dataset.

Highlights

- Presto lacks basic performance optimizations like dynamic partition pruning.
- On a real dataset / workload Presto perform poorly without full re-writes.
- Example: Query 55 without re-writes = 185.17s, with re-writes = 16s. LLAP = 1.37s.



	query13	query19	query27	query3	query34	query42	query43	query52	query55	query7	query73	query79	query89
Hive2 (LLAP)	4.38	9.46	3.03	1.93	3.95	1.47	4.37	1.42	1.37	3.38	2.18	8.20	2.72
Presto	494.56	309.97	273.91	218.28	302.5	193.4	226.72	197.15	185.17	283.58	302.11	419.28	212.11

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

Announcing: Druid is Now GA in HDP 2.6.3

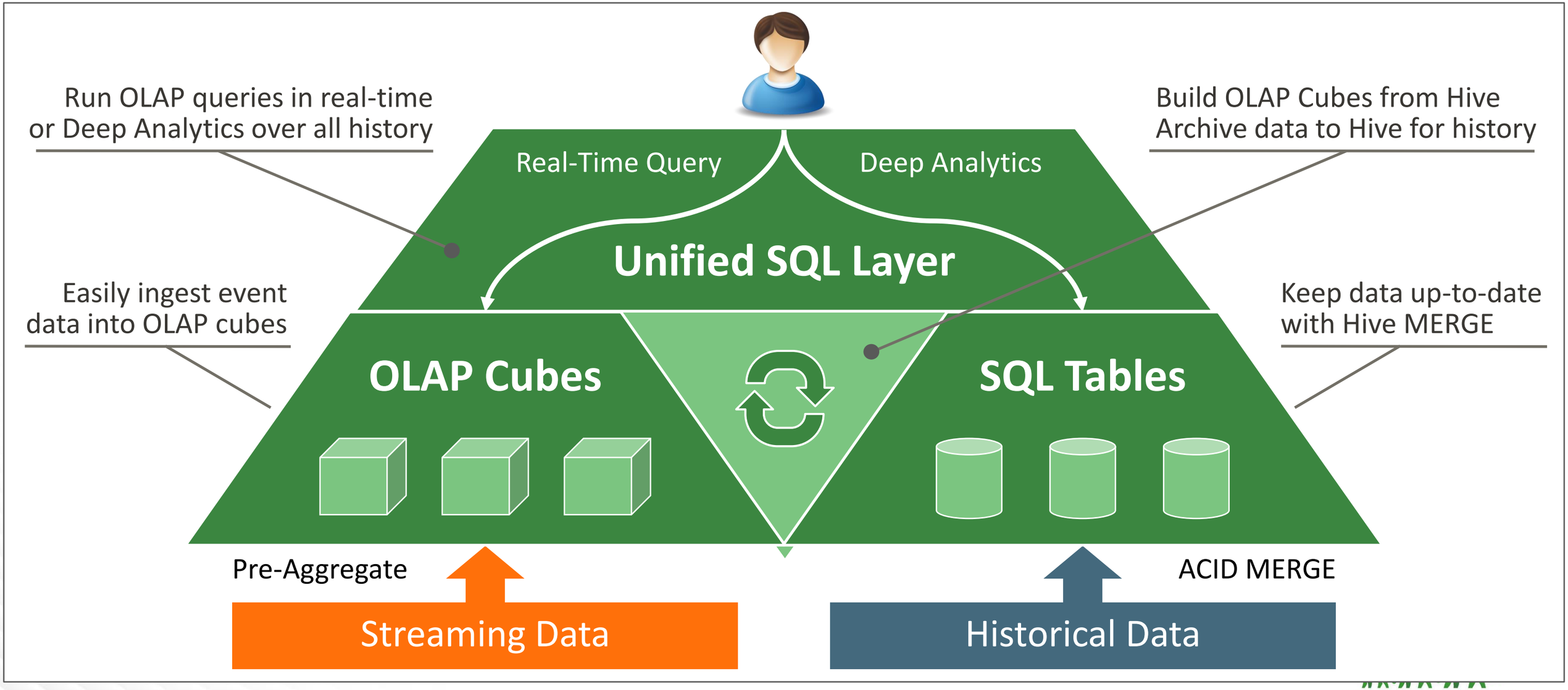
- ✓ Analyze Streaming and Historical Data with SQL
- ✓ Powerful Visualization
- ✓ Simple management and monitoring with Ambari
- ✓ Fine-grained security
- ✓ Integrates with Hortonworks SAM for simple development.

