Analytics on z Systems, what’s new?

Spark on z Systems
  - What is Spark
  - Spark Details
  - The ecosystem for Spark on z/OS
  - Use Cases

New Dimension for the DB2 Analytics Accelerator
  - Introduction
  - Details
  - Support of R functions
  - Positioning
Most Businesses Are Here

- **Data Warehouse Modernization**
  - Data lake
  - Data offload
  - ETL offload
  - Queryable archive and staging

- **Data-informed Decision Making**
  - Full dataset analysis (no more sampling)
  - Extract value from non-relational data
  - 360° view of all enterprise data
  - Exploratory analysis and discovery

- **Business Transformation**
  - Create new business models
  - Risk-aware decision making
  - Fight fraud and counter threats
  - Optimize operations
  - Attract, grow, retain customers
What Spark is (and is not)
What Spark Is, What it Is Not

- An Apache Foundation *open source project*
- Not a product
- An *in-memory compute engine* that works with data
- Not a data store
- Enables *highly sophisticated analysis* on huge volumes of data at scale
- *Unified environment* for data scientists, developers and data engineers
- Radically simplifies the process of developing *intelligent apps* fueled by data
What Spark on z/OS is Not

What it is not

- A data cache for all data in DB2, IMS, IDAA, VSAM …
- Just a different SQL engine or query optimizer
- An effective mechanism to access a **single** data source for analytics

**Why isn’t it the same as a query acceleration / IBM DB2 Analytics Accelerator?**

- Spark does not optimize SQL queries
- Spark is not a mechanism to store data, it rather provides interfaces to uniformly access data & to apply analytics using a unified interface
- DB2 Analytics Accelerator interaction with applications is via DB2; Spark interaction with applications is via Spark interfaces (Stream, MLlib, Graphx, SQL), driven through REST or Java
- Spark analytics can access data in DB2, DB2 Analytics Accelerator, VSAM, IMS, off platform, etc.
Apache Spark – a compute Engine

**Fast**
Leverages aggressively cached in-memory distributed computing and JVM threads
Faster than MapReduce for some workloads

**Ease of use (for programmers)**
Written in Scala, an object-oriented, functional programming language
Scala, Python and Java APIs
Scala and Python interactive shells
Runs on Hadoop, Mesos, standalone or cloud

**General purpose**
Covers a wide range of workloads
Provides SQL, streaming and complex analytics
Over 750 contributors from over 250 companies including IBM, Google, Amazon, SAP, Databricks…
Why does Spark matter to the business

1. Spark makes it easier to access and work with all data
   - Enables new data-based use cases
   - All data: Internal / External, Structured / Unstructured

2. Spark lets you develop line-of-business applications faster
   - Real-time insights, from all data sources
   - Automates analytics with machine learning

3. Spark learns from data and delivers in real-time
   - Clients that lead in data, lead in their industry
Build models quickly. Iterate faster. Apply intelligence everywhere.

**Data Engineer**
- Put right data to work for the job at hand
- Abstract data access complexity
- Enable real-time solutions

**Application Developer**
- Build analytics applications
- Optimize performance
- Leverages machine learning embedded

**Data Scientist**
- Identify patterns, trends, and risks
- Discover new actionable insights
- Build new models

[https://datascientistworkbench.com](https://datascientistworkbench.com)
Spark Details
The Spark Stack, Architectural Overview

- Faster development languages
- Easy of Use Spark Libraries
- High-level API Spark Core
- General purpose Scheduling
- Cluster Manager Scalability
- Fault tolerance
- Data Abstraction

- Java / Python / Scala / R
- Spark SQL Relational Operators
- Spark MLlib Machine Learning
- Spark GraphX Graph Processing
- Spark Streaming Real-Time Streaming
- Spark Core General Execution Engine
- YARN
- MESOS
- Standalone
- DB2 / HDFS / Cassandra / HBase / Oracle / JSON / Parquet / IMS / VSAM...
How to use Spark?

• **Interactive shell:**

![Interactive shell image](image1)

• **API from Scala, Java, Python or R written programs**

```python
val textFile = sc.textFile("hdfs://...")
val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
counts.saveAsTextFile("hdfs://...")
```
What is Spark used for?

