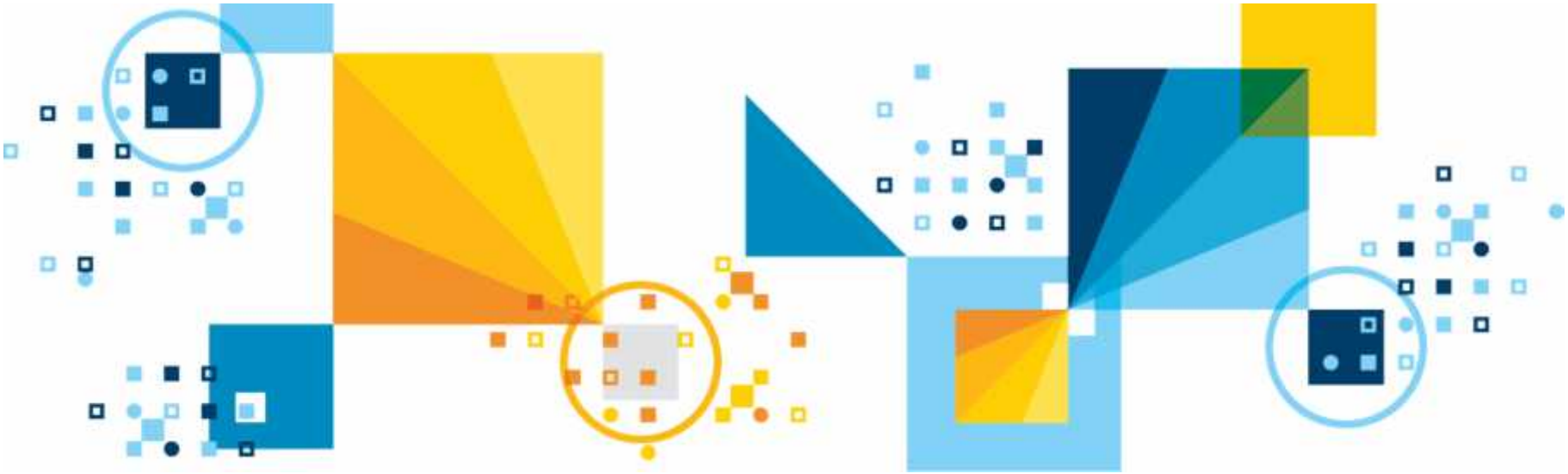
IBM Data Server Strategy Update



Les King

Director, Big Data, Analytics, Database and Cloud Data Solutions

Professor, Big Data, Data Warehousing and DB2, Seneca College

lking@ca.ibm.com
ca.linkedin.com/pub/les-king/10/a68/426

Professional Highlights

- 23 years of Information Management, Database and Analytics
- Technical sales (current)
- Technical customer support
- Software development teams
- Product management
- Taught mathematics at University of Toronto
- Teaching data warehousing, big data and DB2 at Seneca College

Personal Highlights

- English / Irish background
- Sports: squash, down hill skiing
- Certified Advanced Open Water diver
- Two sons: Philip and Richard

Agenda

- Data Server Market and Strategy Update
- Introducing IBM Event Store – BLU Spark
- Introducing the Next Generation Appliance - Sailfish
- Introducing BigInsights 4.3

As A Reminder

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice 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.

Information presented and discussed during this meeting may be both IBM and client confidential. The agreements signed by members of the Technical Advisory Board govern usage of any and all information discussed and shared.

Agenda

- **Data Server Market and Strategy Update**
- **Introducing IBM Event Store – BLU Spark**
- **Introducing the Next Generation Appliance - Sailfish**
- **Introducing BigInsights 4.3**

It is about HYBRID

Its not about Cloud or On-Premises its about Cloud AND On-Premises

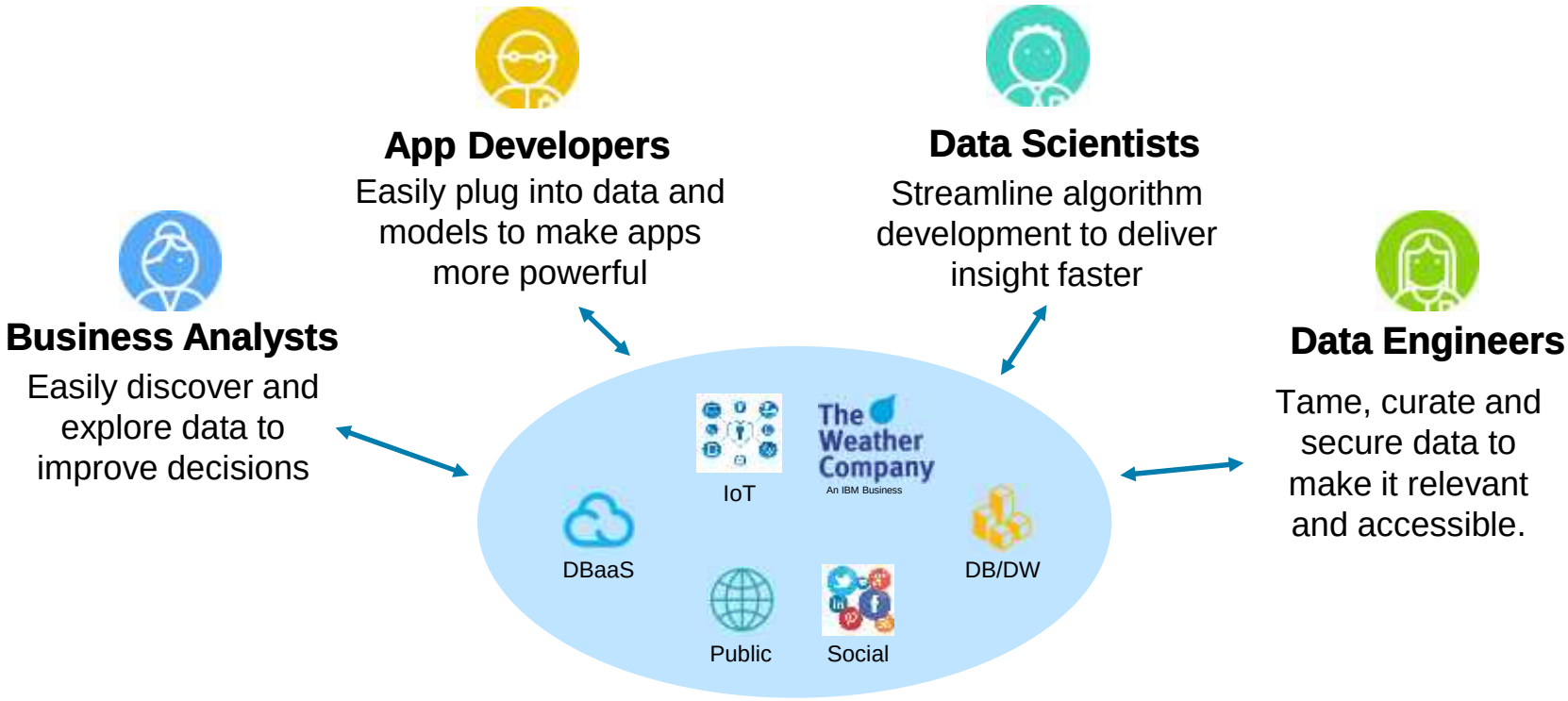
It's About Hybrid

"Through 2020, the most common use of Cloud services will be a hybrid model combining on-premises and external Cloud services."

Gartner, Cloud Computing Innovation Key Initiative Overview

Data Professionals – Evolving Roles

As Data Maturity Increases, So Does the Number of Data Professionals Who Are Hungry to Put Data to Work



Market Observations

1. There is increasing pressure to perform analytics where data gets created

"Point-of-decision HTAP promises to simplify the information infrastructure by removing unnecessary data marts and, potentially, data warehouses." – Gartner

2. Event-driven applications will enable new analytic use cases

"Event-driven real-time digital business is poised to become a priority for mainstream business " Gartner
"In-process HTAP could potentially redefine the way some business processes are executed" Gartner

3. Business applications are leveraging both SQL and NoSQL data in structured repositories for analytics

"Top relational database solutions are now offering a wide range of new features to combine structured and unstructured data types " …. Database decision-makers need to look at investing in these database technologies. – Forrester

4. Hybrid cloud capabilities of software support economies of scope

Public cloud adoption has stalled for the time being, signaling enterprises are moving to the hybridization phase of their IT transformations. TBRI 2H 2016

5. Private cloud needs cloud-scale convenience

IDC "by end of 2016 38% of the IT Market spend will be private hosted or private on Prem Cloud with On-Demand Convenience future growth point within private cloud. Skills, timing or cost to effectively procure, assemble, run, manage disperse infrastructure resource require integrated versatile platform offerings with appliance-like simplicity" – A client

6. Diverse data sources support an ecosystem of innovation

Established vendors…, have continued their cloud-focused innovation around hybrid cloud for both cost and workload optimization. Many have added open-source products to their portfolio — usually by acquisition — in an attempt to capture a new generation of buyers .

IBM's Point of View

1. **There is increasing pressure to perform analytics where data gets created**
   We will lead the market on Hybrid Transaction / Analytical Processing (HTAP) and resilience.

2. **Event-driven applications will enable new analytic use cases**
   We will lead fast data analytics with our analytic warehouse engine and Spark.

3. **Business applications are leveraging both SQL and NoSQL data in structured repositories for analytics**
   We will integrate XML, JSON, and REST seamlessly with our Common SQL engine.

4. **Hybrid cloud capabilities of software support economies of scope**
   We will continue to support 'build once, deploy anywhere' compatibility of our analytic engine, as we enhance the code and offerings supported.

5. **Private cloud needs cloud-scale convenience**
   We will bring public cloud management capabilities to private cloud environments, facilitating economies of scale.

6. **Diverse data sources support an ecosystem of innovation**
   We will provide free access to the capabilities of DB2 to developers, embrace open source DBs as part of our private cloud, and offer easy on-ramps to federate and productize on DB2.

Watson Data Platform

Intelligent by Design
with intelligence services ranging from cognitive APIs to automatic entity analytics infused in every aspect of the making/creation process

Lead the way in collaboration
so data scientists, developers and data engineers are natively supported in their tasks and are working together to deliver an intelligent application

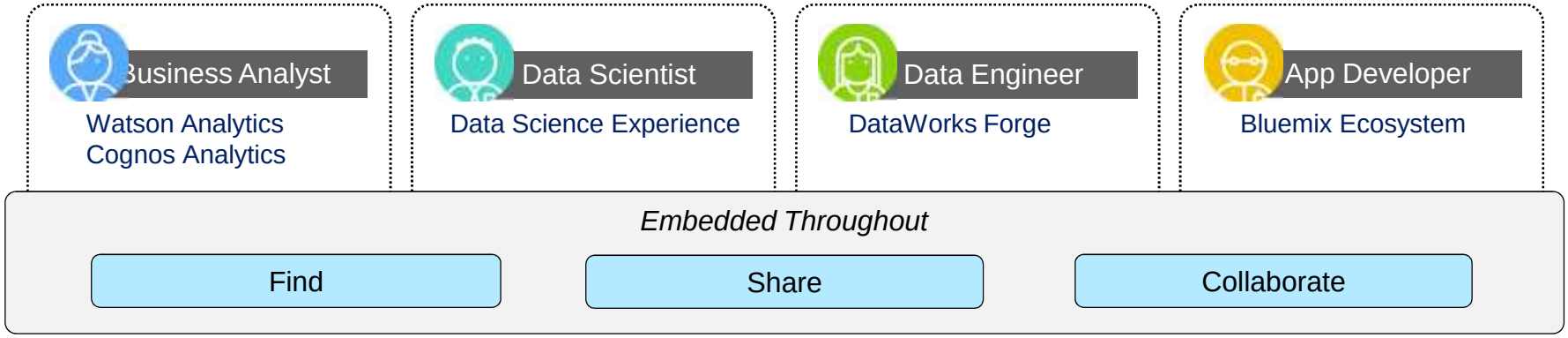
Self-service trusted access to data
giving data professionals the freedom of access to the data they need with the trust that the business expects

Streaming with real-time analytics
to support modern application demands with first class streaming analytics that augments batch

Open and Extensible
achieving scale and the network effect by attracting partners to build on and extend through APIs and toolkits

Be the premier content hub
of open, rich 3rd party content and curated IBM assets that enhances the types of insights users are able to derive

Watson Data Platform – Reference Architecture



Spark as underlying engine

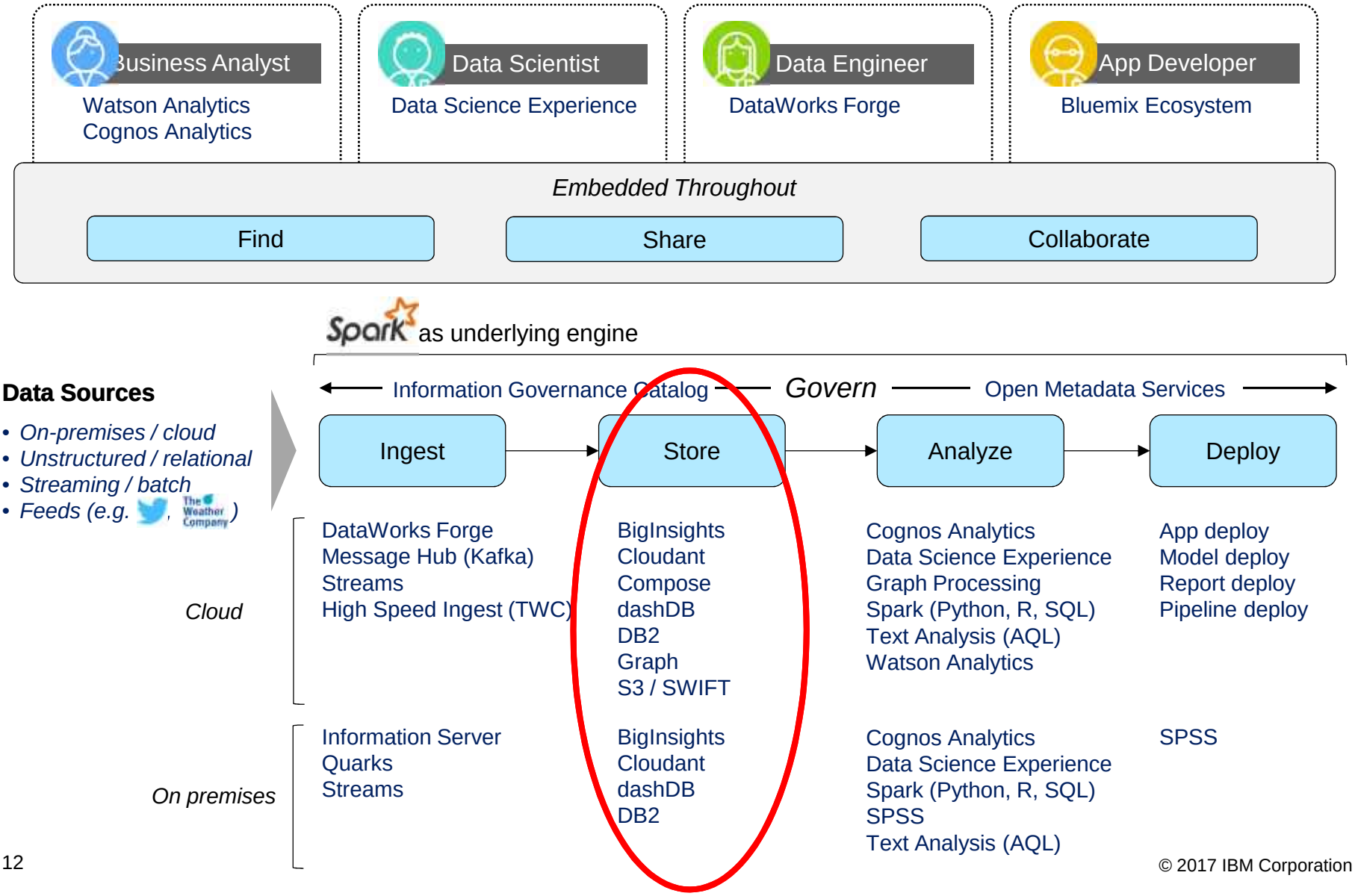
Data Sources

- *On-premises / cloud*
- *Unstructured / relational*
- *Streaming / batch*
- *Feeds (e.g. ,)*

