Live Data Analytics

—

Paul Scott-Murphy,
VP Product Management, WANdisco

Nagapriya Tiruthani
IBM Db2 Big SQL Offering Manager, IBM

think 2018

Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Why Analytics?

Some industry use cases

- Cross sell & Upsell
- Churn
- Predicting lifetime value of customer

- Sentiment analysis
- Digital marketing

- Disease prediction
- Medication effectiveness

- Discount offering
- Demand forecasting

- Dynamic pricing
- Predict flight delays

- Claims prediction
- Fraud and risk detection

Why Live Data?

Rules of the Game have Changed

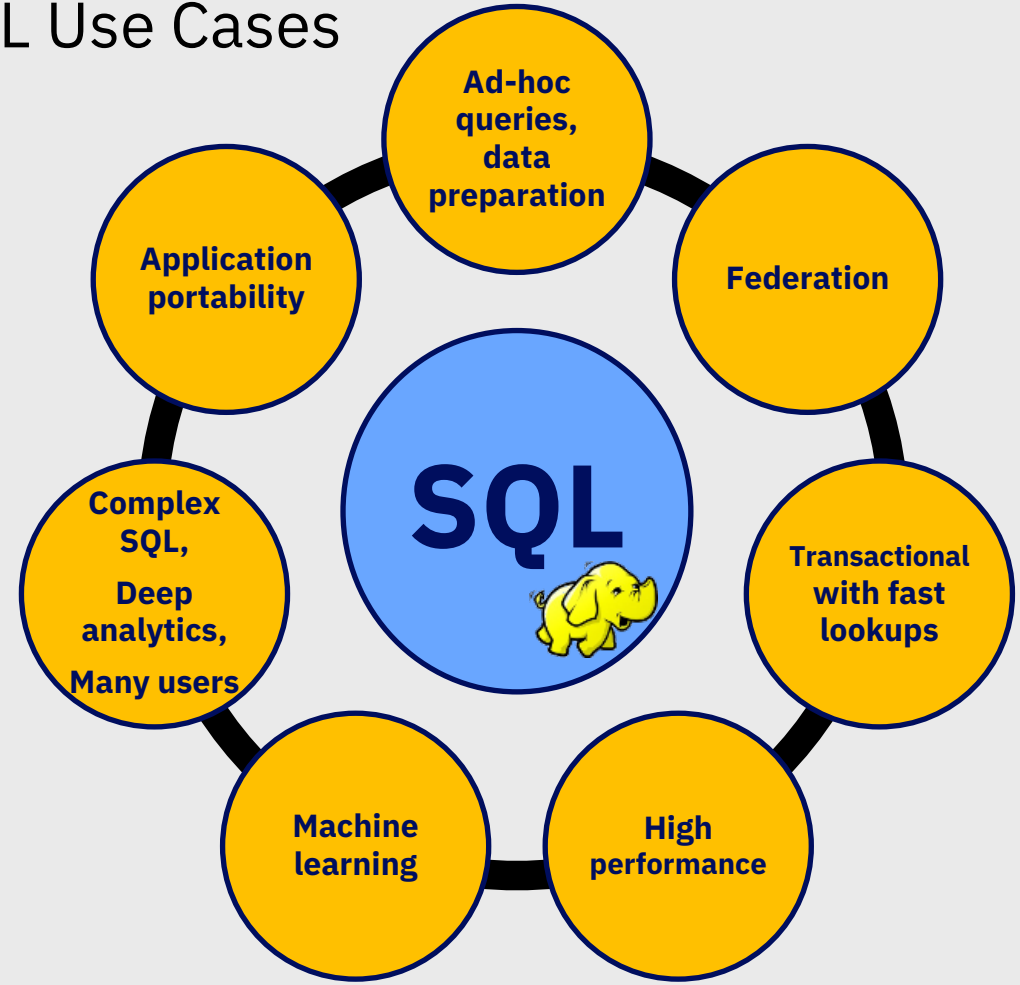
Need to accommodate volume, variety and velocity of data