- **Spark Streaming**: Enables scalable, high-throughput processing of live data streams. Live stream 'chopped' into batches based on time window.
- **Spark SQL**: Provides capability to perform relational queries via SQL (subset of HiveQL). Mix SQL queries with Spark applications.
- **Spark MLlib**: Provides scalable machine learning library, has common machine learning functions. Provides classification, regression, clustering, filtering, etc.
- **Spark GraphX**: Spark APIs for graph style processing and iterative graph computations.
- **Spark Core**: Foundation providing task dispatching, scheduling, i/o. Representation of Spark's basic unit of data: RDD.

From [http://spark.apache.org](http://spark.apache.org)
Why Apache Spark

Existing approach: MapReduce for complex jobs, interactive query, and online event-hub processing involves lots of disk I/O

New Approach: Keep more data in-memory with a new distributed execution engine
How Does Spark Process Your Data - RDDs

Resilient Distributed Data (RDD):
- Spark's basic unit of data
- Immutable, modifications create new RDDs
- Caching, persistence (memory, spilling to disk)

Fault tolerance
- If data in memory is lost it can be recreated from their original lineage

RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes

An RDD is physically distributed across the cluster, but manipulated as one logical entity:
- Spark will automatically “distribute” any required processing to all partitions
- Perform necessary redistributions and aggregations as well

Names

<table>
<thead>
<tr>
<th>Partition 1</th>
<th>Partition 2</th>
<th>Partition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael</td>
<td>Cindy</td>
<td>Dirk</td>
</tr>
<tr>
<td>Jacques</td>
<td>Dan</td>
<td>Frank</td>
</tr>
<tr>
<td>Dirk</td>
<td>Susan</td>
<td>Jacques</td>
</tr>
</tbody>
</table>
The ecosystem Spark on z Systems
Spark and z Systems

**Spark processing of z data**
Access to z data
- JDBC access to DB2 z/OS, IMS
- Rocket Mainframe Data Service for Apache Spark (MDSS)

**Spark running on z Systems**
- z/OS
- Linux on z Systems

<table>
<thead>
<tr>
<th>Operating system</th>
<th>Hardware</th>
<th>Bitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>z/OS 2.1</td>
<td>z Systems</td>
<td>64</td>
</tr>
<tr>
<td>SUSE Linux Enterprise Server 12</td>
<td>z Systems</td>
<td>64</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 7.1</td>
<td>z Systems</td>
<td>64</td>
</tr>
<tr>
<td>Ubuntu 16.04 LTS</td>
<td>z Systems</td>
<td>64</td>
</tr>
</tbody>
</table>

Downloads
These are the available downloads of the IBM Packages for Apache Spark:
- Linux on Power Systems
- Linux on System x
- Linux on z Systems
- z/OS

https://www.ibm.com/developerworks/java/jdk/spark/
IBM z/OS Platform for Apache Spark

- OLTP Applications: CICS, IMS, WAS...
- Analysis routines
- Spark Analytic Applications:
  - Customer Provided
  - IBM Provided
  - Partner Provided
- Spark SQL
- Spark Stream
- MLib
- GraphX
- RDD
- RDD
- RDD
- RDD
- RDD
- DB2 z/OS
- IMS
- VSAM
- SMF
- And more...
- Mainframe Data Service for Apache Spark
- z/OS
- Java z/OS
- Type 2
- Type 4
- Twitter
- Oracle
- HDFS
- JSON/XML
- DB2 LUW
- Tera
- IBM DB2 Analytics Accelerator
- External Data Sources
- IBM z Systems
- © 2016 IBM Corporation
IBM z/OS Platform for Apache Spark ...