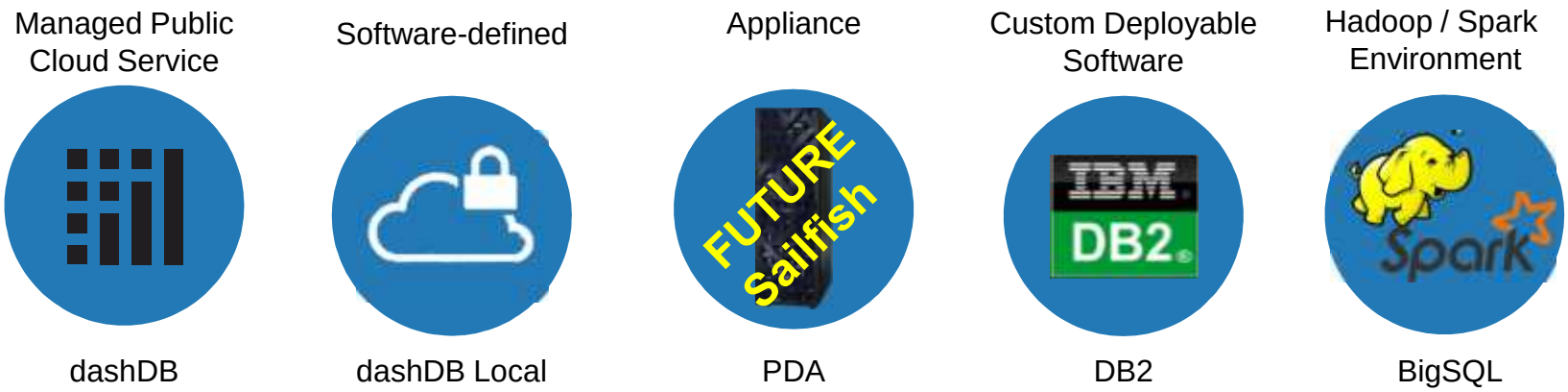
← Information Governance Catalog — *Govern* — Open Metadata Services →

	Ingest	Store	Analyze	Deploy
Cloud	DataWorks Forge Message Hub (Kafka) Streams High Speed Ingest (TWC)	BigInsights Cloudant Compose dashDB DB2 Graph S3 / SWIFT	Cognos Analytics Data Science Experience Graph Processing Spark (Python, R, SQL) Text Analysis (AQL) Watson Analytics	App deploy Model deploy Report deploy Pipeline deploy
On premises	Information Server Quarks Streams	BigInsights Cloudant dashDB DB2	Cognos Analytics Data Science Experience Spark (Python, R, SQL) SPSS Text Analysis (AQL)	SPSS

Watson Data Platform - Reference Architecture



Common SQL Engine



A Common SQL Engine enabling true HYBRID data store solutions for both SQL and NoSQL requirements for any type of workload

- **Application compatibility:** **Write once, run anywhere**
- **Operational compatibility:** **Reuse operational and housekeeping procedures**
- **Licensing:** **Flexible entitlements for business agility & cost-optimization**
- **Integration:** **Common Fluid Query capabilities for query federation and data movement**
- **Standardized analytics:** **Common programming model for in-DB analytics**
- **Ecosystem:** **One ISV product certification for all platforms**
- **Skills:** **Leverage existing skill base including tools, processes, interfaces across the portfolio**

Common SQL Engine

APPLICATIONS

DB2 | NETEZZA PureData | ORACLE

Managed Public Cloud Service	Software-defined	Appliance	Custom Deployable Software	Hadoop / Spark Environment
dashDB	dashDB Local	PDA	DB2	BigSQL

Data Virtualization

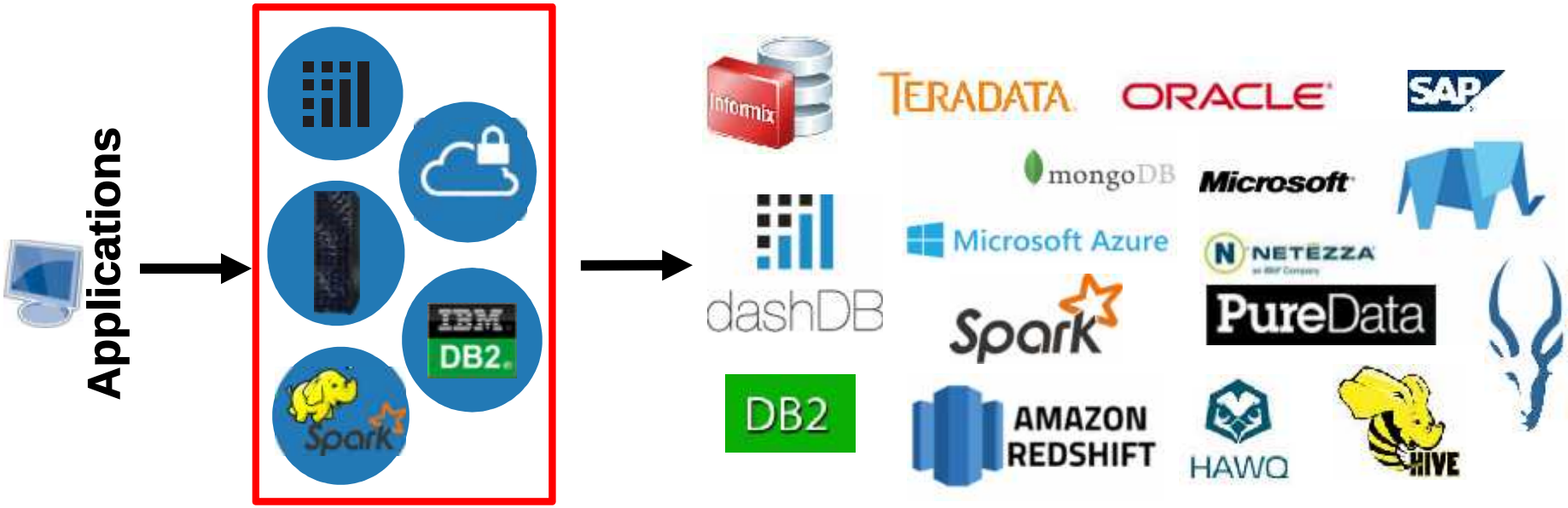
High performance

Enterprise readiness

Heterogeneity

Oracle compatibility

High functionality



You can leverage any of the above offerings as a federation server

This image shows only a subset of all supported data sources

Apache Spark

Spark's core libraries enable analytic processing of data from many sources

- Apache Spark is an **open** source, **in-memory** processing framework
- Distributed data processing & iterative analysis on **massive** data volumes
- Spark's standalone framework goes beyond Hadoop and HDFS
  - Interactive query via **SparkSQL**: Spark is not required to store data (write job outputs) locally
  - Micro-batched event processing via **Spark Streaming**
  - Machine learning libraries via **MLib**
  - Graph processing via **GraphX**

Analytics for Apache Spark

Blends Multiple Data Types, Sources & Workloads

Execute SQL Statements

Streaming Analytics via Micro-batch

M.L. and Statistical Algorithms

Distributed Graph Processing Framework

Spark SQL

Spark Streaming

MLlib Machine Learning

GraphX Graphing

Available on z/Linux

- General compute engine
- Basic I/O functions
- Task dispatching
- Scheduling

Spark Core

Data Sources

IBM Cloud	Public Cloud	Cloud Apps	On-Premises
BigInsights (HDFS), Cloudant (DBaaS), dashDB (Analytics), SQDB (Managed DB2), Swift (Object Storage)	amazon S3, Cassandra, mongoDB, redis, rackspace, Microsoft Azure, MySQL, Hive, PostgreSQL, HDFS, dBase, HBase	NETSUITE, salesforce, CSV, JDBC, {JSON}, Parquet, elasticsearch, Avro	ORACLE, SAP, IBM DB2

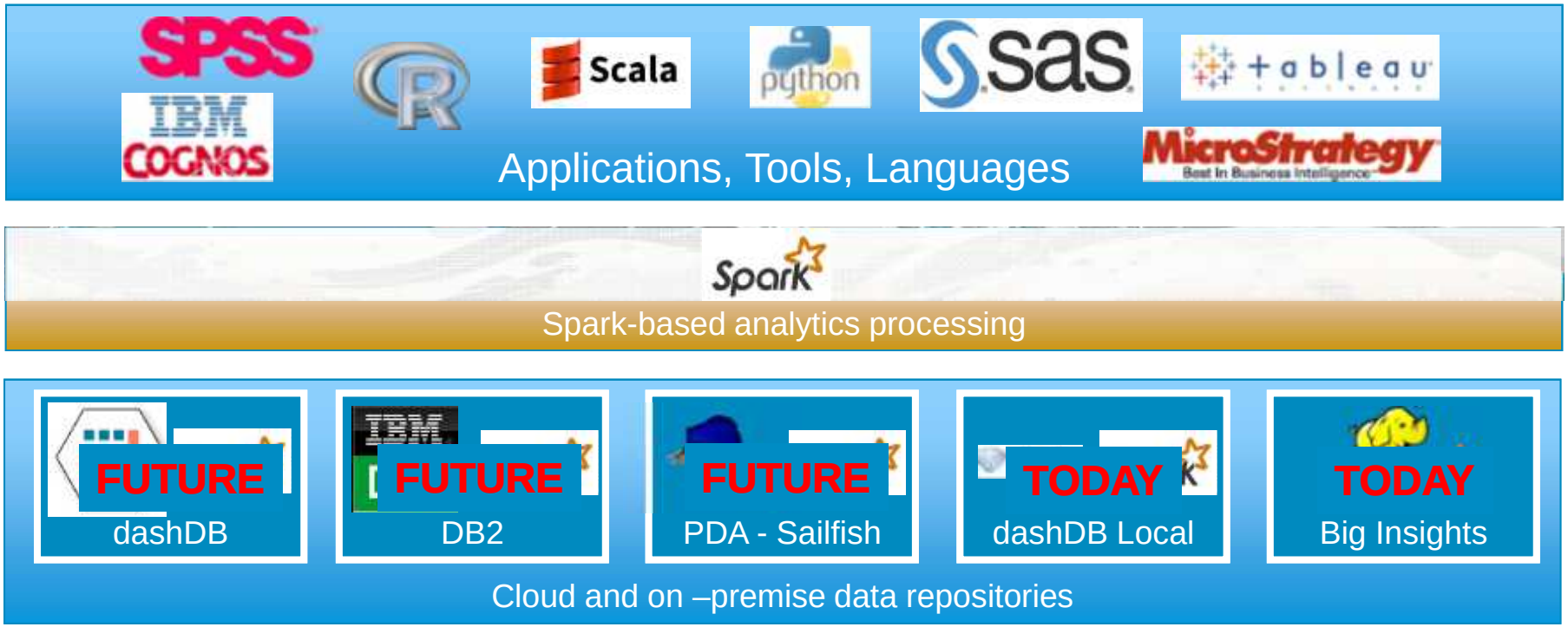
Common SQL Engine with Spark

- Write your applications once and run them against any platform service – with consistent and predictable results
- Building on the power of Spark for widest range of algorithms
- Optimized Spark execution engines alongside database engines in each MPP node
- Leverage a rich and growing ecosystem of IBM and 3rd party BI and Analytics applications
- Exploiting Fluid Query to tap into all necessary data sources

