Db2 –
Ready for Modernized Workloads and Deployments

Les King, IBM
lking@ca.ibm.com

Db2 LUW

The AI Ladder

A prescriptive approach to accelerating the journey to AI

AI

INFUSE – Operationalize AI with trust and transparency

ANALYZE – Scale insights with AI everywhere

ORGANIZE – Create a trusted analytics foundation

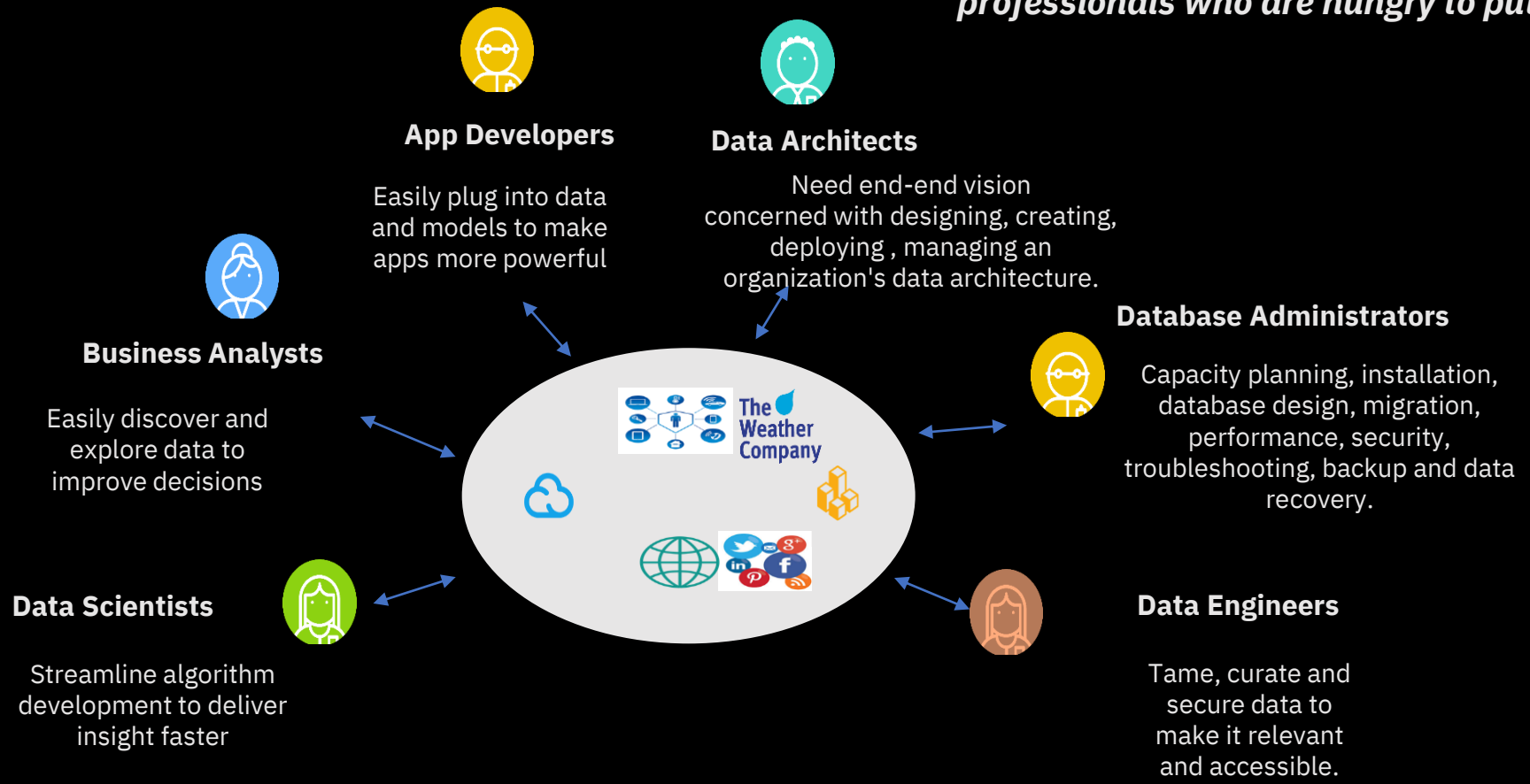
COLLECT – Make data simple and accessible

MODERNIZE your data estate for an AI and multicloud world

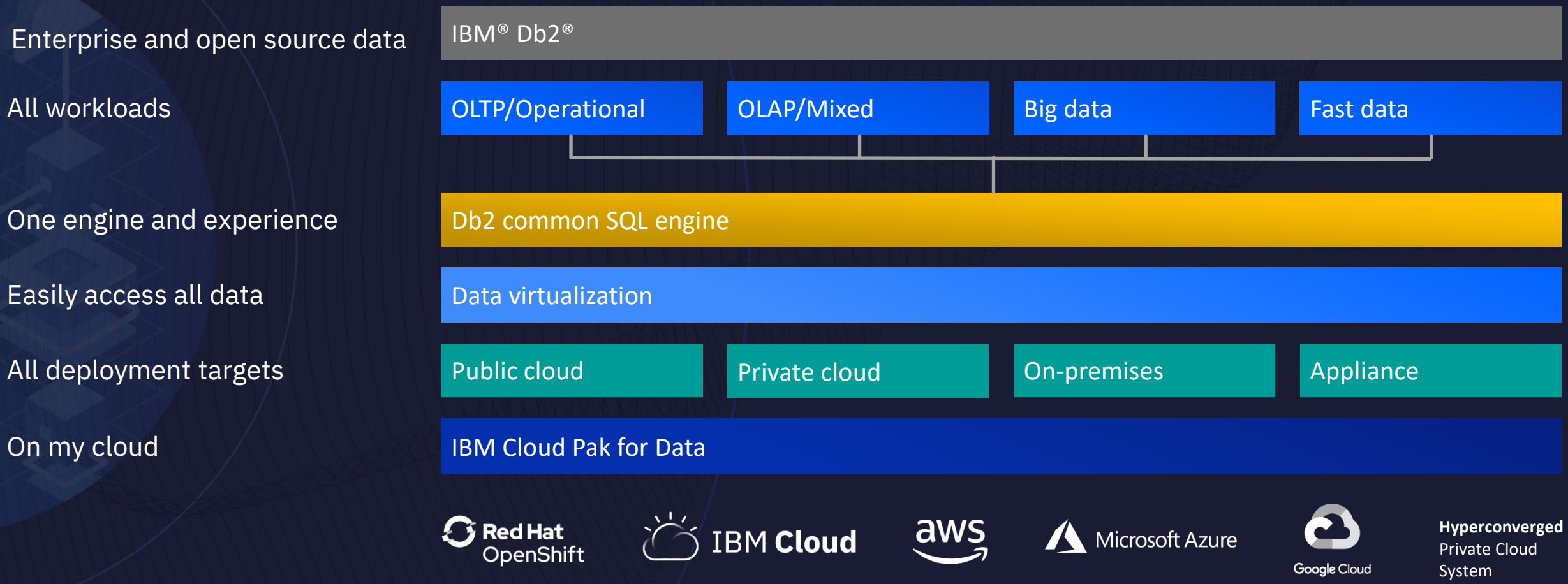
Data of every type, regardless of where it lives

Evolving Needs of all Data Professionals

As data maturity increases, so does the number of data professionals who are hungry to put data to work



Collect: IBM Hybrid Data Management



Two Dimensions of Modernization

- Deployment Form
- Workload Dynamics

Db2 – A Modern Database System

Containerized Deployment Options

The ability to deploy Db2 leveraging containers and a RedHat OpenShift platform

Data Virtualization

Db2 contains a data virtualization component which allows Db2 to be a doorway to all of your business data

Fully Managed DBaaS

The ability to deploy Db2 as a fully managed service for public, multi and hybrid cloud environments

In-Db2 Machine Learning

Allows data scientists and developers to bring machine learning local to the data stored within Db2

Cornerstone of COLLECT in Journey to AI

Db2 is the cornerstone for COLLECT in Cloud Pak for Data to cover your end-to-end needs to leverage AI

Db2 Leverages Machine Learning

Db2 optimizes for your workloads using a machine learning optimizer

Multi-Modal Data Store

Db2 is a multi-modal data store supporting native relational, XML, JSON, BSON and Graph data models

NoSQL / NewSQL Data Store

Db2 supports SQL, Mongo API, XQuery and Gremlin as ways to access data in our multi-modal data store

Db2 – Modernization – Deployment Form

Containerized Deployment Options

The ability to deploy Db2 leveraging containers and a RedHat OpenShift platform

Fully Managed DBaaS

The ability to deploy Db2 as a fully managed service for public, multi and hybrid cloud environments

Cornerstone of COLLECT in Journey to AI

Db2 is the cornerstone for COLLECT in Cloud Pak for Data to cover your end-to-end needs to leverage AI

Multi-Modal Data Store

Db2 is a multi-modal data store supporting native relational, XML, JSON, BSON and Graph data models

Data Virtualization

Db2 contains a data virtualization component which allows Db2 to be a doorway to all of your business data

In-Db2 Machine Learning

Allows data scientists and developers to bring machine learning local to the data stored within Db2

Db2 Leverages Machine Learning

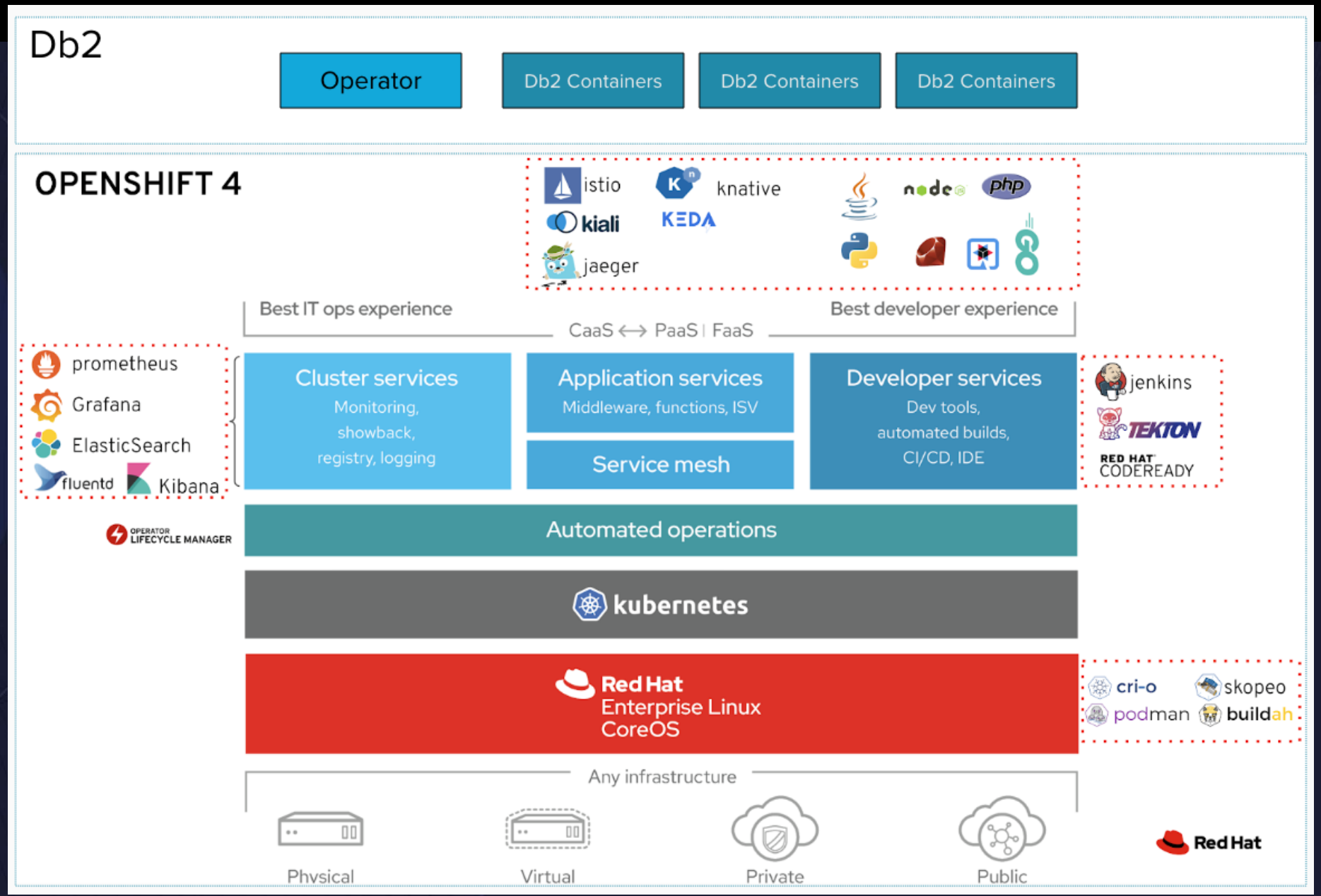
Db2 optimizes for your workloads using a machine learning optimizer

NoSQL / NewSQL Data Store

Db2 supports SQL, Mongo API, XQuery and Gremlin as ways to access data in our multi-modal data store

Db2 & OpenShift - Containerized Deployment

IBM **Db2**
+
RED HAT OPENSHIFT Container Platform

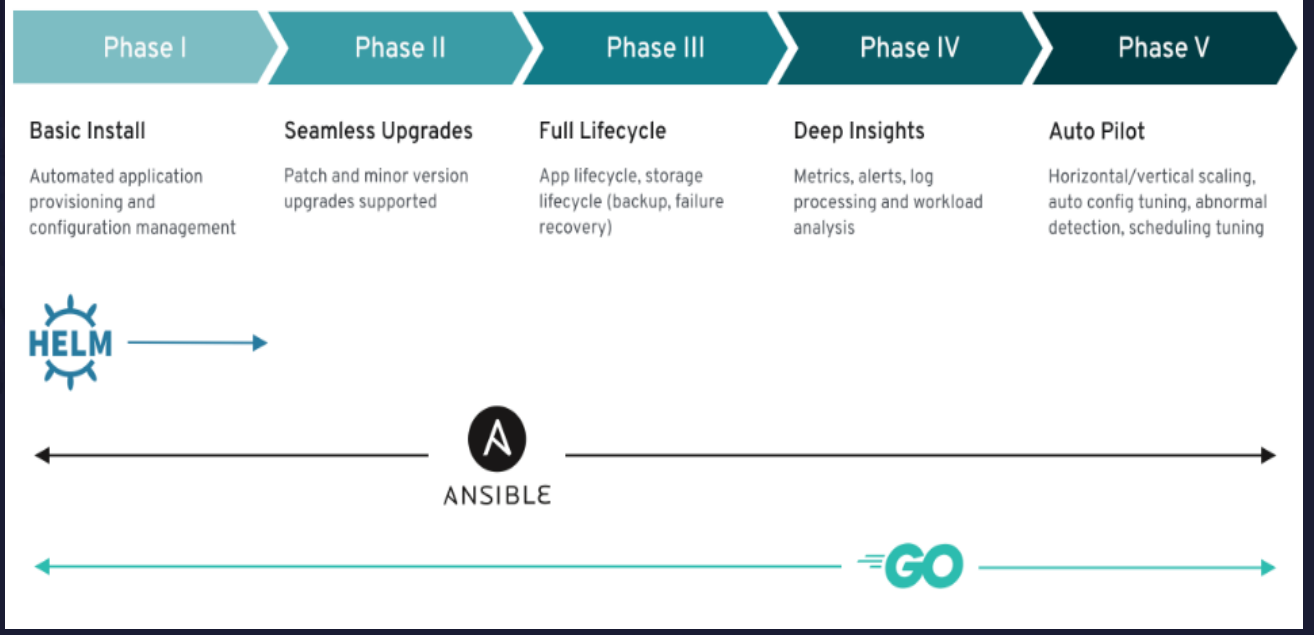


Db2 on Openshift – born in the k8s era…

With a common underlying architecture based on Kubernetes Operators, Db2 delivers a consistent experience across cloud environments:

- **RedHat OperatorHub**: prototyping and production use
- **Cloud Pak for Data**: integration with Governance & AL/ML
- **Db2 Everywhere:** seamless applications migrations to AWS, Azure, SL , Google Cloud
- **"True hybrid" experiences**: transparent data movement between public and private clouds with Common SQL Engine

Red Hat OpenShift



https://marketplace.redhat.com/en-us/products/ibm-db2-on-cloud-pak-for-data

Now Available on Red Hat Marketplace!!

Db2 Operator – Performance Evaluation

All Performance Results for
Db2 on RHOS

Metrics	Db2U on RHOS
Install Time	[Automated & includes instance/Database Creation] Time: ~ 2' to 4' (Db2)
Upgrade Time	[Automated] Time: 20'
HADR Takeover time	[Automated] Time: 2-4' minute
Recovery on Db2 Failure	[Automated] Time: 3 minutes + Partial Crash Recovery time* (observed < 1 min on SMALL)
Recovery on Hardware Failure	[Automated] Time: 6 minutes + Partial Crash Recovery time*
Scaling Compute	Time: 2 minutes

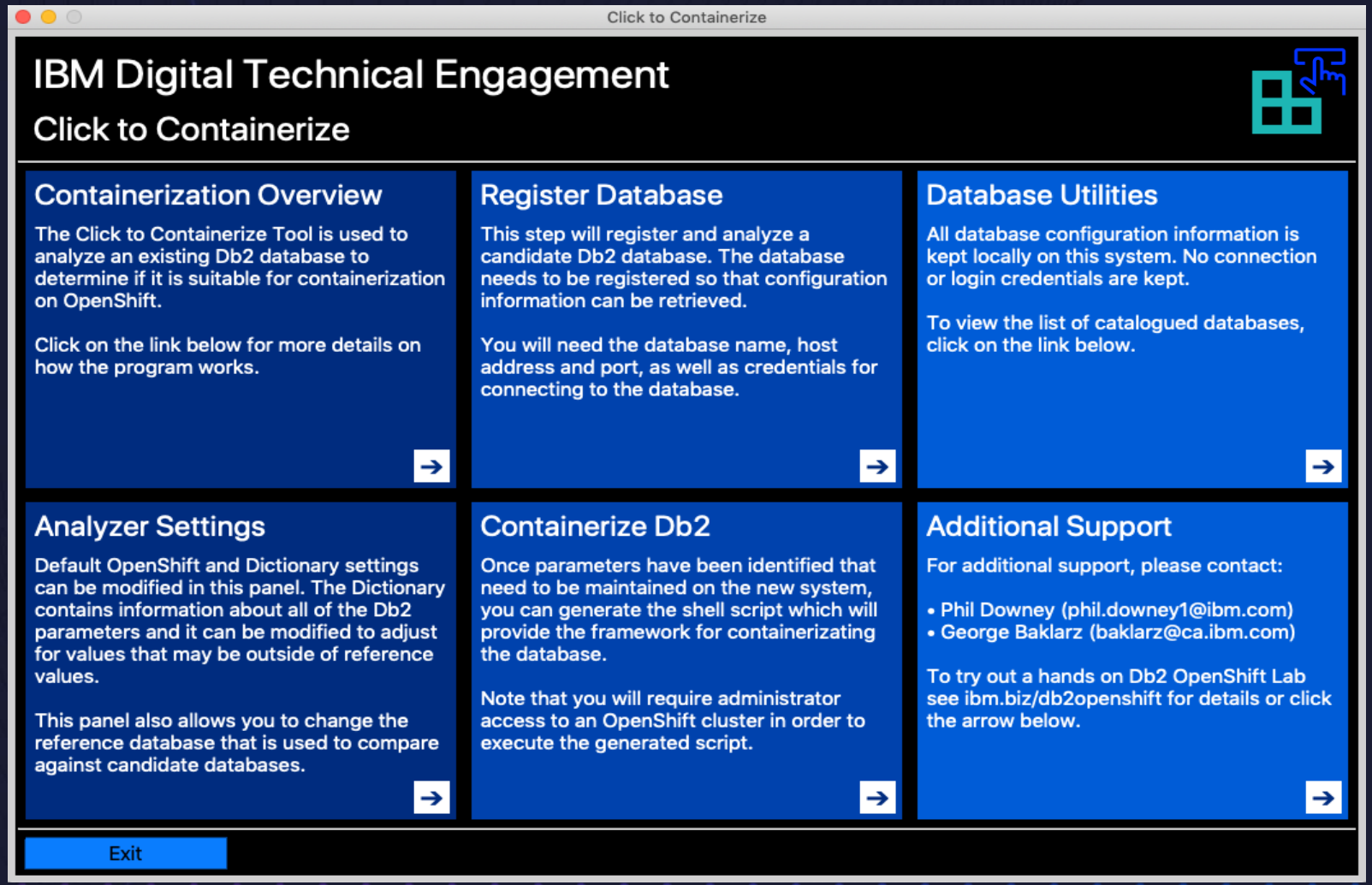
- Partial Crash Recovery time*: Usually complete in minutes with worst case at 20/30 minutes

