Making Enterprise Big Data Small with Ease

Nagapriya Tiruthani, Offering Manager, IBM Db2 Big SQL

Roni Fontaine, Director, Product Marketing, Hortonworks

March 28, 2018

Presenters

Nagapriya Tiruthani

Offering Manager

IBM Db2 Big SQL

Roni Fontaine

Director of Product Marketing

HDP, Operations and Cloud

HORTONWORKS

Agenda

- Data today and its benefits
- Db2 Big SQL Use cases
- Db2 Big SQL's Federation and how it works
- How Hive can Help
- Db2 Big SQL and Hive Working Together
- Resources / Q & A

HORTONWORKS

Rules of the Game have Changed

Need to accommodate volume, variety and velocity of data

Increased adoption of advanced analytics and self-service discovery

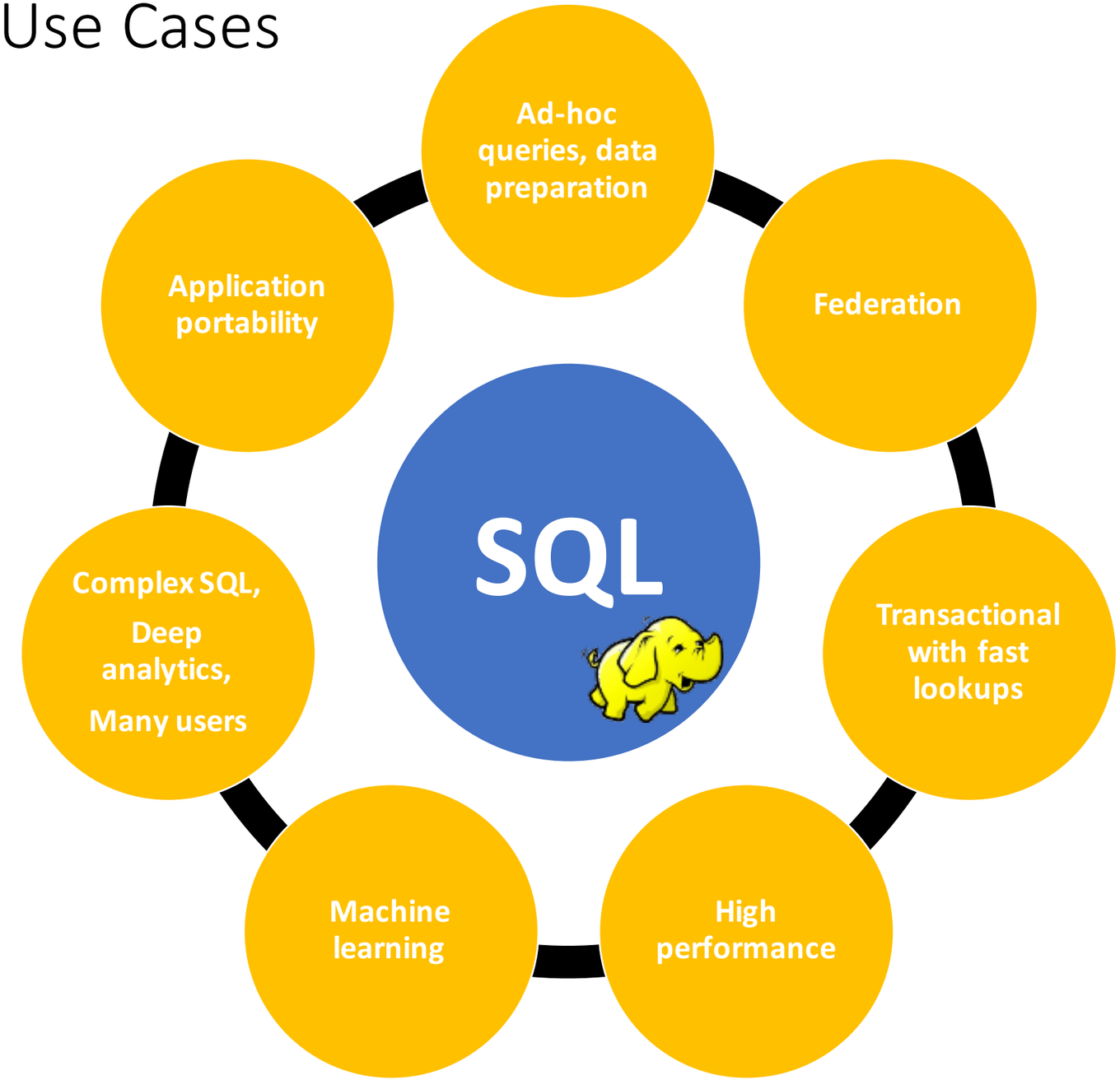
Mobile and IoT are driving need for more real-time, accurate information