18

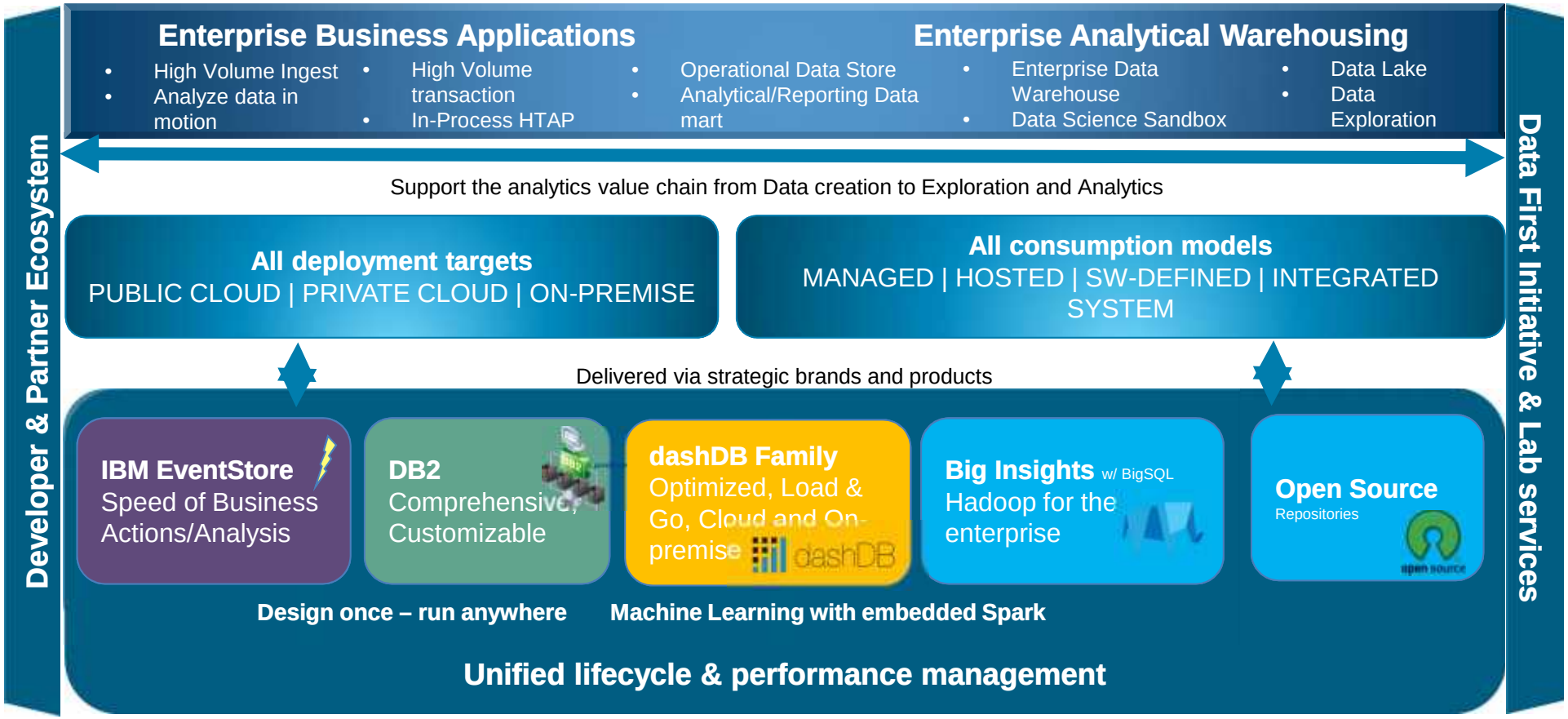
Common SQL Engine with Spark

- Write your applications once and run them against any platform service – with consistent and predictable results
- Building on the power of Spark for widest range of algorithms
- Optimized Spark execution engines alongside database engines in each MPP node
- Leverage a rich and growing ecosystem of IBM and 3rd party BI and Analytics applications
- Exploiting Fluid Query to tap into all necessary data sources



19

Natural Extension of the IBM Database Portfolio



IBM Data Server Manager

Deliver a Simplified User Experience

- **Single installer and integrated repository**

Common integrated web console

- **Provides enterprise view of your environment**
- **Guided workflow and analysis**

Deliver familiar capabilities from Optim Database Tools

- **Performance Management and Database Administration as extensible services**

IBM Data Server Manager

- ✓ ***Simple***
- ✓ ***Scalable***
- ✓ ***Smart***

DB2 for Linux Unix Windows

DB2 LUW Strategic Directions

Extend the core

reliability:availability:scalability:security

Total Business Continuity

- *Planned Events*
- *Unplanned Events*
- *Disasters*

More Leadership Innovation in Analytics
Scalability, Efficiency, "More with Less"
Reliability, Quality, and Security Advances

Empower the future

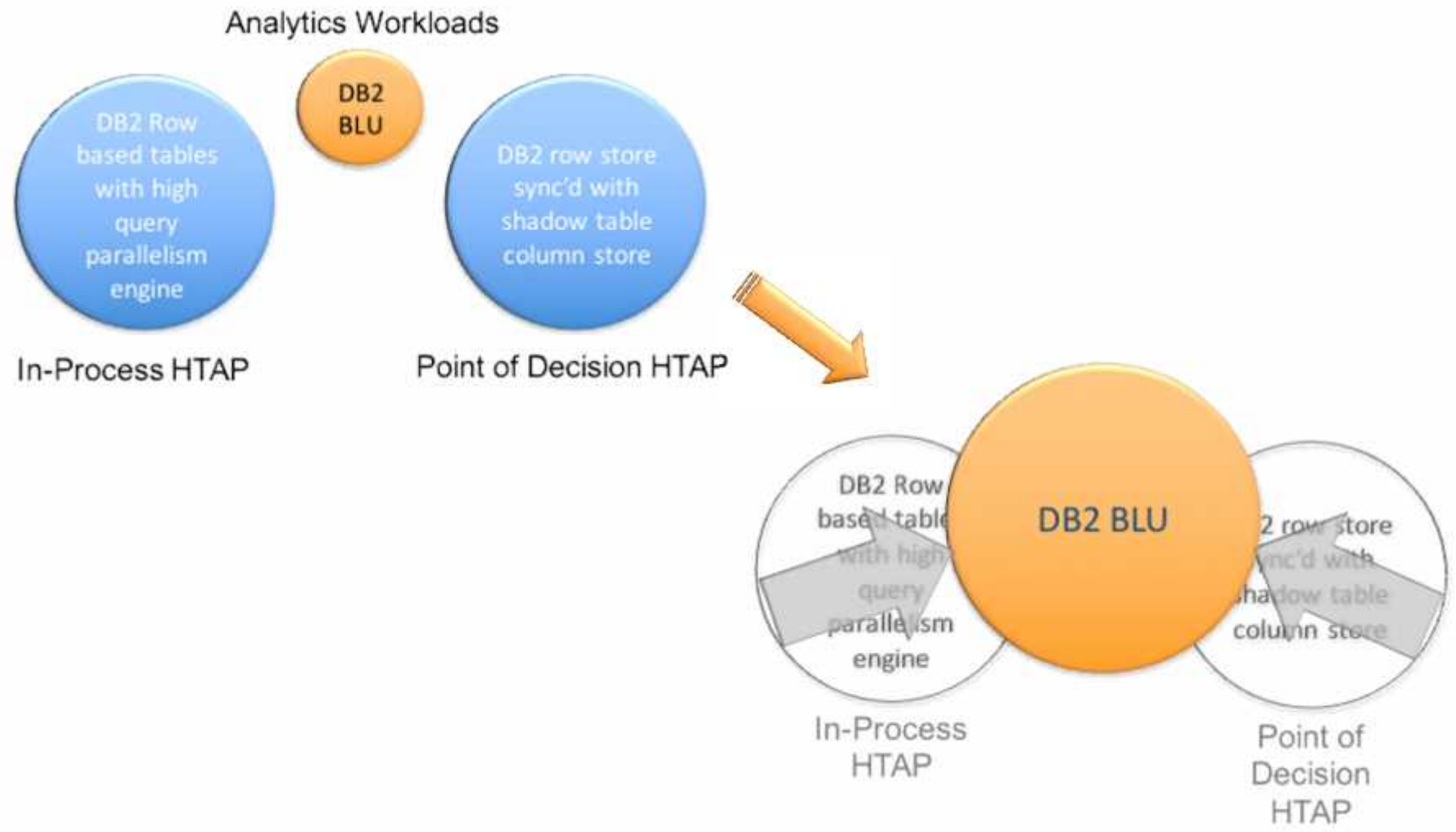
analytics:autonomics:cloud:mobile

Support the Next Wave of Applications

- Hybrid Transaction/Analytical Processing (HTAP)
- Advanced Analytics & Spark Integration
- Hybrid Cloud
- noSQL
- Simplification and Autonomics

Delivered continuously
On your choice of on-premises or cloud deployment, IBM or customer managed
While protecting your application investment in DB2

Where is HTAP Today and Where is it Going ?



DB2 - Never Down Applications – Availability Tiers

Integrated Clustering

- Active/passive
- Hot/cold, with failover typically in minutes
- Easy to setup
- DB2 ships with integrated TSA failover software
- No additional licensing required

HADR

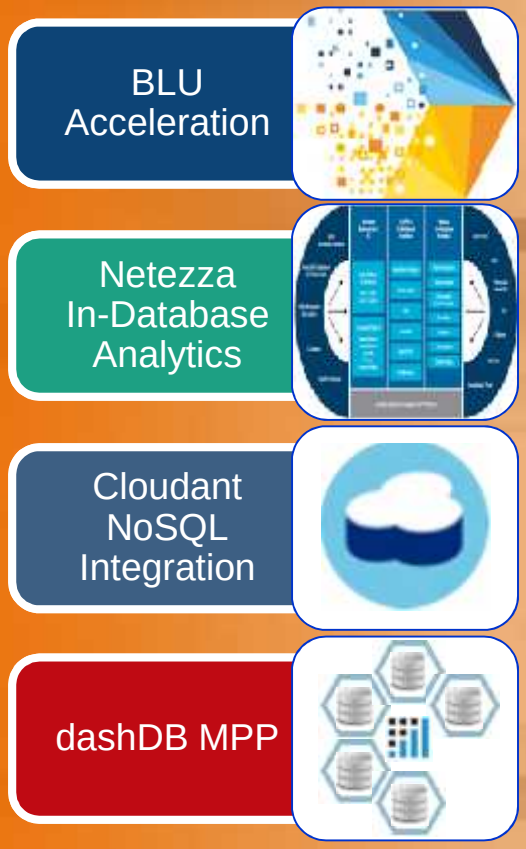
- Active/passive or active/active (with Reads on Standby)
- Hot/warm or hot/hot (with RoS), with failover typically less than one minute
- Easy to setup
- DB2 ships with integrated TSA
- Minimal licensing (full licensing required if standby is active)
- Perform system and database updates without interruption

pureScale

- Active/active
- Hot/hot, with automatic and online failover
- Integrated solution includes CFs, clustering, and shared data access
- Perform system and database updates in rolling online fashion
- Also works with HADR (single target)
- Geographically Dispersed Cluster for multi-site deployment

dashDB for Analytics

In-database analytics capabilities for best performance atop a fully-managed warehouse



dashDB for Analytics

- Fully-managed data warehouse on cloud
  - Choice of SoftLayer or Amazon Web Services
- **BLU Acceleration** columnar technology + **Netezza in-database analytics**
  - BLU in-memory processing, data skipping, actionable compression, parallel vector processing, “Load & Go” administration
  - Netezza predictive analytic algorithms
  - Fully integrated RStudio & R language
- Oracle compatibility
- **Massively Parallel Processing** (MPP)
- On disk data encryption and secure connectivity

dashDB for Transactions

Transactional database capabilities for best performance atop a fully-managed instance

Excellent Transactional Performance

Oracle Compatibility

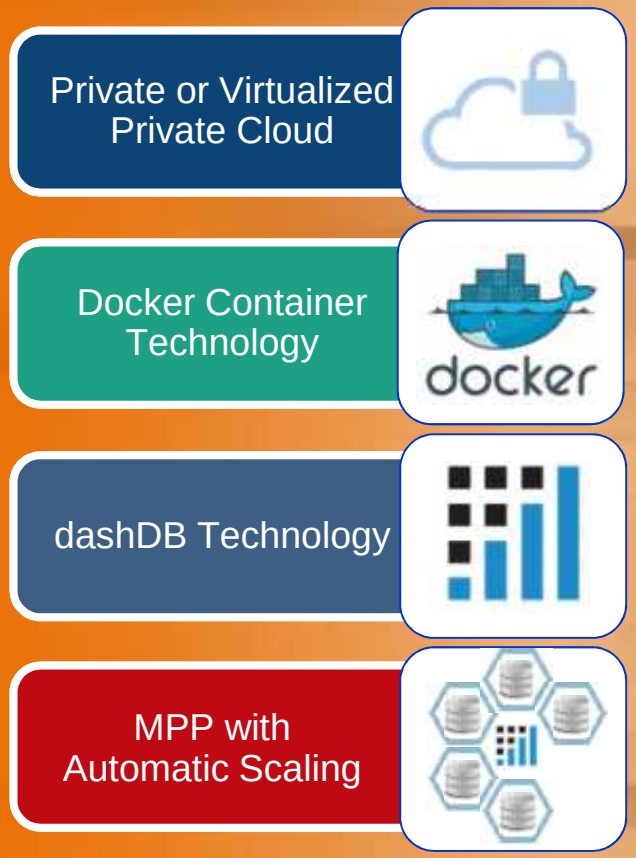
Robust Security

dashDB **for Transactions**

- Fully-managed transactional database as a service
- **Row-organized tables for high transactional performance**
- **Oracle compatibility**
- On disk data encryption and secure connectivity
- Two Enterprise plans plus HA versions
  - 2 cores, 8 GB memory, 500 GB SAN
  - 12 cores, 128 GB memory, 1.4 TB SSD

dashDB Local for Analytics

Benefits of dashDB Technology with Fast Deployment into Private Cloud Environment



- Highly flexible data warehouse
- Optimized for fast and flexible deployment into **private or virtual private clouds**
- Uses **Docker** container technology
- Built on top of **dashDB technology**, it shares the benefits of
  - BLU Acceleration in-memory columnar technology
  - Netezza In-database Analytics
  - Oracle Compatibility
- **Massively Parallel Processing** (MPP) with automated scaling capabilities to increase infrastructure efficiency

PureData System for Analytics – Netezza (Mako)

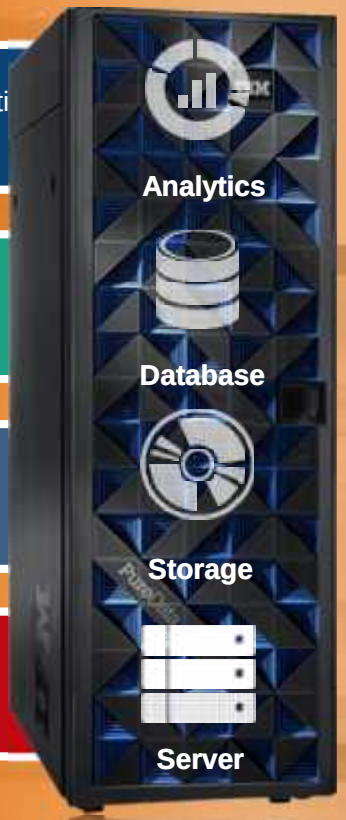
Changing the game for data warehouse appliances (again)

Purpose built Analyti Appliance

Integrated DB, Server and Storage

Standard Interfaces

Low Total Cost of Ownership



Powered by Netezza Technology

Big Data and Business Intelligence ready
with capabilities to unlock data's true potential

Advanced security in an insecure world
at no extra cost

An even broader family of appliance models
to fit a broad range of data capacity needs

What makes it different?