Increased adoption of advanced analytics and self-service discovery

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

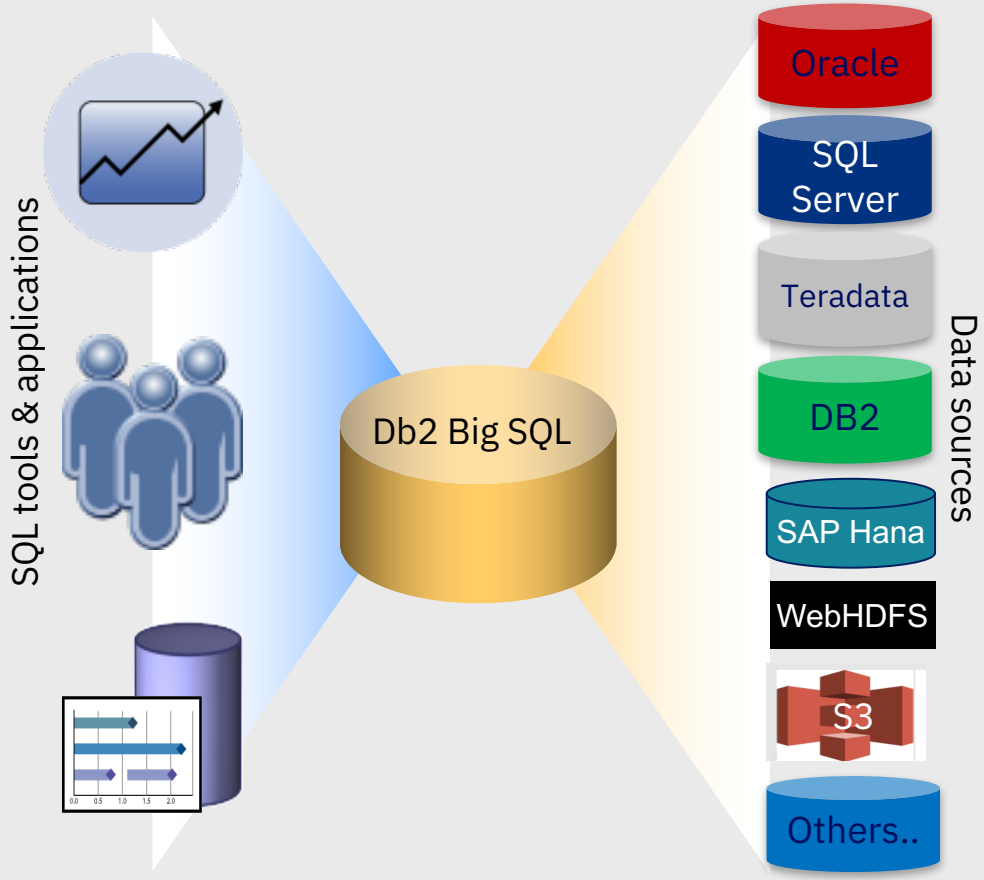
Need for agile data services with high levels of security

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

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

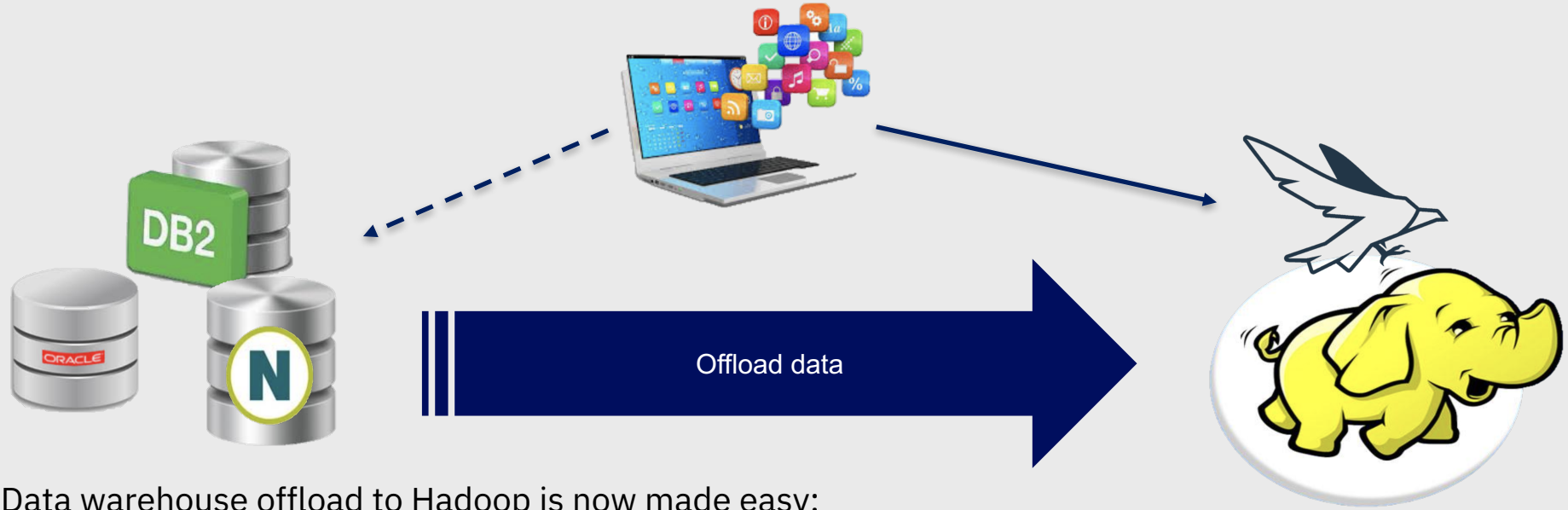
Monitor

- 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

Application Portability: Move Applications without Re-tooling



Data warehouse offload to Hadoop is now made easy:

- Write one, run anywhere…
- Easy porting of applications
- Reuse skills of DBAs/ developers who know ANSI SQL

Db2 Big SQL is the best platform for offloading Oracle Data Marts and Warehouses to Hadoop

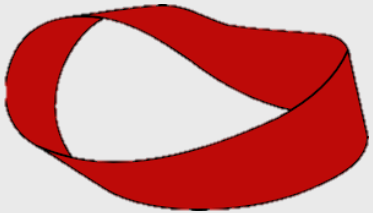
Query Execution

Here's why Db2 Big SQL can get you the best execution for complex queries and many concurrent users with high performance

