Modernize Your Existing EDW With IBM Big SQL & Hortonworks Data Platform

Nagapriya Tiruthani, Offering Manager, IBM Big SQL
Carter Shanklin, Sr. Director, Product Management, Hortonworks
Roni Fontaine, Director, Product Marketing, Hortonworks

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Agenda

- Partnership
- Modernizing your existing EDW – Big SQL
- How Hive helps in the modernization
- Using Hive & Big SQL
- Resources / Q & A

Announcement –June 13, 2017

IBM and Hortonworks Deliver Data Science and Big SQL

Focus on extending **data science, machine learning and Big SQL to** analyze the data in **Apache Hadoop** systems

1. **IBM standardizes on HDP by leveraging Hortonworks Data Platform** as the core Hadoop distribution for Big SQL and DSX

2. Hortonworks introduces New Product Offerings:
   - **IBM Data Science Experience (DSX)**
   - **IBM Big SQL**

IBM + HORTONWORKS

Provides Data Science, Machine Learning & Big SQL

Provides Open Hadoop Data Platform

Make our clients competitive in their markets using advanced analytics faster and at scale

Hortonworks Data Platform & IBM BIG SQL the Bridge for Customers

HORTONWORKS

- #1 Pure Open Source Hadoop Distribution
- 1000+ customers and 2100+ ecosystem partners
- Compatibility
- Capabilities & Overlap between Hive (HWX) and Big SQL (IBM)

IBM

- Leader in SQL technology for Hadoop
- Leader in on premise and hybrid cloud data and analytics solutions
- Federation
- High Performance
- Improved Concurrence
- Complex queries
- Enterprise security features

Challenges with Traditional EDW

Companies have been building Enterprise Data Warehouses for over a decade.

These platforms have become unsustainably expensive and inadequate in compute performance, storage scalability and types of data stored.

Staging data consumes expensive disk space (typically 70%)

ETL and **data load** consumes valuable CPU (up to 90%)

Aged data is rarely queried because it is archived to cold storage

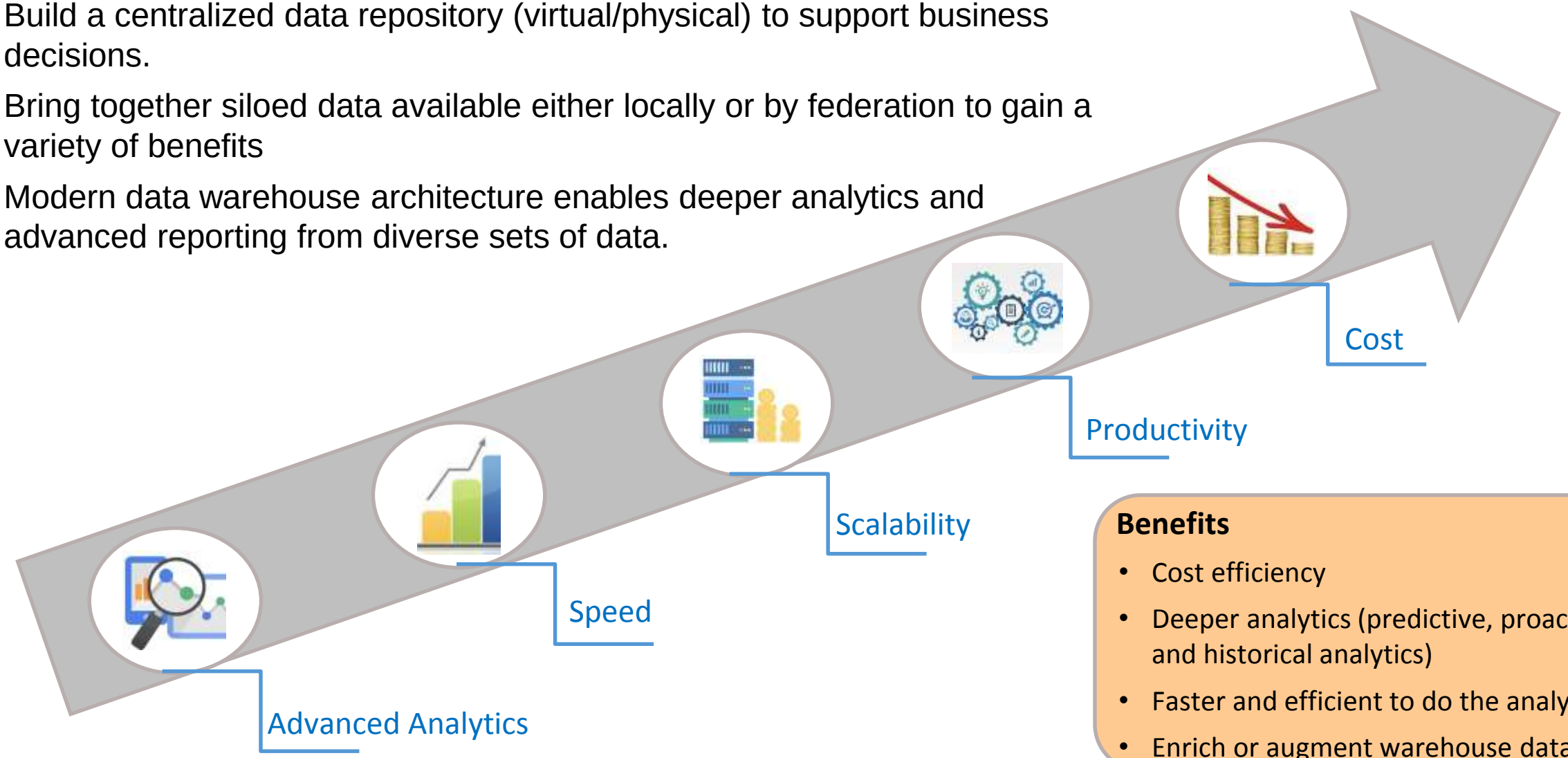
Other data never makes it to the EDW because of cost and project pressures

Schemas are challenged by modern data sources like social

Modernizing your EDW

What is EDW Modernization?