Reference architecture - https://www.redhat.com/en/resources/IBM-Db2-Warehouse-MPP-on-OpenShift-Container-Storage-detail

Db2 C2C (Click 2 Containerize)

Accelerated Pathway for Customers to hybrid Cloud

- It's **Containerization**
  - *NOT Migration*
  - No Backup Restore
  - No Export / Import
  - No Reconfiguration
  - Data Handled Securely



Db2 C2C (Click 2 Containerize)

Accelerated Pathway for Customers to hybrid Cloud

Db2 for OpenShift / Cloud Pak for Data

Db2 On Prem

Db2 on Linux

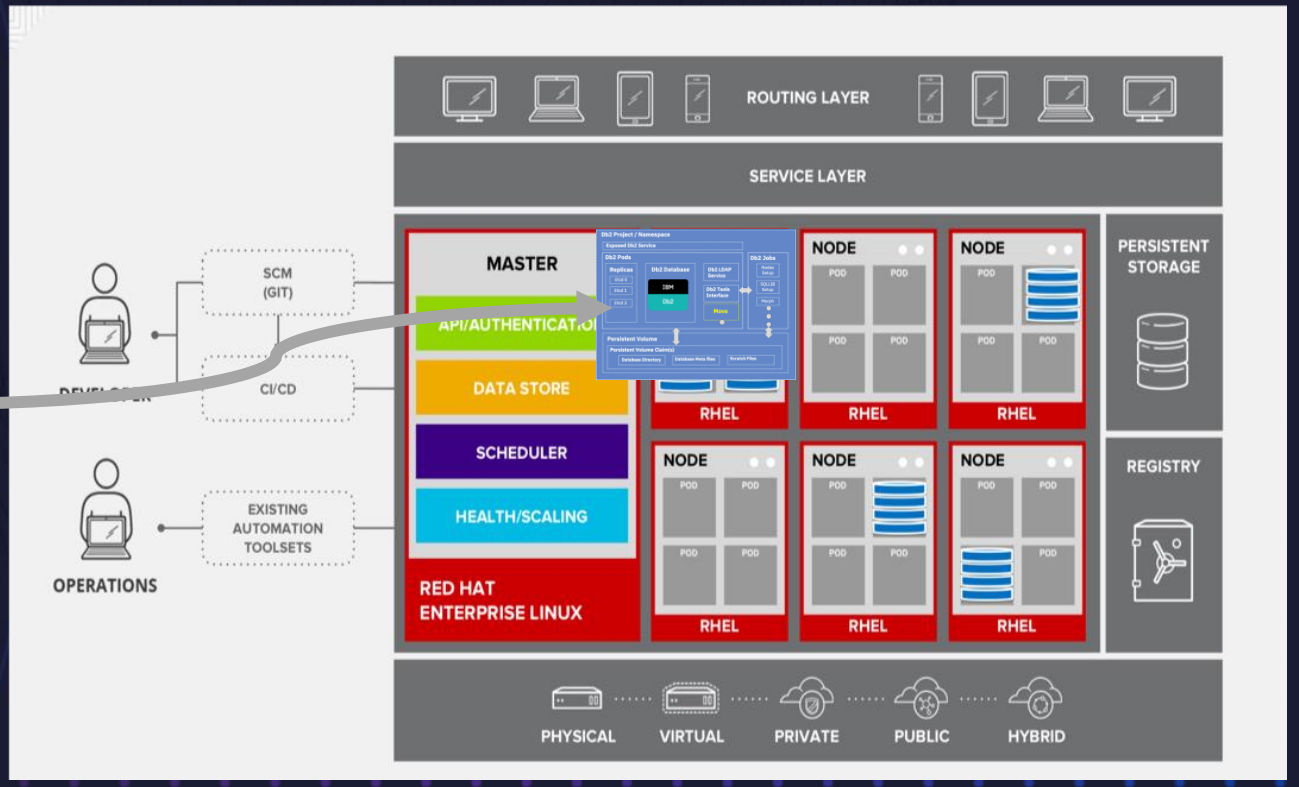
Oc Client

Internet/ Network

- Logs
- Mirror Logs
- Storage Path
- Storage Path X
- External UDFs
- DB Config Files
- KeyStore

Tested over 15,000 km Internet connection

Tunable Parallel Data Transfer
Data Compression
Fully Auditable



Containerized Db2 Hybrid Cloud Service

Cloud Pak for Data

Foundational “out of the box” multi-cloud Data & AI services

The Ladder to AI

InfuseAI

App Developers | Data Engineers | Data Stewards | Data Scientists | Business Users & Analysts

Collect

- Data virtualization
- Data warehousing
- Databases on-demand
- Data source ingestion
- Distributed processing

***Powered by**: Db2 and Db2 Warehouse technologies*

Organize

- Discovery & search
- Data transformation
- Data cataloging
- Business glossary
- Policies, rules & privacy

***Powered by**: Infosphere, Data Stage and IGC/WKC*

Analyze

- Data visualization
- Machine learning learning
- Model build & deploy
- Model management
- Dashboards

***Powered by**: Watson Studio open source and Cognos*

Multicloud Services

- Logging
- Monitoring
- Metering
- Persistent Storage
- Kubernetes
- Security
- Identity Access Mgmt.
- Docker Registry / Helm

Cloud Pak for Data

IBM Cloud | RED HAT OPENSHIFT | openstack | amazon web services | Azure

Cloud Pak for Data + Db2

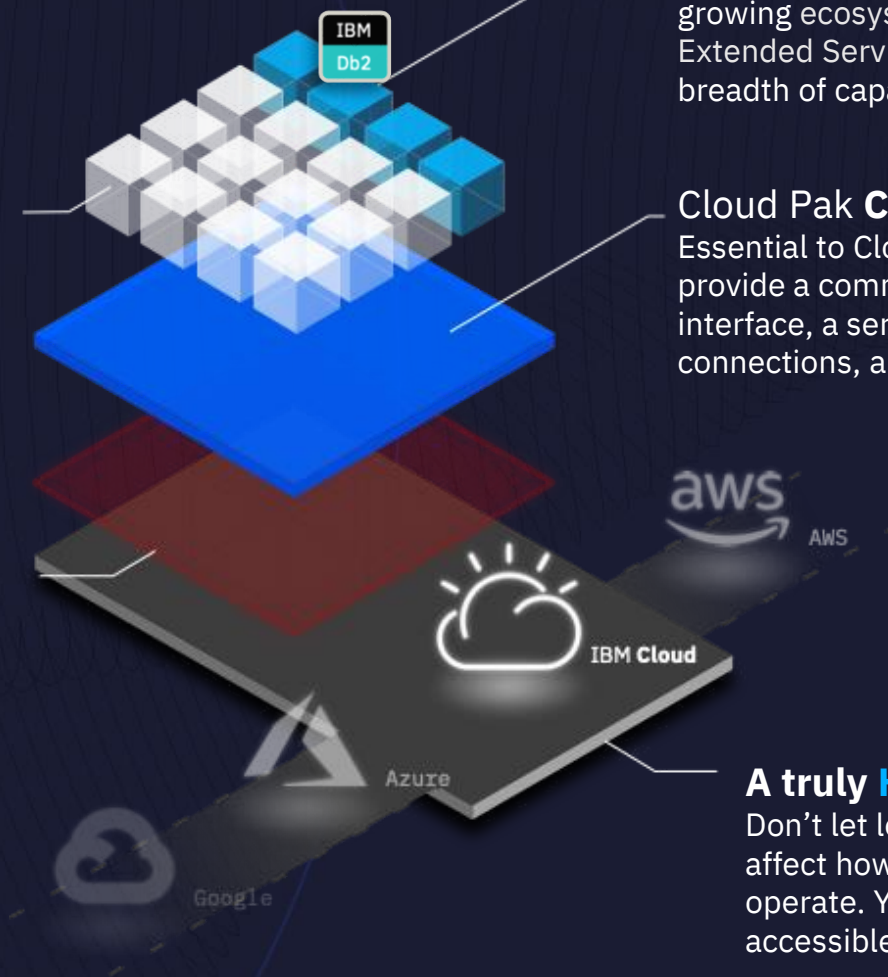
Simplifies, unifies and automates your journey to AI

Extended Services

Beyond its Base Services, Cloud Pak for Data has a growing ecosystem of OpenSource, Partner, and IBM Extended Services **– including Db2** to expand the breadth of capabilities for teams.

Base Services

Core to Cloud Pak for Data, these Base services fastrack organizations on their **end-to-end data and AI journey** with comprehensive continuum of capabilities.



Cloud Pak **Control Plane**

Essential to Cloud Pak for Data, Foundational Services provide a command-line interface, an administration interface, a services catalog, a central list of connections, and the central user experience.

Red Hat® OpenShift®

A streamlined hybrid-cloud foundation to take advantage of the underlying resource and infrastructure optimization and management. OpenShift license dedicated to run Cloud Pak for Data.

A truly Hybrid-Cloud world

Don’t let lock-in or major IT decisions affect how your Data and AI temas operate. Your work and data should be accessible in any Hybrid-Cloud strategy.

IBM Db2 Cloud Offerings

Fully-managed, high-performance, elastic cloud services, built for the modern enterprise.

IBM Db2 on Cloud

- Transactional / Operational database as a service

- **IBM Db2 Warehouse** on Cloud
- Cloud data warehouse as a service

- Comes with all the features you need to securely run your enterprise data workloads:

- *Role-Based and Row/Column Access Control*
- *Auditing* • *Query Federation* • *In-Database ML* • *Geospatial Analytics* • *JSON* • *Graph Query* • *Workload Management* • *Oracle Compatibility*

Tens of thousands of customers worldwide

Hundreds of billions of activities monthly

Petabytes of data under management

Db2 – Modernization - Workloads

Containerized Deployment Options

The ability to deploy Db2 leveraging containers and a RedHat OpenShift platform

Fully Managed DBaaS

The ability to deploy Db2 as a fully managed service for public, multi and hybrid cloud environments

Cornerstone of COLLECT in Journey to AI

Db2 is the cornerstone for COLLECT in Cloud Pak for Data to cover your end-to-end needs to leverage AI

Multi-Modal Data Store

Db2 is a multi-modal data store supporting native relational, XML, JSON, BSON and Graph data models

Data Virtualization

Db2 contains a data virtualization component which allows Db2 to be a doorway to all of your business data

In-Db2 Machine Learning

Allows data scientists and developers to bring machine learning local to the data stored within Db2

Db2 Leverages Machine Learning

Db2 optimizes for your workloads using a machine learning optimizer

NoSQL / NewSQL Data Store

Db2 supports SQL, Mongo API, XQuery and Gremlin as ways to access data in our multi-modal data store

Workload Dynamics

- Defining New & Modern Workloads
- Db2 and Mixed Workloads
- Db2 a Multi-Modal Database System
- Data Virtualization

New and Modern Workloads Definition

Different Kinds of Workloads – Where Db2 Plays

	OLTP	Mixed Workloads		Traditional Enterprise Data Warehouse (EDW)			
Workload	OLTP	Operational	Extreme Analytics	Operational Data Store	Operational Analytics	Analytics	Data Marts
Use	Transactions	Transactions with operational analytics	All types of analytic workloads and federated sources	Simple queries	Single-record look-up / IUD	Deep analytics – bulk scan	Computationally heavy / mining
Data Types	Traditional structured	Traditional structured + events + JSON + Graph + …	Traditional structured + events + JSON + Graph + …	Traditional structured	Traditional structured	Traditional structured	Traditional structured
Performance	1000s TPS	1000s TPS + 10s QPS	1000 QPS, many complex	100s QPS	10s-100s of S/IUD per second	100s QPS – 1000 QPS	Lower volume long running
Inserts	1000/s	1000/s	Continual Data Ingest	snapshots Seconds delay	Continual Data Ingest (Trickle feed) – OR - Batch - hourly/n-times daily/daily/etc		
Schema	Normalized	Normalized	De-normalized	Normalized	De-normalized	De-normalized	De-normalized
Applications	SQL	SQL + NoSQL + NewSQL	SQL + NoSQL + NewSQL	SQL	SQL	SQL	SQL

Different Kinds of Workloads – New Workloads

	OLTP	Mixed Workloads		Traditional Enterprise Data Warehouse (EDW)			
Workload	OLTP	Operational	Extreme Analytics	Operational Data Store	Operational Analytics	Analytics	Data Marts
Use	Transactions	**Transactions with operational analytics**	**All types of analytic workloads** and **federated sources**	Simple queries	Single-record look-up / IUD	Deep analytics – bulk scan	Computationally heavy / mining
Data Types	Traditional structured	**Traditional structured + events + JSON + Graph + …**	**Traditional structured + events + JSON + Graph + …**	Traditional structured	Traditional structured	Traditional structured	Traditional structured
Performance	1000s TPS	1000s TPS + 10s QPS	1000 QPS, many complex	100s QPS	10s-100s of S/IUD per second	100s QPS – 1000 QPS	Lower volume long running
Inserts	1000/s	1000/s	**Continual Data Ingest**	snapshots Seconds delay	Continual Data Ingest (Trickle feed) – OR - Batch - hourly/n-times daily/daily/etc		
Schema	Normalized	Normalized	De-normalized	Normalized	De-normalized	De-normalized	De-normalized
Applications	SQL	**SQL + NoSQL + NewSQL**	**SQL + NoSQL + NewSQL**	SQL	SQL	SQL	SQL

New and Modern Workloads
Mixed Workloads

ML Optimizer

ML based optimizer improves query performance using predictive estimations to improve query access plans.

The Db2 Machine Learning Optimizer provides an additional level of intelligent optimization to deliver query execution strategies that improve on traditional cost-based query optimization

Basic workload cost optimizers can suggest query execution strategies, but they aren't sensitive to recent changes in the database, and they can't learn from experience

The Db2 Machine Learning Optimizer, by contrast, incorporates feedback from actual query performance to recommend execution strategies that may deliver improved results

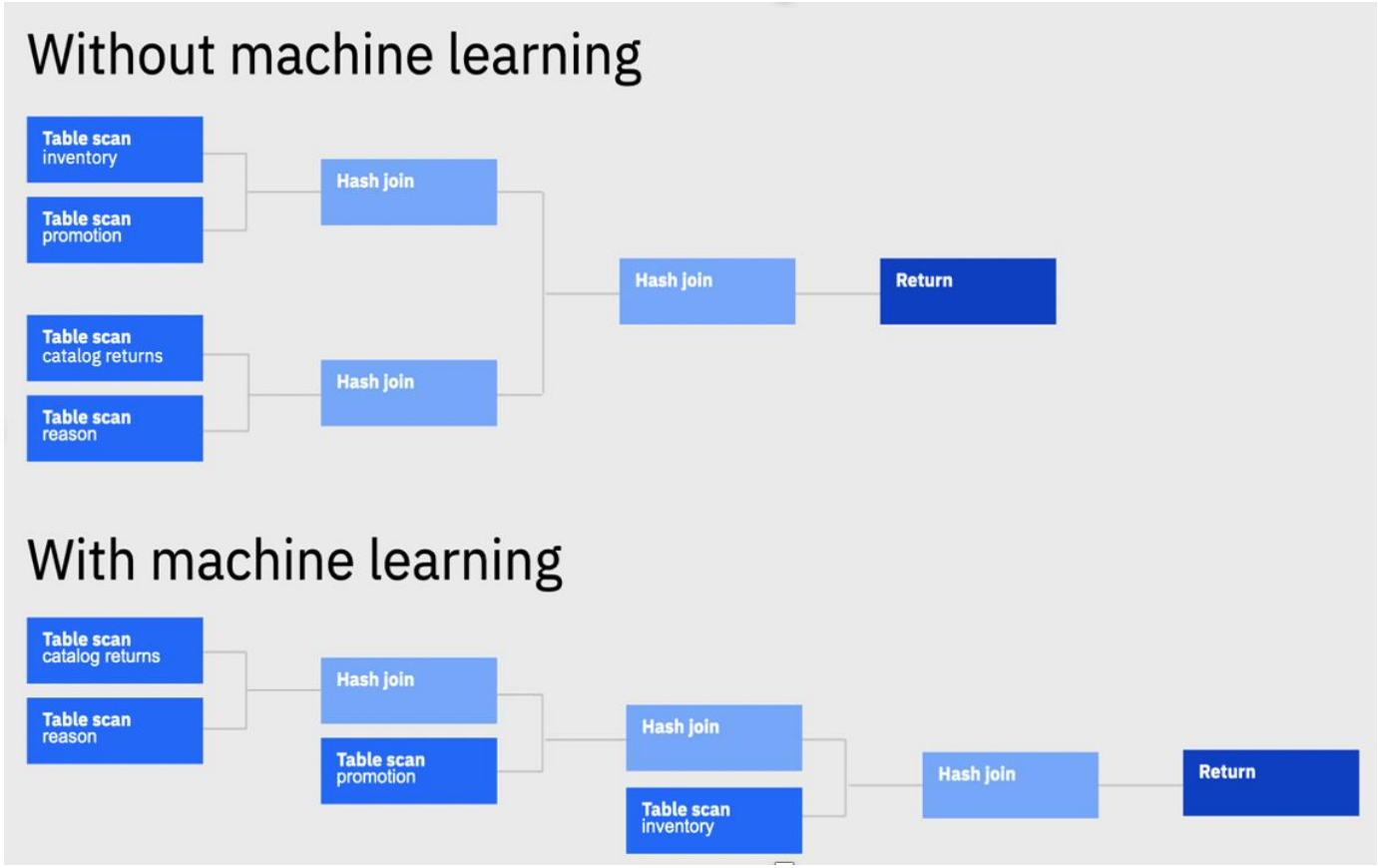
Up to **10x** fast query performance!

ML based Query Optimization

Get intelligent query processing with ML-based optimizer that uses Neural networks to learn from experience

- Uses ML to learn from experience and **improve decisions using feedback** from actual query performance, in line with the changes in the database
  - **Speed up** query and database execution with improved decision making and accuracy
- Applies ML to better **estimate predicate filtering**, which is key to better query performance
- **Confidence-based querying** adds machine learning extensions to SQL expressions using neural networks, to find similarities, patterns and matches
- **Optimizes** memory and compute capacity resources and database resources based on specific workload needs

Get up to 10x better query performance with Db2



ML Optimizer

Benefits for the end user

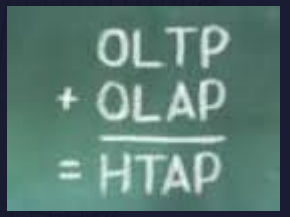
By dynamically improving query performance, the Db2 ML Optimizer helps to:

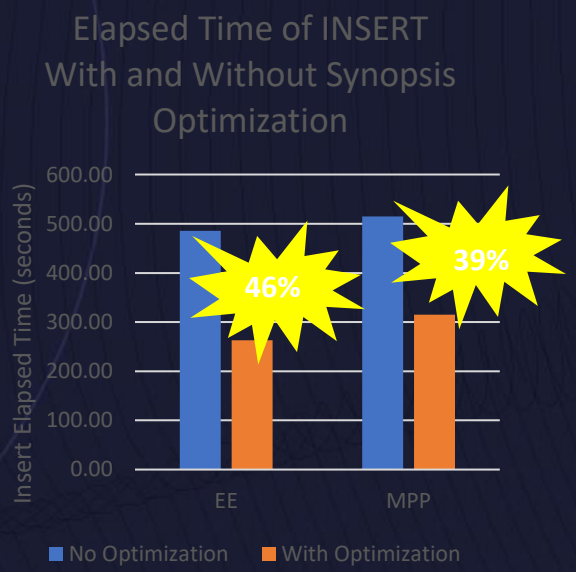
- **Speed-up e-commerce transactions** and improve customer satisfactions, especially during major sales events and holidays.
- **Improve customer experience in support calls,** prevent churn, and create better loyalty in order to protect and grow revenue.
- Do more with your existing infrastructure, **meet and exceed the SLA** expected by the business.
- Quickly **identify fraud patterns** and improve fraud detection.