Speed - 10-100x faster than traditional custom systems[1]

Simplicity - minimal administration and tuning

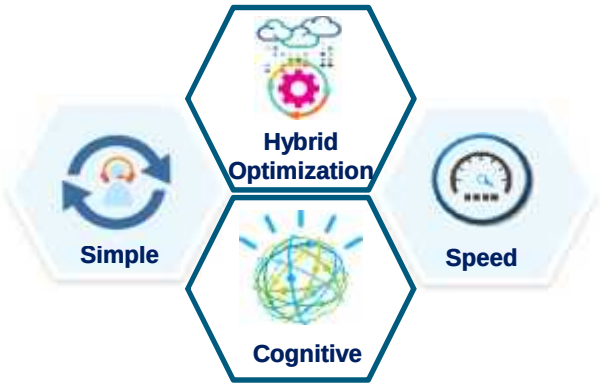
Scalability - petabyte+ scale user data capacity

Smart - high performance, advanced analytics

[1] Based on IBM customers' reported results. "Traditional custom systems" refers to systems that are not professionally pre-built, pre-tested and optimized. Individual results may vary.

dashDB Cognitive Integrated Platform for Analytics

Combining extreme performance and simplicity for Advanced Analytics and Hybrid Data Warehouse Optimization



Sailfish *is IBM's industry-leading* ***Cognitive Integrated Platform*** *for Data Management and Analytics that will* ***integrate*** *seamlessly with other ground and cloud data services, delivering ultra* ***fast & scalable*** *performance, cloud* ***elasticity*** *together with end to end* ***security*** *and the ultimate in* ***simplicity*** *across all dimensions of the client's experience.*

Cloudant

Cloudant delivers a fully-managed database in service to the Analytics, App, and API economy

IBM Cloudant®

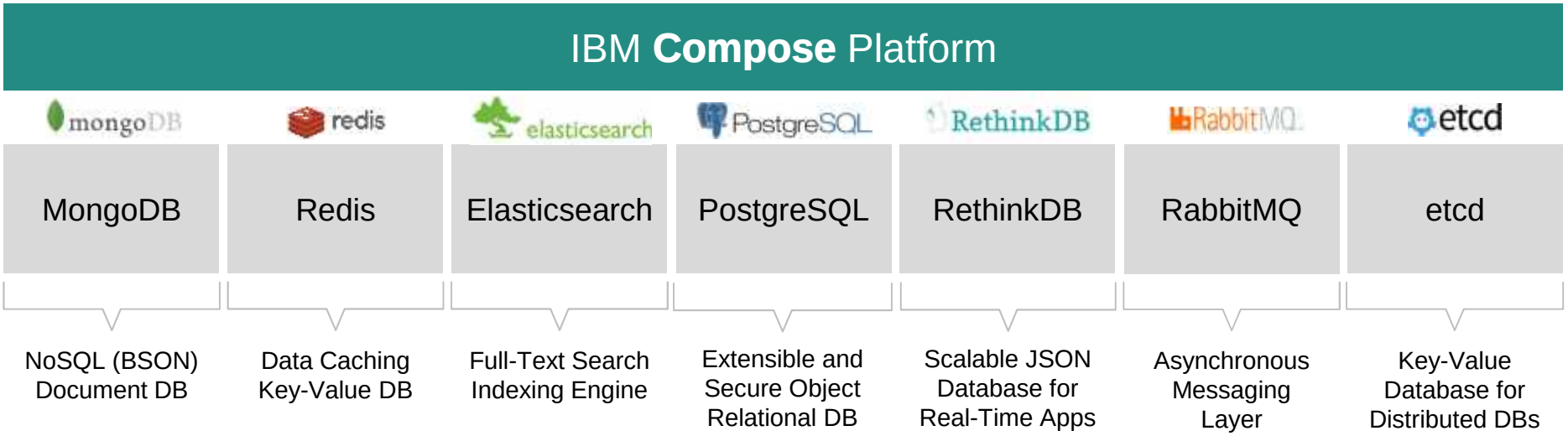
A fully-managed NoSQL database layer that can be **developed & deployed in days**

Spark Integration (Spark SQL)

dashDB Integration (Analytics)

- Operational NoSQL JSON store
- Master-less architecture for maximum **scalability** & **availability**
- Advanced APIs
  - REST (HTTPS) API
  - Replication & synchronization
  - Geo-load balancing
  - Incremental MapReduce indexes
  - Military-grade Geospatial indexes
  - Lucene full-text search
- Offline access to mobile apps & data

IBM Compose Open Source Stack



- **Compose is a managed platform for open-source DBaaS**
  - Services can be adopted individually via **Public** multi-tenant deployments
  - Entire catalogue can be licensed & deployed a la carte via **IBM-Managed** and **Self-Hosted** single-tenant configurations
- **Best-practice delivery & configuration of open source technologies**
  - All services are production-ready and configured for HA out of the box
  - Automated (no-cost) backups, elastic scale-out, intuitive dashboards
- **New services ScyllaDB and MySQL now available for Compose customers**

DB2 on Cloud

- **Provides a hosted DB2 environment that is**
  - Hosted on IBM SoftLayer (virtual private nodes/bare metal) or AWS (virtual private nodes) cloud platform
  - Administered by your organization's DBAs (all software including OS & database)
  - Paid on a month-to-month basis via subscription model (support included)

- **Benefits include**
  - Convenience without the loss of control on cost effective infrastructure
  - Five high performance hardware configurations and two database software tiers to match capability and affordability needs
  - BLU Acceleration
  - Native encryption support included in all configurations ensuring data remains secure in the cloud
  - HADR for high availability and disaster recovery
  - Unlimited ability to create databases to fully utilize the cloud infrastructure
  - pureScale available for cloud deployments

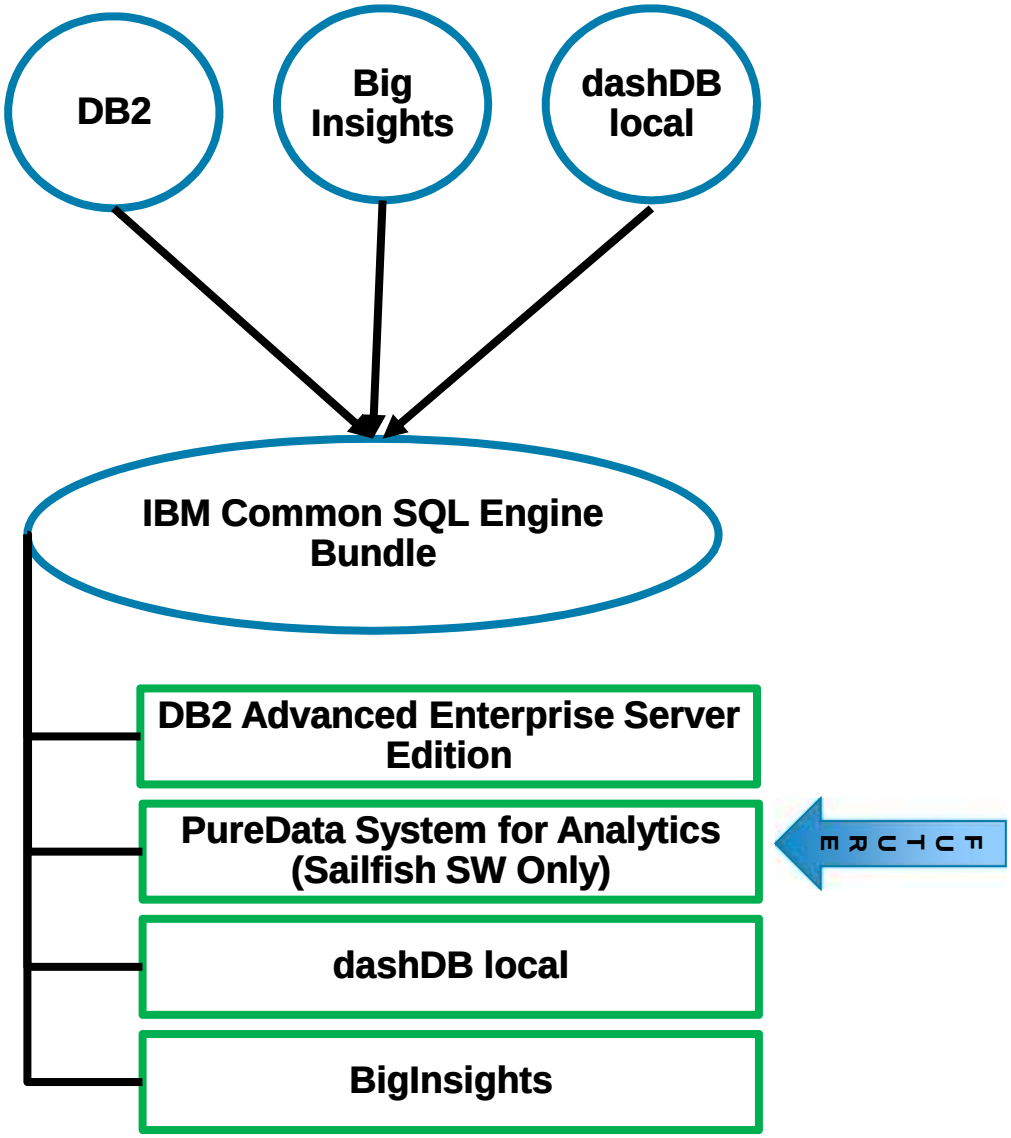
- **Deployment options**
  - Choice cloud provider: SoftLayer or AWS
  - Five t-shirt sized configurations: Small, Medium, Large, X-Large, 2X-Large
  - Two versions of DB2 available: Workgroup Server Edition (Standard) and Advanced Enterprise Server Edition (Advanced)

Our Offerings – Who Does What?

Activity	DB2 Software (on-premises)	DB2 Software (BYOL on IaaS)	DB2 on Cloud	PDA (on-premise)	dashDB Local for Analyitcs	dashDB for Analytics	dashDB for Transactions
Provision hardware infrastructure	Customer	IaaS Vendor	IBM	Built into appliance	Customer or IaaS Vendor	IBM	
Manage hardware infrastructure	Customer	IaaS Vendor	IBM	Built into appliance	Customer or IaaS Vendor	IBM	
Database manger software installation	Customer		IBM	Built into appliance	IBM *	IBM	
Database manager instance creation	Customer		IBM	Customer	IBM *	IBM	
Database creation	Customer		Customer	Customer	IBM *	IBM	
Database configuration	Customer		Customer	Pre-set in appliance	IBM *	IBM	
Manage database environment	Customer		Customer	Customer	Customer	IBM	
Database manager software maintenance	Customer		Customer	Customer	IBM *	IBM	
OS maintenance	Customer		Customer	Customer - Included in Firmware updates	Customer	IBM	
Setup encryption	Customer		Customer	Customer – Two Choices	IBM	IBM	
Database backup and restore	Customer		Customer	Customer	Customer	IBM	

Introducing – IBM Data Server Bundle

- A single offering which provides deployment flexibility
- Includes solutions for all deployment options: private cloud, public cloud, traditional relational and Hadoop
- Investment protection as shifts between deployment choices are required
- Leverages the Common Analytics Engine for Application Portability and Data Virtualization



Agenda

- Data Server Market and Strategy Update
- **Introducing IBM Event Store – BLU Spark**
- Introducing the Next Generation Appliance - Sailfish
- Introducing BigInsights 4.3

Agenda

- Data Server Market and Strategy Update
- Introducing IBM Event Store – BLU Spark
- **Introducing the Next Generation Appliance - Sailfish**
- Introducing BigInsights 4.3

Next Generation Appliance – Maintain Core Values

PureData System for Analytics

World's fastest and "greenest" analytical platform

World's First Analytic Data Warehouse Appliance

World's First Petabyte Data Warehouse Appliance

World's First 100 TB Data Warehouse Appliance

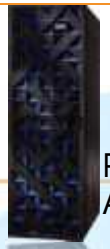
World's First Data Warehouse Appliance

NPS® 8000 Series

NPS® 10000 Series

TwinFin™

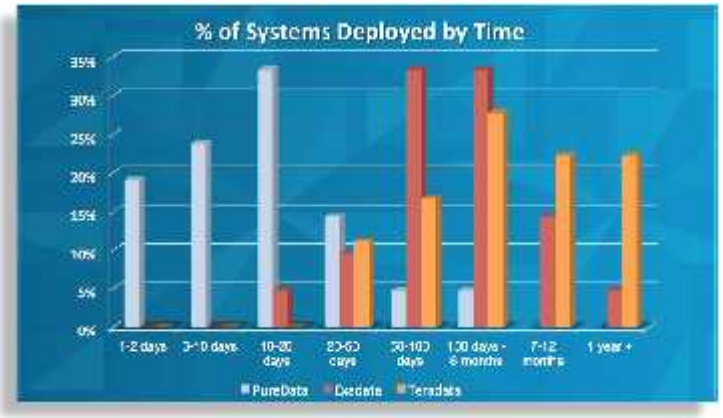
TwinFin™ with i-Class™ Advanced Analytics



PureData System for Analytics N2000

PureData System for Analytics N3000

-Reduced administration
-Performance portal
-Lower end starting point
-More scale-out



2003 2006 2009 2010 2012 2014 2016+

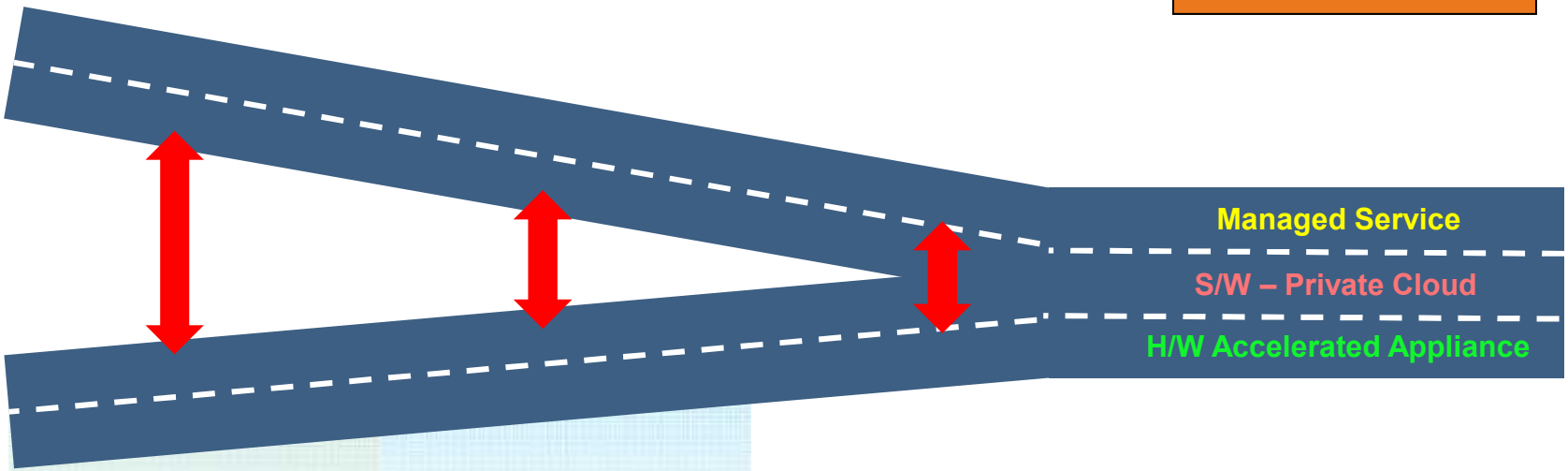
dashDB Local for Analytics convergence with PDA

Cloud first, in-memory, SIMD acceleration (BLU) for analytics

Enhanced elasticity, UDX integration, and managed service

Further UDX integration and optimization

SQL, Analytic, Application, ETL & Client **Convergence**



Latest generation Netezza Appliance

N3001

Integration with Hadoop for high speed query routing and bulk data movement

Continued enhancements to caching algorithms, optimizer, workload optimizations, etc.