Performance

Materialized Query Tables

Advanced Statistics



Elastic Boost

World Class Cost Based Optimizer

Concurrent users

Self Tuning Memory Manager

Advanced Workload manager

Complex query

SQL Compatibility

Hardened runtime

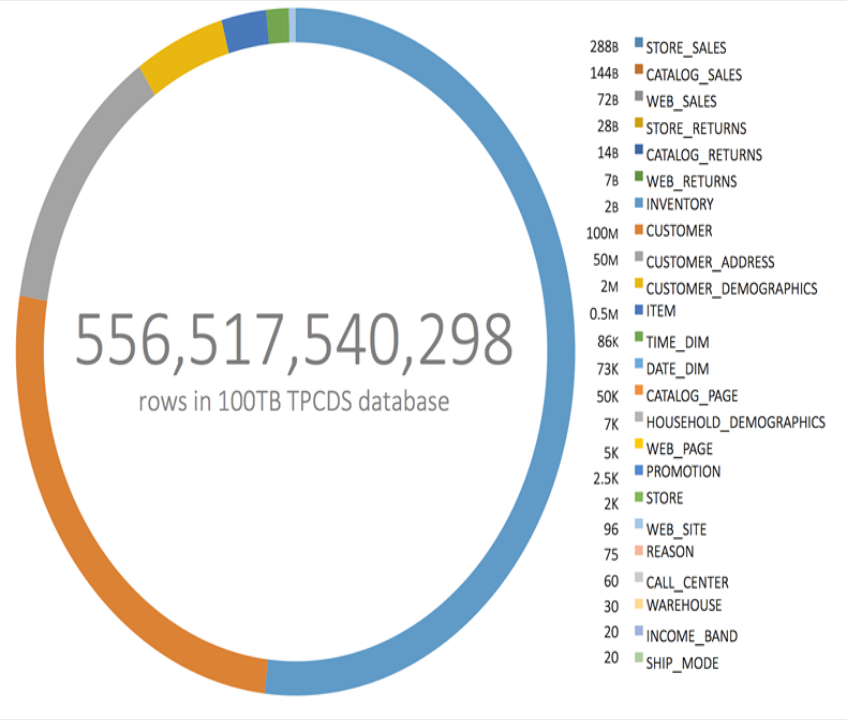
Native Row & Columnar stores

Query rewrite

Query Performance at a Glance – Db2 Big SQL & Spark SQL

Leads performance metrics on high volumes of data and concurrent streams

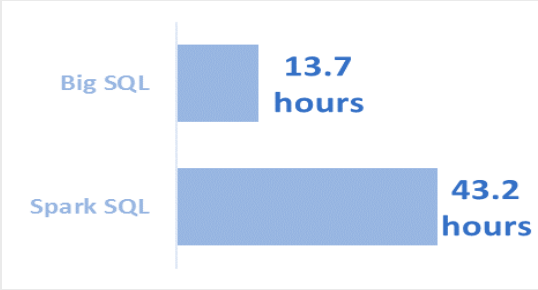
SNAPSHOT OF 100TB HADOOP-DS



PERFORMANCE

Db2 Big SQL 5.0 is **3.2x** faster than Spark SQL 2.1

(4 Concurrent Streams)



COMPRESSION

60%

SPACE SAVED WITH PARQUET

AVERAGE CPU USAGE

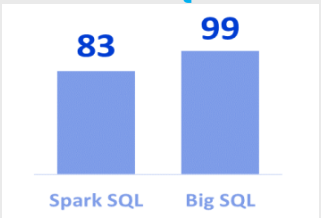
76.4%

MAX I/O THROUGHPUT

READ 4.4 GB/SEC

WRITE 2.8 GB/SEC

WORKING QUERIES



I/O (vs Spark)

Db2 Big SQL reads **12x** less data

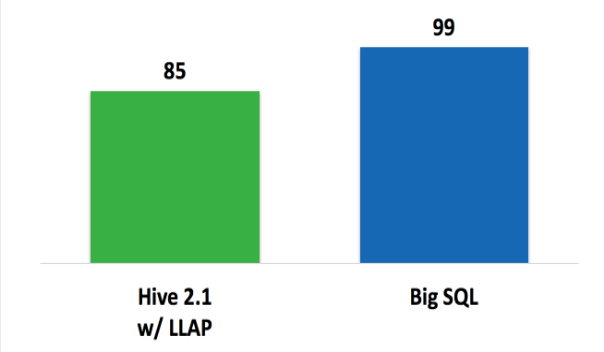
Db2 Big SQL writes **30x** less data

Blog on benchmark: https://developer.ibm.com/hadoop/2017/02/07/experiences-comparing-big-sql-and-spark-sql-at-100tb/