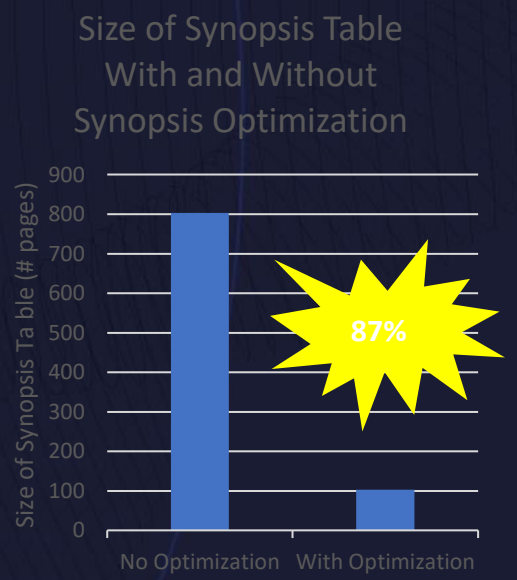
Get things done up to 10x faster!

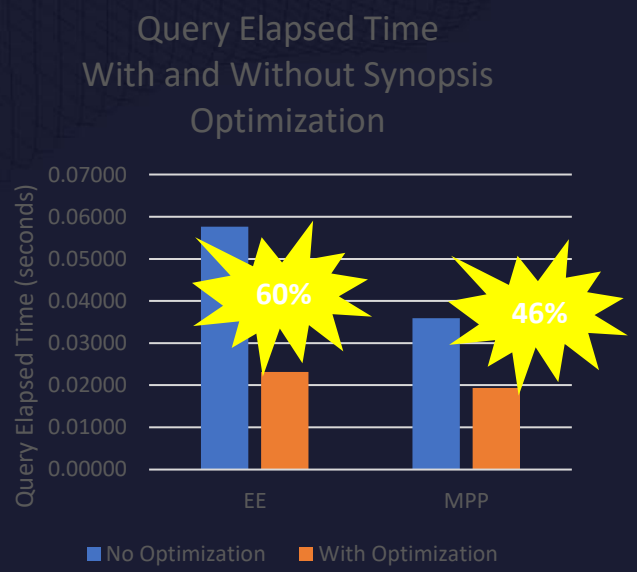
Performance: Synopsis Table Maintenance

- Buffer tuples in memory and write them to synopsis table every 1024 rows
  - Up to 46% improvement in insert performance!
  - Up to 87% reduction in synopsis table size!
  - Up to 60% reduction in elapsed time with queries with range predicates!









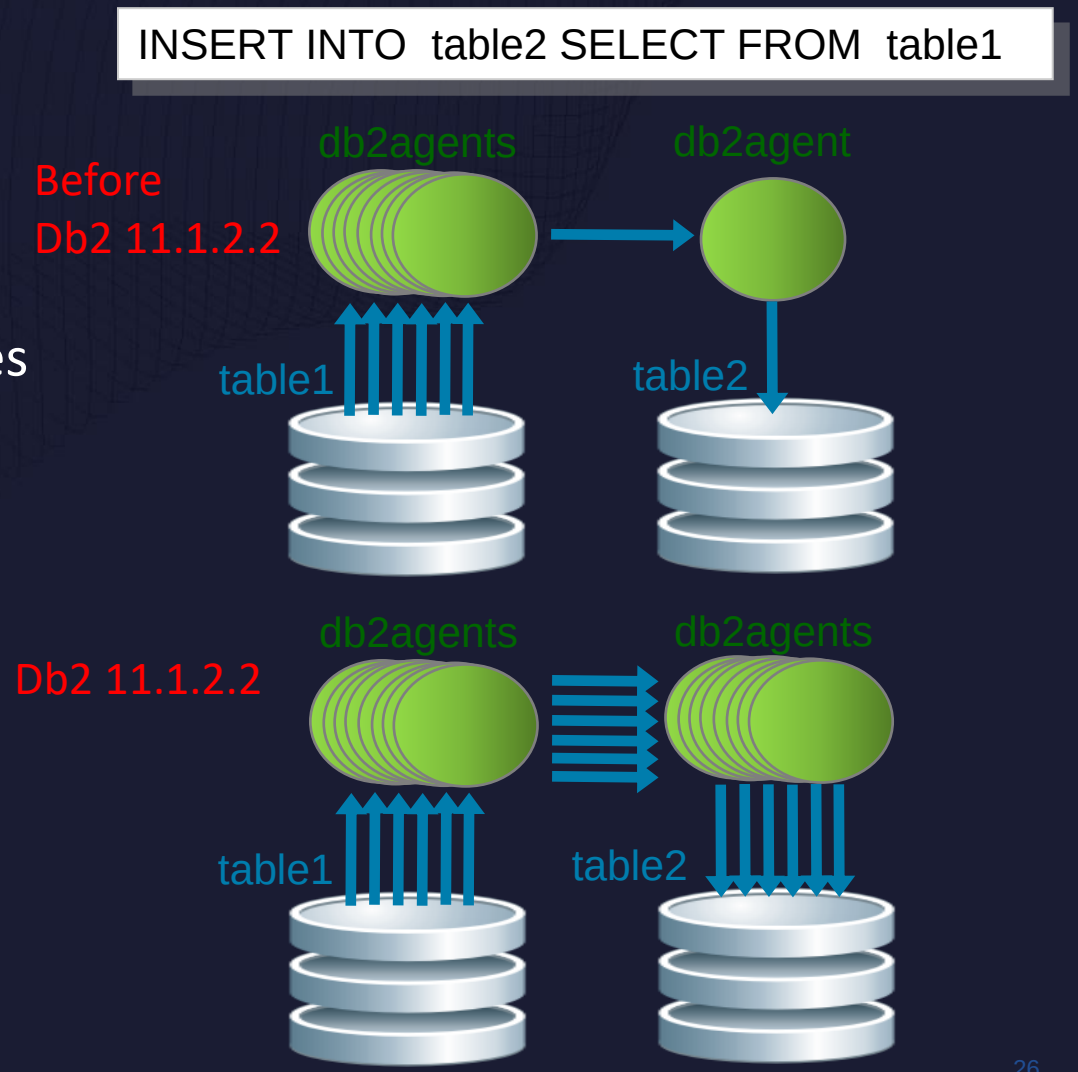
- IBM p760 / POWER7+ 32 cores / 1TB RAM
- Table with 50 columns.

- Single user
- Insert test 100K rows with commit count = 1
- Select query with 10 pairs of range predicates

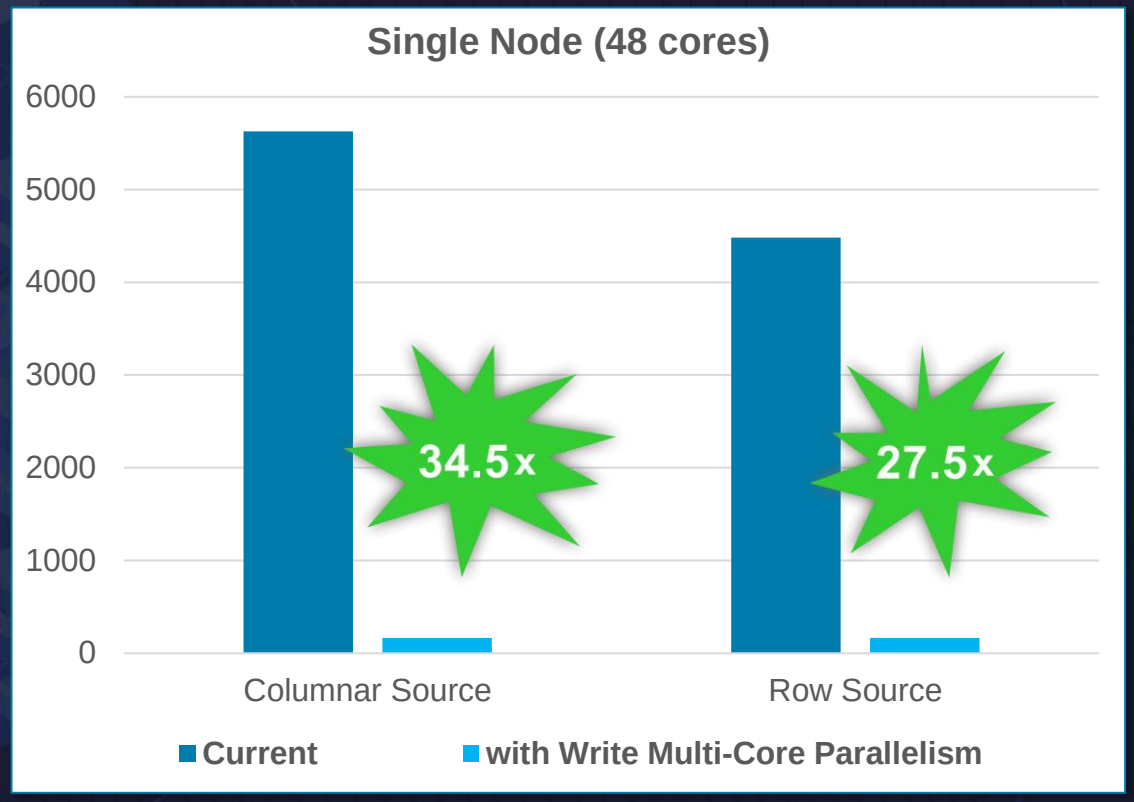
Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, 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.

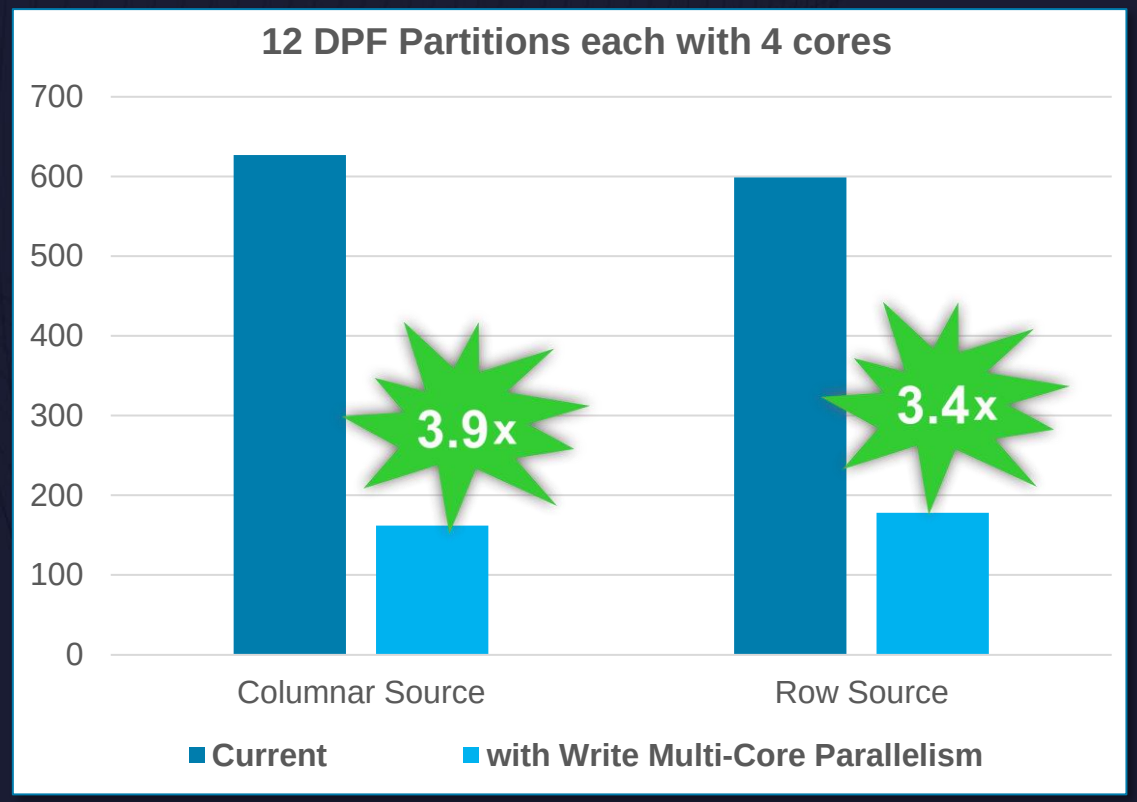
Multi-core parallelism – INSERT/UPDATE/DELETE

- Db2 has exceptional multi-core scalability for queries:
  - Combined MPP and SMP parallelism
  - Includes columnar and row-based tables
  - Applies to sub-select of INSERT
- New multi-core parallelism for INSERT on columnar tables
  - Includes a variety of data population methods (i.e., INSERT, INGEST, etc.)
  - Extends to index maintenance
  - Also applicable to MPP deployments
- Very significant reduction in time for:
  - ETL/ELT batch jobs
  - Data population/ingest jobs
  - Move/copy data from one table into another



Multi-core parallelism: Internal test results



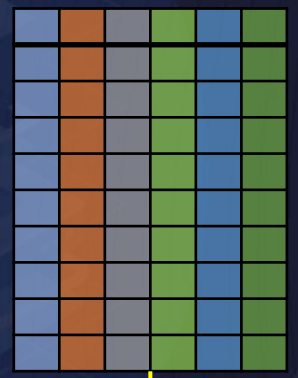


- 100 Million Rows
- 47 GB

- DPF tests used 12 logical DB partitions, & co-located tables
- Same 48 core server used for single-node and DPF tests

Trickle Feed INSERT: Baseline - Bulk INSERT

Bulk Insert



Column Vectors

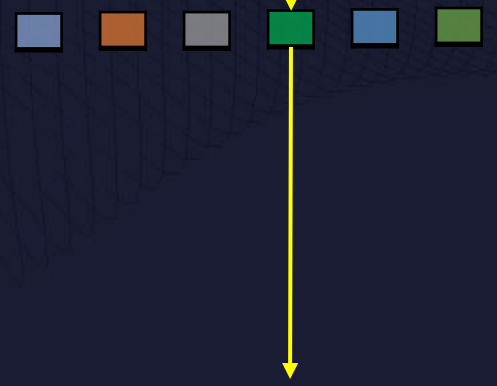
Column Insert Buffer Writer

Logging is done for each Insert Buffer when full

Less dirty pages to process in the bufferpool

Trickle Feed INSERT Prior To Db2 11.5.6

Insert row split



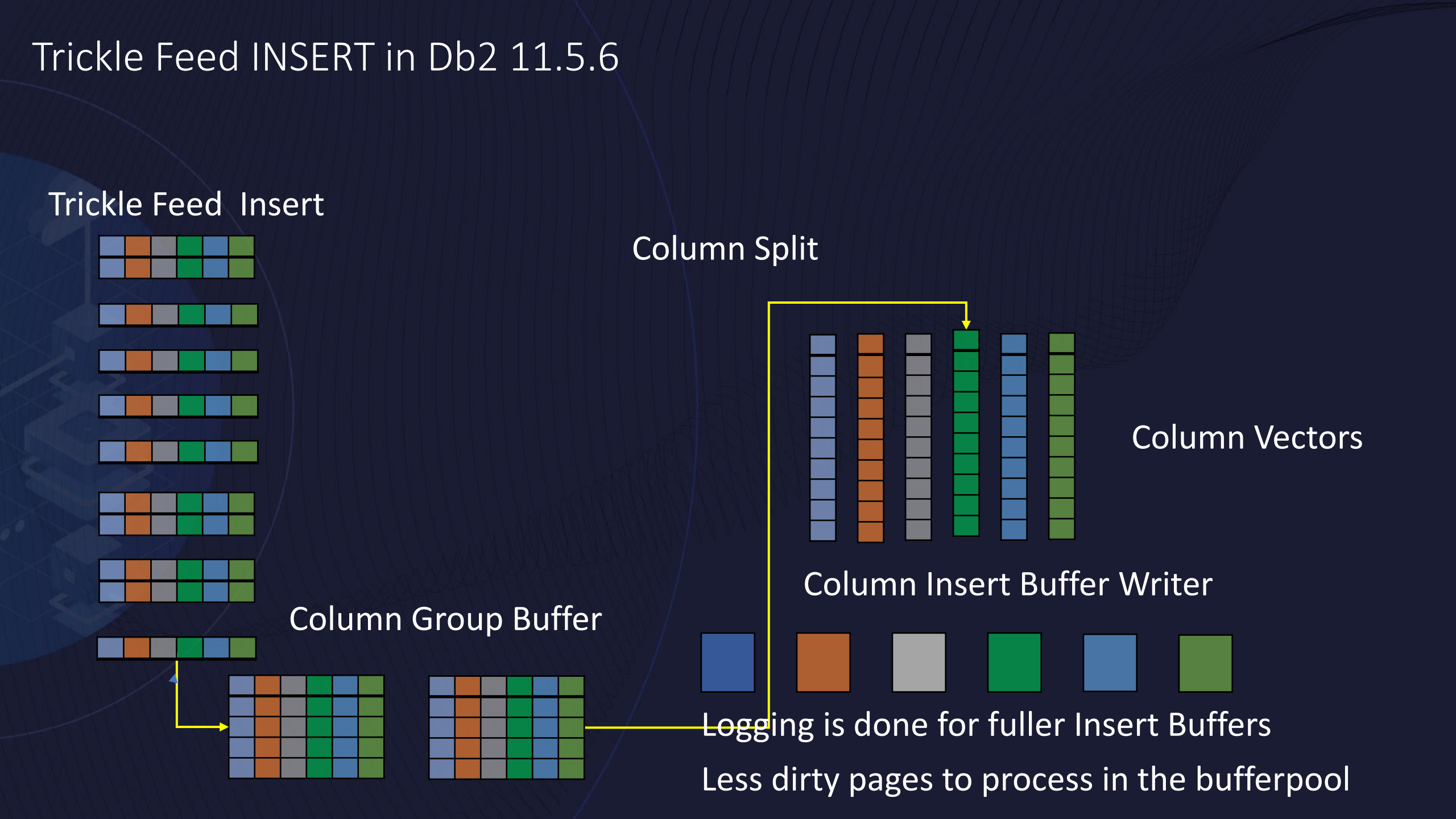
Column Insert Buffer Writer

Trickle Feed Insert

Logging done for each sparse Insert Buffer

More dirty pages to process in the bufferpool

Trickle Feed INSERT in Db2 11.5.6



Benefits of Trickle Feed Enhancements

- Reduced storage space for small tables
- Significant reduction in log space usage (~ 50%-75% with Integer columns)
- Reduction in bufferpool dirty pages
- Modest performance improvement
- Benefits to UPDATE with the INSERT portion

Columnar – Index Support

- DB2 11.1. currently implicitly creates unique indexes to support PRIMARY or UNIQUE KEY constraints
  - A unique index can be used to access data if at most 1 row qualifies and every key in the index must have an equality predicate OR the FETCH FIRST 1 ROW ONLY clause is specified

- As of Db2 11.1.3.3, Db2 now supports CREATE INDEX for column-organized tables:
  - Unique and non-unique indexes
  - DROP and ALTER also supported
  - ALTER INDEX only supports COMPRESS [YES|NO]

- These new indexes can be used to access column-organized data using most of the same methods as row-organized data and without restrictions on the number of qualifying rows

New and Modern Workloads
Multi-Model

Db2 Multi-Model Support – The Value

1

*Don't **migrate or duplicate** your data in another store*

- Leverage Db2 as a native data store for any supported model of data.
- Query your federated data sources as you would query your local Db2 tables

2

*Keep **Data consistent** and **Transactions fast***

- Inserts made through RDBMS and the data are reflected in appropriate model in real-time
- Db2 scales up to petabyte scale
- ACID Properties apply. All models of data have transactional awareness.