- Build a centralized data repository (virtual/physical) to support business decisions.
- Bring together siloed data available either locally or by federation to gain a variety of benefits
- Modern data warehouse architecture enables deeper analytics and advanced reporting from diverse sets of data.



Benefits

- Cost efficiency
- Deeper analytics (predictive, proactive and historical analytics)
- Faster and efficient to do the analytics
- Enrich or augment warehouse data

Do you have any of these challenges?

Want to modernize your EDW without long and costly migration efforts

Need to query, optimize and integrate multiple data sources from one single endpoint

Require skill set to migrate data from RDBMS to Hadoop/Hive

Operationalize machine learning

Offloading historical data from Oracle, Db2, Netezza because reaching capacity

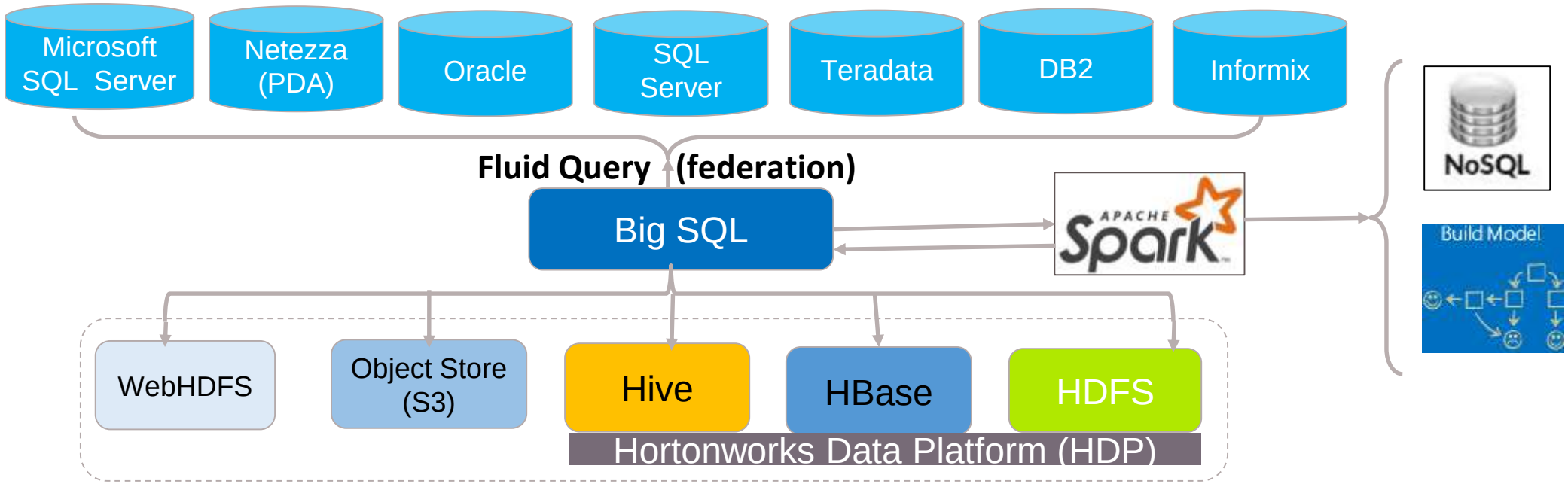
Slow query performance for SQL workloads

Here's How Big SQL Addresses These Challenges

- **Compatible** with Oracle, Db2 & Netezza SQL syntax
  - Modernizing EDW workloads on Hadoop has never been easier
  - Application portability (eg: Cognos, Tableau, MicroStrategy,…)
- **Federates** all your data behind a single SQL engine
  - Query Hive, Spark and HBase data from a single endpoint
  - Federate your Hadoop data using connectors to Teradata, Oracle, Db2 & more
  - Query data sources that have Spark connectors
- **Addresses** a skillset gap needed to migrate technologies
- **Delivers** high performance & concurrency for BI workloads
  - Unlock Hadoop data with analytics tools of choice
- **Provides** greater security while accessing data
  - Robust SQL based row filtering, column masking with role-based access control and Ranger integration
- **Operationalize** machine learning through integration with Spark
  - Bi-directional integration with Spark exploits Spark's connectors as well as ML capabilities