Query Performance at a Glance – Db2 Big SQL & Hive LLAP with Tez

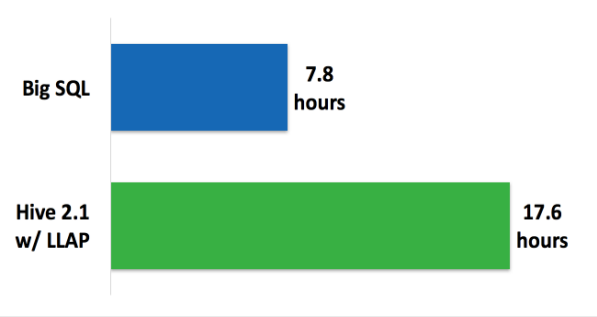
HADOOP-DS @ 10TB
85 COMMON QUERIES

WORKING COMPLIANT QUERIES: 6-streams



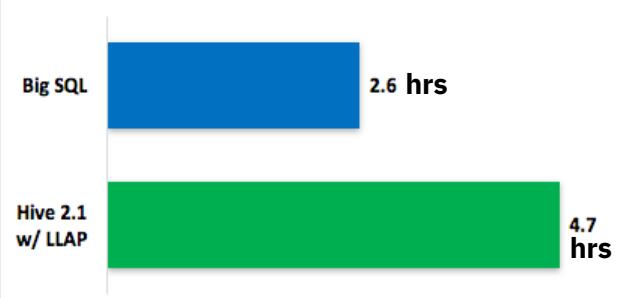
PERFORMANCE: 6-streams

Db2 Big SQL **2.3x** FASTER



PERFORMANCE: 1-stream

Db2 Big SQL **1.8x** FASTER



RESOURCE UTILIZATION:

6-STREAMS

1.5x FEWER CPU CYCLES USED

WORKLOAD

SCALE FACTOR: **10 TB**
FILE FORMAT: **ORC** (ZLIB)
CONCURRENCY: **6 STREAMS**
QUERY SUBSET: **85 QUERIES**

INTERESTING FACTS

FASTEST QUERY

5.4x FASTER (Db2 Big SQL: 1.5 SEC, HIVE: 8.1 SEC)

SLOWEST QUERY (QUERY 67)

1.7x FASTER (Db2 Big SQL: 6827 SEC, HIVE: 11830 SEC)

Db2 Big SQL FASTER FOR **80%** OF QUERIES RUN

STACK

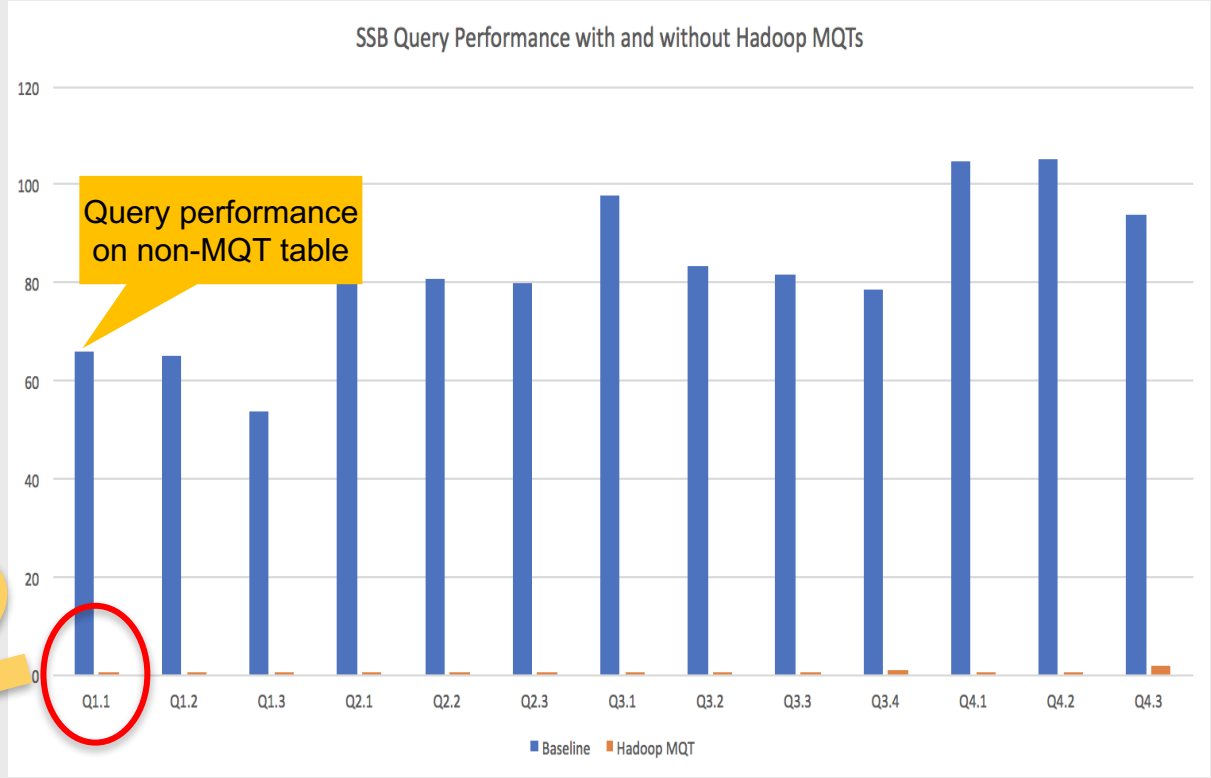
HDP 2.6.1
Db2 Big SQL 5.0.1
HIVE 2.1 LLAP ON TEZ

Performance using MQTs on Star Schema Benchmark Queries

Using Scale Factor 1000, tested 13 queries that join 1 fact with 4 dimension tables
6 Billion Lineitems, 30 Million Customers rows