3

*Keep database **Secure and Scalable***

- Db2's sophisticated security and access controls and bitemporal support are available for all supported models of data.
- Choose between row-organized, column-organized, pureScale, DPF (MPP) flavors of Db2 – leverage HADR. All models are supported in all Db2 topologies.

4

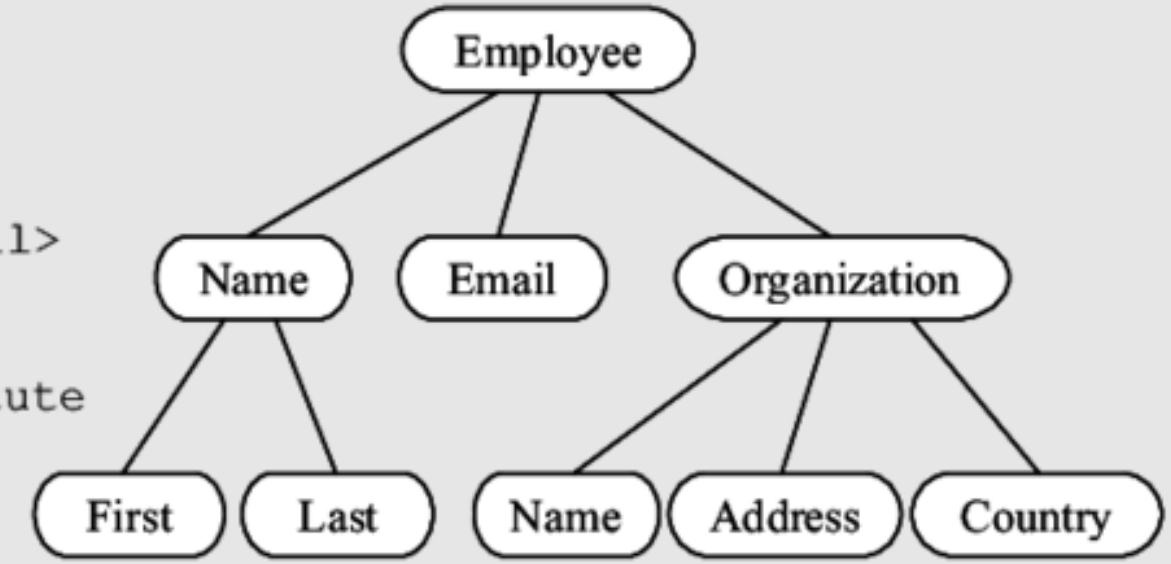
Developers can access multiple** models, all from the same **data stored once

- Developers see a SQL, NoSQL and NewSQL model for accessing any of the supported models in Db2.
- Users are able to visualize results sets in whatever model of data they choose.

New and Modern Workloads
XML

Db2 Supporting XML (pureXML)

```
<Employee>
    <Name>
        <First>Lassi</First>
        <Last>Lehto</Last>
    </Name>
    <Email>Lassi.Lehto@fgi.fi</Email>
    <Organization>
        <Name>
            Finnish Geodetic Institute
        </Name>
        <Address>
            PO Box 15,
            FIN-02431 Masala
        </Address>
        <Country CountryCode="358">Finland</Country>
    </Organization>
</Employee>
```



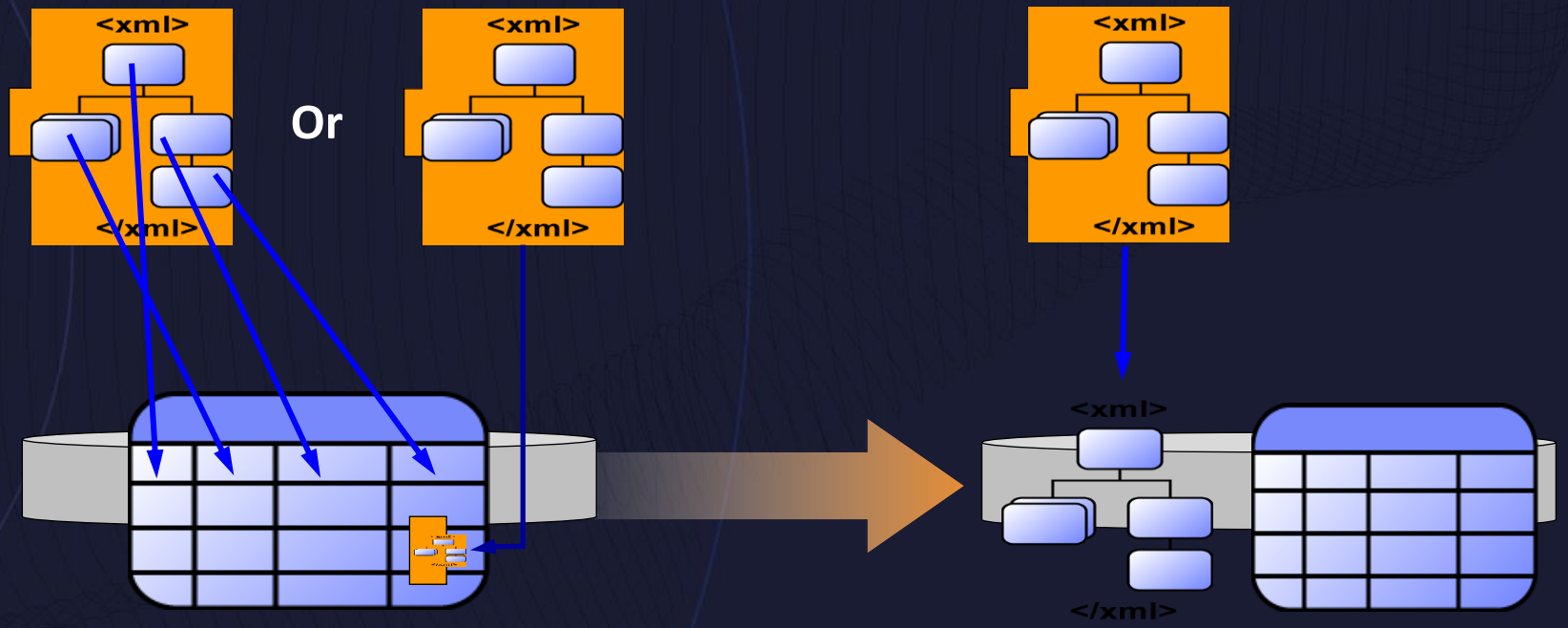
Db2 – XML Capabilities – Storage View

High cost development

Poor performance

Streamlined development

High performance

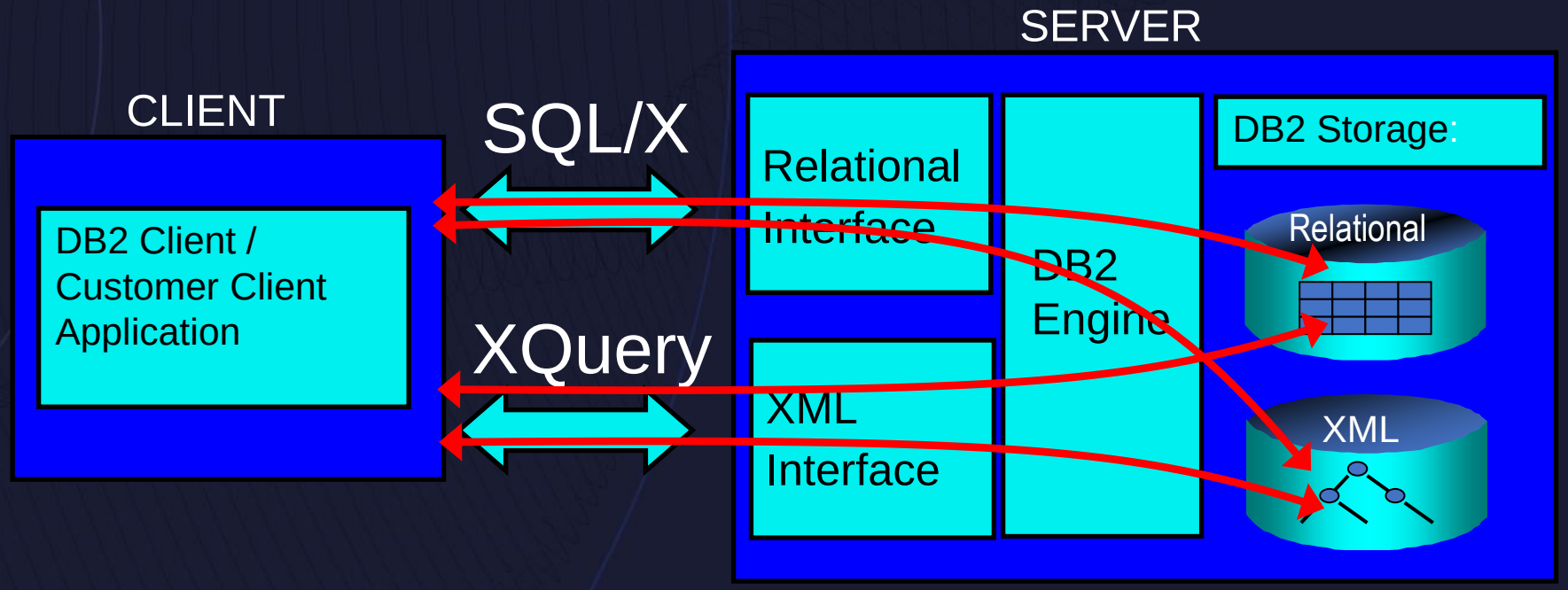


Business data in XML form
managed in relational database

Business data in XML form
managed with Db2 pureXML™

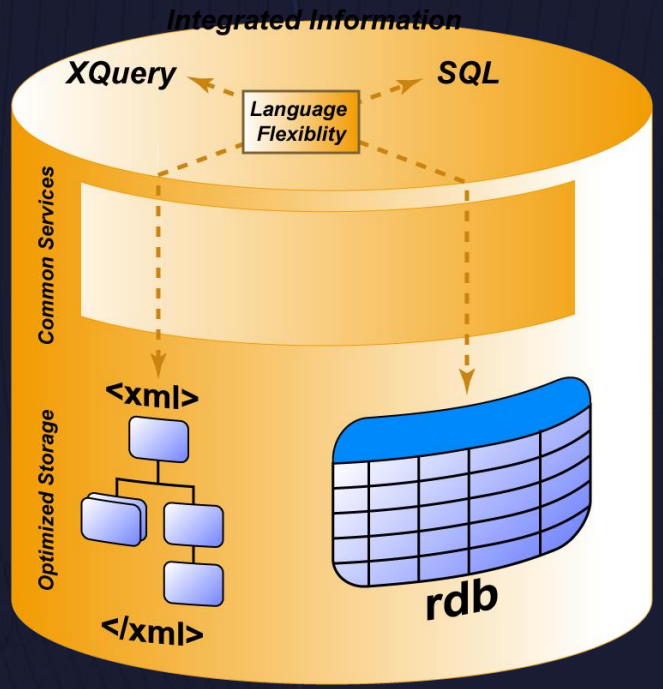
Integration of XML & Relational Capabilities

- **Native XML data type (server & client side)**
- **XML Capabilities in all Db2 components**
- **Applications combine XML & relational data**



Db2 – XML Capabilities – Retrieval View

SQL Person… "I see a world class RDBMS that also supports XML"



XML Person… "I see a world class XML repository that also supports SQL"

From a user and developer view – you have natively stored XML – accessible using XQuery

XML – From a SQL User perspective

- Data Definition

  create table dept(deptID int, deptdoc **xml**);

- Indexing

  Create index deptindex on dept(deptdoc) generate key using **xmlpattern '/Person/Dept' as sql double**;

- Insert

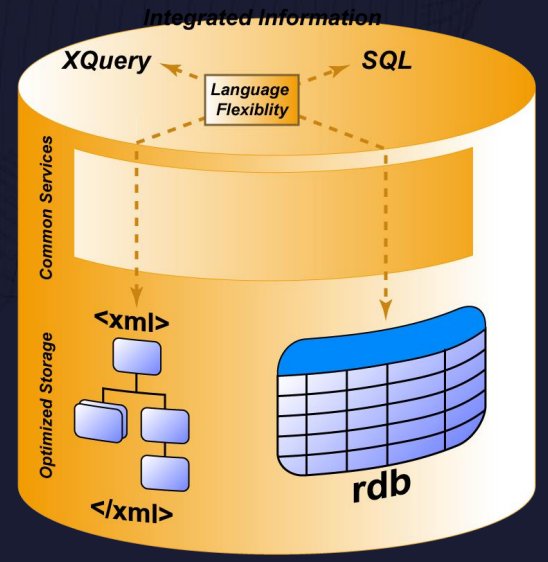
  insert into dept(deptID, deptdoc) values (?,?)

- Retrieve

  select deptID, deptdoc from dept

- Query

  select deptID, **xmlquery**('**$d/dept/name**' passing deptdoc as "d") from dept where deptID <> "PR27";



SQL/XML Functions

- Scalar functions
  - **XMLELEMENT** - generates an XML element
  - **XMLATTRIBUTES** - used within XMLELEMENT, specifies attributes
  - **XMLFOREST** - produces a forest of XML elements from SQL values
  - **XMLCONCAT** - concatenates a variable number of XML values
  - **XMLNAMESPACE** – produces a namespace declarations in an XML
- Aggregate function
  - **XMLAGG** - to group or aggregate XML data
- Cast functions
  - **XMLCAST** - converts between XML data type and standard relational types
  - **XML2CLOB** – converts XML data type to serialized XML as a char/varchar/clob
- Additional functions
  - **XMLPARSE** – parses character/BLOB data, produces XML value
  - **XMLVALIDATE** – validates XML value against an XML schema
  - **XMLEXISTS** - determines if an XQuery returns a results (i.e. a sequence of one or more items)
  - **XMLQUERY** - executes an XQuery and returns the result sequence
  - **XMLTABLE** - executes an XQuery, returns the result sequence as a relational table (if possible)
  - **XMLSERIALIZE** – converts XML data type to serialized XML as a char/varchar/clob/blob

XML – From an XML Developer Perspective – The FLWOR Expression

- FOR: iterates through a sequence, bind variable to items
- LET: binds a variable to a sequence
- WHERE: eliminates items of the iteration
- ORDER: reorders items of the iteration
- RETURN: constructs query results

```
FOR $movie in xmlcolumn('movies.doc')
LET $actors := $movie//actor
WHERE $movie/duration > 90
ORDER by $movie/@year
RETURN   <movie>
                {$movie/title, $actors}
              </movie>
```

```
<movie>
  <title>Chicago</title>
  <actor>Renee  Zellweger</actor>
  <actor>Richard  Gere</actor>
  <actor>Catherine  Zeta-Jones</actor>
</movie>
```

SQL/XML: Choice of output formats

```
SELECT
   XMLELEMENT(NAME "Department",
        XMLATTRIBUTES (e.dept AS "name" ),
        XMLAGG( XMLELEMENT(NAME "emp", e.firstname)  )
                                ) AS "dept_list"
FROM employee e
WHERE .....
GROUP BY  e.dept;
```

Start With

Produce

Available Functions:

- XMLELEMENT
- XMLATTRIBUTES
- XMLFOREST
- XMLCONCAT
- XMLAGG
- XML2CLOB
- XMLNAMESPACES
- XMLCAST

firstname	lastname	dept
SEAN	LEE	A00
MICHAEL	JOHNSON	B01
VINCENZO	BARELLI	A00
CHRISTINE	SMITH	A00

```
dept_list
```

```
<Department name="A00">
    <emp>CHRISTINE</emp>
    <emp>VINCENZO </emp>
    <emp>SEAN</emp>
</Department>
```

```
<Department name="B01">
    <emp>MICHAEL</emp>
</Department>
```

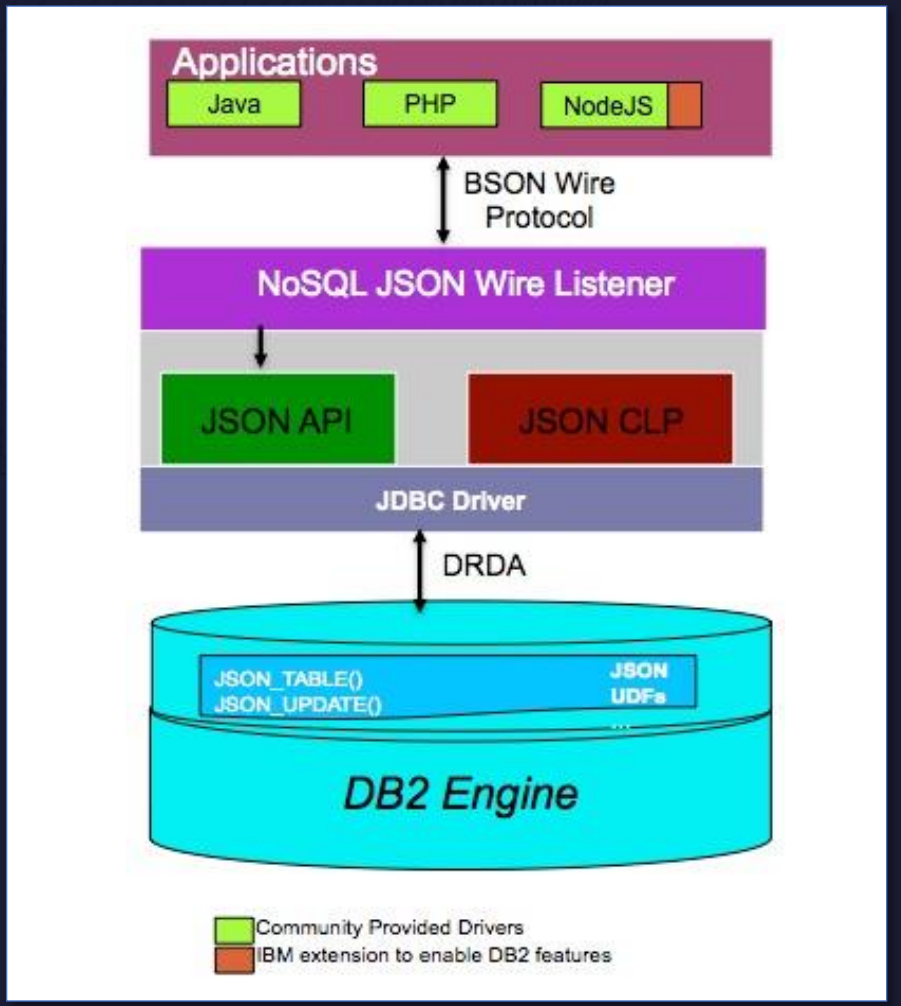
New and Modern Workloads JSON/BSON

Db2 – Support of JSON and BSON

- Db2 10.5 FP1 introduced JSON NoSQL support
  - Focused on allowing Db2 to participate in the NoSQL paradigm
  - Support of Mongo API through wire listener

- Our customers began to ask for native SQL support of JSON in Db2
  - Legacy applications accessing new JSON data
  - New JSON applications wanting to access legacy data

➢Support of INDEX on JSON through INDEX ON EXPRESSION to provide “tier 1” performance



Proprietary JSON SQL functions integrated with Db2 11.1.2.2

- To help simplify our customer's experience with the proprietary JSON SQL functions, we have now officially included them as part of Db2