Data Virtualization

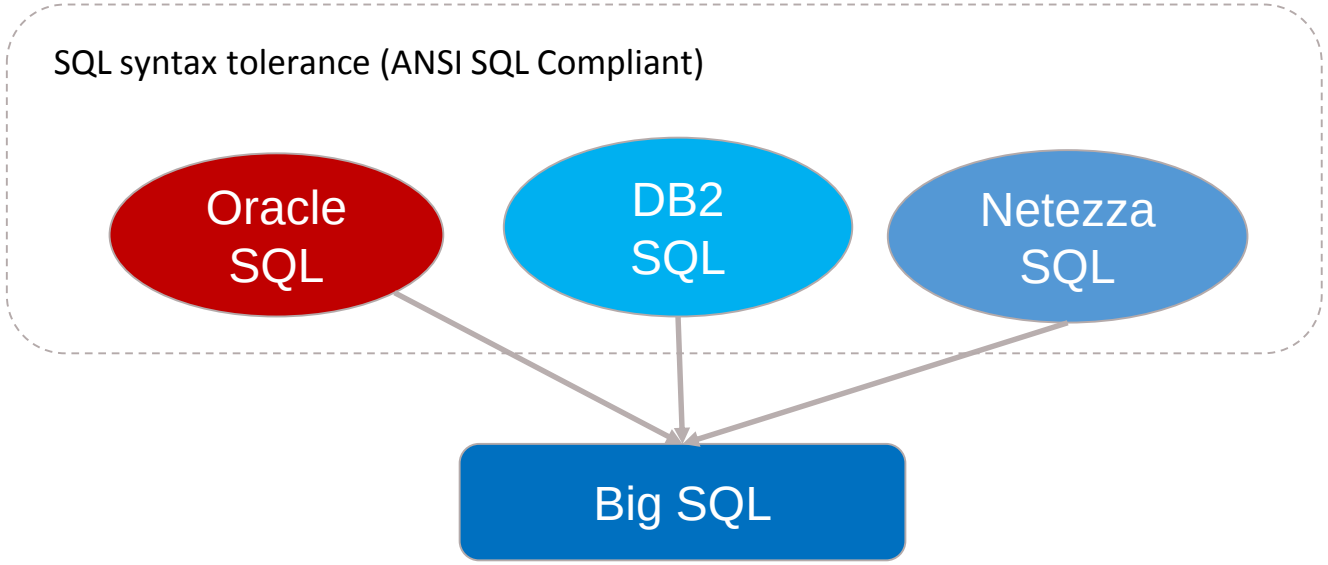
Big SQL allows query federation by virtualizing data sources and processing where data resides



Big SQL queries heterogeneous systems in a single query - only SQL-on-Hadoop that virtualizes more than 10 different data sources: RDBMS, NoSQL, HDFS or Object Store

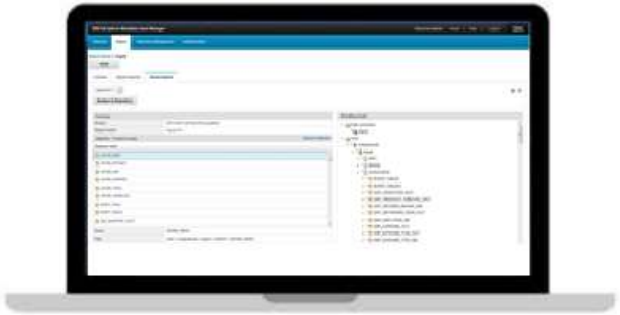
Data Offloading and Analytics

Big SQL is a synergetic SQL engine that offers SQL compatibility, portability and collaborative ability to get composite analysis on data



Cognos Analytics

InfoSphere Metadata Asset Manager



- **Automatic memory management** and **workload management** handles complex query execution without failures
- Executes **all** 99 TPC-DS queries in single stream and multi-stream environment
- Easy porting of **enterprise applications**
- Ability to work seamlessly with Business Intelligence tools like **Cognos, Tableau, etc.** to gain insights

Data Security

Big SQL offers row and column level access control (RBAC) among other security settings

Role Based Access Control enables separation of Duties / Audit

BRANCH_A

BRANCH_B

FINANCE (security admin)

Row Level Security

EMPNO	FIRST_NAME	SALARY	BRANCH_NAME
2	Chris	29007.57	Branch_A
3	Paula	14987.06	Branch_A
5	Pete	19114.22	Branch_A
8	Chrissie	24922.36	Branch_A

Total: 8 Selected: 0 ‹ 1 › 10 | 25 | 50 | 100

Row and Column Level Security

EMPNO	FIRST_NAME	SALARY	BRANCH_NAME
1	Steve		Branch_B
4	Craig		Branch_B
6	Stephanie		Branch_B
7	Julie		Branch_B

Total: 8 Selected: 0 ‹ 1 › 10 | 25 | 50 | 100

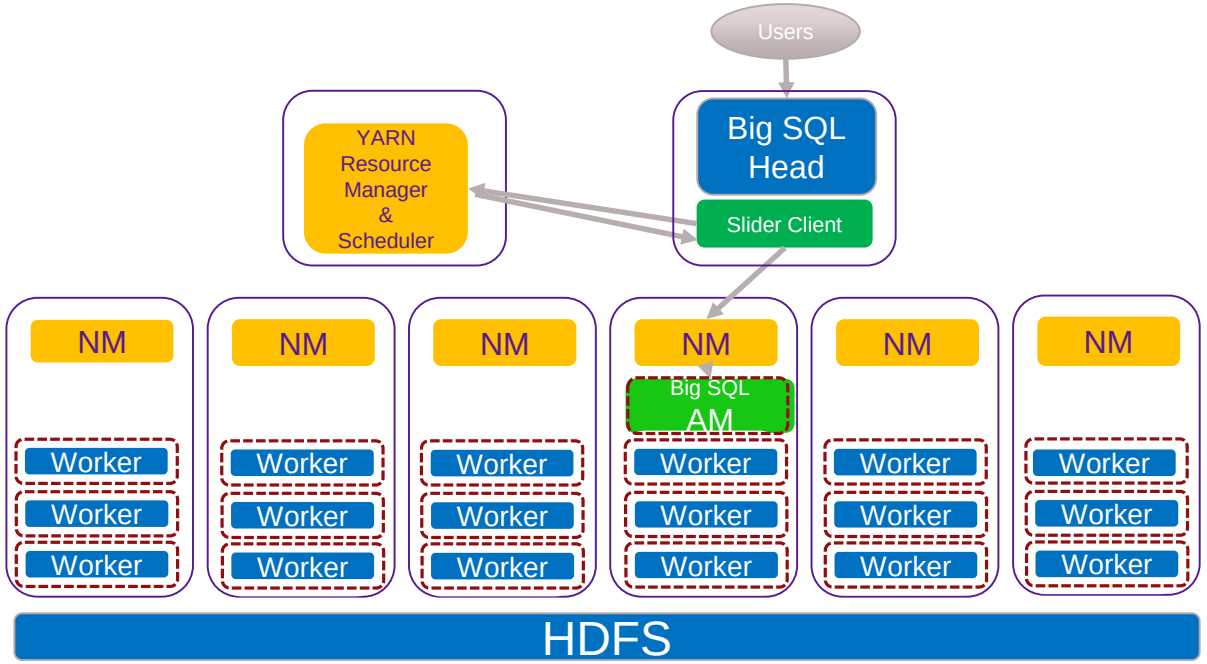
EMPNO	FIRST_NAME	SALARY	BRANCH_NAME
1	Steve	25970.38	Branch_B
2	Chris	29007.57	Branch_A
3	Paula	14987.06	Branch_A
4	Craig	22518.93	Branch_B
5	Pete	19114.22	Branch_A
6	Stephanie	26183.81	Branch_B
7	Julie	13629.91	Branch_B
8	Chrissie	24922.36	Branch_A