Need for agile data services with high levels of security

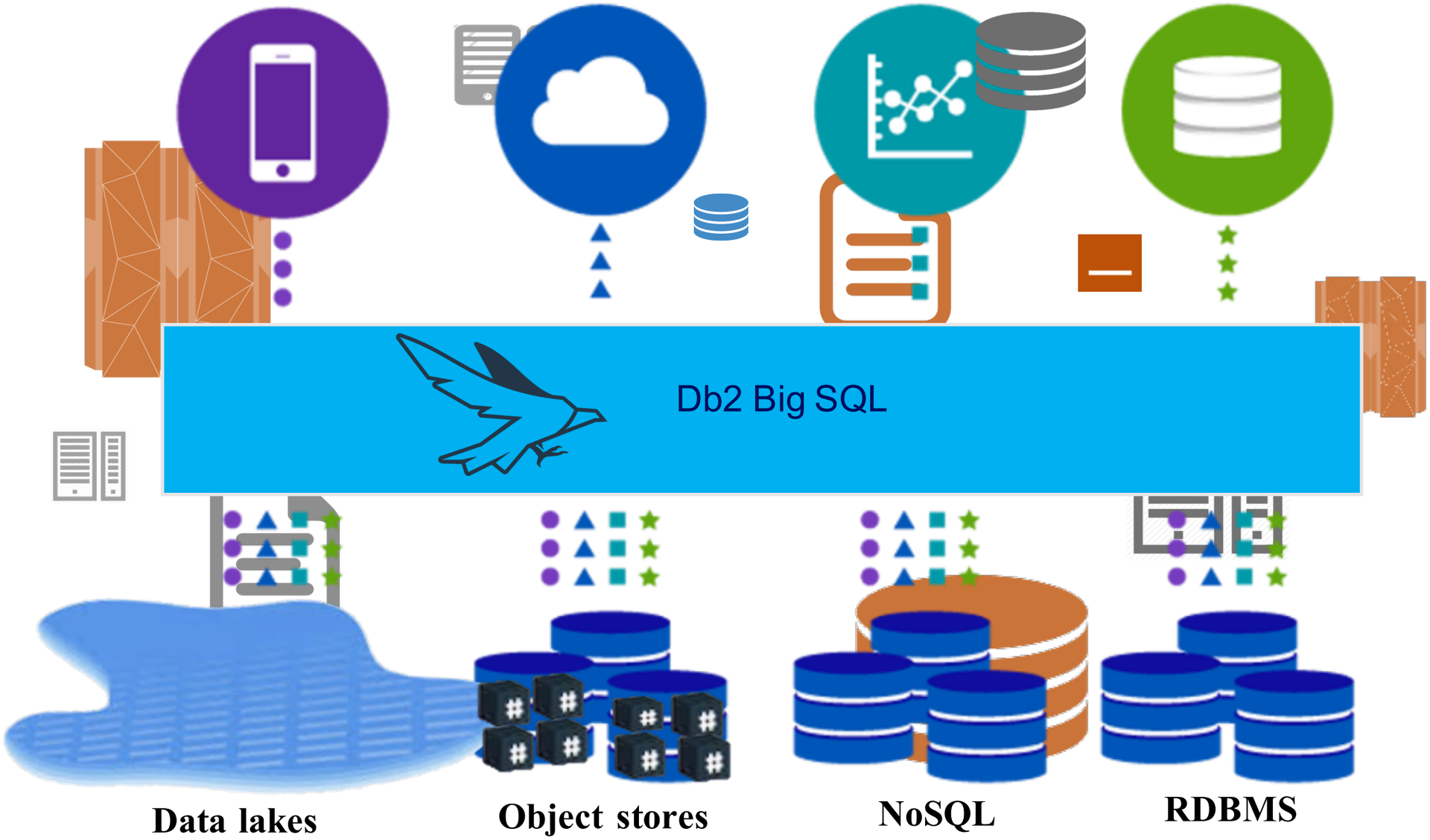
Benefits of Bringing all the Data Together