➢ The functions are automatically created in SYSTOOLS schema for a new database or added to an existing one when updated (via db2updv111)

➢ The functions are now documented in the Db2 knowledge center under a section called "SQL access to JSON documents"

Insert, Retrieve, and Document Validation

- BSON2JSON – Convert BSON into JSON
- JSON2BSON – Convert JSON into BSON
- BSON_VALIDATE – Ensure LOB is valid BSON

Field Retrieval

- JSON_VAL – Extract data from JSON into SQL data types
- JSON_TYPE – Returns data type of a JSON field

Array Retrieval

- JSON_TABLE – Returns a table of values from a JSON array
- JSON_LEN – Returns #elements in a JSON array
- JSON_GET_POS_ARR_INDEX – Find a value within an array

Document Update

- JSON_UPDATE – Update a field or document using set syntax

New built-in JSON SQL functions

- New implementations of JSON SQL functions based on recent ISO report on SQL support for JavaScript Object Notation (JSON)

- The (proprietary) SYSTOOLS functions will be de-emphasized but will continue to be supported

Schema	Name	Comments
SYSIBM	BSON_TO_JSON	Convert BSON formatted document into JSON strings
SYSIBM	JSON_TO_BSON	Convert JSON strings into a BSON document format
SYSIBM	JSON_ARRAY	Creates JSON array from input key value pairs
SYSIBM	JSON_OBJECT	Creates JSON object from input key value pairs
SYSIBM	JSON_VALUE	Extract an SQL scalar value from a JSON object
SYSIBM	JSON_QUERY	Extract a JSON object from a JSON object
SYSIBM	JSON_TABLE	Creates relational output from a JSON object
SYSIBM	JSON_EXISTS	Determine whether a JSON object contains the desired JSON value

New JSON NoSQL Wire Listener Update

- Enhance IBM NoSQL JSON Driver and IBM NoSQL Wire Listener APIs to MongoDB version 3.6.3 for better adoption
  - Mongo client of v4.x works with wire listener now
- Kerberos Authentication support for IBM NoSQL JSON Wire Listener
- Logging Enhancement for IBM NoSQL JSON Wire Listener

JSON & BSON Support in Db2 (V11.5.5)

- Natively store JSON or BSON data in Db2
- Get top tier performance with INDEX support
- Under the covers we leverage:
  - VARCHAR / VARBINARY for smaller documents
  - CLOBS / BLOBS for larger documents
- Full ACID and enterprise level availability, scalability, performance, security and recoverability available
- SQL Support
  - Proprietary UDFs with many JSON/BSON functions
  - SQL:2016 Standards (report) – 70% complete
- NoSQL Support
  - Mongo API wire listener
- Data Ingestion
  - LOAD, IMPORT, INSERT, etc

Columnar: Compact Varchar (11.5.4)

- Memory optimized efficient handling of VARCHARs in Columnar **Vectors and Workunits**
  - Store actual width, pad free in Vectors and Workunits
- Primary target : VARCHAR datatype
- Reduce memory consumption, spill I/O and OOMs primary impact focus
- Improved stride size and concurrency within query
- Performance improvement in some query is secondary benefit
- Where Applied:
  1. Columnar Vectors and Workunits
  2. Columnar Sort (both key and payload vectors)
  3. Columnar OLAP
  4. Columnar Group By payload vectors only
- But not in Group by and Join keys and payload

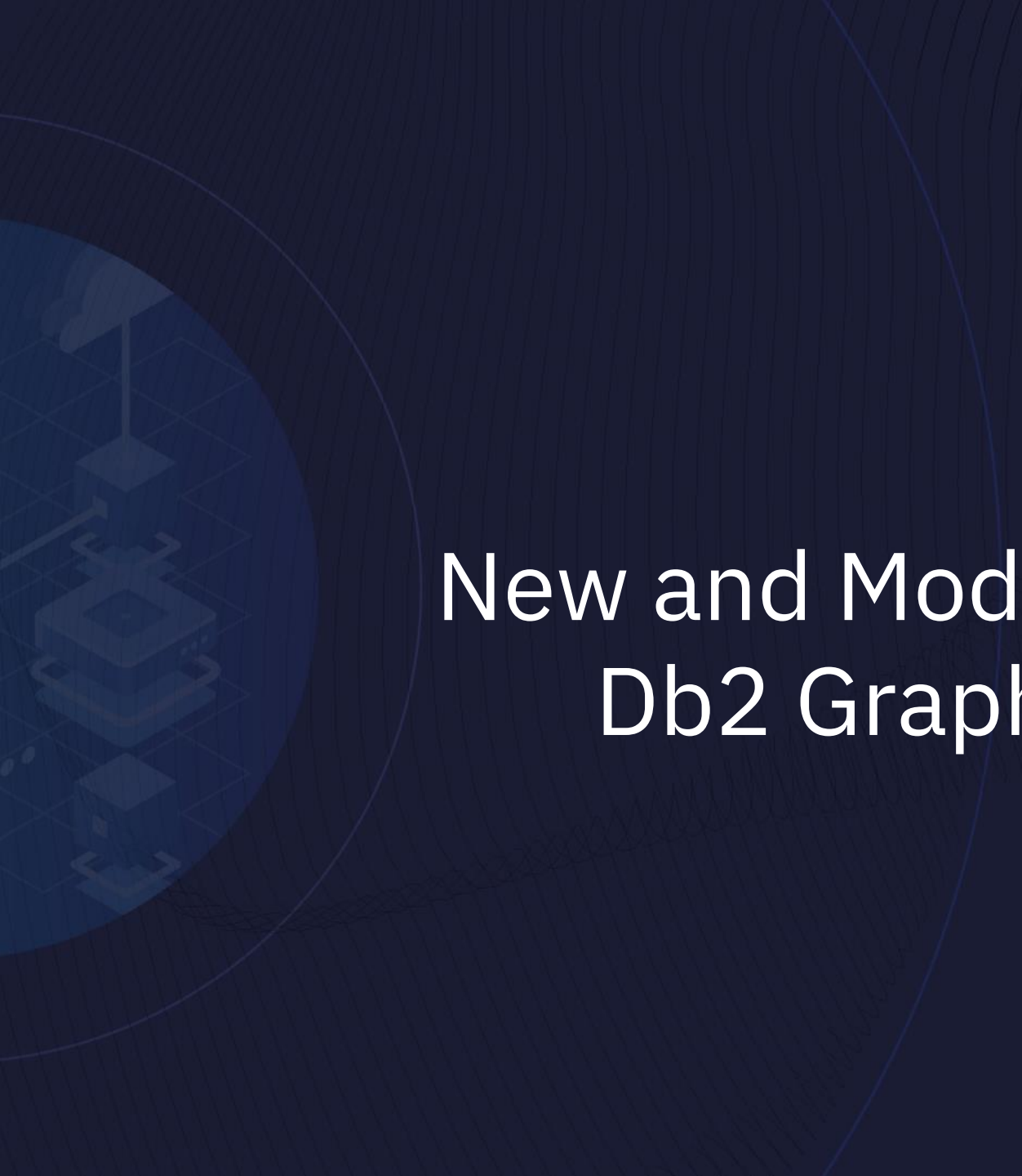
Columnar: Compact Varchar (11.5.5)

Impact

- Improved memory efficiency for wide VARCHARs in Columnar **Group By** and **Join**
- Reduce memory consumption, spill I/O and Out of Memory errors
- Performance improvements
- Increase in concurrency within Group By and Join operator

Results (from internal PQA workload)

- Overall workload elapse time, memory footprint and spilling greatly improved
  - Performance: Up to **2.9X overall** workload, **17.6X individual** query
  - Memory reduction: Up to **1.1X overall** workload, **2.5X** individual query
  - Spilling reduction: Up to **5.6X overall** workload**, >1200X** individual query

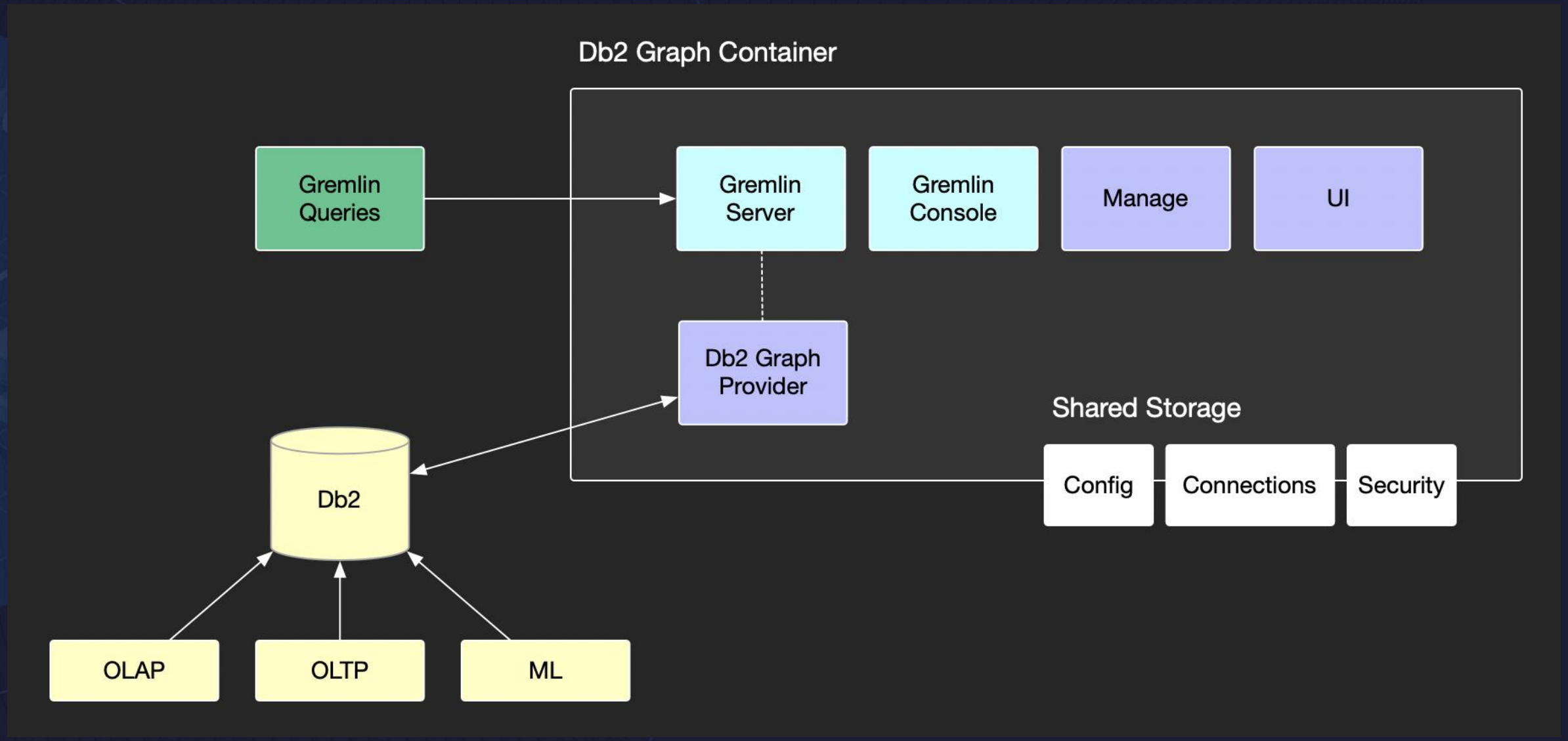


New and Modern Workloads
Db2 Graph Database

What is Db2 Graph?

- Based on Apache TinkerPop
  - TinkerPop is an open-source graph framework
  - Gremlin is the graph query language of TinkerPop
  - Db2 Graph is a provider plugin for TinkerPop
- Includes the TinkerPop Gremlin Server and Console
- Provides a model builder and visualization of query results
- Available as a standalone container on IBM Cloud Container Registry and within Db2 on CP4D 4.0

Db2 Graph Components



How does Db2 Graph work?

- Defines a virtual graph model on top of Db2 tables
  - Uses referential information to automatically create a graph schema
  - Maps tables or views to graph vertices and edges
- Users create the graph model using the Db2 Graph user interface or with the container's manage command
- Db2 Graph uses the model to convert Gremlin queries to SQL
- Db2 remains untouched. No change in data, structure or performance.
- Existing applications are uninterrupted.
- Data is fetched by Db2 Graph at time of execution. Data consistency and updates are reflected in real-time.

Db2 Graph - Mapping relational data to a graph model

- Persists the identification of vertices and edges in a json model.
  - Vertices represent tables with a primary key
  - Edges represent relationships between tables identified by a foreign key
- Views are supported, but not included in the auto-generation and must be manually added.

Patient Table

patientID	name	address	subscriptionID
1	Alice	…	115
…			…

HasDisease Table

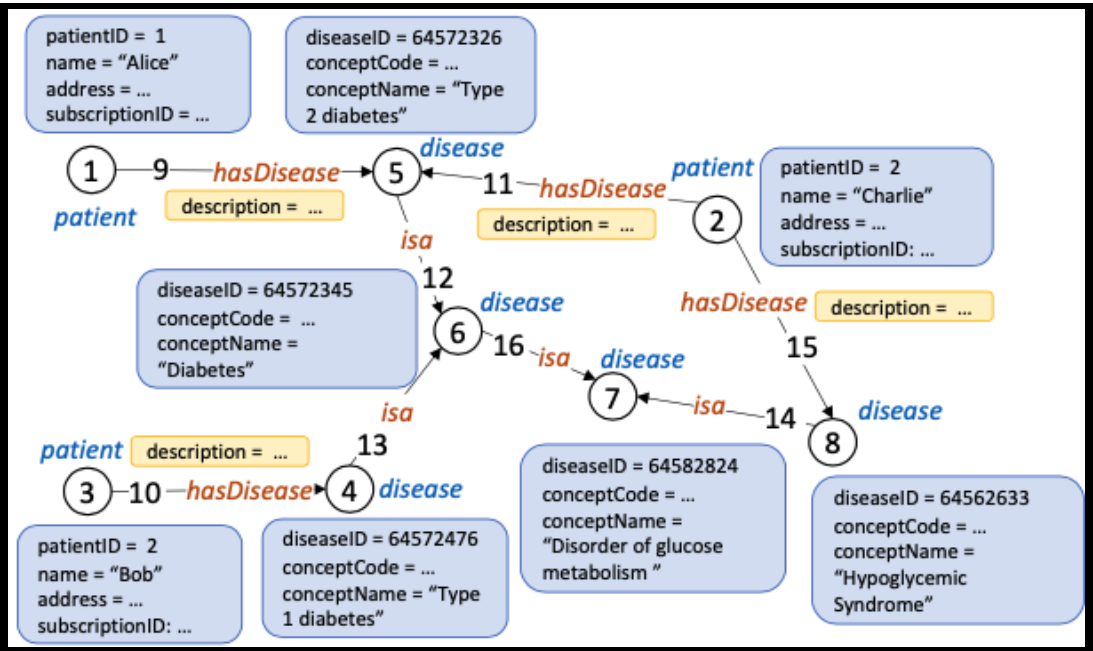
patientID	diseaseID	description
1	64572326	…
…	…	…