- Almost **any data source** from **any location** can be processed in the Spark environment on z/OS
- **Mainframe Data Service for Apache Spark** (MDSS) is key to providing a single, optimized view of heterogeneous data sources
- MDSS can integrate **off-platform data** sources as well
- Large majority of cycles used by MDSS are **zIIP-eligible**
- Possible to use Spark on z/OS without it, but MDSS is recommended
- Integration in Online Transaction Processing (OLTP) is possible, but response time may be an issue
- Spark is the **high performance solution** for processing big data
- IBM DB2 Analytics Accelerator optimization with DB2 for z/OS can be integrated
z Systems Analytics and Spark: Well-matched

- Faster framework for in-memory analytics, both batch and real-time
- Federated data-in-place analytics access while preserving consistent APIs
- Why Spark on z Systems
  - Optimized performance with zEDC for compression
  - SMT2 for better thread performance
  - SIMD for optimized floating point performance and accelerate analytics processing for mathematical models via vector processing
  - Leverage IBM Java optimizations and enhancements
  - zIIP eligibility
Why and when Use Spark on z/OS?

The environment where Apache Spark z/OS makes sense:

- Running real-time or batch analytics over a variety of heterogeneous data sources
- Efficient real-time access to current and historical transactions
- Where a majority of data is z/OS resident
- When data contains sensitive information
- Don't scatter across several distributed nodes to be held in memory for some unknown period of time
- When implementing common analytic interfaces that are shared with users on distributed platforms

z/OS strengths are valuable in a Spark environment:

- Intra-SQL and intra-partition parallelism for optimal data access to:
  - nearly all z/OS data environments
  - distributed data sources
The Ecosystem for Spark on z/OS – relevance to the new consumers

- **Data Scientist**
  - The primary customer
  - Creates the spark application(s) that produce insights with business value
  - Probably doesn’t know or care where all of the Spark resources come from
  - "The convincer"

- **App Developer**
  - Helps the Data scientist assemble and clean the data, write applications
  - Probably better awareness about resource details, but still is primarily concerned with the problem to solve, not the platform
  - "The builder"

- **Data Engineer**
  - Close to the platform, probably a Z-based person
  - Works with the IM specialist to associate a view of the data with the actual on-platform assets
  - "The wrangler"
Use Cases
Business Use Case illustrated in a showcase

Financial Institution:
offers both retail / consumer banking as well as investment services
Business Critical Information owned by the organization

Credit Card Info
Trade Transaction Info
Customer Info

Produce right-time offers for cross sell or upsell, tailored for individual customers based on: customer profile, trade transaction history, credit card purchases and social media sentiment and information

Spark zOS Client Insight Use Case

STANDARD &POOR’S
Stock Price History – public data

Social Media Data: Twitter
Spark z/OS Showcase: Reference Architecture

IBM z Systems

- Cloudant
- Linux on z
- REST API
- CICS Transaction system
- Spark Job Server
- Spark
- DB2 z/OS
- JDBC
- IMS Transaction system
- JDBC
- IMS
- JDBC
- VSAM
- Customer Info
- Credit Card Info
- Trade Transaction Info

Open Source Visualization Libraries

© 2016 IBM Corporation
Analyzing SMF Data with Spark

Spark application is agnostic to data source and number of sources.

MDSS required on at least one system, MDSS agents required on all systems. No IPL required for installation.

Logstream recording mode required for real time interfaces.

Mainframe Data Service for Spark z/OS (MDSS)
New Redbook on Apache Spark implementation on IBM z/OS
Want to see and know more about Spark on z Systems? Then be our Guest for the Spark on z Systems Event

Apache Spark on IBM z Systems Workshop

November 23, 2016
IBM Client Center, Boeblingen, Germany

Apache Spark – the “Operating System for analytics” – greatly simplifies the programming of analytics and allows the analysis of all Enterprise Data in place, in real time. Available on IBM z/OS and for Linux on z Systems, Apache Spark ensures that IBM z Systems data sources can be accessed and analyzed by data scientists using their favorite languages and tools.