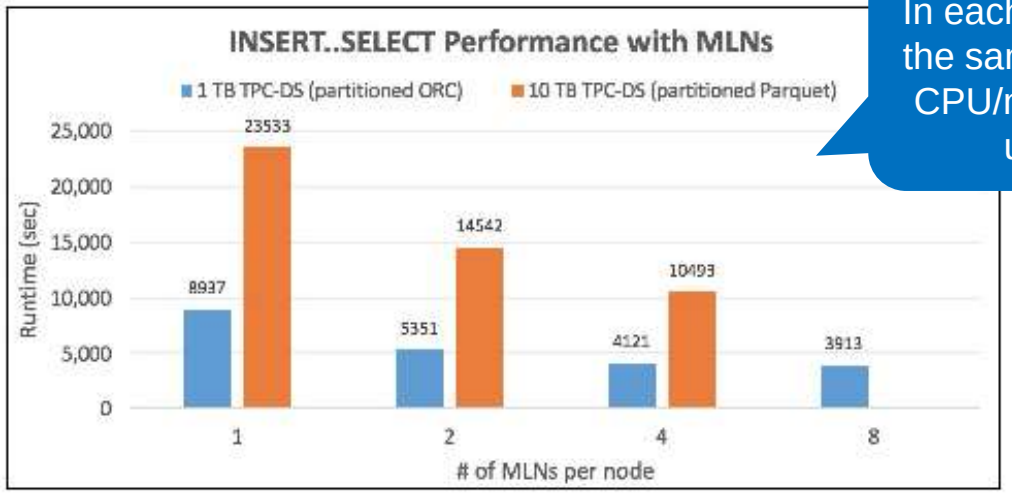
Total: 8 Selected: 0 ‹ 1 › 10 | 25 | 50 | 100