Disease Table

diseaseID	conceptCode	conceptName
64572326	44054006	"Type 2 diabetes"
…	…	

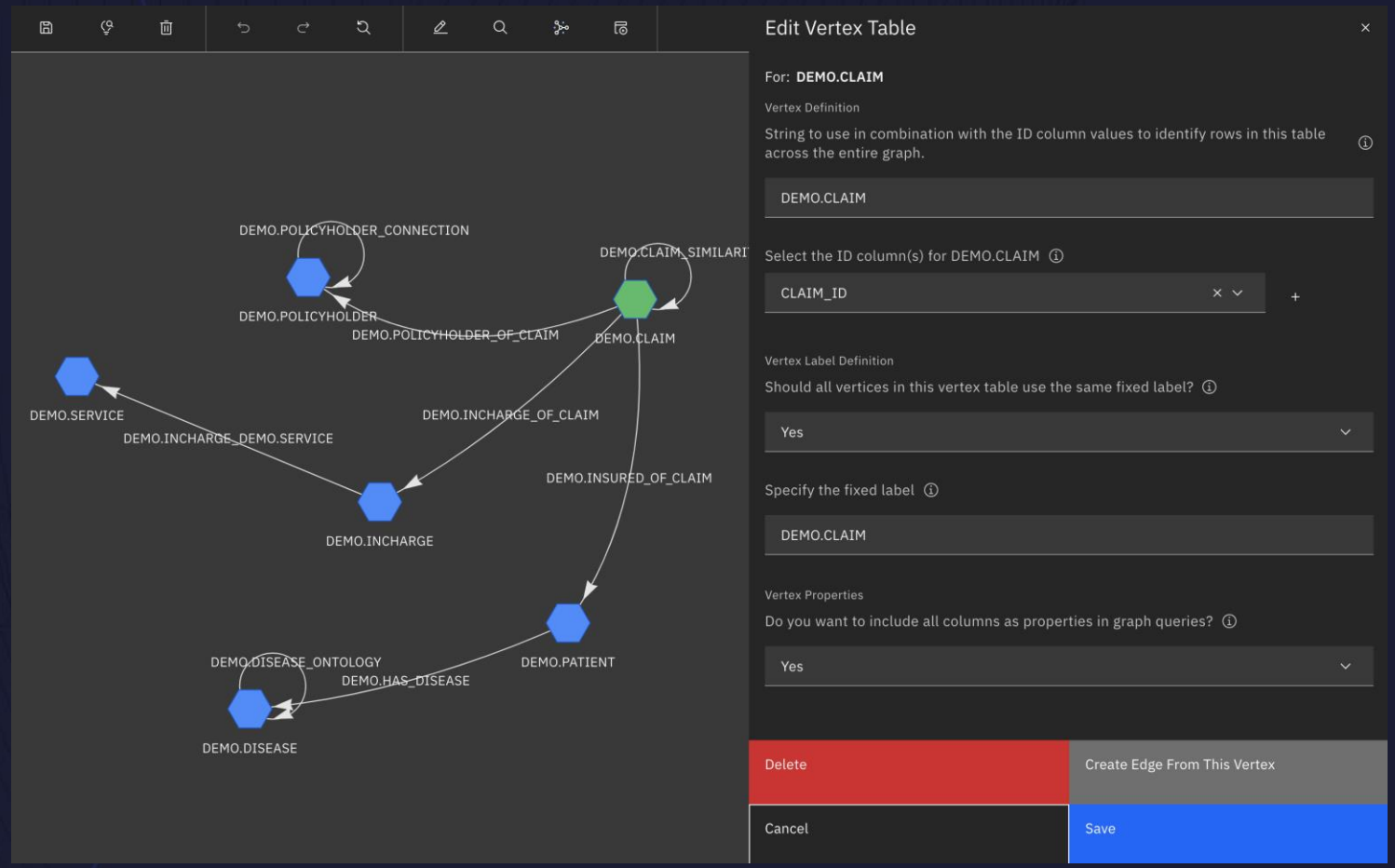
DiseaseOntology Table

sourceID	targetID	type
64572326	73211009	"isa"
…	…	



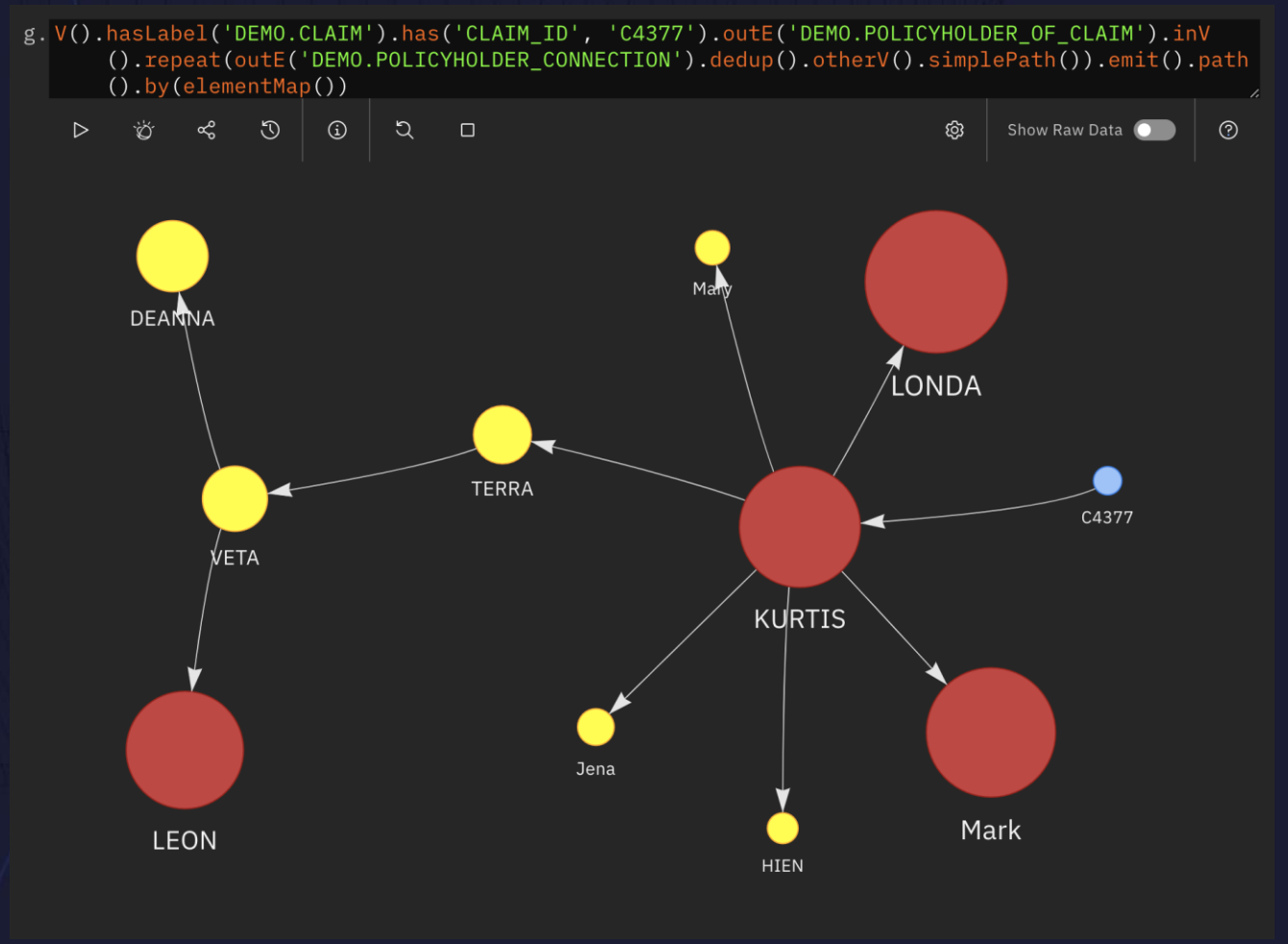
Db2 Graph - Creating a graph model

- Using the Db2 Graph user interface, you can:
  - easily auto-generate a model by dragging and dropping schemas (or tables)
  - add tables without primary keys
  - add views
  - fully customize how vertices and edges are represented



Db2 Graph - Visualizing query results

- Execute queries directly in the user interface with:
  - query syntax highlighting
  - interactive visualizations
  - customization of labels, colors and sizes using visual insights that help identify patterns
  - re-runnable query history



Db2 Graph – Performance

- LinkBench graph & queries
  - 10 million node and 100 million node sample graphs
  - 32 core CPU, 256GB memory
  - Db2 and Db2 Graph running on same machine

LinkBench Query	Gremlin
getNode(id, lbl)	g.V(id).hasLabel(lbl)
countLinks(id1, lbl)	g.V(id1).outE(lbl).count()
getLink(id1, lbl, id2)	g.V(id1).outE(lbl).filter(outV().id() == id2)
getLinkList(id1, lbl)	g.V(id1).outE(lbl)

Real-time access to data, no export required to run queries with Db2 Graph

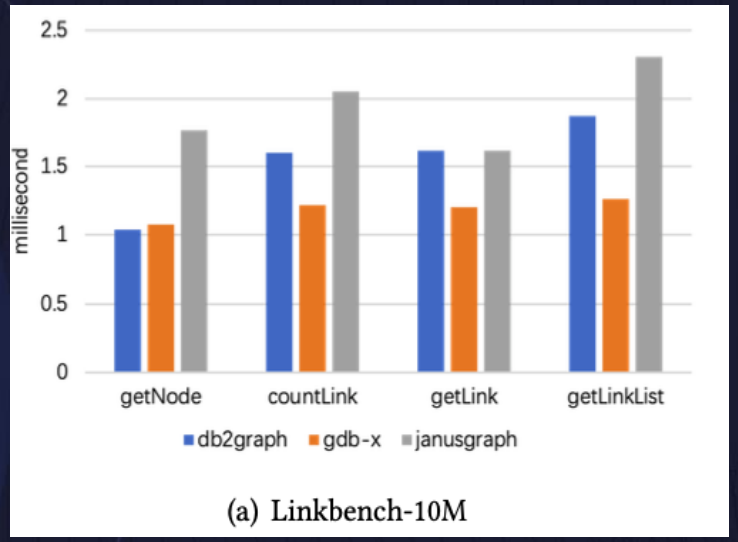
Opening the graph takes ~9x longer, because of aggressive caching or complex data formats

	Db2 Graph		Export	GDB-X			JanusGraph		
Linkbench Dataset	Disk Usage	Open Graph	From DB	Disk Usage	Load Data	Open Graph	Disk Usage	Load Data	Open Graph
10M	4.6GB	1.4 sec	5 min	28GB	42 min	14 sec	29GB	65 min	15 sec
100M	45.8GB	2.1 sec	32 min	327GB	8 hr	15 sec	326GB	13.5 hr	17 sec

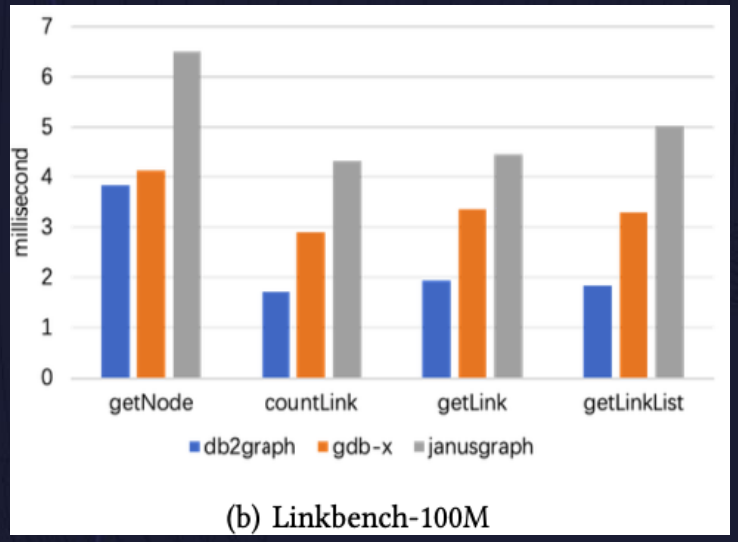
Need 6-7x more space to store data in the format

Db2 Graph - Performance

Latency – lower is better



(a) Linkbench-10M

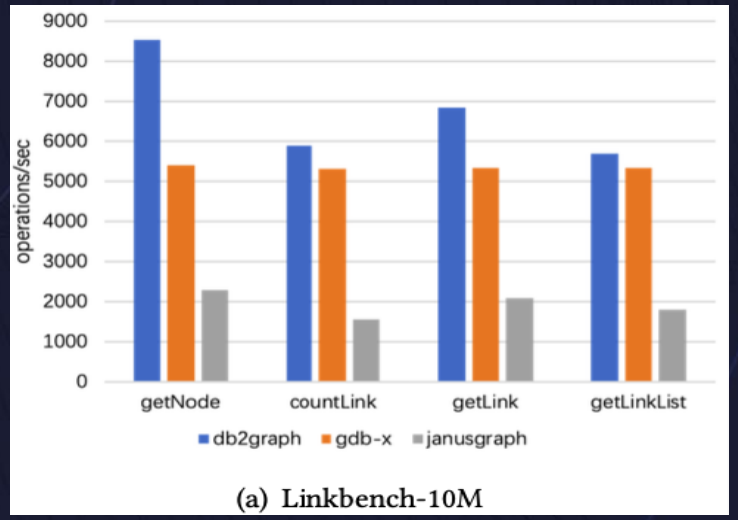


(b) Linkbench-100M

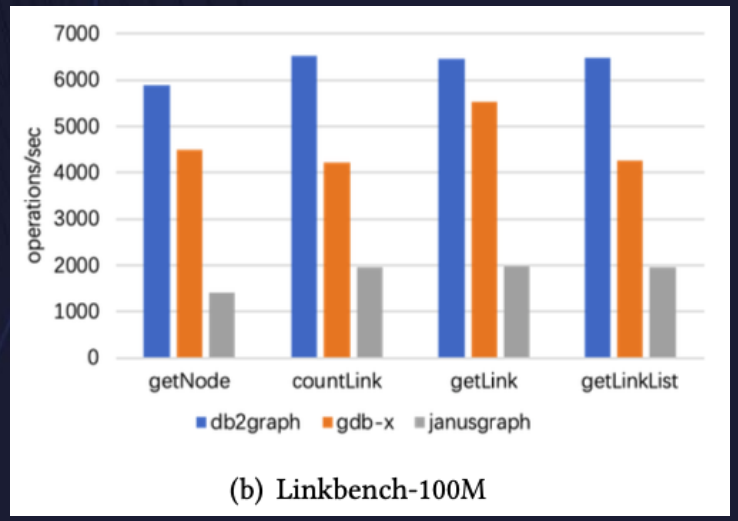
gdb-x performs well for smaller graphs because of their caching mechanism

gdb-x does not perform well when the size increase. For Db2 Graph, the data fits in Db2's buffer pool

Throughput – higher is better



(a) Linkbench-10M



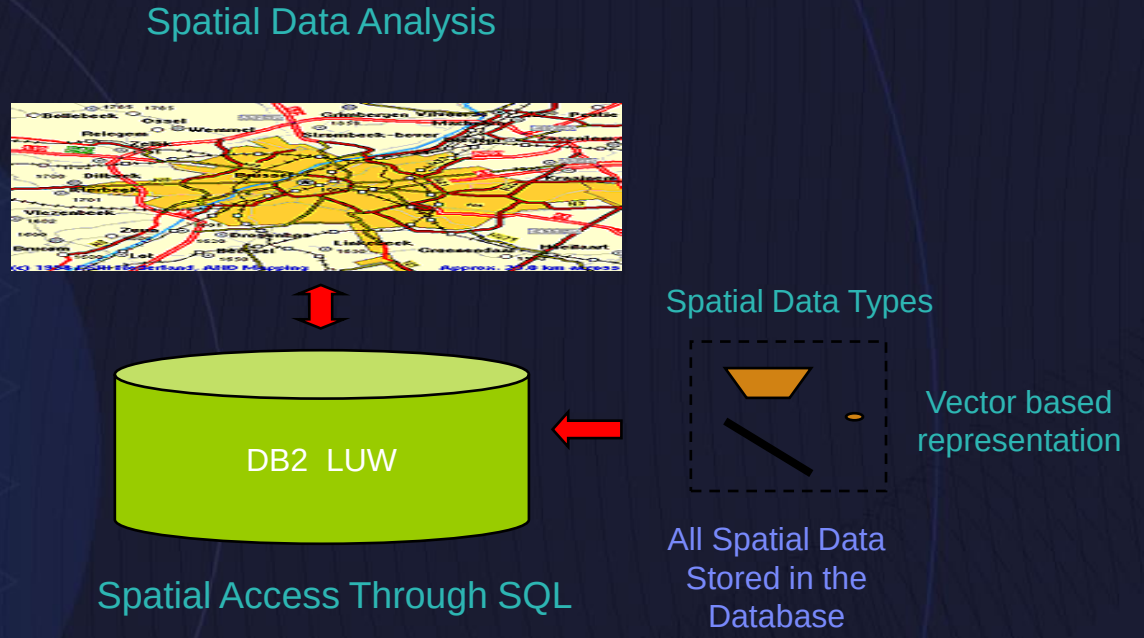
(b) Linkbench-100M

Db2 Graph is the clear winner, the underlying Db2 engine is good at handling concurrent queries.

New and Modern Workloads
Spatial Analytics

Spatial Basics

On-Premise (single-node, MPP, pureScale)
Db2 WHoC (public, dedicated, local), IIAS



- Implements
  - **Open Geospatial Consortium (OGC) Simple Features for SQL using Types and Functions**
  - **ISO SQL/MM part 3: spatial**
  - **Geography Markup Language (GML) for geometries**
  - **Well-known Text/Binary (WKT/WKB) for geometries**
  - **Shapefile import/export (de-facto)**
  - **WKT for coordinate systems (>5000 predefined)**

```
SELECT a.road_id, a.time, i.id, ST_Distance(a.loc, i.loc,'METER') as distance
FROM accidents a, intersections i
WHERE ST_Distance(a.loc,i.loc,'METER') < 10000 AND a.weather = 'RAIN'

SELECT a.name, a.type
FROM highways a, floodzones b
WHERE ST_Intersects(a.location,b.location) = 1 AND b.last_flood > 1980
```

Spatial Components

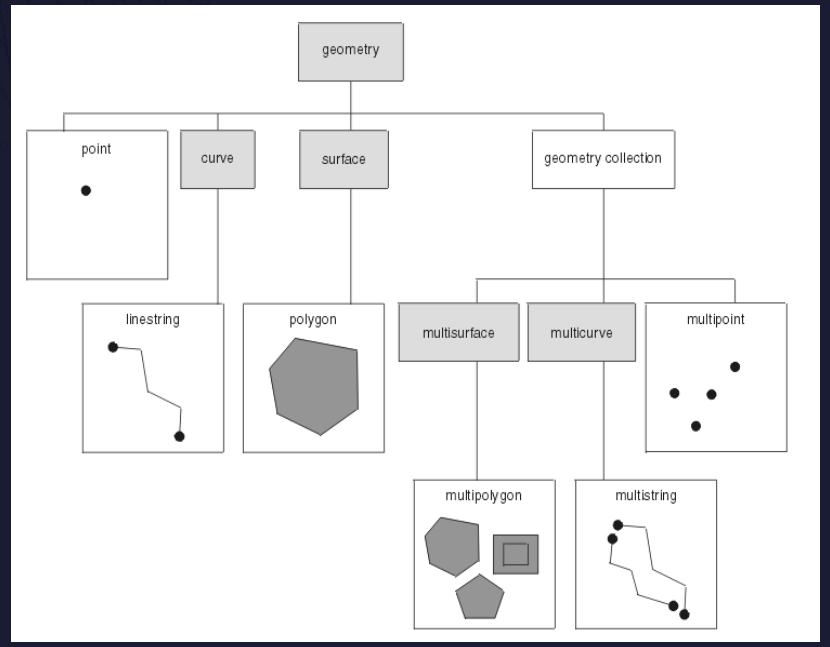
- Spatial Extender (DB2GSE) and Spatial Analytics (SYSGEO)
- ~350* functions/ methods
- Spatial Metadata
  - catalog tables and views
- Data Types
  - Hierarchical geometries
  - SQL Types (ST_Geometry)
  - Geometries stored as BLOBs

ST_Area
ST_AsBinary
ST_AsText
ST_Boundary
ST_Buffer
ST_Centroid
ST_Contains
ST_ConvexHull
ST_CoordDim
ST_Crosses
ST_Difference
ST_Dimension
ST_Disjoint
ST_Distance
ST_Endpoint
ST_Envelope
ST_Equals
ST_ExteriorRing
ST_GeomFromWKB
ST_GeometryFromText
ST_GeometryN
ST_GeometryType
ST_InteriorRingN
ST_Intersection
ST_Intersects
ST_IsClosed
ST_IsEmpty
ST_IsRing
ST_IsSimple
ST_IsValid
ST_Length
ST_LineFromText
ST_LineFromWKB
ST_MLineFromText
ST_MLineFromWKB
ST_MPointFromText
ST_MPointFromWKB
ST_MPolyFromText
ST_MPolyFromWKB
ST_NumGeometries
ST_NumInteriorRing
ST_NumPoints
ST_OrderingEquals
ST_Overlaps
ST_Perimeter
ST_Point
ST_PointFromText
ST_PointFromWKB
ST_PointN
ST_PointOnSurface
ST_PolyFromText
ST_PolyFromWKB
ST_Polygon
ST_Relate
ST_SRID
ST_StartPoint
ST_SymmetricDiff
ST_Touches
ST_Transform
ST_Union
ST_WKBToSQL
ST_WKTToSQL
ST_Within
ST_X
ST_Y