	SSB Query	Baseline	Hadoop MQT
Quick metric *queries*	Q1.1	66.154	0.532
	Q1.2	65.047	0.459
	Q1.3	53.712	0.443
Product insight queries	Q2.1	82.068	0.457
	Q2.2	80.701	0.45
	Q2.3	79.942	0.452
Customer insight queries	Q3.1	97.854	0.474
	Q3.2	83.449	0.549
	Q3.3	81.514	0.642
	Q3.4	78.563	0.917
	Q4.1	104.917	0.475
	Q4.2	105.051	0.45
	Q4.3		

Notice the negligible performance when using Hadoop MQTs



Self-service Analytics: Democratize Data Science and ML

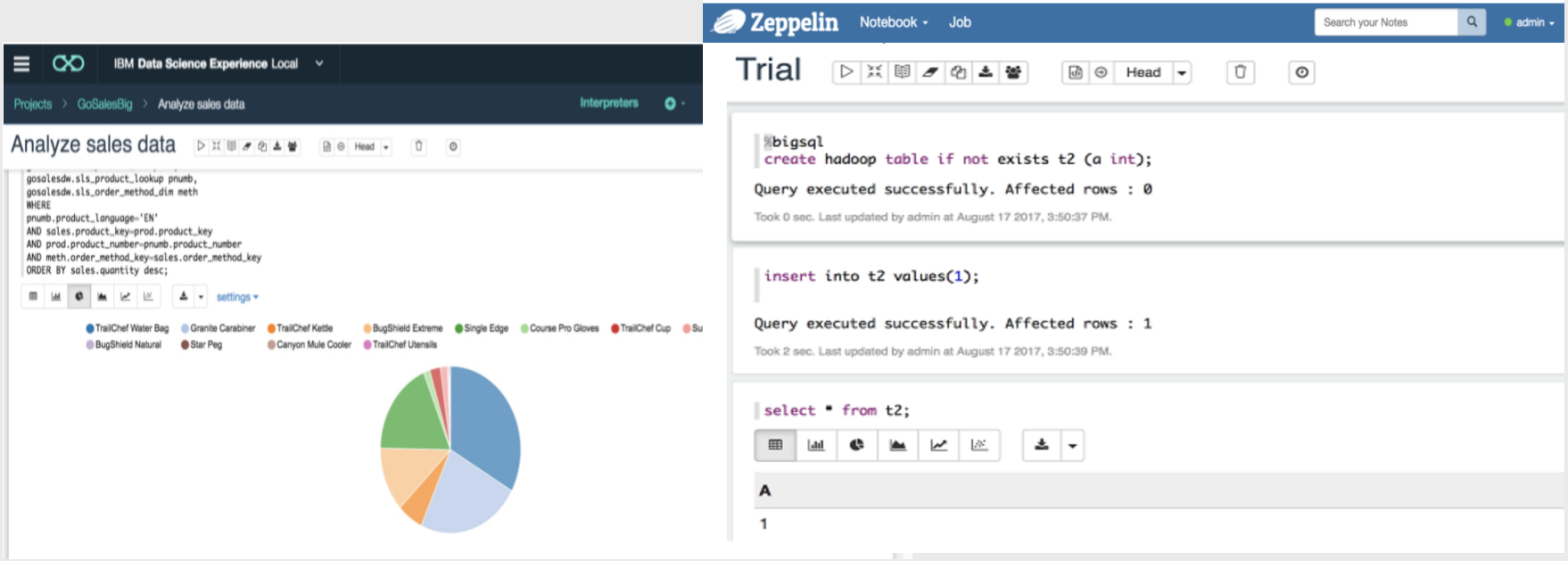
*Leverage **Db2 Big SQL** throughout your journey*

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

Training Machine Learning Models using Db2 Big SQL Made Easy



For more details check the blog: https://developer.ibm.com/hadoop/2017/11/07/ibm-big-sql-machine-learning-demo/

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

- Intelligent Query Routing
- Cost-based optimizer
- SQL pushdown
- ANSI-compliant SQL
- Quick access to Data value
- Query with open source Hadoop file formats

Augment

- 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

Monitor

- 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

Security

- Granular SQL level access control for row filtering and column masking
- Define policies in Ranger for centralized security management

Performance

- 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


Are your data capabilities scaling at the same pace as your data?


Unless you can be sure of continuous availability of your data your business is seriously at risk

Broken issues in business today

Data availability

Existing technology does not meet the demands of large data volumes

Data loss risk

Big data strategy

Advanced analytics capability

Reputation, relevance, innovation, competitiveness

What IBM Big Replicate does

IBM Big Replicate gives you LIVE DATA — consistent data everywhere, spanning platforms and locations, even for changing data at petabyte scale.

Critical data is always available, and accessible from anywhere.