Next Generation PDA

- Big Data optimized
- Elastic
- Mixed workloads
- Tiered storage
- Extreme concurrency
- Real time data ingest

IDAA on Cloud V1.1

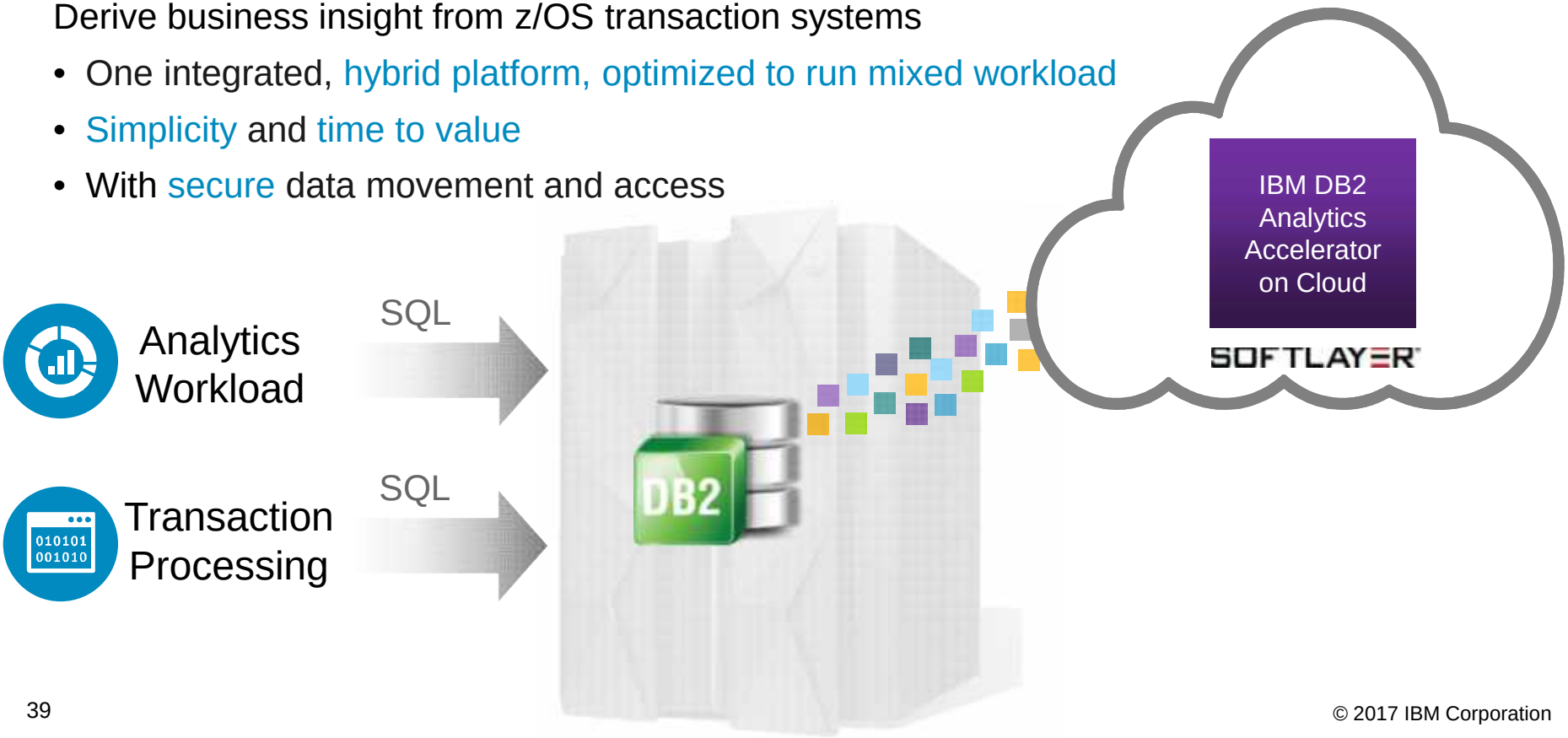
The Foundation for Open and Easy Access to DB2 Data – with dashDB under the covers!

Rapid acceleration of existing business-critical queries

- Improve performance and enable new insights while retaining z Systems security and reliability

Derive business insight from z/OS transaction systems

- One integrated, hybrid platform, optimized to run mixed workload
- Simplicity and time to value
- With secure data movement and access



IBM Database Conversion Workbench (DCW)

- **Helps you migrate your databases to dashDB**
- **Two options**
  - DCW (full version) – free of charge plugin for Data Studio
  - Database Harmony Profiler – standalone application (specifically for dashDB and DB2 on Cloud as targets)
    - Formerly known as DCW Lite
- **Supported migration paths into dashDB:**
  - **Oracle Database**
  - **PureData System for Analytics (PDA / Netezza)**
- **Compatibility Evaluation reports on the estimated compatibility ratio of customer's PDA or Oracle Database code with dashDB**
  - Outlines conversion issues, code that can be converted, and code that must be refactored manually
- **Code Conversion converts SQL statements to dashDB-compatible syntax**
  - Substantially cuts down on the time spent by customers refactoring their code

- **Additional reading:**
  - DCW developerWorks page: http://ibm.co/1NL1Fme
  - PDA to dashDB conversion guide: https://ibm.biz/PDA-to-dashDB-Conversion-Guide

Agenda

- Data Server Market and Strategy Update
- Introducing IBM Event Store – BLU Spark
- Introducing the Next Generation Appliance - Sailfish
- **Introducing BigInsights 4.3**

IOP 4.3 - Component Currency

Component	IOP 4.2 Version	IOP 4.3 Version
Ambari	2.2.0	**2.4.2**
Avro	1.7.7	1.7.7
Flume	1.6.0	**1.7.0**
Hadoop	2.7.2	**2.7.3**
HBase	1.2.0	**1.2.4**
Hive	1.2.1	**1.2.1**
Zookeeper	3.4.6	3.4.6
Knox	0.7.0	**0.11.0**
Oozie	4.2.0	**4.3.0**
Titan	1.0.0	1.0.0
SystemML	0.10	**0.13**

Component	IOP 4.2 Version	IOP 4.3 Version
Parquet	2.2.0	**2.2.0**
Parquet-mr	1.6.0	**1.6.0**
Pig	0.15.0	**0.16.1**
Ranger	0.5.2	**0.6.2**
Sqoop	1.4.6	**1.4.7**
Slider	0.90.2	0.91
Phoenix	4.6.1	**4.8.1**
Spark	1.6.1	**2.1**
Kafka	0.9.0.1	**0.10.1.0**
Solr	5.5	**6.3**

IOP 4.3 - Platform Support

Hardware	Operating System(s)	JDK's
X86_64	RHEL 6.8	OpenJDK 1.8
	RHEL 7.2, 7.3	OpenJDK 1.8
Power LE	RHEL 7.2, 7.3	OpenJDK 1.8

IOP 4.3 - Upgrade Support

- **IOP 4.3 will support the following upgrade paths**
  - 4.1 to 4.3
  - 4.2 to 4.3
- **Express Upgrade support (Fast off-line)**
  - An automated upgrade process, but require admin interaction
  - Very few pre-requisites to perform
  - Incurs cluster downtime but executes relatively quickly
- **Rolling Upgrade support (Slower on-line)**
  - An automated upgrade process, but require admin interaction
  - Some pre-requisites to perform
  - Require longer upgrade time window than Express Upgrade, but
  - HDFS, Yarn, HBase will function in limited mode (read-only) during upgrade process

IOP 4.3 – Technical Highlights

- **IOP 4.3 will be compliant with ODPi 2.0 runtime and ODPi 1.0 Operational specs**
- **Ambari 2.4.2**
  - Log search capability on IOP (Preview Feature)
  - Dynamic stack extensions (Install, configure, and upgrade support for custom services)
  - Role based access control beyond today's Ambari Admin, Operator and Read-Only permissions
  - Customized alerts, Grafana dashboards (thresholds, repeat tolerance, sort and filtering)
  - Improve usability of Blueprints (Exclude services or components when exporting a blueprint)
  - Workflow View (Preview Feature)
- **Hadoop : Enhance cgroups and CPU scheduling feature.**
- **Hadoop : YARN node label support which allows grouping nodes and specifying where the application will run**
- **Spark : Core API enhancements, Spark structured streaming and Spark SQL.**
- **Spark Integration with Notebook: Jupyter Notebook service is available on IOP 4.3.**
- **R4ML: R package to support SystemML on top of SparkR.**
- **Upgraded SystemML through SparkML**
- **Kafka Connectors (HDFS, JDBC) – ship part of product**
- **HBase Spark Connector**
- **HBase MOB (medium object files) support.**
- **Phoenix – Phoenix query server for remote access**
- **Knox - Ambari, Ranger support through Knox authentication, Knox SAML (SSO) provider support**
- **Ranger - Kafka and Solr plug-ins, PAM authentication, Kerberos support, tag based policies**
- **Titan – Add Titan server and security support to Titan**
- **Solr : Parallel SQL interface, Near real time processing, Solr UI, Indexing on HDFS files.**
- **Hive 2.x (Preview feature)**
- **Hive on Spark (Preview feature)**

IOP 4.3 – Technical Highlights

- **IOP 4.3 will be compliant with ODPi 2.0 runtime and ODPi 1.0 Operational s**
- **Ambari 2.4.2**
  - Log search capability or ure)
  - Dynamic stack extensions (Install, configu and upgrade support for custom services)
  - Role based access control beyond today's Ambari Admin, Operato permissions
  - Customized alerts, Graf (thresholds, repeat tolerance, sort and ring)
  - Improve usability of Blueprints (Exclude services or components when exporting a blueprint)
  - Workflow View (Preview
- **Hadoop : Enhance cgroup ling feature.**
- **Hadoop : YARN node label support which allows grouping nodes and specifying where the application will run**
- **Spark : Core API enhancements, Spark structured streaming and Spark SQL.**

Distributed streaming platform

Gateway providing perimeter security

Distributed Graph Database

- **Spark Integration with Notebook: Jupyter Notebook service is available on IOP 4.3.**
- **R4ML: R package to support SystemML on top of SparkR.**
- **Upgraded SystemML through SparkML**
- **Kafka Connectors (HDFS t of product**
- **HBase Spark Conne**
- **HBase MOB (medium object files) support.**
- **Phoenix – Phoenix query server for remote access**
- **Knox - Ambari, Ranger su ox authentication, Kno support**
- **Ranger - Kafka and Solr plug-ins, PAM authentication, Kerberos support, tag based policies**
- **Titan – Add Titan server a rt to Titan**
- **Solr : Parallel SQL interfa processing, Solr UI, Indexing on HDFS files.**
- **Hive 2.x (Preview feature)**
- **Hive on Spark (Preview feature)**

OLTP engine using Hbase as the data store

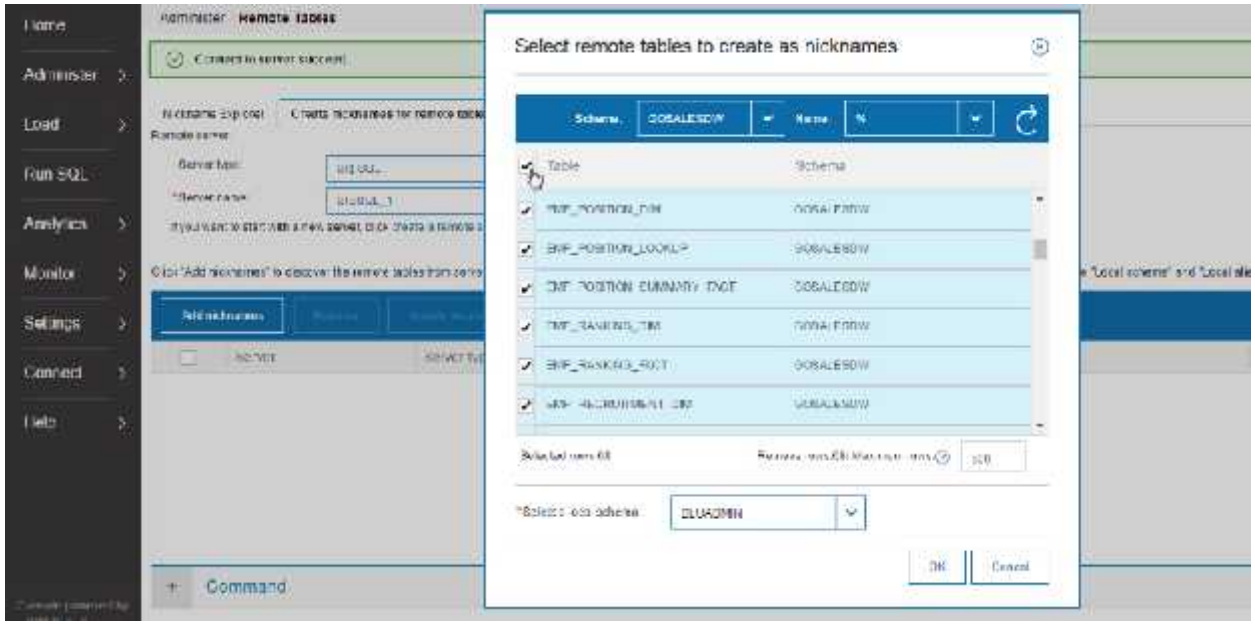
Define, administer and manage security policy over Hadoop

Search, indexing, replication, load balancing, failover

IOP 4.3 – New Free Maintenance & Migration Offer

- **For existing and new clients who implement IOP**
- **Free maintenance !!**
- **No purchase of BigInsights required**
- **Only pay if you are leveraging high value Big Data capabilities**
- **Register at : http://resources.Aberdeen.com/ibm-opensource/**
- **Switch to IBM and save 50% off your current Hadoop bill**
  - ***For a limited time, for half the cost of your current Hadoop bill, IBM will provide IBM® BigInsights®, Big SQL, and IBM Lab Services Migration assistance***