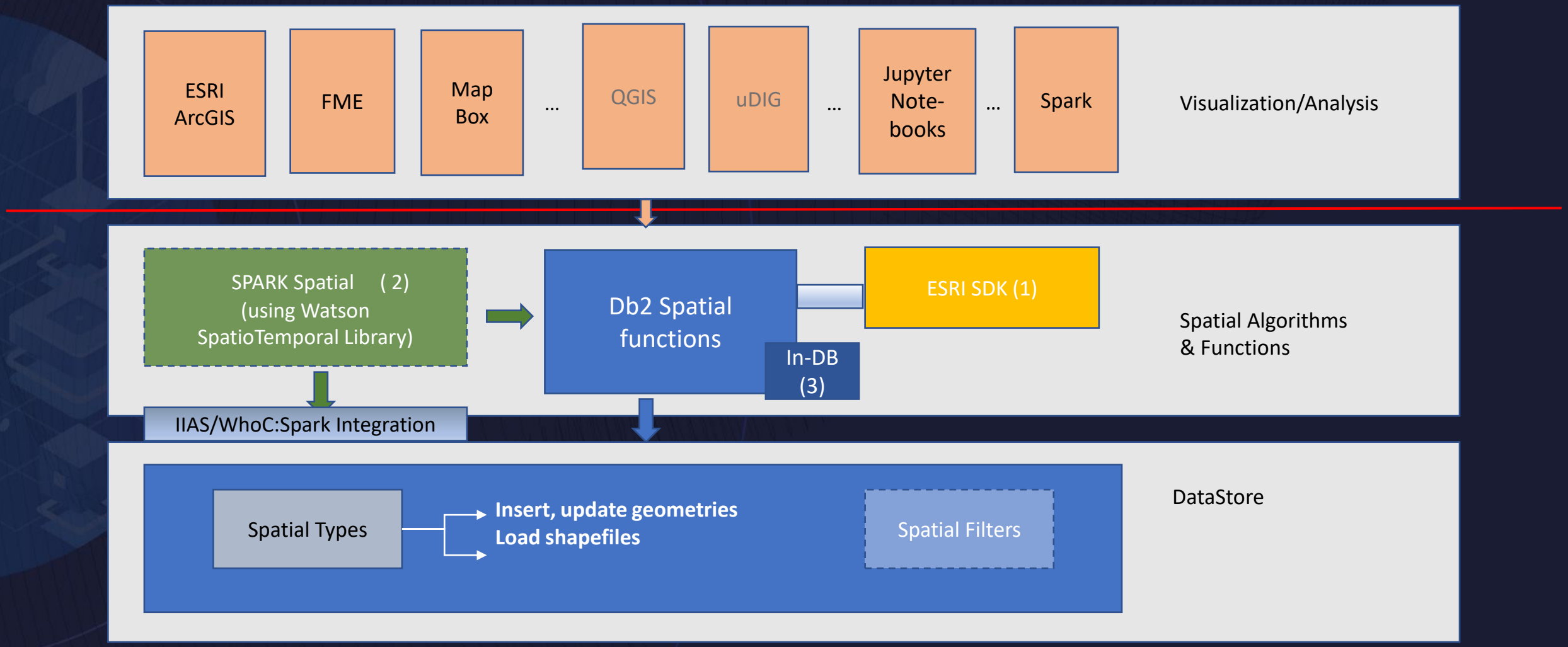
And more…

Simplified Constructors from
- x,y
- WKT
- WKB
- GML
- shape

Linear referencing
Spatial aggregation
ST_AsGML
ST_AsShape



Spatial Layers



Db2 11.5.6 - Spatial Analytics Summary

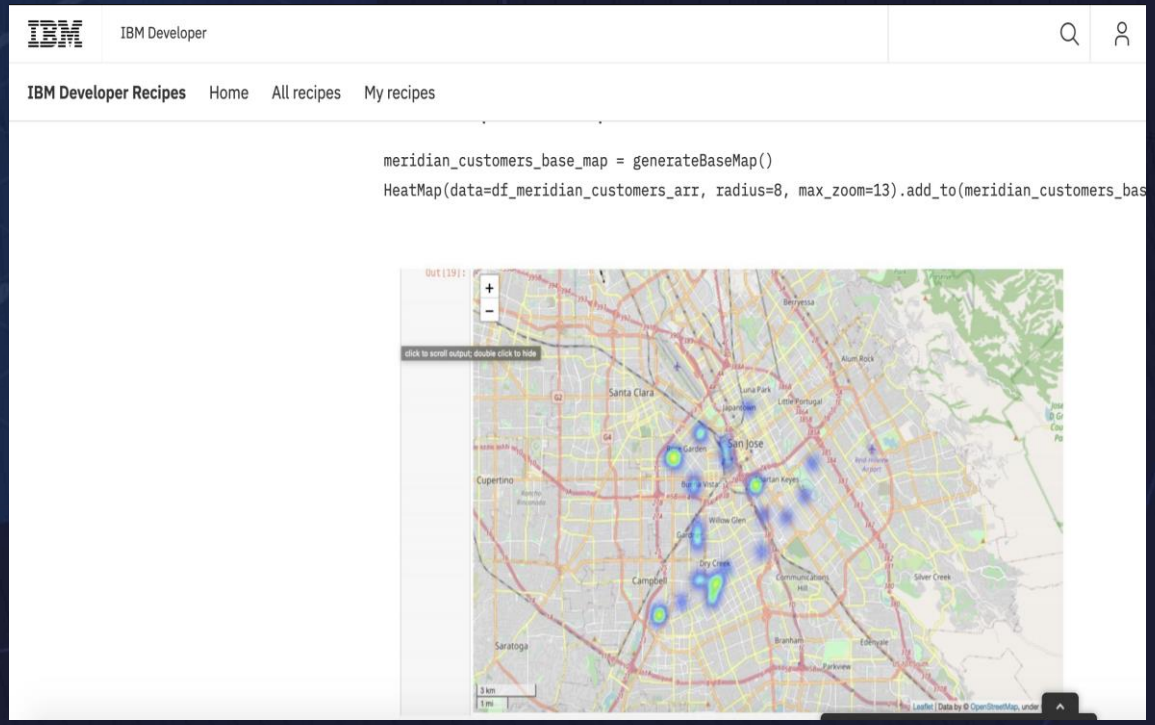
Spatial Processing in Db2	Spatial Extender	Spatial Analytics	Spatio-Temporal Toolkit
Processing Method	In-Database	In-Database	Spark distributed processing using data frames
Data Organization	Row-Store	Column-Store Row-Store	n/a
Index Type	Spatial Grid	N/A	Geohash
Spatial Joins	Yes	Yes	Yes
Spatio-Temporal Joins	No	No	Yes
Function Type	Planar (with few exceptions)	Planar (with few exceptions)	Geodetic
Support for custom Coordinate Systems	Yes	Yes	No
Support for Spatial Reference Systems	Yes, default = 0, undefined	Yes, default = 4326, WGS84	No, all data in WGS84
Maximum Shape Size (compressed)	4 MB	4 MB	n/a

Spatial Transport Formats	Spatial Extender	Spatial Analytics	Spatio-Temporal Toolkit
GeoJSON	No	Yes*	Yes
GML	Yes	Yes*	No
KML	Export only	Export only	No
SDE	Yes	Yes	No
Shapefiles	Yes	Yes	No
WKB	Yes	Yes	Yes
WKT	Yes	Yes	Yes

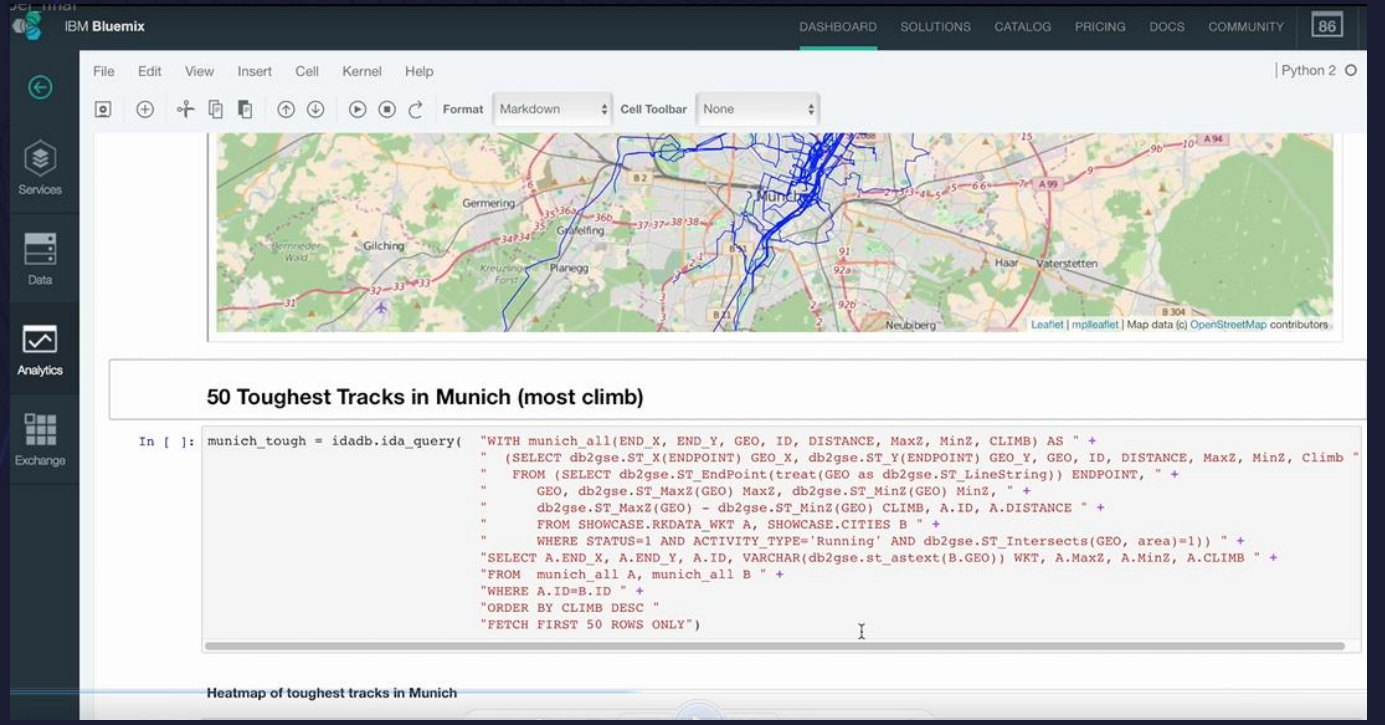
Spatial Analytics – Demos and More Information

Documentation:
https://www.ibm.com/docs/en/db2/11.5?topic=spatial-data

Banking location usage insight scenario
https://developer.ibm.com/recipes/tutorials/ibm-db2-spatial-analytics-bank/



Runkeeper Scenario
https://ibm.app.box.com/s/kb7baqbchmfptyf903q9fyqvpe3w093n



New and Modern Workloads In-Db2 Machine Learning

Existing Challenges with ML Projects

ML Development:

- ❑ **Talent gap** (62% faces this))
- ❑ **Sensitive data** (e.g., GDPR)
- ❑ **Collecting data** (19% efforts))
- ❑ **Cleaning data** (60% efforts)

ML Deployment:

- ❑ **Infrastructure** (scalable)
- ❑ **Scoring Performance**
- ❑ **Integration**

66% of ML projects leverage relational data

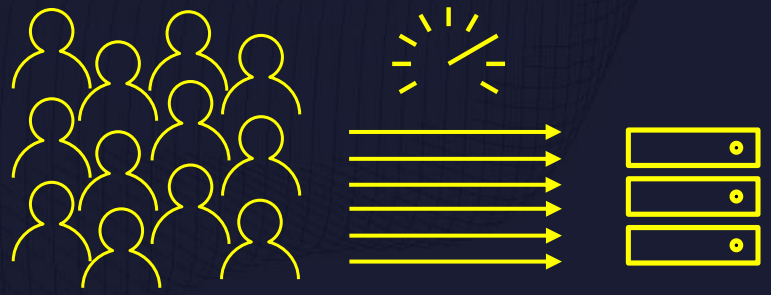
55% of ML projects don’t go beyond experiments

In-Db2 ML - In-database Scoring

Latency-sensitive Decisions

Instantaneous predictions

Examples:
- Payment processing
- Fraud detection
- Loan/claim pre-approval

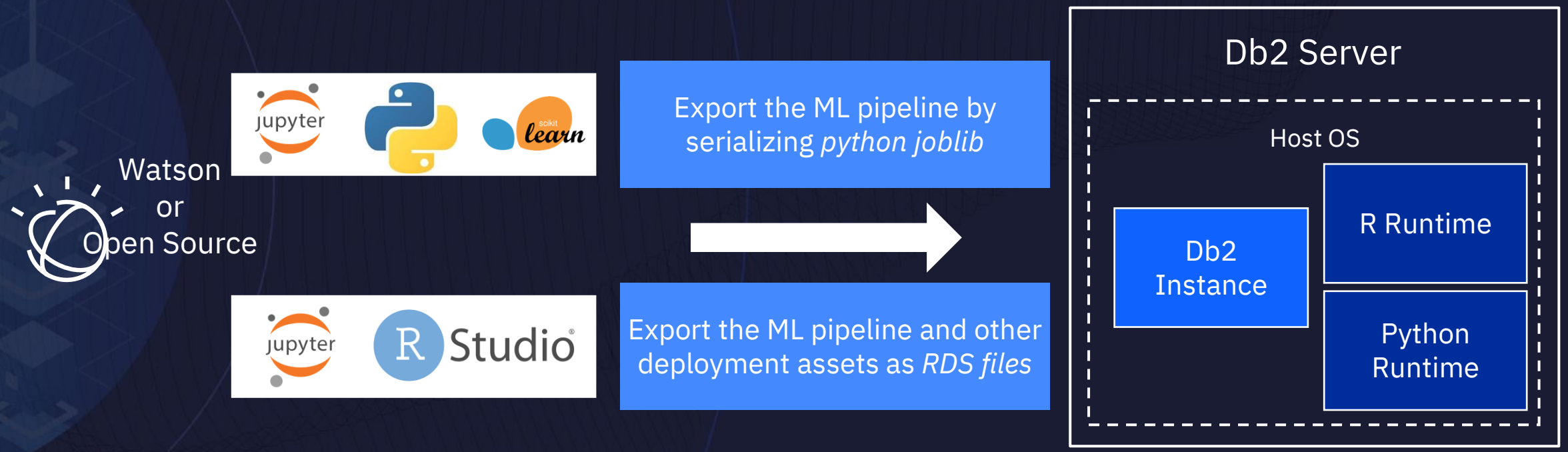


Large Batch Predictions

Near real-time prediction using “fresh” operational data

Examples:
- Anomaly detection
- Escalation risk prediction
- Intra-day forecasting
- Dynamic price optimization

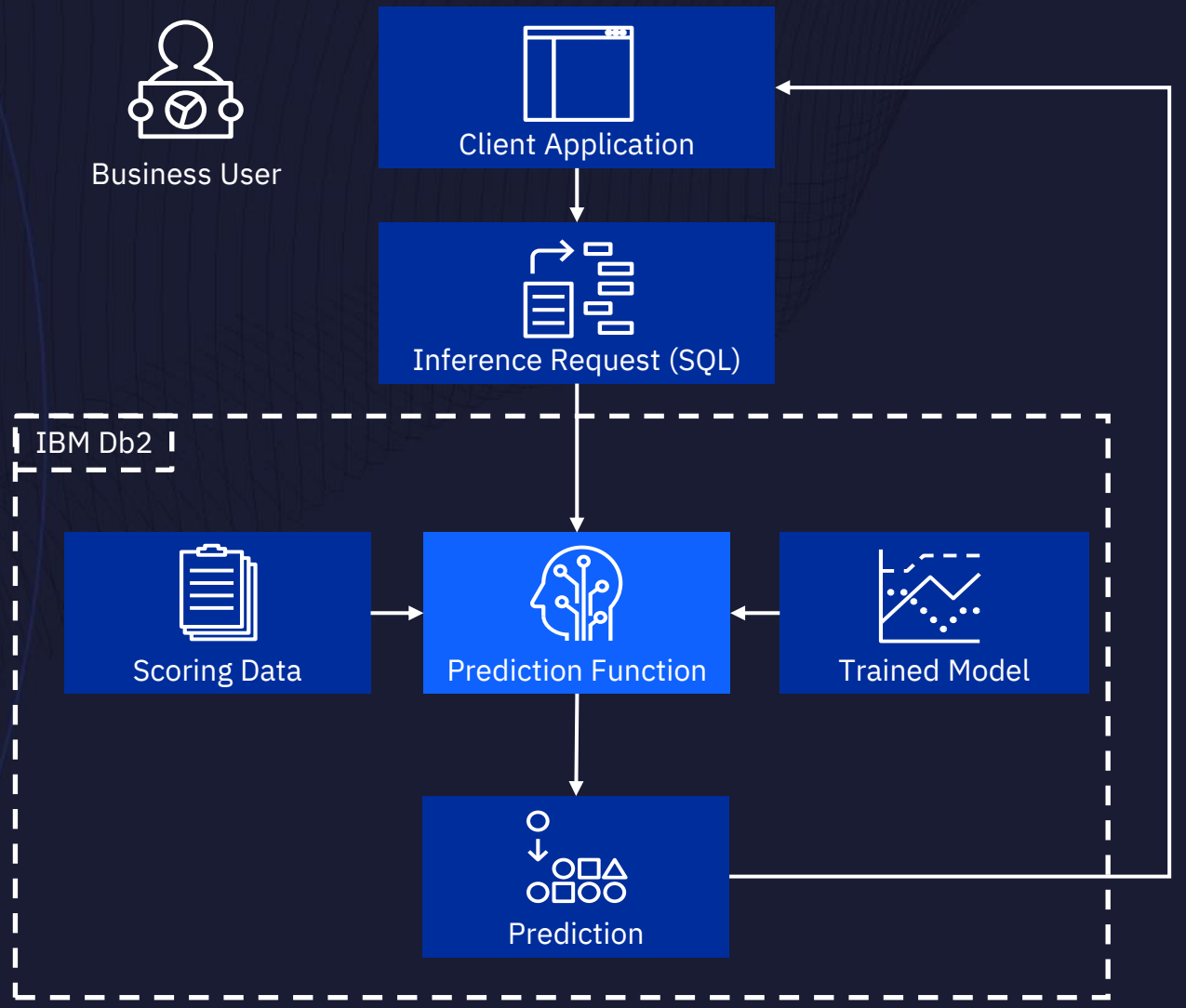
R/Python UDF: Scoring R Models via Db2



In-Db2 ML - Inferencing

Benefits:

– ML Infrastructure
– Low-latency
– High-throughput
– Simpler Integration



In-Db2 ML - 5x Speed up of Inferencing with IBM Db2

Model & Dataset

Logistic Regression (scikit-learn) Model

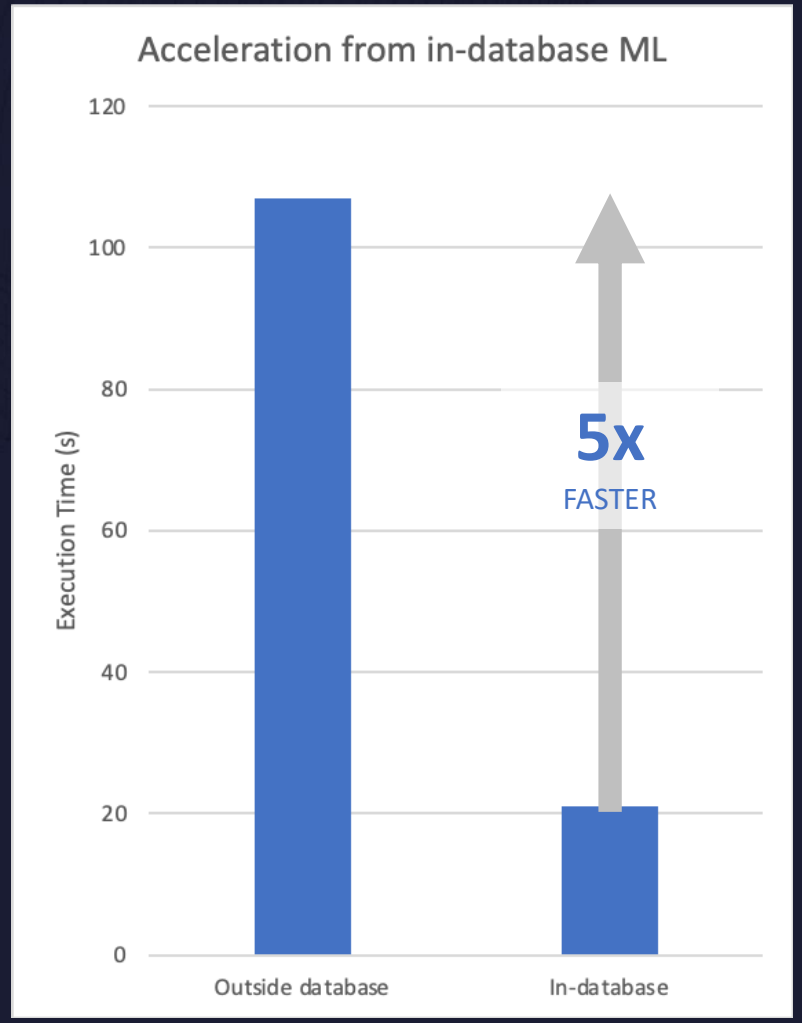
Batch size: 800k rows

Scoring on a Separate System

Data retrieved (over network) from Db2, scored, and written back to Db2: **1m47s**

Scoring with Db2

Data retrieved, scored, and written back inside Db2: **21s**



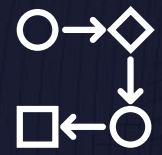
In-Db2 ML - AutoAI

Fast Model Selection

Select top-performing models in only minutes.

Start Quickly

Get started with experimentation, evaluation, and deployment.



AI Lifecycle Management

Enforce consistency and repeatability of end-to-end ML and AI development.

Integrated UI

Automate data prep, feature engineering and hyper-parameter optimization.