- Business Intelligence
  - Extending a warehouse to include outside data
  - Prototyping new applications using disparate data
  - Handling mergers and acquisitions as well as staged migrations between databases
  - Warehouse or ODS maintenance (trickle-feed)
- Business Integration
  - Real-time, targeted access to multiple operational databases
  - Replicate/consolidate distributed data to a central store
  - Portal development – access disparate data and make available for Web deployment
- Warehouse Augmentation
  - Capacity Relief
  - Queryable archive
  - Multi-temperate business models
  - Migrate dimensioned data from 3rd party RDBMS
  - Discovery & Exploration

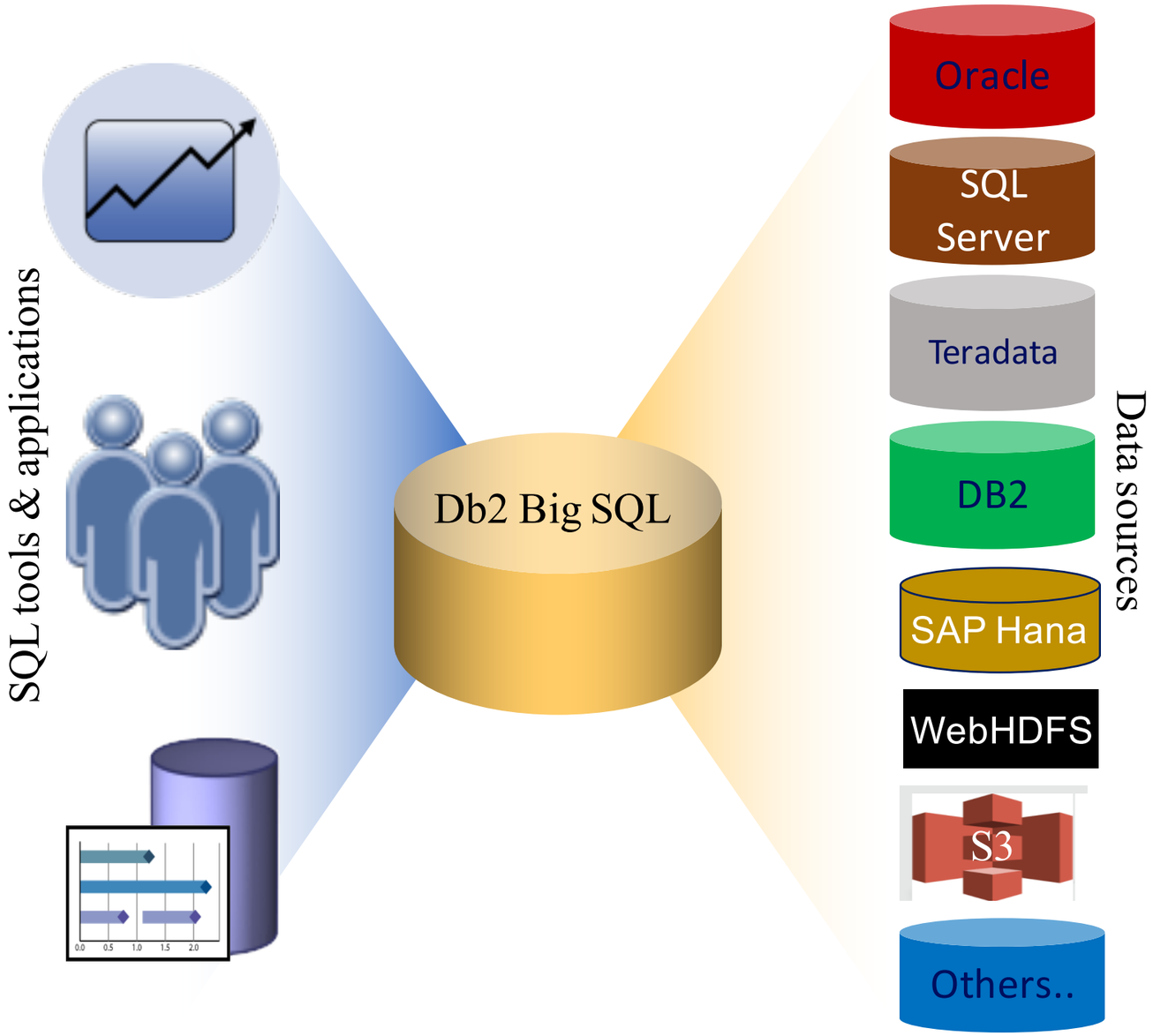
Db2 Big SQL Use Cases



Federation: Eliminating the Information Bottleneck



What does Db2 Big SQL Federation get you?