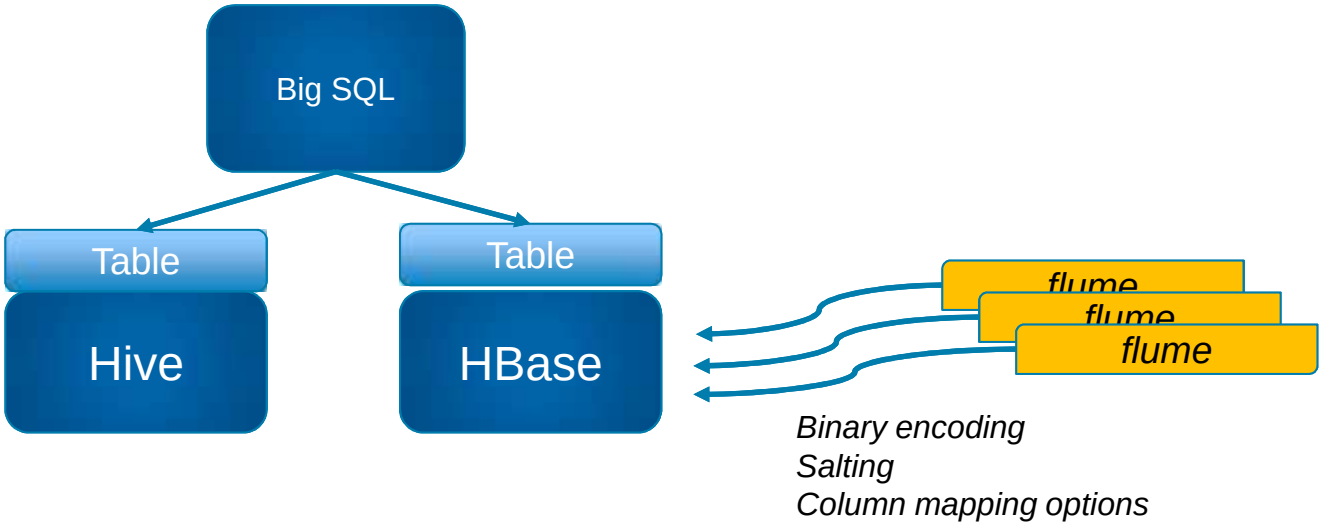
Big SQL 4.3 - Fluid Query Improvements / Usability



- **Big SQL V4.2**
  - Many obscure commands
  - Customer installs drivers
  - Hand coded SQL required to define servers, wrappers, and nicknames

- **Big SQL V4.3**
  - Drivers Pre-installed
  - Point-and-Click Wizard in Data Server Manager
  - Opportunity to optimize generated DDL

Big SQL 4.3 - Flume Connector for Big SQL (HBase Tables)



- **New Hbase Connector to load data that is flowing through Apache Flume into Big SQL tables stored in HBase.**
- **Apache Flume already has a default sink implementation for HBase. But, the implementation does not support Big SQL/HBase tables using different encodings, salting, and column mapping options.**
- **New connector supports all settings that may be defined on a Big SQL table stored in HBase, and that loads data in the required format.**

Big SQL 4.3 - Tables over Object Store

Protocols Supported: Swift, S3

```
CREATE HADOOP TABLE staff ( …)
LOCATION
'swift://swifttables.softlayer/staff';
```

```
CREATE HADOOP TABLE staff ( … )
LOCATION
      's3a://s3atables/staff';
```

- **Create Tables over Data residing in Object Store directly (no copy required into Hadoop)**
- **Once configured, Object Store tables work like any other table in Big SQL**
- **Benefits:**
  - No need to copy data into Hadoop first! Query data where it resides.
  - Partitioning supported!
- **Tradeoff:**
  - Expect reduced performance relative to Hive tables

LOAD FROM Object Store also supported!

Big SQL 4.3 - Tables over WebHDFS (Technical Preview)

```
CREATE HADOOP TABLE staff ( … )
   PARTITIONED BY (JOB VARCHAR(5))
LOCATION 'webhdfs://namenode.acme.com:50070/path/to/table/staff';
```

- **Transparently access data on any platform implementing WebHDFS**
  - Examples: Microsoft Azure Data Lake
- **Once setup, WebHDFS tables work like any other table in Big SQL**
- **Technical Preview Limitations:**
  - WebHDFS via Knox not supported
  - Performance not well understood. Reduce performance expected.

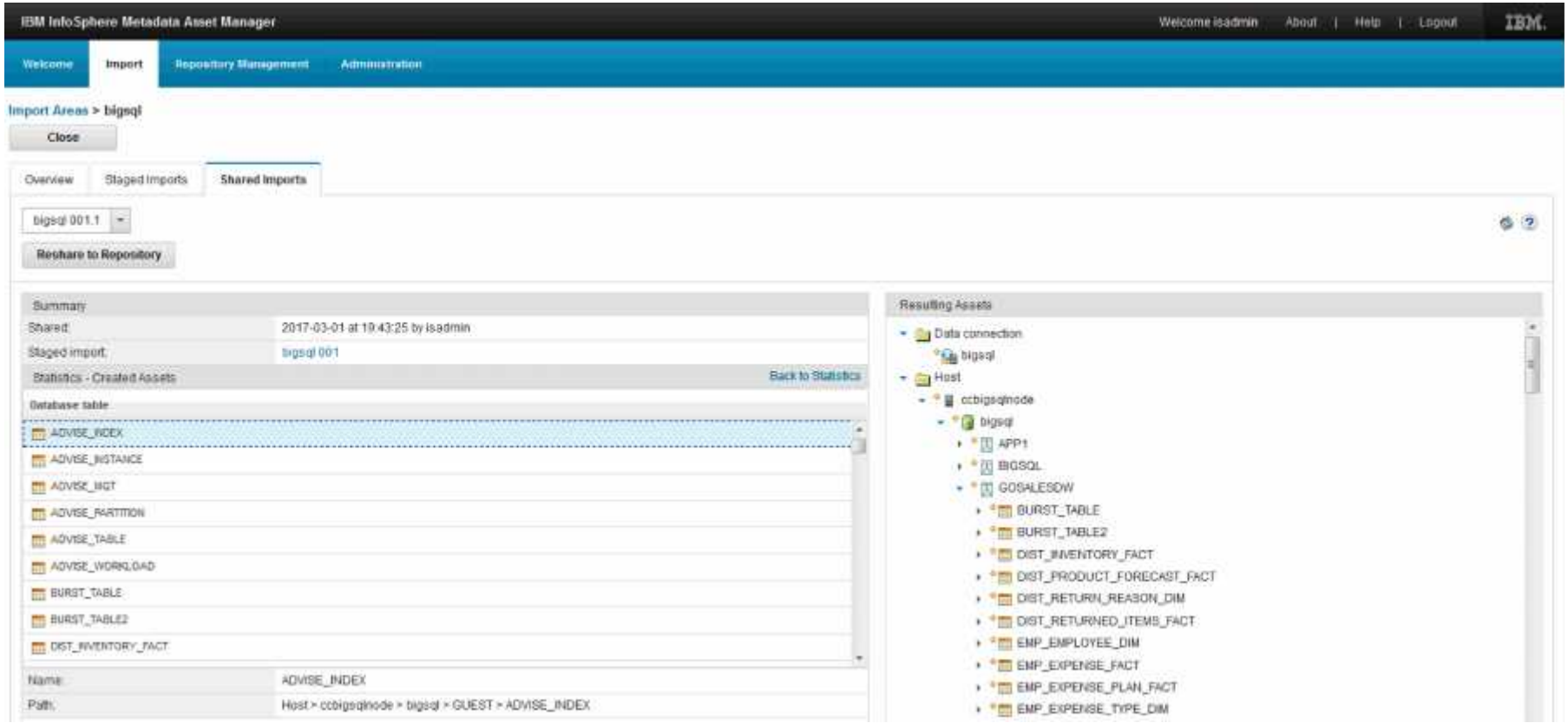
LOAD FROM WebHDFS also supported!

Big SQL 4.3 - Apache Ranger Integration

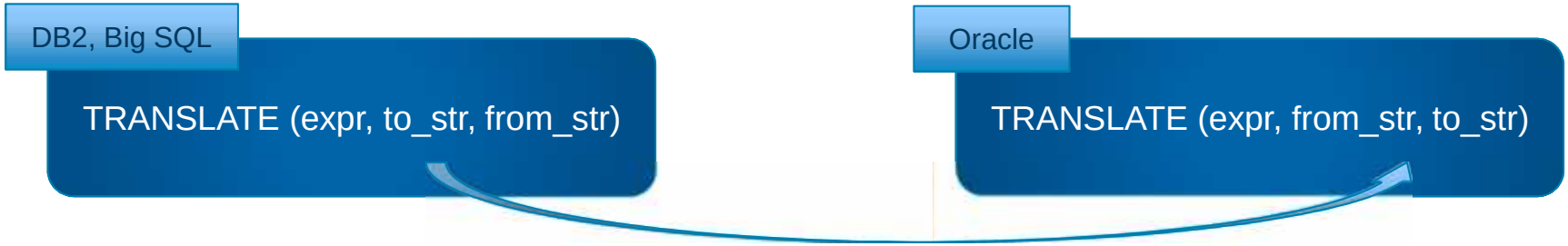
- **Setup ACLs for access to Big SQL tables:**
  - create, alter, analyze, load, truncate, drop, insert, select, update, and delete.
- **Supports Ranger Audit**
  - Big SQL access audit records written to HDFS and/or Solr
- **If also using Ranger Plugin for Hive – operates independent of Big SQL plugin**

Big SQL 4.3 - Information Governance Catalog Integration

- **Metadata Asset Manager discovers Big SQL objects to support information governance**



Big SQL 4.3 – Oracle Compatibility - SET sql_compat='ORA'



Same function, parameters reversed!

- **SQL_COMPAT global variable lets Big SQL support multiple vendors SQL syntax**
- **Enables PL/SQL support! (New for V4.3)**
- **SQL data-access-level enforcement**
  - enforce data access levels at run time rather than at compile time.
- **Oracle database link syntax (@ symbol)**
- **Setting of the DB2_COMPATIBILITY_VECTOR registry variable (inherited from DB2) is not recommended in Big SQL. Custom compatibility features should be enabled only by using the SQL_COMPAT global variable.**

Big SQL 4.3 - Oracle PL/SQL Support

```
set sql_compat='ORA'
```

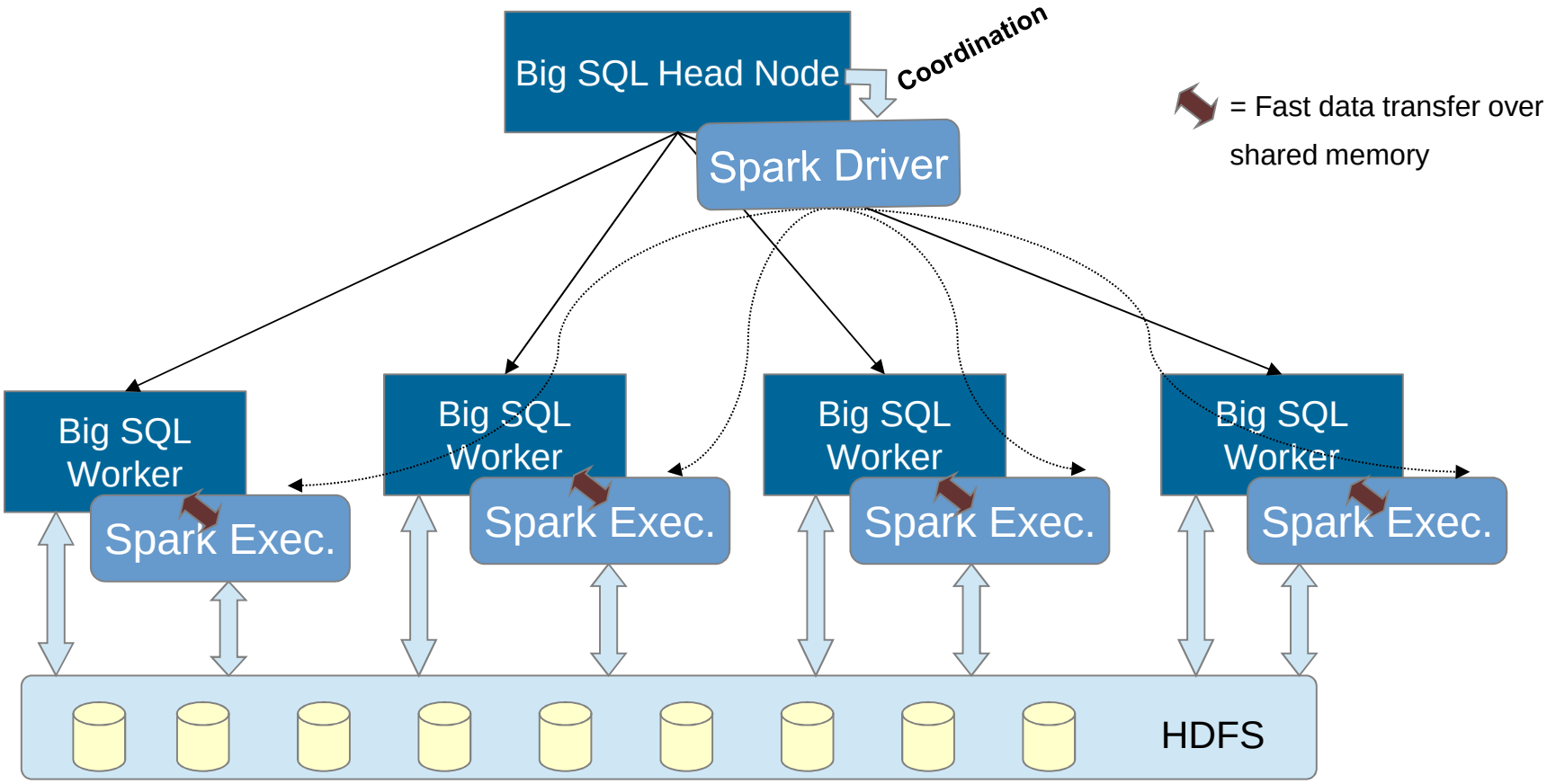
Easy session variable to switch modes!

```
create or replace procedure plsql_proc (fetchval out integer)
as

cursor cur1 is
select count(*) from syscat.tables ;

-- begin

open cur1 ;

fetch cur1 into fetchval ;
close cur1 ;

end
```

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

- **Pre V4.2, Big SQL already supposed high degree of Oracle SQL compatibility**
- **Big SQL V4.3 adds support for Oracle PL/SQL procedural language**

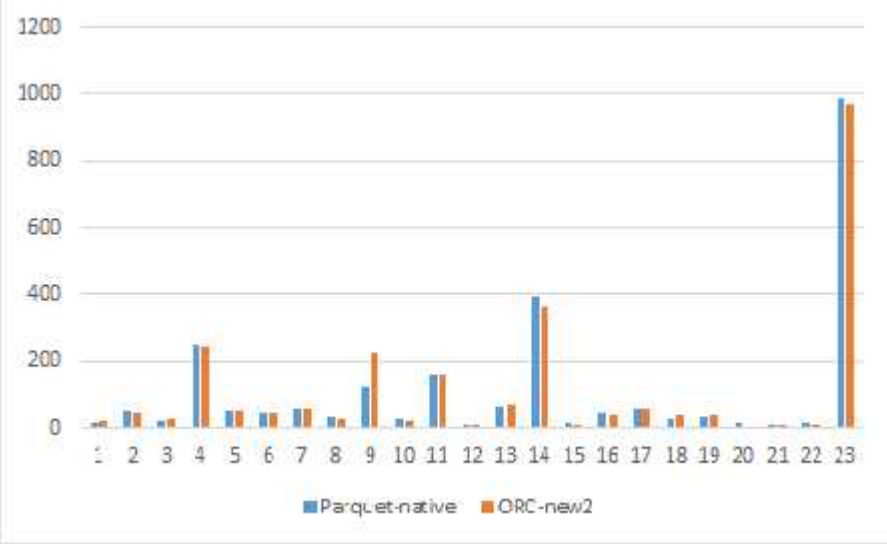
Big SQL 4.3 – Deep Integration with Spark