Performance: Big SQL Elastic Boost

Launch Multiple Workers per Host

More Granular Elasticity





In each scenario, the same TOTAL CPU/memory is used

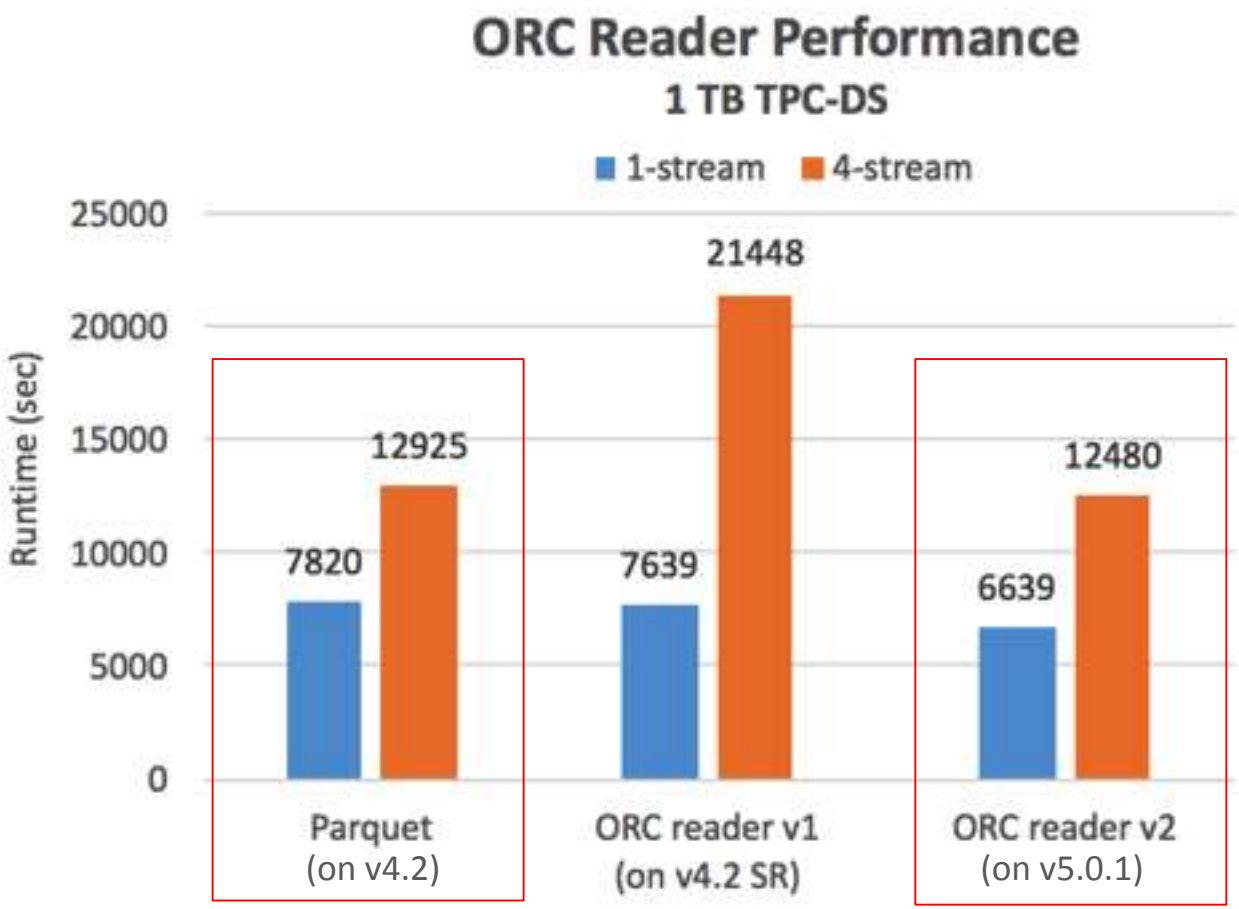
For both 1 and 10 TB TPC-DS dataset

2 Workers/Node: **1.6x speedup**

4 Workers/Node: **2.2x speedup**

Performance: Parquet vs ORC File Format

Enhancements to ORC file format has improved performance when compared to previous releases and it at par with Parquet file format



Spark Integration

Big SQL is a self-tuning memory management SQL engine that integrates with Spark 2.1

Spark 2.1 is a powerful analytic co-processor that complements the rich SQL functionality of Big SQL

Bi-directional integration allows Spark jobs can be executed from Big SQL

Tight integration with Spark enables Big SQL worker and Spark Executor to communicate in memory without writing to disk

Democratize Data Science and Machine Learning

*Leverage **Big SQL** throughout your journey*

Data Ingestion → Data Transformation/ Data Science/ Machine Learning → Data Visualization

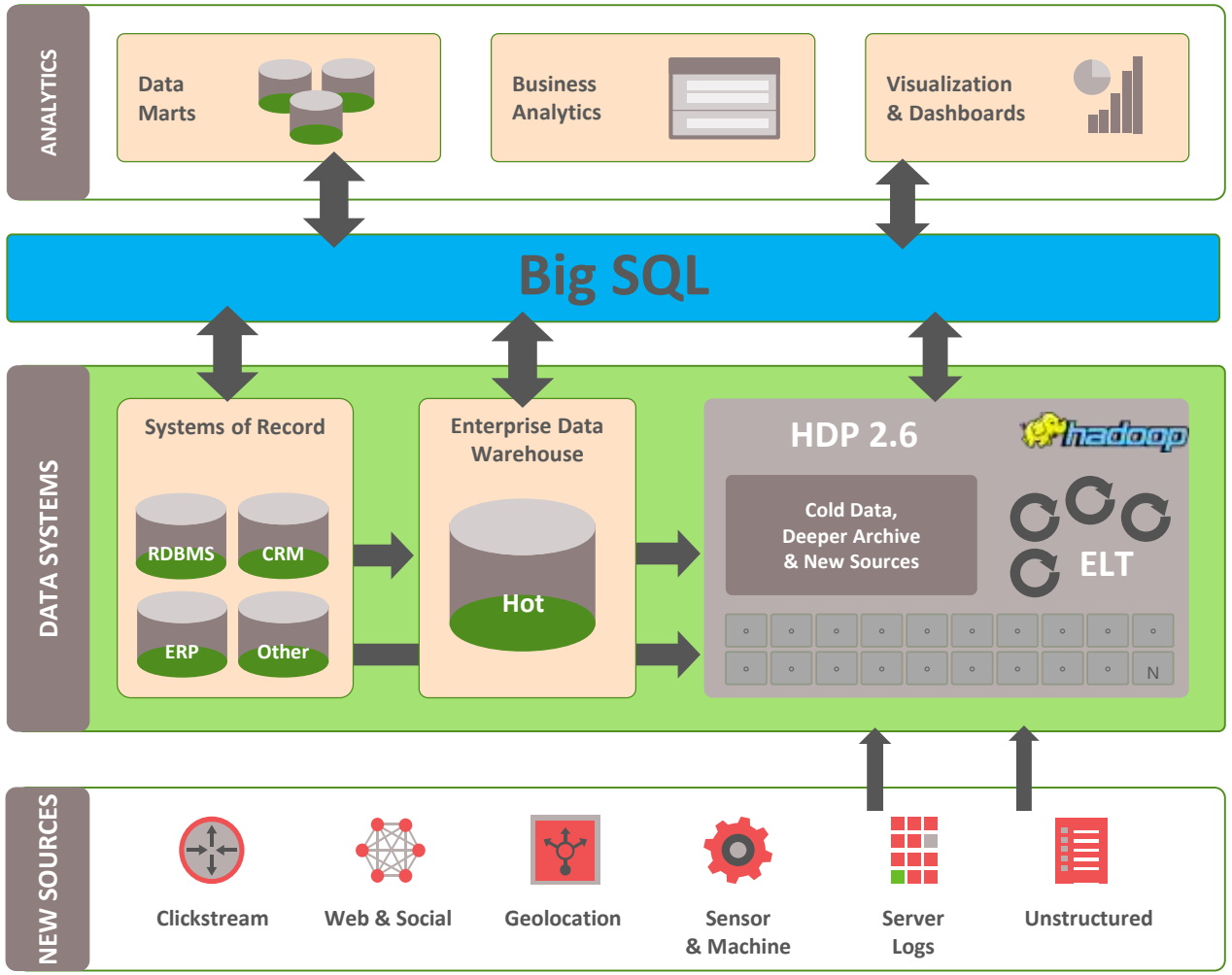
Virtualize disparate data sources like Hadoop, RDBMS, and Object Stores (S3) to join data in a single query