Transparent

- Appears to be one source
- Programmers don't need to know how / where data is stored

Heterogeneous

- Accesses data from diverse sources

High Function

- Full query support against all data
- Capabilities of sources as well

Autonomous

- Non-disruptive to data sources, existing applications, systems.

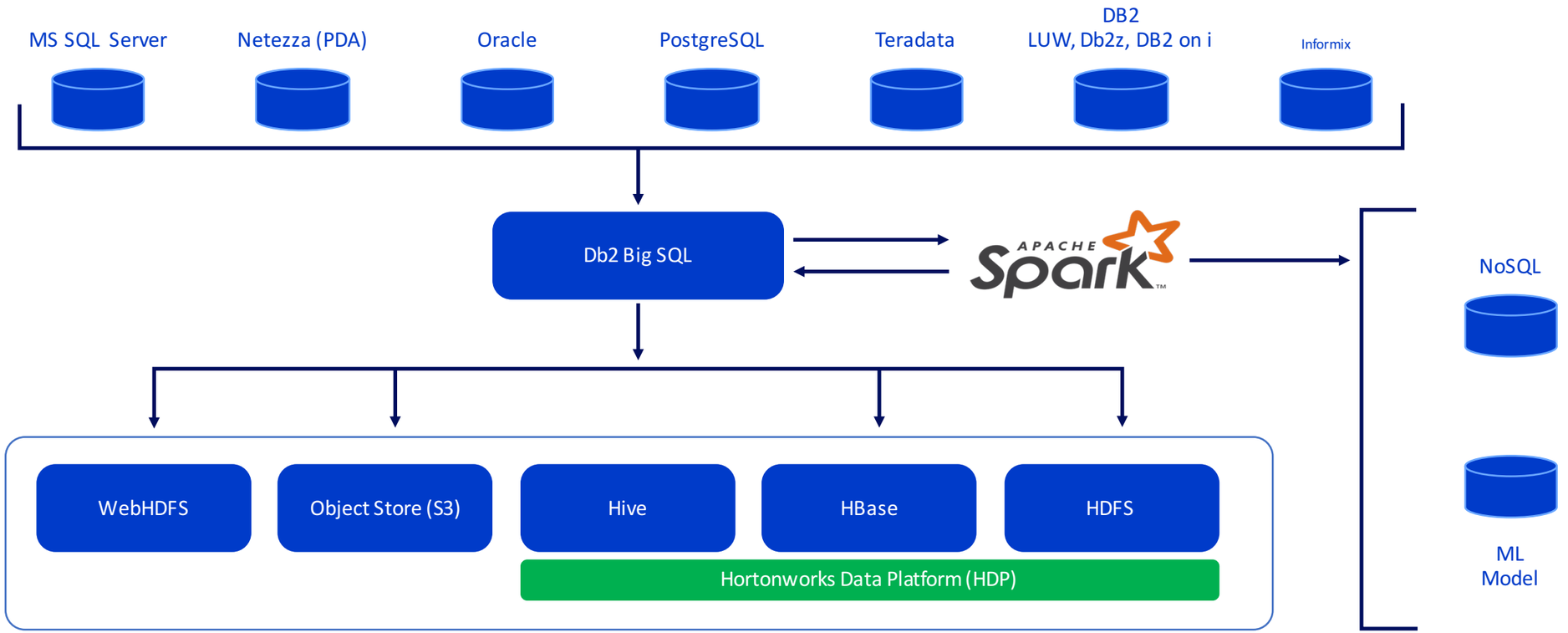
High Performance

- Optimization of distributed queries

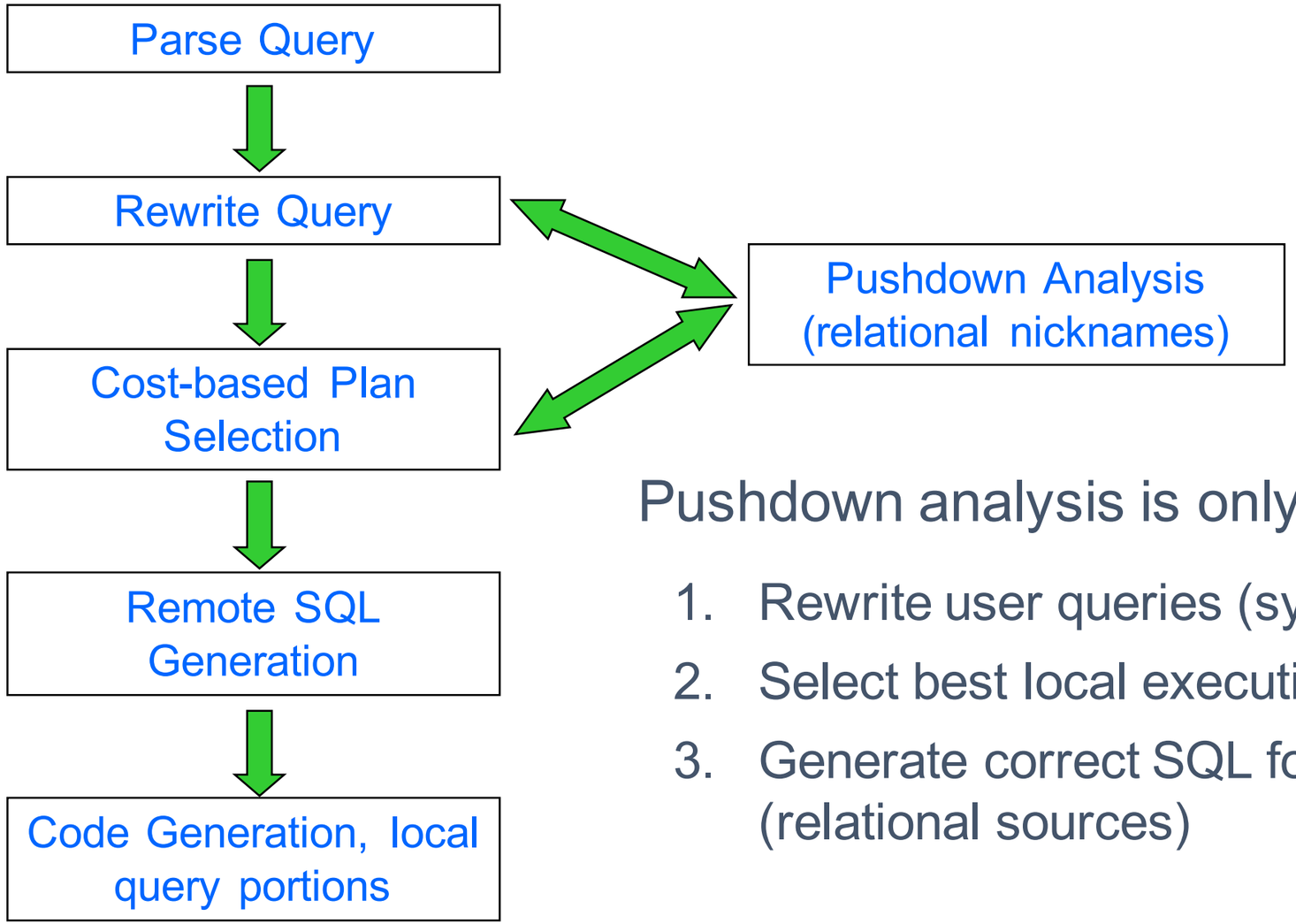
Federation – Virtualize Heterogeneous Data

Db2 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



Optimizer Flow for federated queries



Pushdown analysis is only part of the optimizer's job.