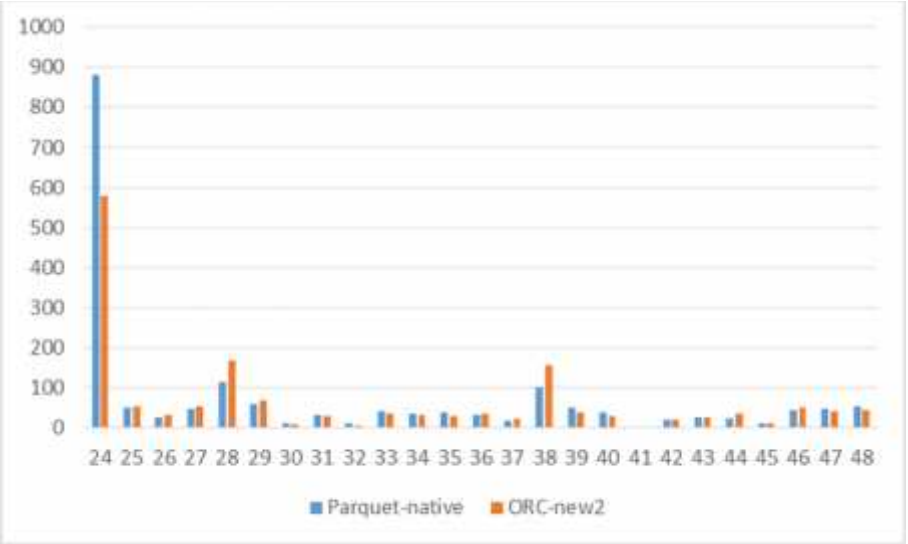


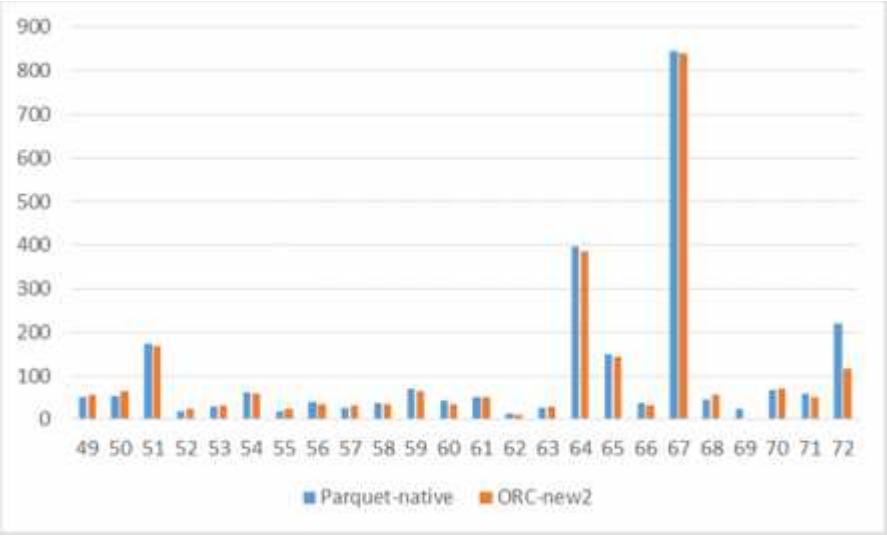
Big SQL 4.3 - Improvements to Big SQL ORC Readers

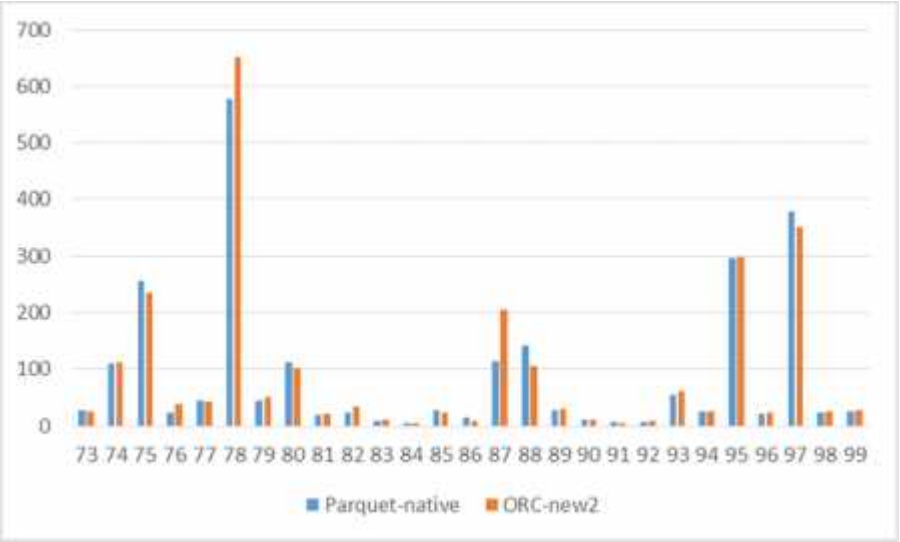
- **Battle of Storage formats:**
  - Hortonworks prefers ORC
  - Cloudera prefers Parquet
- **Parquet format was preferred format in V4.2 and prior due to Big SQL native C/C++ implementation of readers**
- **Big SQL on Hortonworks → Big SQL needs to work better with ORC**
  - Still Java Implementation, but now has similar performance to Parquet Native Readers
- **IBM long term objective: Optimize performance for both storage formats.**

Big SQL 4.3 - Parquet (Native) vs. ORC (Java) Performance TPC-DS queries



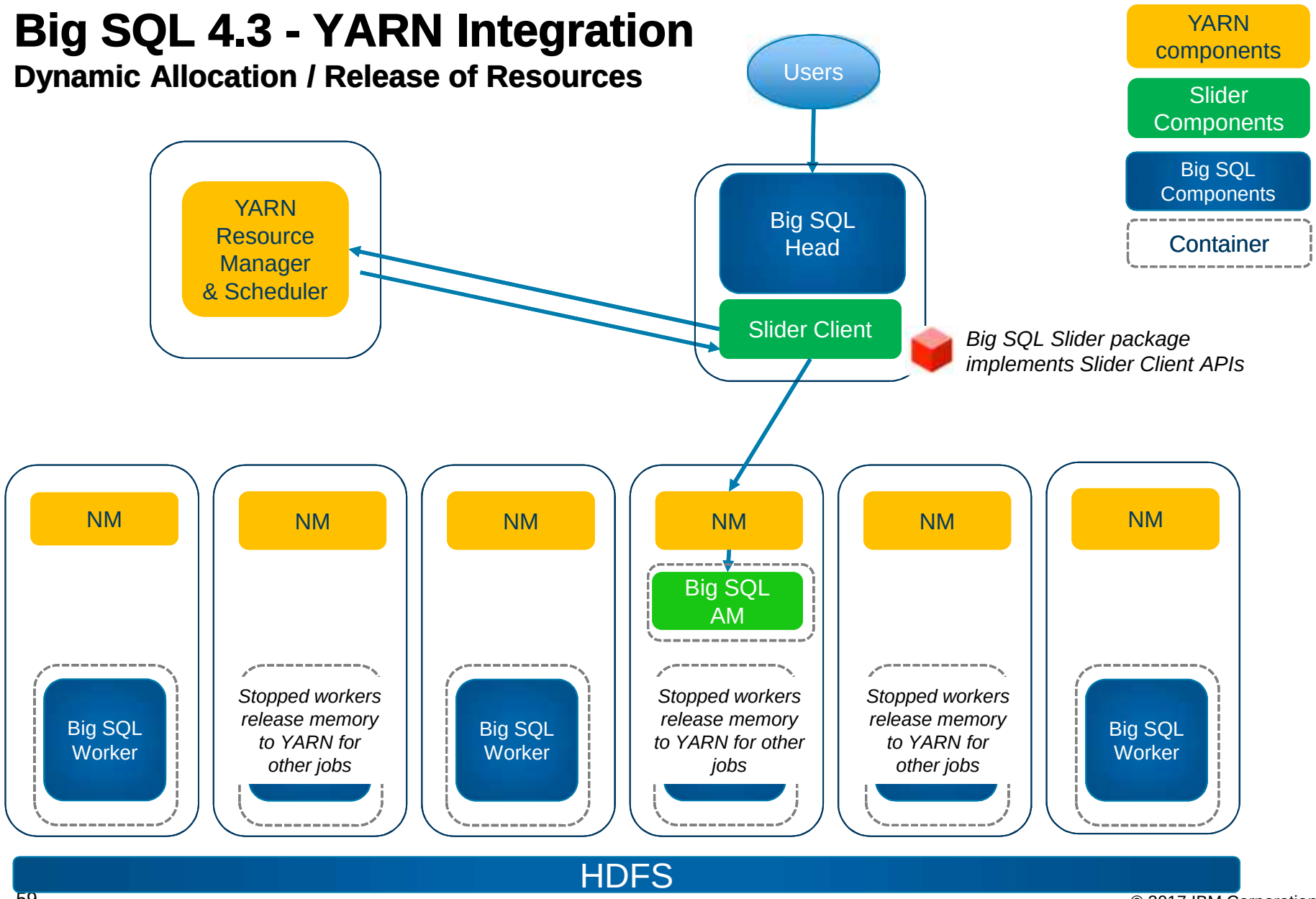






Big SQL 4.3 - YARN Integration

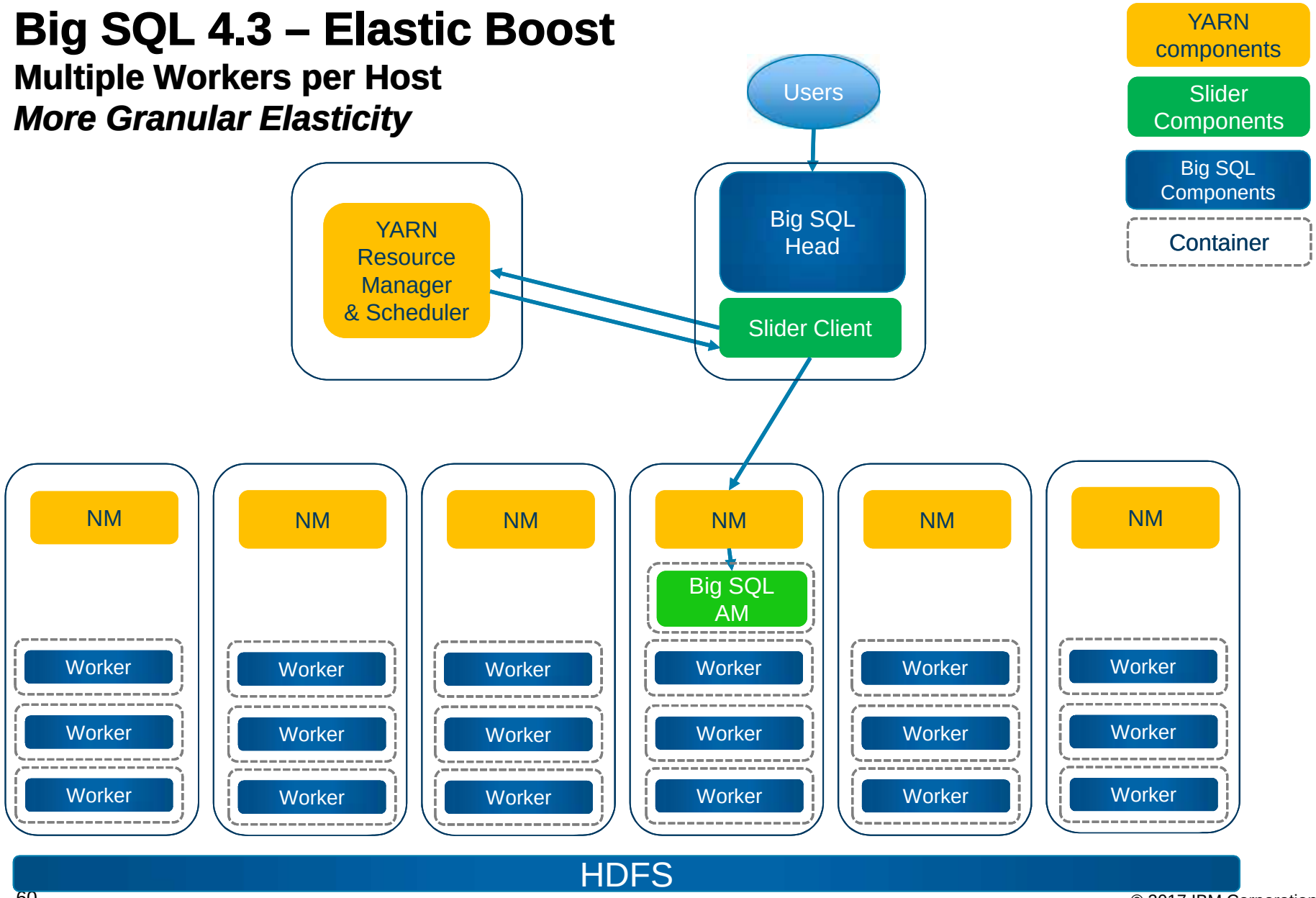
Dynamic Allocation / Release of Resources