LIVE DATA

Always available and accurate, anywhere

Your market problems

Global data consistency accessible from anywhere

Petabyte scale data movement with continuous replication

Data disaster protection

Data lake optimization (increase value of data lake investment)

Real-time, multi-location, multi-data source analytics

LIVE DATA use cases: analytics, data discovery, and experimentation

Multi-location and data source analytics

Easy and rapid offloading of real- time advanced analytics to the cloud

Data consistency for traditional BI to experimental ML models

All parties have access to the same data

IoT and Edge devices

Coalesced and aggregated data for multi-location /multi-source IoT and Edge analytics

Data governance capabilities across multiple edges (cross- geography)

LIVE DATA
Fully functional copies of data accessible from anywhere

All IBM Big Replicate LIVE DATA scenarios

Data movement/
data lake migration

Disaster recovery,
high availability, and
data governance

Advanced analytics,
data discovery, and
experimentation

LIVE DATA
Data and applications continuously available and in sync even if the data is changing – regardless of platform, location, or scale

LIVE DATA use cases: Data movement and data lake migration

On-premise

Consistency of data to support critical business applications and analytics

Hybrid-cloud

Sharing on-prem data with the cloud:

Simplified and risk-free data lake ingestion (moving data into a central data lake)

Enable experimentation and exploration using new cloud services without breaking existing workflows

Multi-cloud

Data consistency across multi-region environments

Connect any version of Hadoop to multiple clouds

No vendor lock-in

LIVE DATA
Migrate transactional production system with no downtime or data loss, even when data is changing

LIVE DATA use cases: DR, HA, and data governance

RPO/RTO SLA

Near zero RPO/RTO replication capabilities regardless of platform or geography

Self-healing – Automatic repair of missing data in the event of a failure/dramatically reduced workload

Data sovereignty regulation compliance

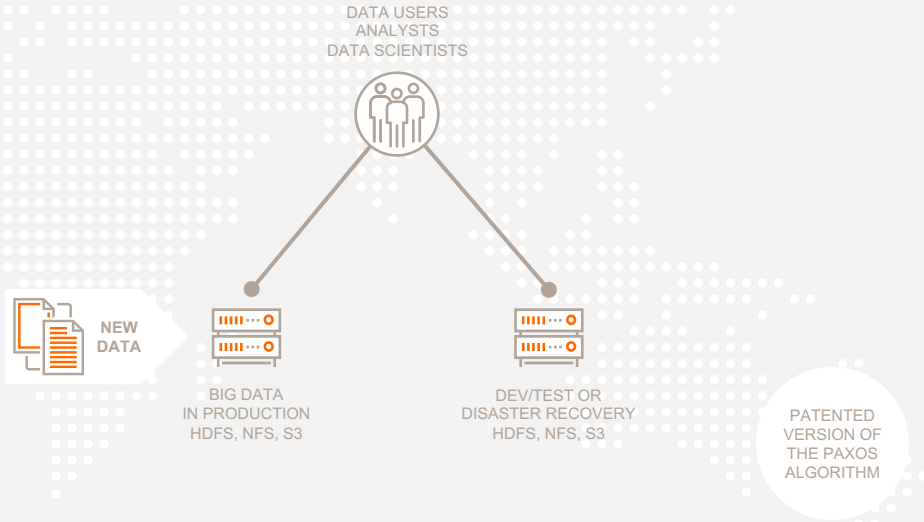
Complete control of data consistency and location to satisfy strict regulatory requirements

Supports audit trail management

Selective replication for regulation compliance

LIVE DATA
Automatic recovery with guaranteed data consistency

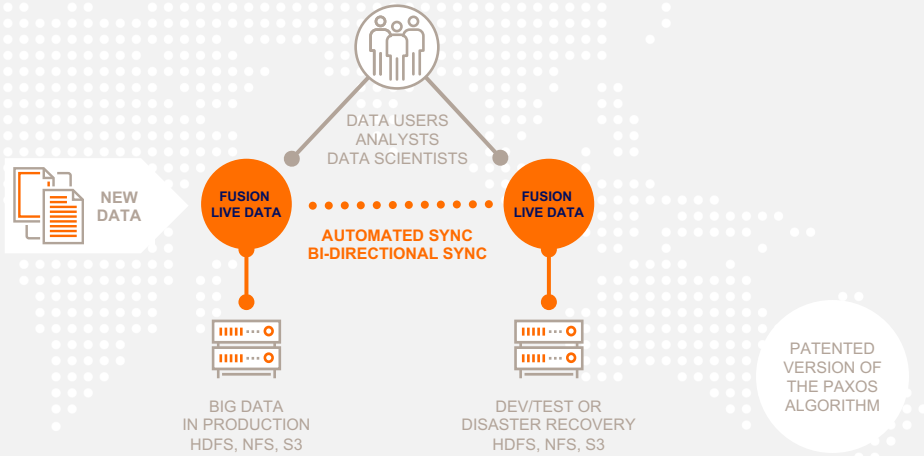
Today production and development/DR Hadoop clusters are not in sync



With WANdisco Fusion, production and development/DR Hadoop clusters are always in sync with LIVE DATA

WANdisco Live Plugins include:

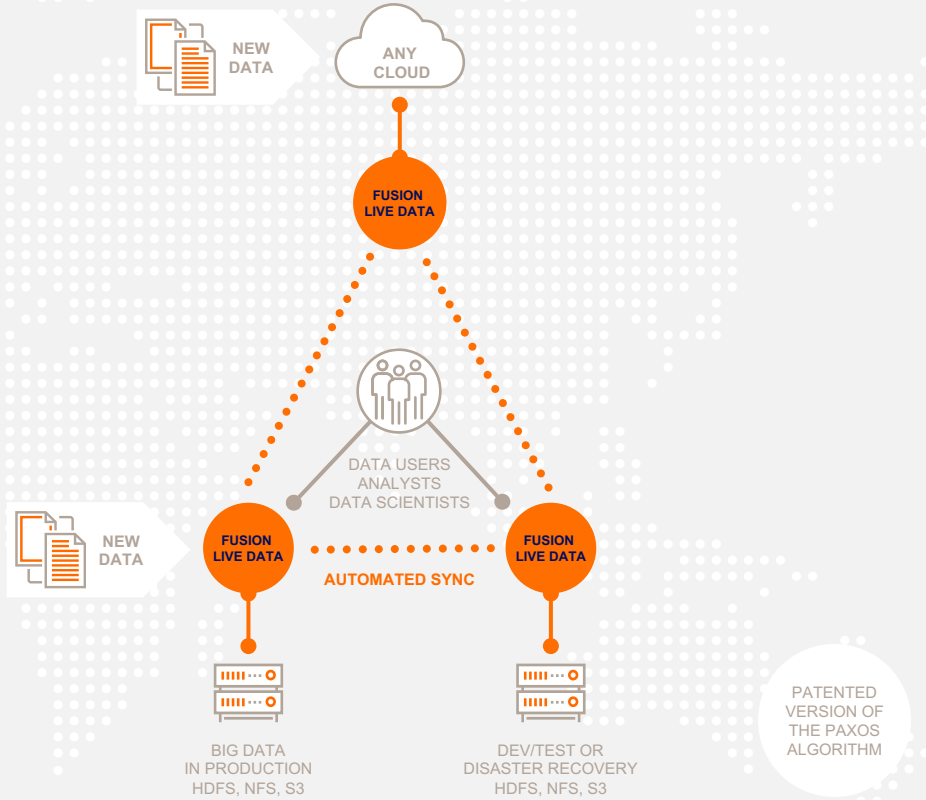
- Hadoop HDFS
- Azure HD Insights
- Cloudera CDH
- Hortonworks HDP
- Security: Sentry and Ranger
- S3 Object Storage
- AWS EMR
- NetApp
- OpenStack Swift
- Oracle Big Data Appliance



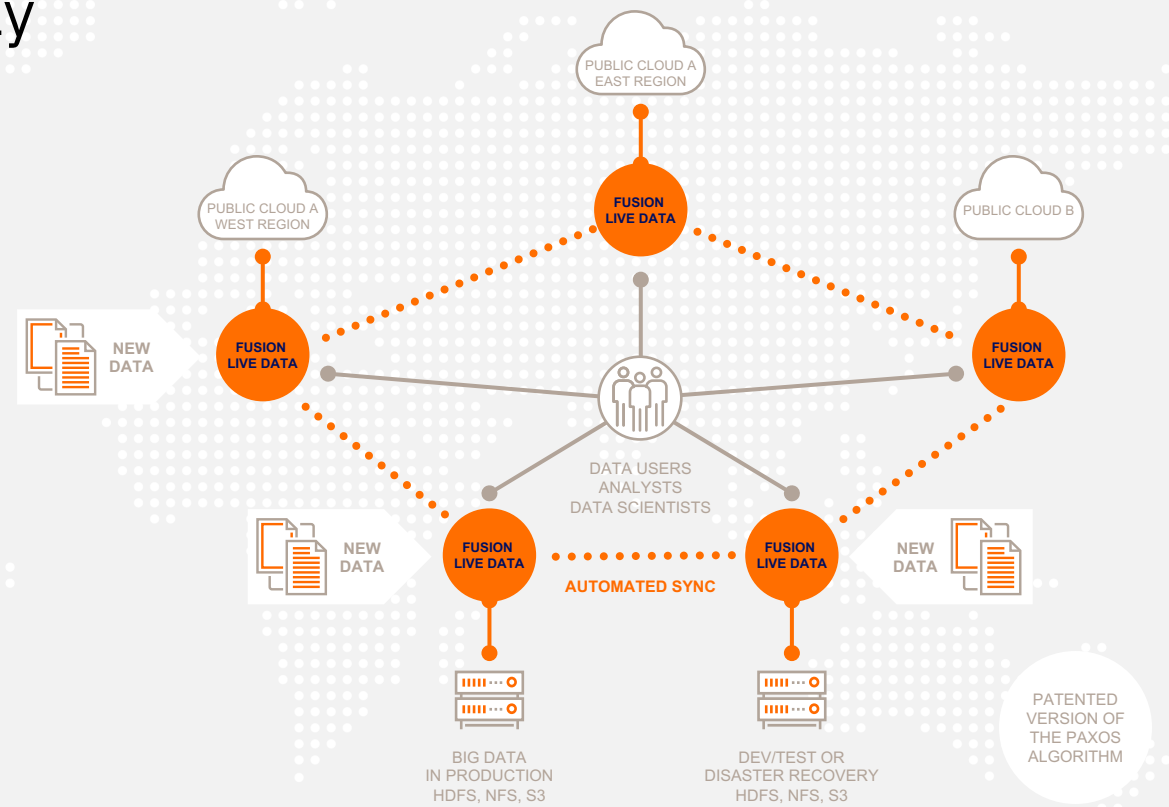
Big Replicate can replicate big data to any cloud