Manipulate data and operationalize data science models written in various languages

Perform data discovery, analyze, and visualize business results in notebooks or other BI tools

Top Use case – EDW Optimization

Realize Cost Savings Faster with Big SQL



Archive Data away from EDW
- Move cold or rarely used data to Hadoop as active archive
- Store more of data longer

Offload costly ETL process
- Free your EDW to perform high-value functions like analytics & operations, not ETL
- Use Hadoop for advanced ETL

Optimize the value of your EDW
- Use Hadoop to refine new data sources, such as web and machine data for new analytical context
- Access old data in traditional RDBMS using Big SQL's federation
- Combine data in different silos without duplication

Exploit Spark to avail latest technologies
- Spark integration brings data from NoSQL databases and other non-traditional sources
- Make Spark's use cases secure with granular security defined in Big SQL

Reduce the migration effort & skillset gap
- Use existing investment in Oracle, Db2 and Netezza technology skills
- Big SQL allows you to migrate applications with out major code rewrites and additional SQL development

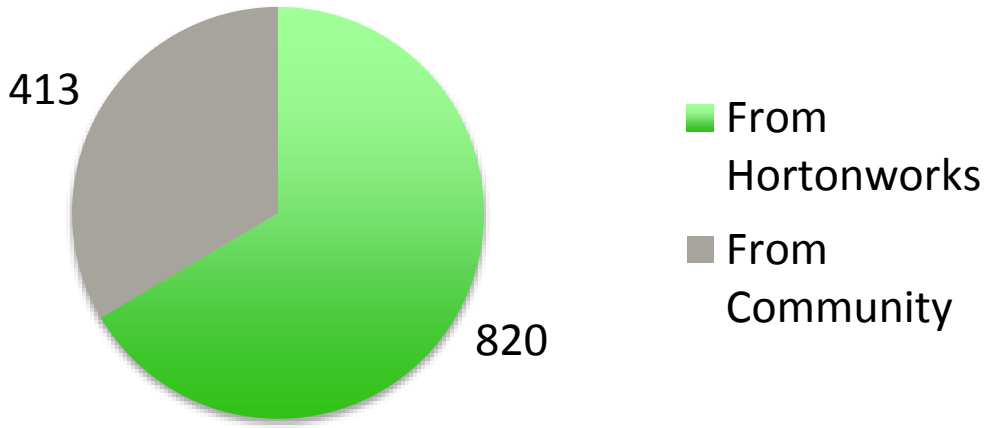
How Hive Helps in the Modernization

HORTONWORKS®
POWERING THE FUTURE OF DATA™

HDP 2.6: A Major Milestone for Apache Hive

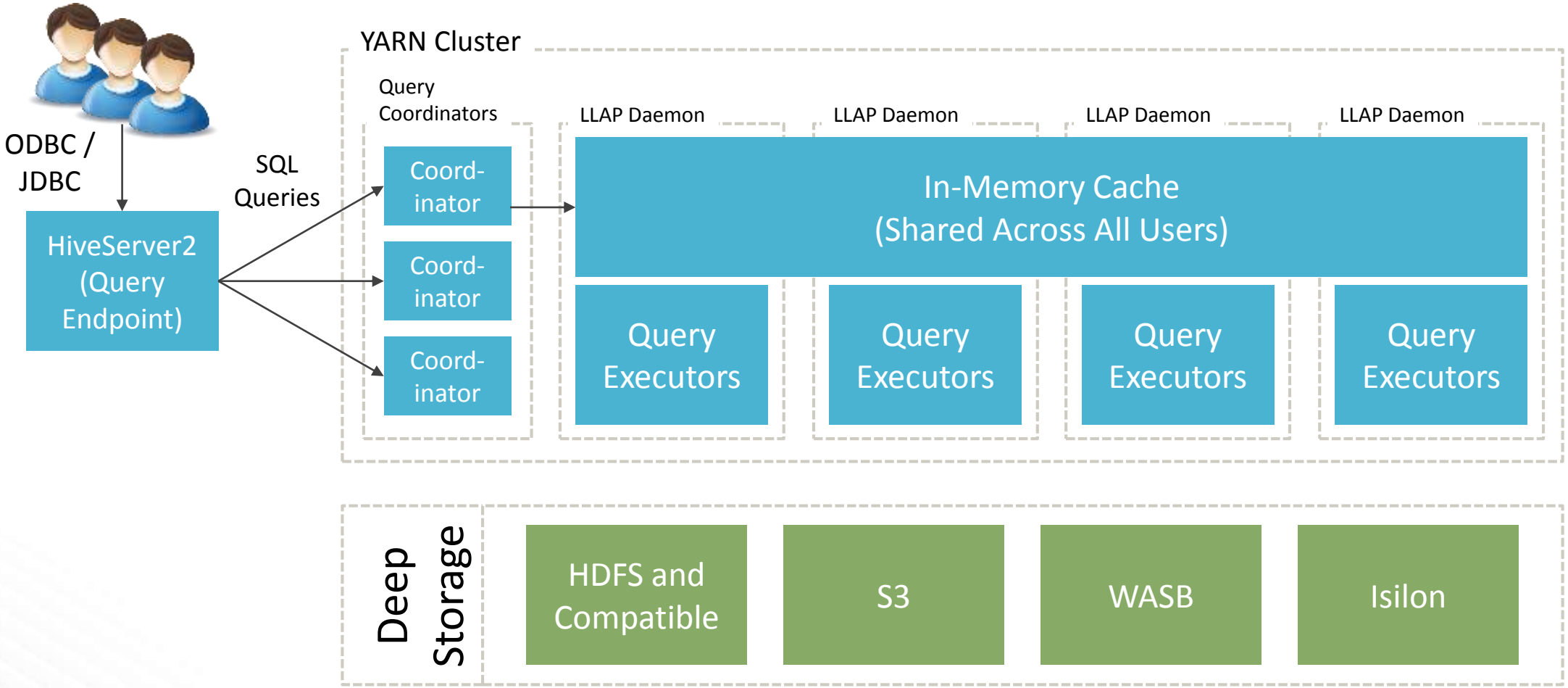
- Major Improvements:
  - Hive LLAP Now GA
  - ACID MERGE
  - SQL: All 99 TPC-DS out-of-the-box with only trivial rewrites
  - Hive View 2.0: Great Features for DBAs
  - Diagnostics: Tez UI Total Timeline View
  - Hive OLAP Indexes powered by Druid
- At a High Level:
  - 1200+ features, improvements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