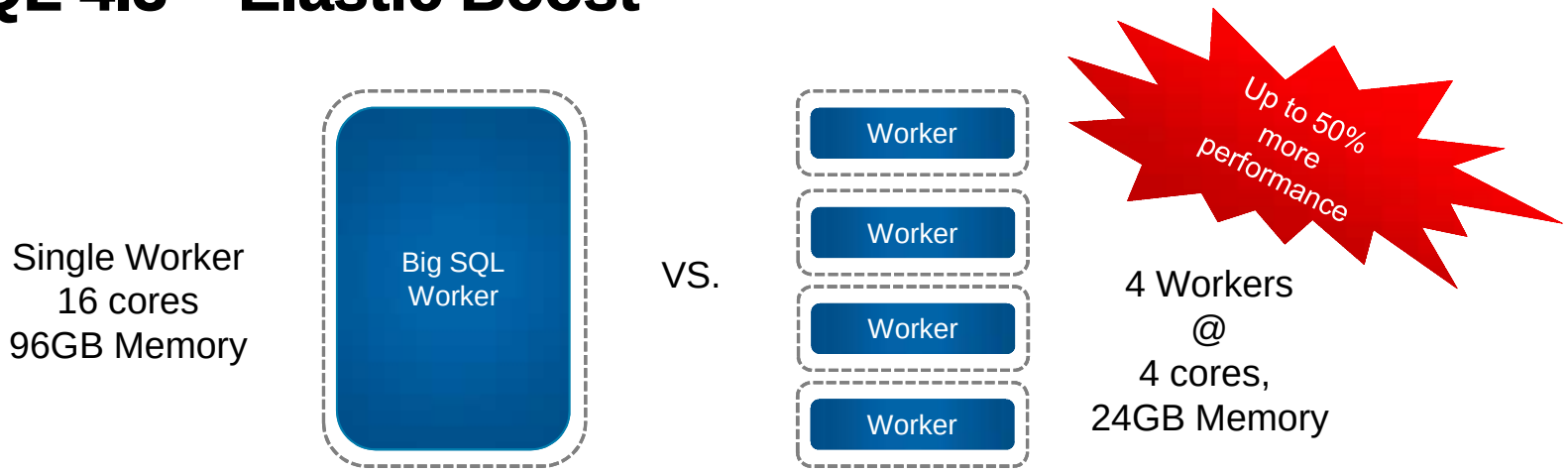
Big SQL 4.3 – Elastic Boost

Multiple Workers per Host

More Granular Elasticity



Big SQL 4.3 – Elastic Boost

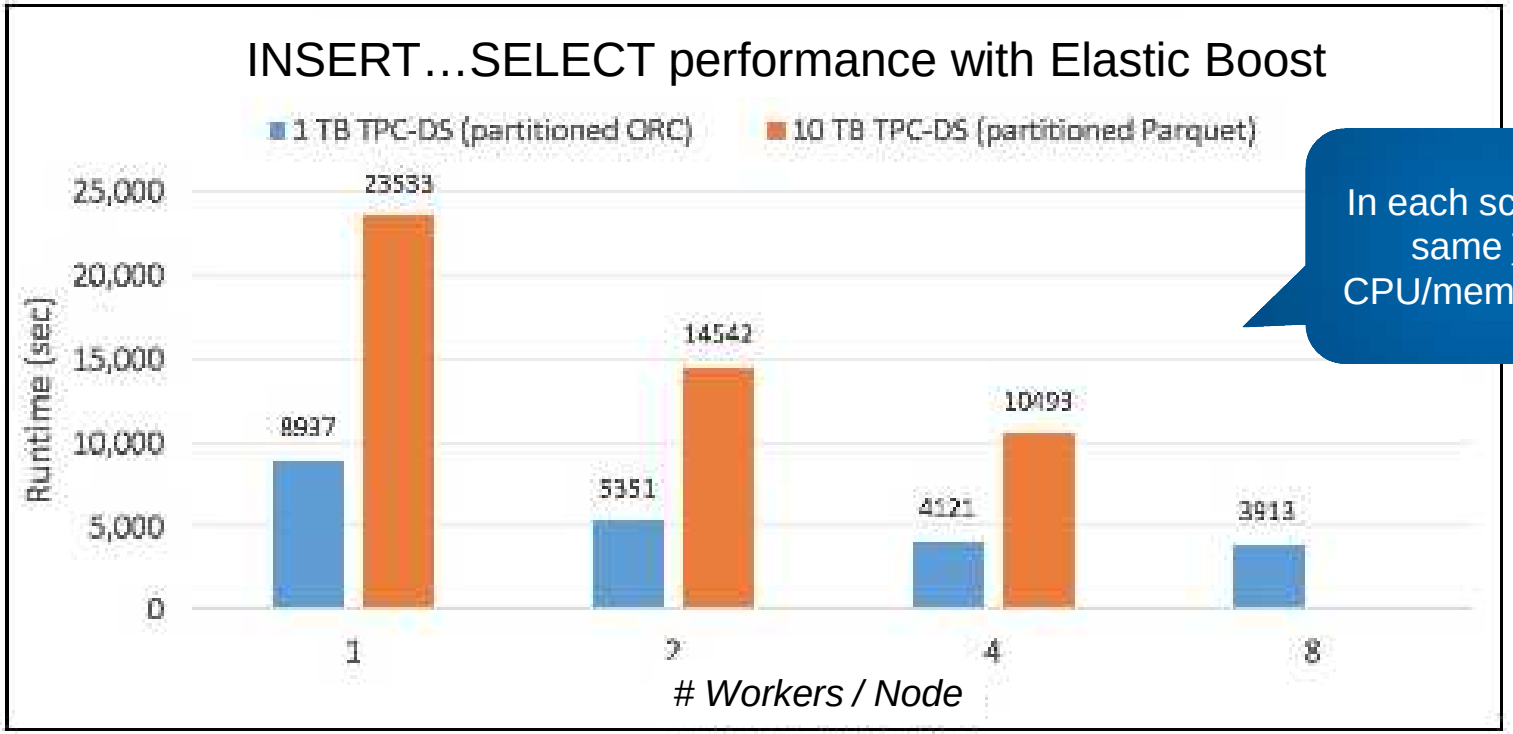


- **For large SMP servers (> 8 cores, 64GB memory) – multiple workers per host yields up to 50% more performance* given the same resources (memory, CPU, disk, network)**
  - World-Record Result: Big SQL Hadoop-DS running TCP-DS queries at 100TB scale used 12 workers/host

- **Usual constraints:**
  - Elastic boost will result in greater memory and CPU exploitation, but bottlenecks may show up in other areas of the shared host (workers still share network, disk, etc..)
  - Assumes relatively balanced activation of workers across all nodes (YARN decides)
  - Minimum recommended worker resources (2 cores, 24GB memory) still applies.

Big SQL 4.3 - Elastic Boost

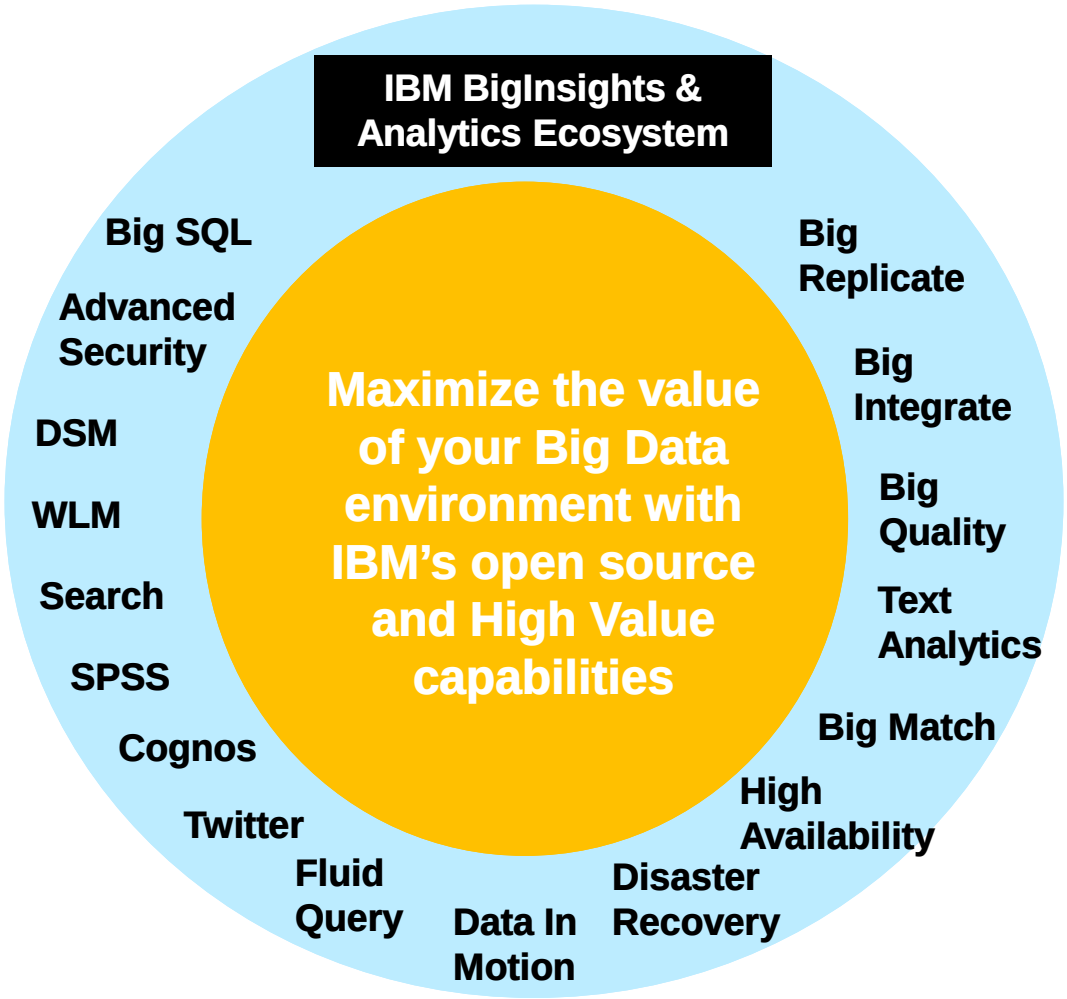
Improves INSERT… SELECT FROM Performance



- **For both 1 and 10 TB TPC-DS dataset**
  - 2 Workers/Node: **1.6x speedup**
  - 4 Workers/Node: **2.2x speedup**

High Value Capabilities for your Big Data Environment

Available and Supported on all key distributions



Business Analyst

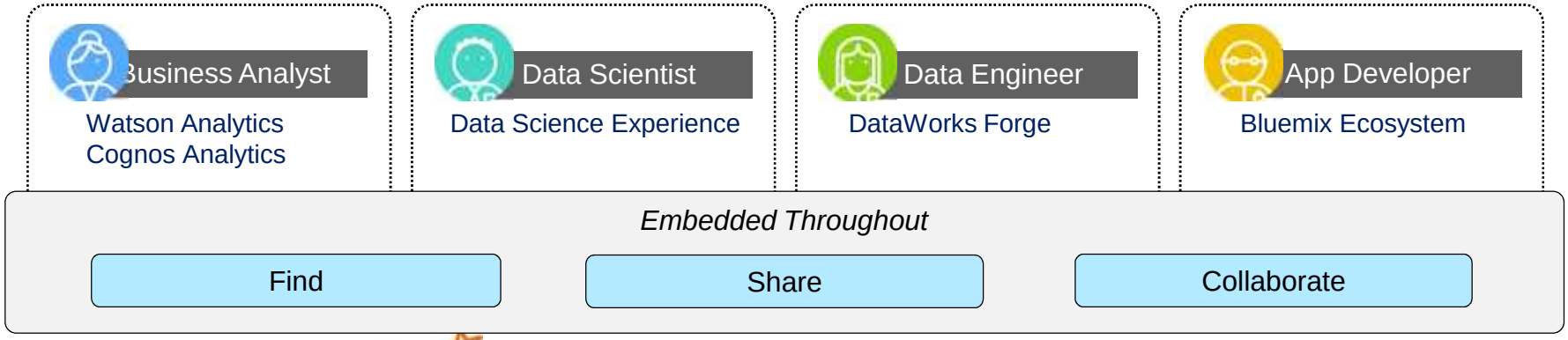
Data Scientist

Administrator

- **IBM Hadoop Distribution:** Free and is composed entirely of open source Apache Hadoop related projects.
- **IBM Hadoop Ecosystem**: IBM enterprise Hadoop product ecosystem for Data Science, Management, Security, and Integration

Watson Data Platform – Reference Architecture (Hadoop)

BigInsights includes high value capabilities which provide solutions for all aspects of our Watson Data Platform. This provides the ability to include a Hadoop environment, with full Ingest, Store, Analyze and Deploy capabilities, as part of a broader Hybrid cloud ecosystem.



Spark as underlying engine

Data Sources

- *On-premises / cloud*
- *Unstructured / relational*
- *Streaming / batch*
- *Feeds (e.g. ,)*

← Information Governance Catalog — *Govern* — Open Metadata Services →

Ingest	Store	Analyze	Deploy
Streams	BigInsights **	Cognos Analytics	Data Server Manager
Data Replication	Hortonworks *	Data Scientist Tools	Workload Management
Big Integrate	Cloudera *	Graph Processing	High Availability
Big Quality	MapR *	Spark (Python, R, SQL)	Advanced Security
Big Match	GPFS	Text Analysis (AQL)	Governance
Governance Catalog	Common Analytic Engine	Watson Explorer	Big Replicate
			Spark

* Supported open source framework distributions

** An open source framework is included with BigInsights for a complete single vendor solution

Store

BigInsights IOP Distribution	**Non-proprietary** apache open source framework included; Maximum value option
BigInsights and Spectrum Scale (GPFS)	**POSIX Compliant with encryption** option for your Hadoop cluster
BigInsights and Common Analytic Engine	**Common Analytic SQL MPP Engine** for Hybrid Cloud and Hybrid Warehouses
Hortonworks	Many **IBM Big Data high value capabilities** supported on Hortonworks
Cloudera	Many **IBM Big Data high value capabilities** supported on Cloudera
MapR	Some **IBM Big Data high value capabilities** supported on MapR

Ingest

BigInsights BigIntegrate	**Ingest, transform, process and deliver** any data into & within Hadoop
BigInsights BigQuality	**Analyze, cleanse and monitor** your big data
BigInsights BigMatch	**Customer Matching** natively within Hadoop
Information Governance Catalog	**Understand and manage** metadata for **your most critical information**
Data Replication	**Real-time ingest data into Hadoop** with lowest impact to sources
Streams	**Ingest and Analyze data in motion** for high volumes of data and low latency

Analyze

Business Analytic Tools	**Cognos, SPSS, Watson Analytics, Twitter** all available for Hadoop data
Spark API	**SQL, Python, R, Streaming, ML and Graphing** analytics with Spark
BigSQL and Fluid Query	**Fully functional SQL engine** for Hadoop with built-in **Federation Server**
Data Scientist Tools	**Data Server Manager** and **Data Scientist Experience** for collaboration
Text Analytics	**High Speed** text analytics leveraging **in-memory Spark engine**
Watson Explorer	**High Speed** sophisticated **Search Engine** for data in Hadoop

Deploy

Spark	**In-memory framework for analytic processing** of all types of data
Data Server Manager	**Management and Monitoring** of data in a Hadoop cluster
Workload Management	Ability to **control, prioritize and manage resources** in Hadoop cluster
High Availability and Disaster Recovery	**Protection** against **outages and disasters** for mission critical applications
Advanced Security	Be assured of your **Data Protection** in a Hadoop cluster
Governance	**Data Life Cycle Management** for data in a Hadoop cluster

IBM Data Server Strategy Update