Note: Superset remains in Technical Preview in HDP 2.6.3

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

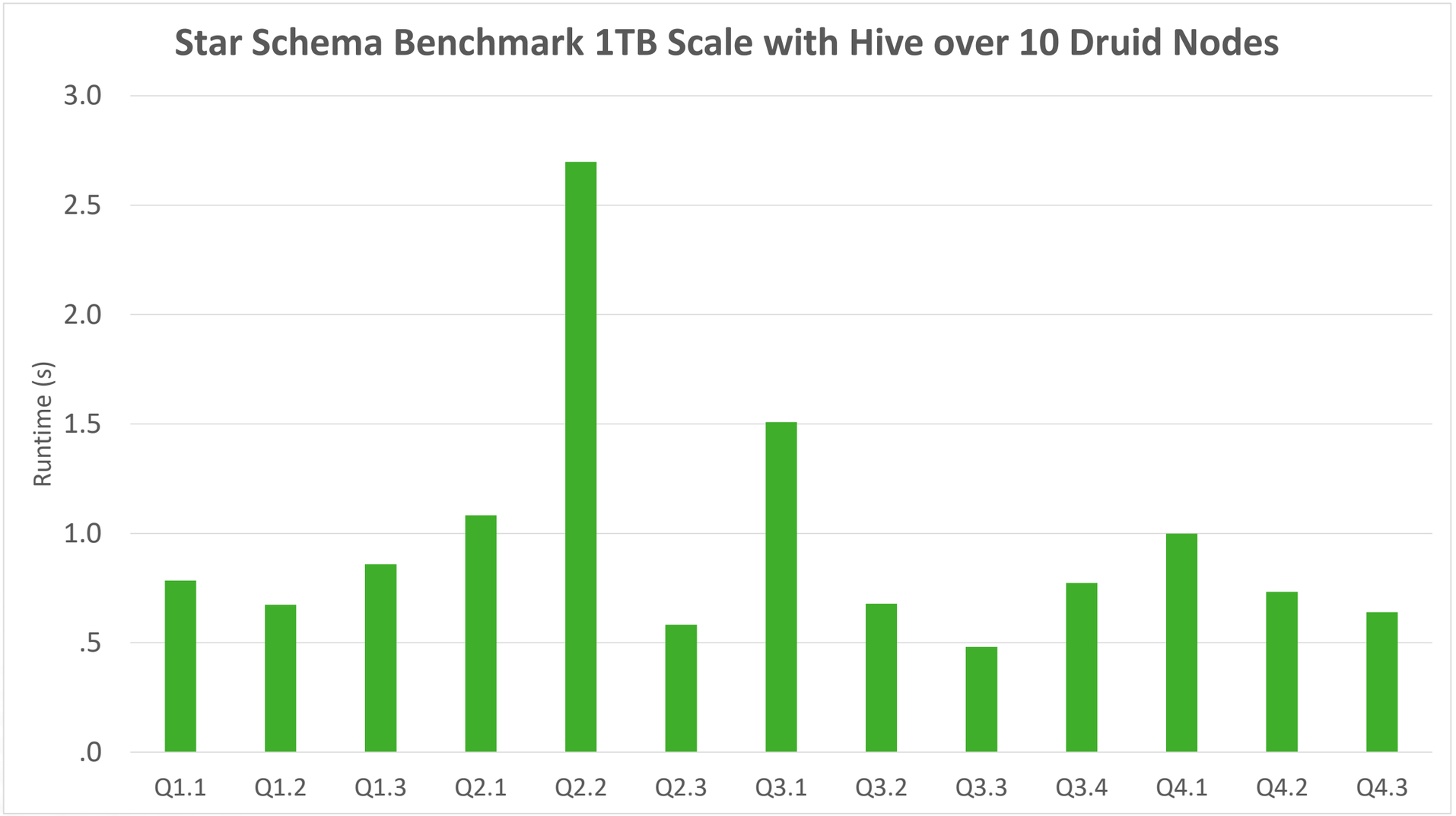
HORTONWORKS

Hive + Druid = Insight When You Need It



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

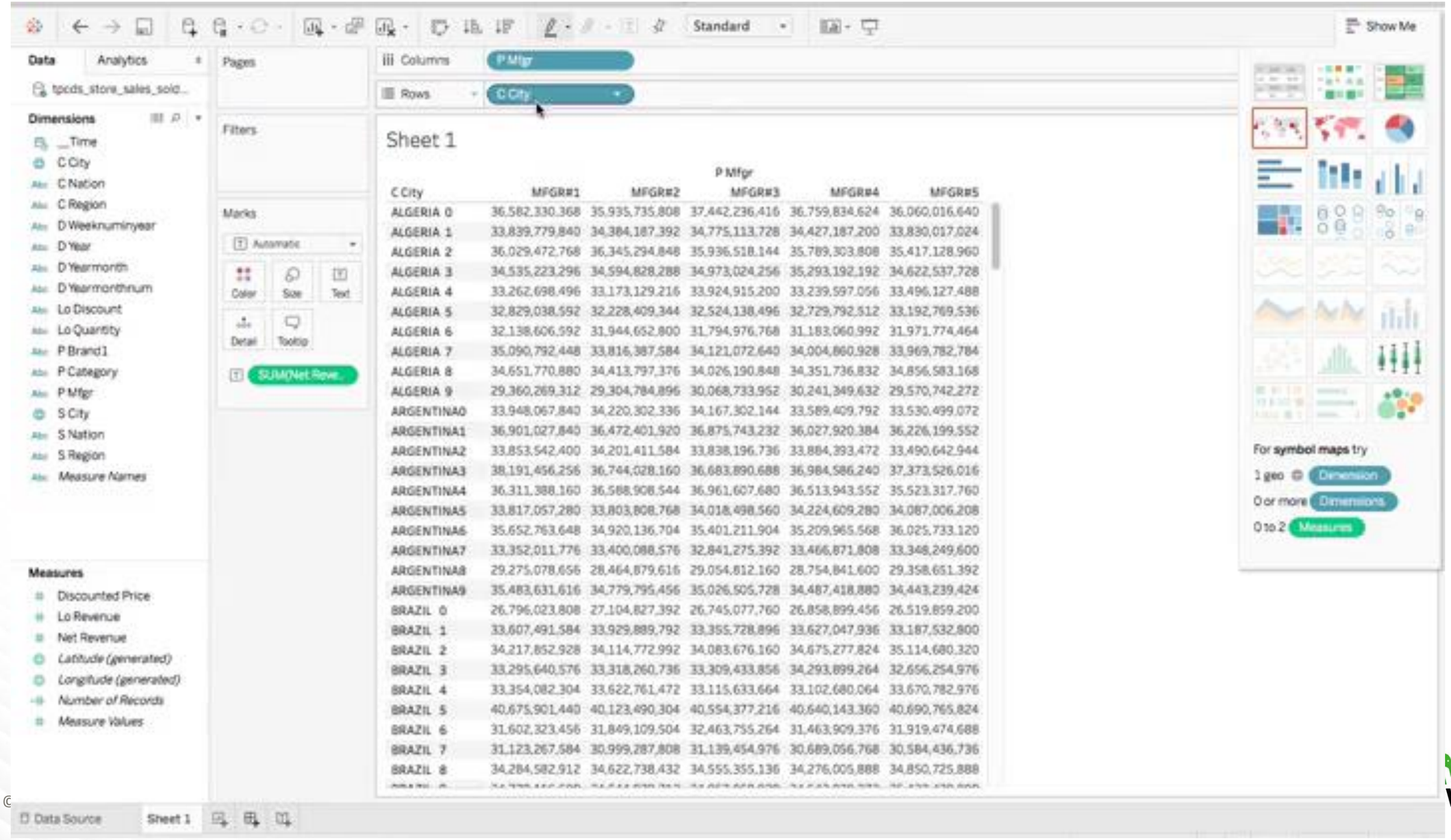
OLAP Analytics in Milliseconds with Hive over Druid



© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS

OLAP Analytics in Milliseconds with Hive over Druid



Agenda

Scalable Data Warehousing on Hadoop

Overview

Solution Architecture

Customer Use Cases

What's New

Roadmap

© Hortonworks Inc. 2011 – 2016. All Rights Reserved

HORTONWORKS

Roadmap At A Glance

Current

Scalable DW in Hadoop			
	HDP 2.6 (Current)	HDP 3.0 (Future)	Beyond HDP 3.0
Fast BI	• LLAP GA • Vectorized Decimal • SSD Cache • Text / JSON Cache LLAP • Tech Preview: Hive / Druid	• Fine-grained Resource Management Policies • View navigation for Druid tables • Transparent Parquet cache • Intermediate Results Spooling	• Query Cache • Admission Controls
SQL / EDW	• ACID MERGE • SQL: Cross Product, Multi Subquery, TPC-DS Complete	• Tables default to ACID capable • Column NOT NULL / Defaults • Surrogate Key Generation • Better Unicode support	• Multi-Statement Transactions • Improved HPL/SQL
Cloud	• LLAP Template for Hortonworks Data Cloud	• Replication / DR	• Full ACID support for S3 / WASB
Operations	• Hive View 2.0: Tools for DBAs • Tez UI: Hive Power Tools	• Hive Studio: Single pane of glass for Hive development and debugging	• Usage Reports • Schema Recommendations

Note: Roadmap is forward-looking and subject to change without notice.

© Hortonworks Inc. 2011 – 2017. All Rights Reserved

HORTONWORKS