HDP 2.6 Improvements



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

Hive LLAP: MPP Performance at Hadoop Scale



HDP 2.6 Makes Hadoop Data Management a Reality with SQL MERGE

- Hive implements ANSI-standard SQL MERGE.
- MERGE makes data maintenance 8x simpler with 5x higher performance.
- Legacy Hive or Spark approaches don't protect applications against dirty reads or partial failures.

Complexity of a Type 2 SCD Update with and without MERGE				
	Number of Queries	Number of Full Table Scans	Isolation	Applications Protected From Partial Failures
Hive MERGE	1	1	Yes	Yes
Old Techniques	8	5	No	No

Comprehensive SQL in Hive Including All 99 TPC-DS Queries

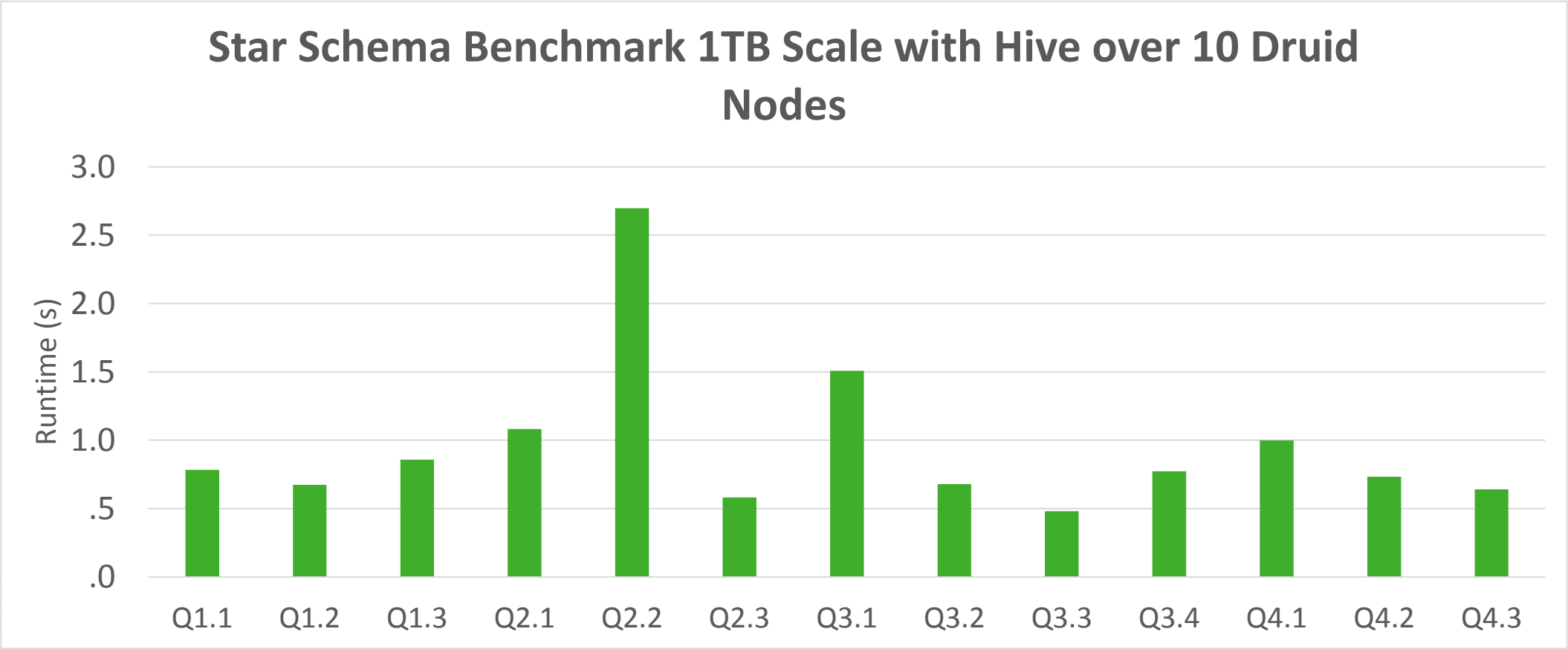
Highlights

- 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.