1. Rewrite user queries (syntactic transformations)
2. Select best local execution plan
3. Generate correct SQL for remote query fragments (relational sources)

“Pushdown” of Query Operations

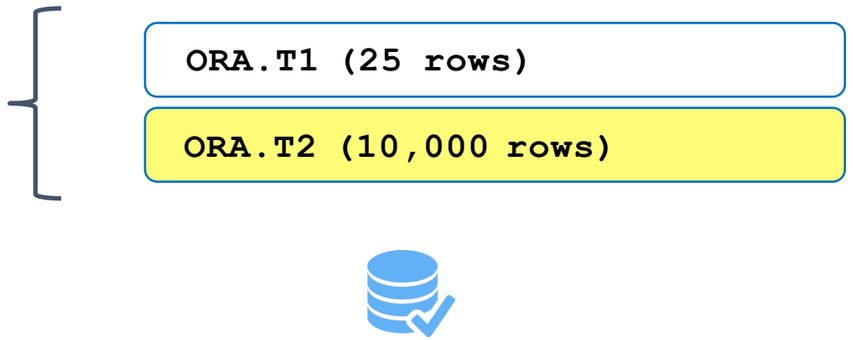
- Db2 Big SQL decides whether some or all parts of a query can be "pushed-down", i.e. processed at the remote data source(s). Pushdown-ability depends on
  - availability of needed functionality at remote source
  - server options (example: is collating sequence at Federated server and remote source the same?)
- Example: A remote source that can handle an equality predicate, but not count(*)....

Application	IBM database	Non-IBM database
`Select count(*) from t1 where col = 27`	`Select count(*) from t1 where col = 27`	`Select ‘1’ from t1 where col = 27`

Actual Pushdown is Cost Based

- Just because processing can be pushed down doesn't mean it will be. Decision influenced by estimates of rows processed/returned.
- Consider a join of two nicknames ORA.T1 and ORA.T2 on a single remote source that is "nearly" a Cartesian product. May be better to do the join at the Db2 Big SQL engine to avoid retrieval of many rows.
- Retrieving (10,000 + 25) rows to do a local join is probably faster than retrieving (10,000 * 25) = 250,000-row remote join result

```
Select … from ORA.T1, ORA.T2
where T1.a = T2.b;
```



Rich Capabilities that Brings Data Together

- ✓ Easily access information on demand
- ✓ Combine data in Hadoop with disparate sources to form a data lake
- ✓ Quickly extend your data warehouse by enriching it

Connect

- Quick access to Data value
- Common Framework
- ODBC/JDBC
- Spark integration enables new data sources
- Connect all data sources in single query

Query

- Intelligent Query Routing
- Cost-based optimizer
- SQL pushdown
- Local data caching
- ANSI-compliant SQL

Monitor

- Easily define & manage through a common UI
- Simple point & click to discover and query
- Monitor and visualize active queries

Data Placement

- Schema conversion when moving data
- Bulk data copy to Hadoop
- Filtered subsets of data

IBM Db2 Big SQL

One engine for all enterprise needs on Hadoop

- ✓ Executes complex queries with high performance
- ✓ Combine data in Hadoop with disparate sources to form a data lake
- ✓ Enables reusing application and skills

Query	Augment	Monitor	Security	Performance
▪ Intelligent Query Routing ▪ Cost-based optimizer ▪ SQL pushdown ▪ ANSI-compliant SQL ▪ Quick access to Data value ▪ Query with open source Hadoop file formats	▪ Access data in RDBMS & NoSQL data sources ▪ Operationalize ML models ▪ Spark integration for in-memory data exchange ▪ Local data caching for federated queries using MQTs	▪ Automatic memory manager manages queries to completion ▪ Manage workloads and prioritize ▪ Audit queries ▪ Simple point & click to discover and query ▪ Monitor and visualize active queries	▪ Granular SQL level access control for row filtering and column masking ▪ Define policies in Ranger for centralized security management	▪ Create MQTs to cache data aggregate for fast response ▪ Enable elastic boost to maximize resources consumption and parallel execution

To Summarize

- With Db2 Big SQL, users can focus on **what they want to do**, and not worry about how it is executed
- Data Scientists/Business Analysts can be **3-4 times** more productive using Db2 Big SQL compared to Spark SQL.
- Proof points:
  - Able to successfully run all 99 TPC-DS queries @ 100TB in 4-concurrent streams
  - Performance leadership
  - Uses fewer cluster resources
  - Simpler configuration with mature self-tuning and workload management features