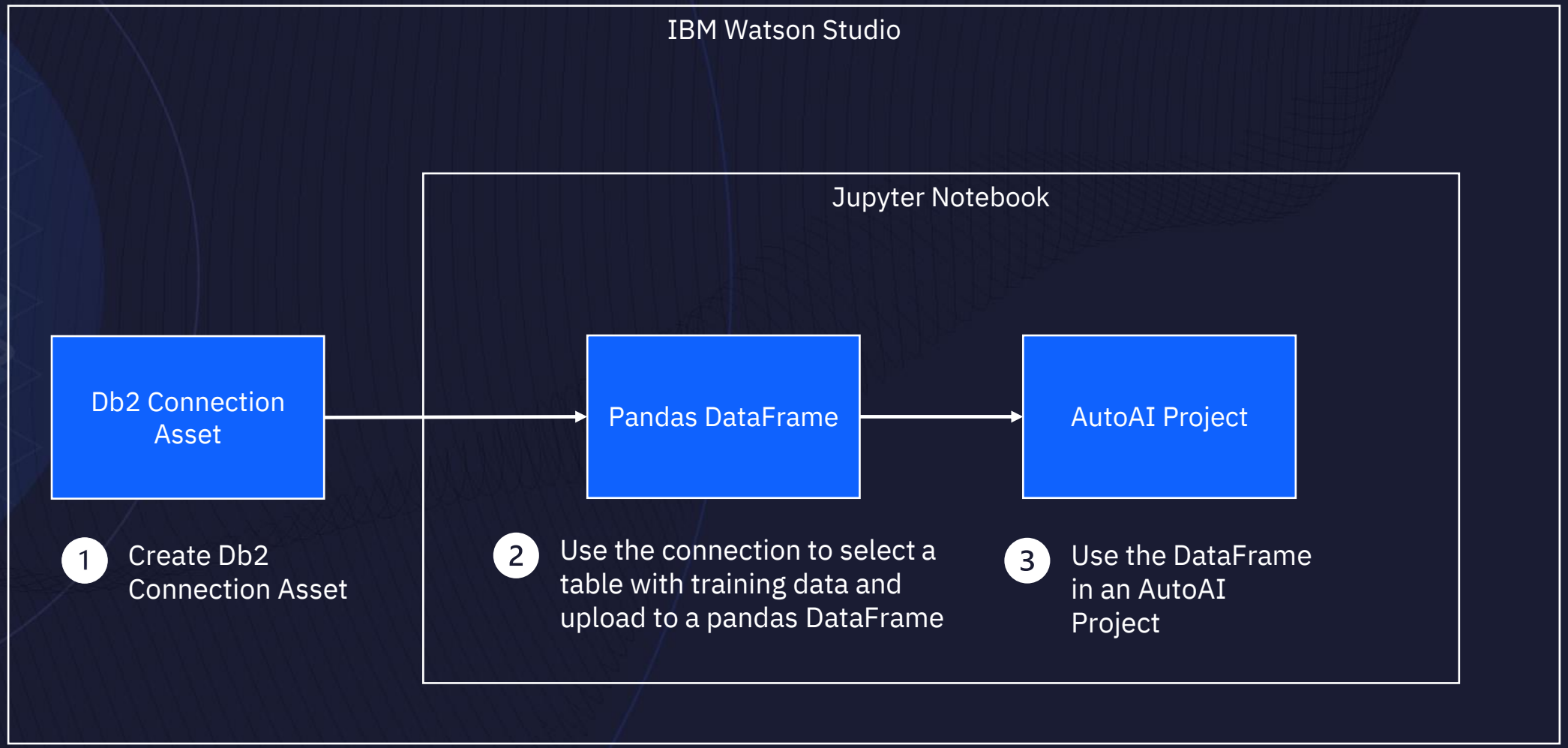
Pipeline Leaderboard

Watch and compare the top performing models on the leaderboard.

One-Click Deployment

When you are ready, save and select services to deploy with Watson Machine Learning.

In-Db2 ML - AutoAI & Db2: Workflow



In-Db2 ML - Supported Db2 Databases and ML Frameworks

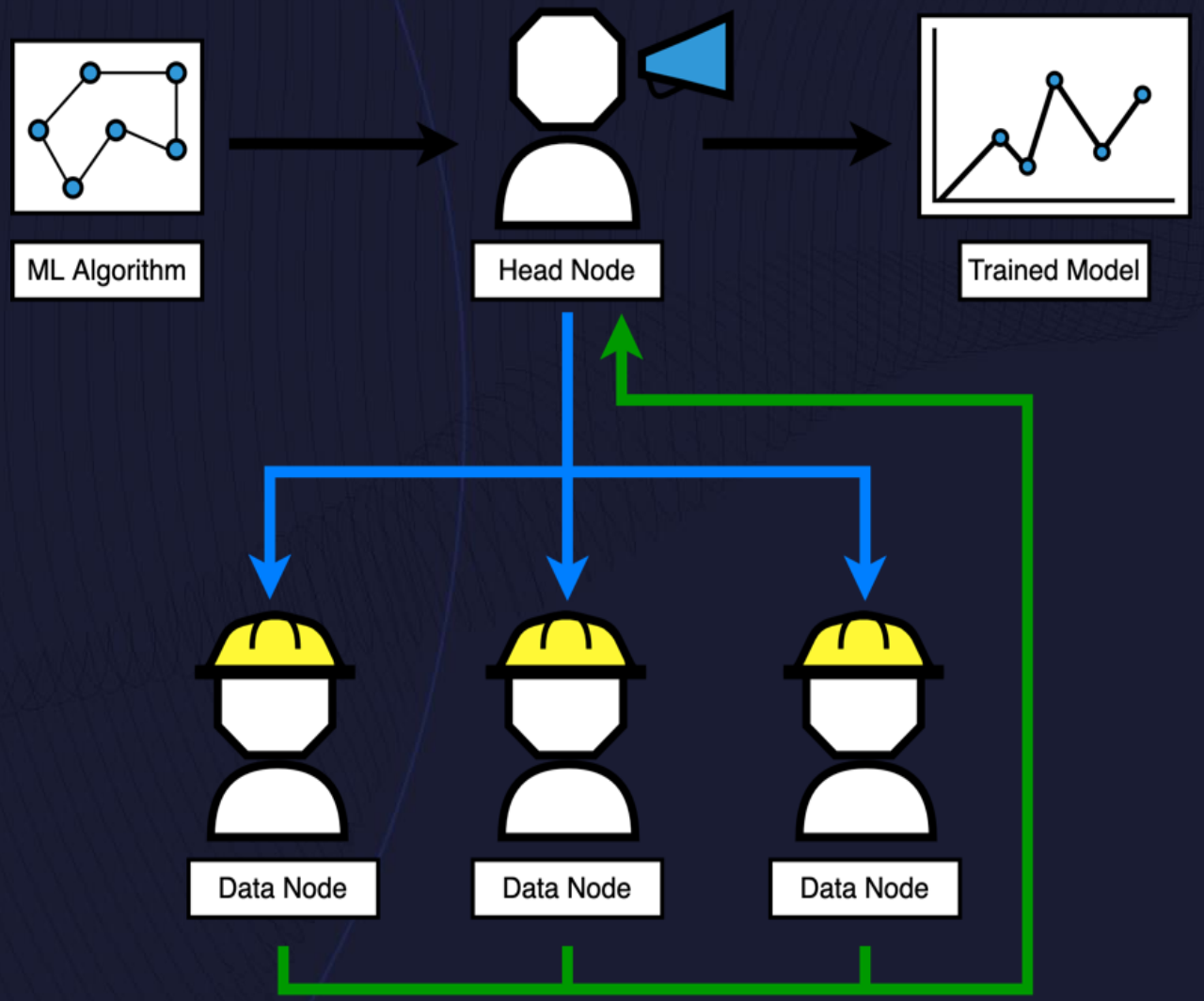
Supported Databases

- Db2
- Db2 on Cloud
- Db2 Hosted
- Db2 zOS
- Db2 Warehouse

Supported ML Frameworks

- scikit-learn
- lightGBM
- XGBoost

Distributed Model Training



Accelerating and Optimizing AI Lifecycle – With IBM Db2

Snap ML: Accelerated and Distributed Machine Learning Algorithms in Db2

Popular ML Algorithms

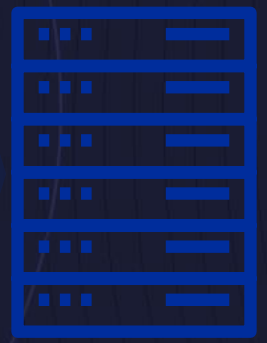
- Linear Regression*
- Logistic Regression
- SVM
- SnapBoost
- Decision Tree
- Random Forest

23X

Decision Tree

Db2 11.5.5 ML Capabilities

- Speed
- SQL Interface
- Distributed Training
- Security
- Scalability



Platforms
x86 Linux
PPCLE
zLinux (Linear Reg. unsupported)

Data Exploration	Data Preprocessing	Model Training	Model Evaluation	Model Deployment
Summary Statistics	Missing Values	Decision Trees	Confusion Matrix	In-Db2 Prediction
Virtualized DataFrame	Normalization	Naïve Bayes	Accuracy	Python UDF
Feature Selection	Standardization	Linear Regression	Precision	
DataFrame Operations	Discretization	K-Means Clustering	Recall	
	Train/test Split	Association Rules	MSE	
			MAE	

Challenges	Benefits of ML with Db2
Talent Gap	SQL interface for ML Deployment
Sensitive data	Virtualized data access and model building
Data transfer costs	In-Db2 ML – no data transfer
Infrastructure	Secure and scalable Db2 Infrastructure
Inference performance	In-Db2 scoring of Db2-native and open source models
Integration of ML	SQL-based Inference Queries

In-Db2 ML - Demos and Tutorials

Demos:

Build a Customer Segmentation Model with Db2 (K-Means Clustering)
Build a Classification Model with Db2 (Decision Tree)
Build a Regression Model with Db2 (Linear Regression)
Integrate a Db2-native model with a Cognos Dashboard
Deploying a ML Model Trained on Cloud Pak for Data to Db2

Hands-On:

Tutorials and Jupyter Notebooks
Pre-configured Hands-on Environment

Documentation:

Db2 11.5 Knowledge Center

In-Db2 ML - 4-hour Instructor-Led or self-paced Hands-on Training

- Gentle Introduction to Machine Learning Concepts
- Hands-on: build and deploy ML models using Db2's built-in ML Stored Procedures
- Hands-on: build and integrate open-source Python models with IBM Db2

The Workshop is delivered via requestable VMs available for both IBM employees and customers.

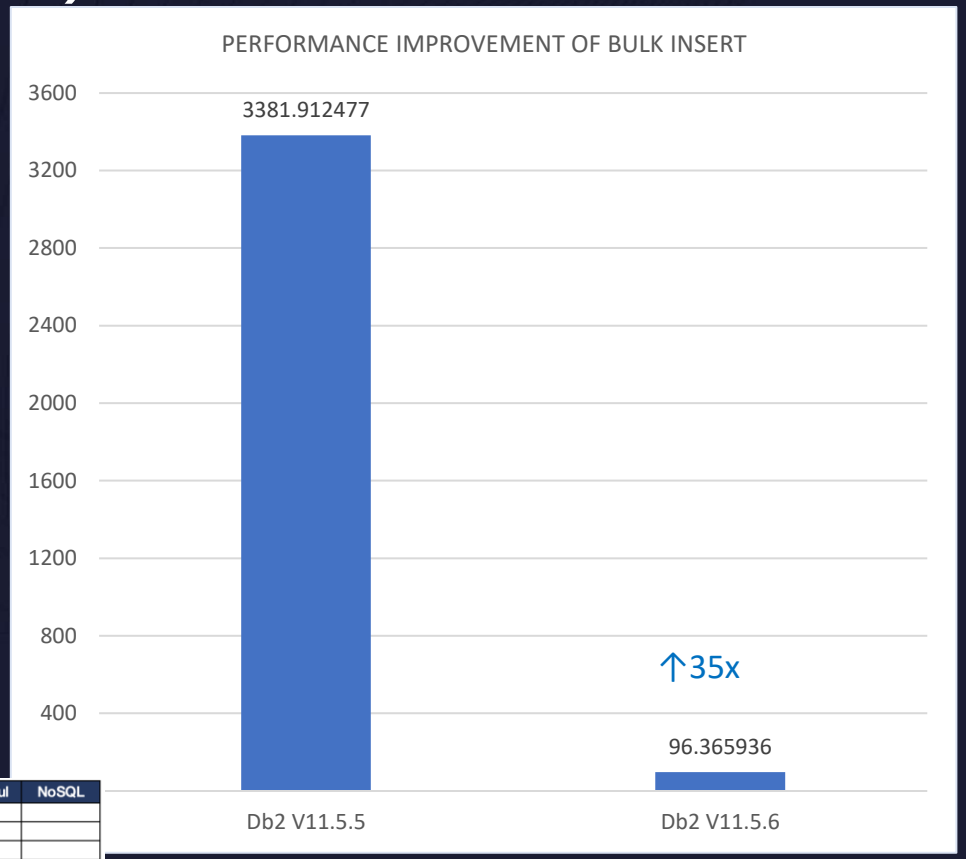
Contact: Shaikh Quader
Email Address: shaikhq@ca.ibm.com

Available as part of IDUG NA (Virtual) in June

New and Modern Workloads
Data Virtualization

Federation – New Enhancements (11.5.6)

- Connectivity – Spark JDBC Connectivity Support
- Functionality – Column Length Variation for Code Page Conversion
- Functionality – Nickname Hidden Column Support
- Performance – Federation DRDA Bulk Insert for Db2 Family Data Sources



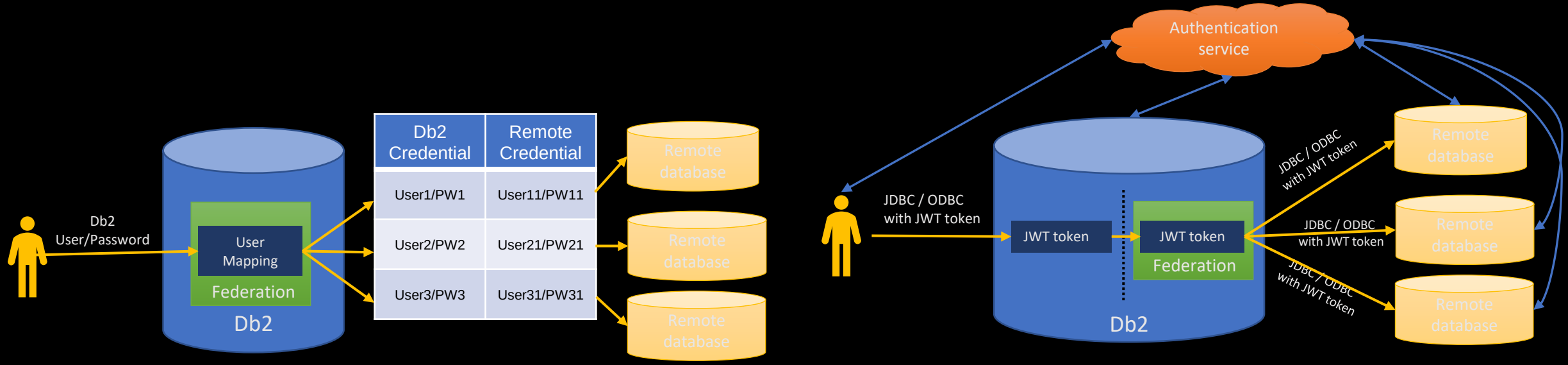
Category	Data Source	Native	ODBC	JDBC	RESTful	NoSQL
Relational	Db2 LUW	Yes		Yes		
	Db2 for IBM i	Yes				
	Oracle	Yes	Yes	Yes		
	MS SQL Server	Yes	Yes	Yes		
	Informix	Yes				
	Sybase	Yes				
Warehouse / Appliance	IIAS	Yes		Yes		
	Netezza		Yes	Yes		
	Teradata	Yes		Yes		
	SAP HANA		Yes	Yes		
	Greenplum		Yes	Yes		
Open Source	MySQL Community		Yes	Yes		
	MySQL Enterprise		Yes	Yes		
	PostgreSQL		Yes	Yes		
	MariaDB		Yes	Yes		
	Derby			Yes		
Hadoop	IBM Db2 BigSQL	Yes		Yes		
	Hive		Yes	Yes		
	Spark		Yes	Yes		
	Impala		Yes			
Files	Delimited	Yes				
	Excel	Yes	Yes			
	XML	Yes				
	JSON					Yes
	CSV	Yes				
Mainframe	Db2 for z/OS	Yes		Yes		
	IBM DVM for z/OS			Yes		

Category	Data Source	Native	ODBC	JDBC	RESTful	NoSQL
Message Queue	MQSeries	Yes				
Cloud	Db2 Warehouse	Yes		Yes		
	MS Azure SQL		Yes			
	Oracle Cloud		Yes			
	Amazon AWS Redshift			Yes		
	Google BigQuery			Yes		
	Amazon AWS S3			Yes		
	Salesforce			Yes		
	Snowflake		Yes	Yes		
NoSQL	Hyperledger Fabric					Yes
	MongoDB					Yes
	CouchDB					Yes
	Hbase HDFS					Yes
	Cassandra					Planning
	Redis					Planning
	Jira				Yes	
	Aha!				Yes	
	GitHub				Yes	
	HubSpot				Yes	
	TeamCity				Yes	
	api.spacexdata.com				Yes	
	earthquake.usgs.gov				Yes	
	Google Calendar API				Yes	
	groupkt.com				Yes	
	Yelp				Yes	

Supported Before v10.5
Supported In v11.1
Supported In v11.5 GA
Supported in v11.5.4
Supported in v11.5.5
Supported in v11.5.6
Working / Planning

Federation JSON Web Token (JWT) Single Sign-On (SSO) Support

- Federation User Mapping manages the relationship between **different credentials** of the local Db2 & remote databases
- Federation SSO shares the **same credential** between Db2 and remote database
- Federation determines the authentication mechanism to be
  - User Mapping when both are using user/password authentication
  - SSO when both are using JWT authentication, and get/pass JWT from Db2 to remote
- Only support Db2-to-Db2 (DRDA) connection



Db2 Federation Performance Improvement (V11.5.5)

- Common Scenario – these performance enhancements will benefit all federated workloads
- Featured Scenario – these performance enhancements will benefit specific use cases and functional items

Item for Common Scenario	Atomic Query		100 TPCDS Query
Federation Inter-partition Parallelism	1x~2x (3 nodes)	2.5x~3.5x (5 nodes)	4.8x (4 nodes)
Federation FMP buffer size enhancement	~0.3x		0.17x
Item for Featured Scenario	**Featured Query**		
Federation FFNR Pushdown	1270x (full pushdown)	56x (partial pushdown)	
Federation NoSQL Pushdown for MongoDB	50x		
Federation Bulk Insert for Oracle	84x		

New and Modern Workloads

Demos and More Information

Db2 JSON, BSON Support
ibm.biz/db2json

Db2 Graph Database and Graph Query Demos:
https://youtu.be/C5vmcYKEN-U
https://youtu.be/5_5UMeGWHV8

In-Db2 Machine Learning Demos:
Build a Customer Segmentation Model with Db2 (
Build a Classification Model with Db2 (Decision Tree)
Build a Regression Model with Db2 (Linear Regression)
Integrate a Db2-native model with a Cognos Dashboard

Hands-On:
Tutorials and Jupyter Notebooks

Documentation:
Db2 11.5 Knowledge Center

Db2 Resources

Information Resources:

- Db2 Roadmap - http://ibm.biz/AnalyticsRoadmaps
- Db2 RFE (Idea) Portal - http://ibm.biz/submitdb2idea
- Db2 Recorded Educational Webinars- http://ibm.biz/db2webinar
- Subscribe to Db2 technical newsletter - http://ibm.biz/db2nlsignup
- Connect with the Db2 online community - http://ibm.biz/db2tribe

Developer Resources:

- Db2 Developer Page to get started - http://ibm.biz/db2developer
- For Experienced Db2 developers, get your fav Db2 code sample on github - http://ibm.biz/db2github
- Want to try Machine Learning with Db2, check out - http://ibm.biz/learndb2
- Want details on Db2 Python Driver - http://ibm.biz/db2-drivers-python
- Want Details on Db2 PHP Driver - http://ibm.biz/db2-drivers-php
- Want Details on Db2 Node.js Driver - http://ibm.biz/db2-drivers-node
- Download the free Db2 python e-book - http://ibm.biz/db2pythonbook

Db2 –
Ready for Modernized Workloads and Deployments

Les King, IBM
lking@ca.ibm.com

Db2 LUW