Announcing: Druid is Now GA in HDP 2.6.3

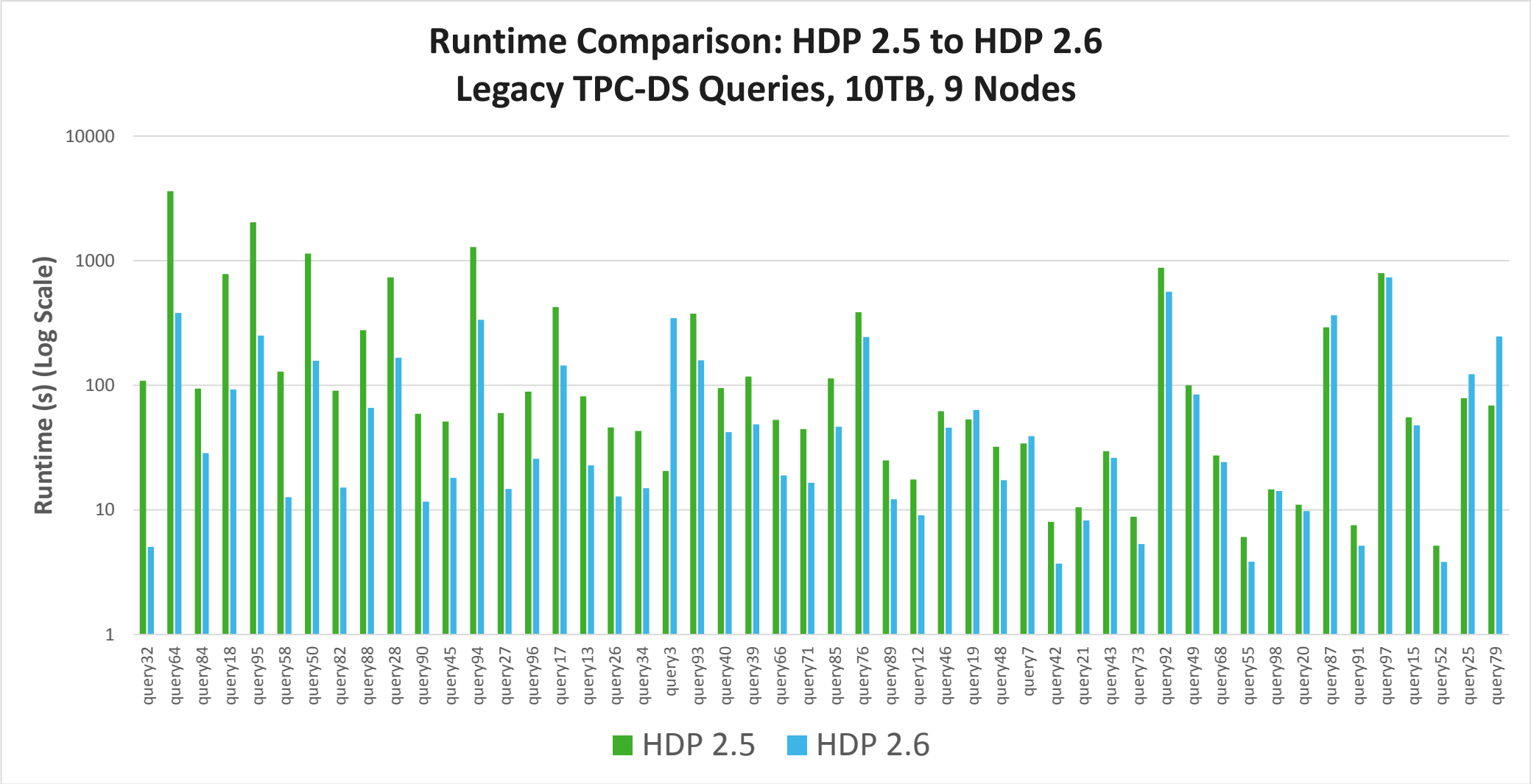
- ✓ Analyze Streaming and Historical Data with SQL
- ✓ Powerful Visualization
- ✓ Simple management and monitoring with Ambari
- ✓ Fine-grained security
- ✓ Integrates with Hortonworks SAM for simple development

OLAP Analytics in Milliseconds with Hive over Druid



https://hortonworks.com/blog/apache-hive-druid-part-1-3/#comment-24833

Runtime Comparison: HDP 2.5 and HDP 2.6 (Lower is Better)



Using Hive & Big SQL

HORTONWORKS®
POWERING THE FUTURE OF DATA™

How Big SQL Complements HDP

Application Portability & Integration

Data shared with Hadoop ecosystem
Comprehensive file format support
Superior enablement of IBM and Third Party software

Performance

Modern MPP runtime
Powerful SQL query rewriter
Cost based optimizer
Optimized for concurrent user throughput
Results not constrained by memory

Rich SQL

Comprehensive SQL Support
IBM SQL PL compatibility
Extensive Analytic Functions

Federation

Distributed requests to multiple data sources within a single SQL statement

Main data sources supported:
DB2, Teradata, Oracle, Netezza,
Informix, SQL Server

Enterprise Features

Advanced security/auditing
Resource and workload management
Self tuning memory management
Comprehensive monitoring

Breaking Things Down: Where IBM Big SQL Shines

- Run Oracle, DB2 or Netezza Workloads on Hadoop
- Federate Hadoop and Non-Hadoop Data
- Complex SQL Workloads

Breaking Things Down: Where Apache Hive Shines

- Fast SQL That Scales from Terabytes to Petabytes
- Easy to Keep Data Fresh with ACID MERGE
- Join Historical and Streaming Data in Real Time

Resources

HORTONWORKS®
POWERING THE FUTURE OF DATA™

Resources

- HWX Big SQL Web Page: https://hortonworks.com/partners/ibm-bigsql/
- Big SQL Solutions Sheet: https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2017/08/IBM-Big-SQL-Solution-Sheet_final.pdf
- Big SQL Data Sheet https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2017/08/IBM-Big-SQL-Datasheet_final.pdf
- Big SQL Resources
  - Big SQL Web Page/Sandbox https://www.ibm.com/us-en/marketplace/big-sql
  - Big SQL Master Class Videos https://www.youtube.com/playlist?list=PL7FnN5oi7Ez9itAnZ6rs9A30YYjVB1wN_
- Big SQL Blog: https://hortonworks.com/blog/big-sql-apache-hadoop-across-enterprise/

Questions?