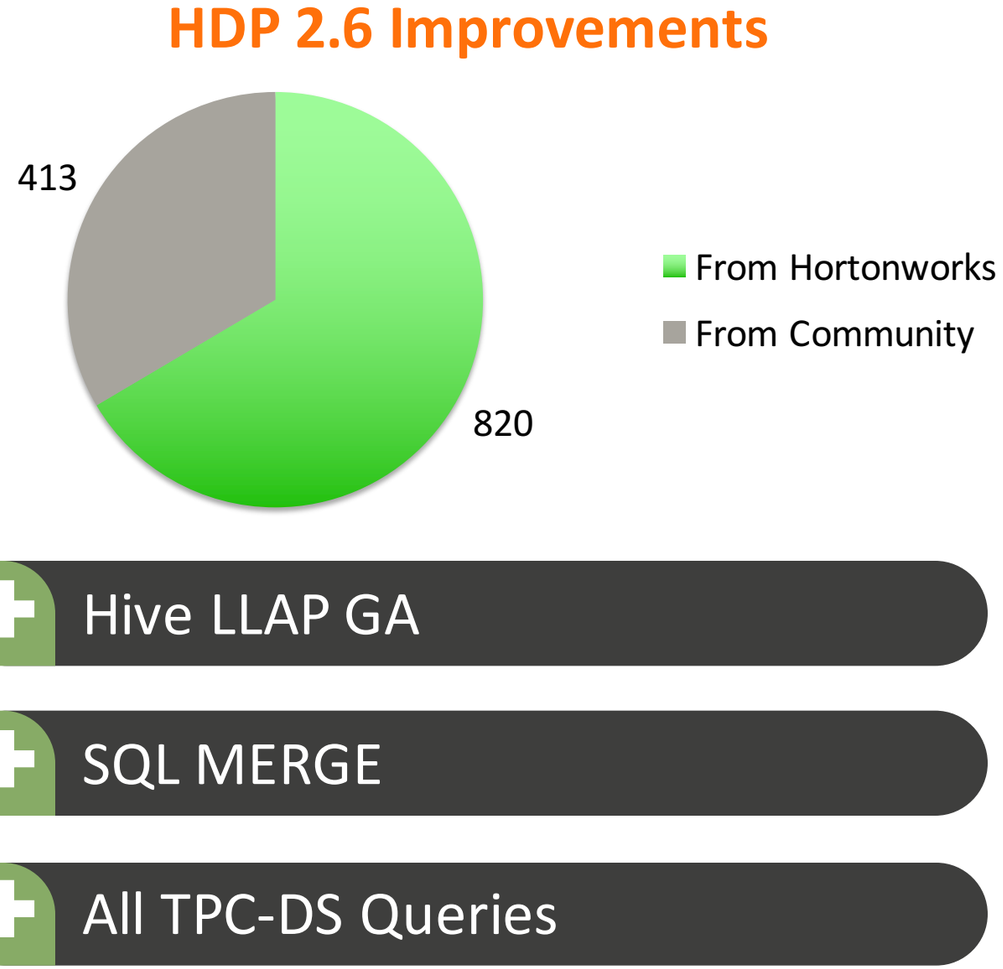
Db2 Big SQL is the best SQL over Hadoop engine for complex analytical workloads

How Hive Helps

HORTONWORKS

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 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

HORTONWORKS

Using Hive & Db2 Big SQL

How Db2 Big SQL Complements Hive

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

HORTONWORKS

Breaking Things Down: Where IBM Db2 Big SQL Shines

HORTONWORKS

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

HORTONWORKS

Resources

Resources

- HWX Db2 Big SQL Web Page: lhttps://hortonworks.com/partners/ibm-bigsql/
- Db2 Big SQL Solutions Sheet: https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2017/08/IBM-Big-SQL-Solution-Sheet_final.pdf
- Db2 Big SQL Data Sheet https://2xbbhjxc6wk3v21p62t8n4d4-wpengine.netdna-ssl.com/wp-content/uploads/2017/08/IBM-Big-SQL-Datasheet_final.pdf
- Db2 Big SQL Resources
  - Db2 Big SQL Web Page/Sandbox https://www.ibm.com/us-en/marketplace/big-sql
  - Db2 Big SQL Master Class Videos https://www.youtube.com/playlist?list=PL7FnN5oi7Ez9itAnZ6rs9A30YYjVB1wN_
- Db2 Big SQL Blog: https://hortonworks.com/blog/big-sql-apache-hadoop-across-enterprise/

HORTONWORKS

Thank you