WANdisco Live Plugins include:

- Hadoop HDFS
- Azure HD Insights
- Cloudera CDH
- Hortonworks HDP
- Security: Sentry and Ranger
- S3 Object Storage
- AWS EMR
- NetApp
- OpenStack Swift
- Oracle Big Data Appliance

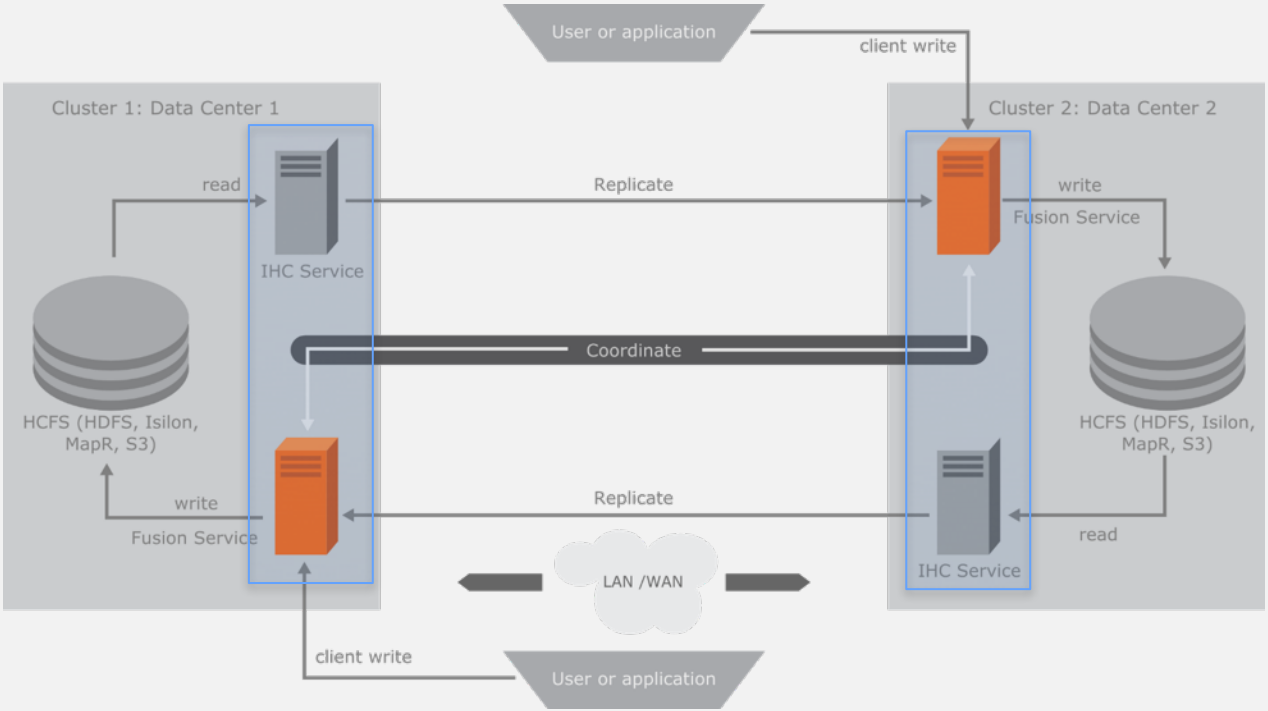


LIVE DATA is available where your data teams are: globally



How does it work?

Coordinating writes



IBM Big Replicate benefits

Meet the most stringent data SLAs

Guaranteed data consistency meeting demanding enterprise SLAs across any combination of Hadoop distributions and cloud storage

Remove risks and complexity of traditional data movement

Managed in real-time – continuous replication – with zero business disruption

Eliminate manual scripting and snapshotting

No business disruption caused by the limitations of one-way batch oriented tools (continuous replication in contrast with a PTO backup)

Leverage all of your compute power

Utilize passive backup clusters and convert these into active production assets (e.g., schedule different jobs on each cluster

Ability to schedule different and more jobs across freed up capacity

Ingest jobs on all sites

Accommodate changes with flexible architecture

Enable multi-vendor Hadoop strategies (no data gravity)

Extensible plugins for multiple storage systems (cloud object storage, big data, NFS, physical devices, security (Ranger, Sentry), etc.)

Seize cost savings opportunities with more capacity

Maximize economies of scale offered by cloud storage

Potentially scale down amount of storage required in your environment

Unparalleled replication

Questions?

Paul Scott-Murphy
VP Product Management, WANdisco

Notices and disclaimers

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.** The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.
.

Thank you

Nagapriya Tiruthani
Offering Manager, IBM Db2 Big SQL
—
ntiruth@us.ibm.com

Paul Scott-Murphy
VP Product Management, WANdisco
—
paul.scott-murphy@wandisco.com
+1-925-399-4065
wandisco.com

Prasad Pandit
Offering Management, Big Data portfolio
—
pprasad@us.ibm.com

Vinayak Agrawal
Offering Manager, IBM Big Replicate
—
vagrawal@us.ibm.com