For more information click [here](#)
Contact: Khadija Souissi, souissik@de.ibm.com
References

- Spark Communities
  - [https://cwiki.apache.org/confluence/display/SPARK/Committers](https://cwiki.apache.org/confluence/display/SPARK/Committers)
  - [https://amplab.cs.berkeley.edu/software/](https://amplab.cs.berkeley.edu/software/)
- Spark SQL Programming Guide:
- IBM SystemML
  - Open Source: June 2015, we announced to open source SystemML
  - SystemML has been accepted as an Apache Incubator project
  - Source code: [https://github.com/apache/incubator-systemml](https://github.com/apache/incubator-systemml)
- Big Data University – Spark Fundamentals course
- IBM paper on Fueling Insight Economy with Apache
- A Deeper Understanding of Spark Internals
  - [https://www.youtube.com/watch?v=dmL0N3gSc8](https://www.youtube.com/watch?v=dmL0N3gSc8)
- IBM z/OS Platform for Apache Spark documentation
DB2 Analytics Accelerator on Cloud
Introduction
Announcing a new dimension for the DB2 Analytics Accelerator . . .

**New Product!**

**DB2 Analytics Accelerator on Cloud Version 1.1**

*High-speed analysis of enterprise data with cloud agility, flexibility and ease of deployment*

**High-speed analysis**
Rapid insight from enterprise data in a secure cloud environment

**Fast and Simple Deployment**
Improved agility and quick time to value

**Secure cloud environment**
Comprehensive data encryption capabilities based on a dedicated, bare-metal deployment

**New Version!**

**DB2 Analytics Accelerator for z/OS Version 6.1**

*Integrated on-prem and cloud solution supporting transactional and analytics workloads for right-time insight*

**New dimension of deployment**
Support for the IBM DB2 Analytics Accelerator on Cloud Version 1.1

**Flexible hybrid cloud**
A hybrid model with tight integration between cloud and on-premise deployment options

**Speed and Simplify**
Quickly deploy new or additional Accelerator instances by deploying applicable workload in the cloud

**Support for Data Science using R**
Support of R functions enabling in-database analytics on DB2 for z/OS using R - the most popular language used by data scientists
IBM DB2 Analytics Accelerator V6.1 introduces a flexible hybrid cloud

Flexible hybrid cloud:
A hybrid model with tight integration between cloud and on-premise deployment options

Chose the feature you need

Cloud Feature

On Premise Feature
DB2 Analytics Accelerator on Cloud and On-Prem

Ways They are Alike and Ways They are Different

**On Cloud**

- **Flexibility**
  Upsize & Downsize on-demand

- **Agility**
  Deploy within hours instead of days

- **Fits the Cloud Strategy**
  For mainframe customers investing on Cloud

**Common Characteristics**

- **DB2 z/OS Control**
  No application change needed

- **Unprecedented Speed**
  From hours/days to minutes/seconds

- **zOS Security & Governance**
  Maximum Security guaranteed by DB2 z/OS

**On - Prem**

- **High Capacity**
  Large data volumes scale to Petabytes

- **High Concurrency**
  Designed for high update volumes

- **Built-in Analytics**
  In-Database Analytics provide native support for data scientists.
Benefits of a Cloud Offering in general

Operational expense (Opex) instead of Capital expense (Capex)
  • Similar to MLC based pricing model

Operations and solution management by IBM
  • Hardware and software
    • Including flexible updates, trouble shooting, monitoring

Fast provisioning without shipping hardware
  • Hours or days instead of weeks
  • On-demand self-service to request new Accelerator systems
  • POC systems on demand

Fast, small and more frequent software updates

Flexible, elastic scaling
  • Elastic scaling for more demand
  • Multiple service instances per customer
DB2 Analytics Accelerator on Cloud
Details
IBM z Systems

DB2 Analytics Accelerator on Cloud - Architecture

Data Studio (DBA)

DB2 for z/OS (on premises)

VPN Client and Gateway

Internet

Secure (VPN)

Fast

VLAN

Accelerator Server

dashDB

(local, RAID, encrypted)

SOFTLAYER®

• Hardware: One physical server running in Softlayer
  • 256 GB main memory
  • 24 cores
  • ~ 4 TB uncompressed data (assuming 4x compression)

• Software: Setup based on container technology (Docker). Container contains dashDB and Accelerator Server software

All data on cloud is encrypted, when in motion between the VPN endpoints and at rest
A new acceleration engine for DB2 Analytics Accelerator on Cloud

Using BLU acceleration
Incorporated in dashDB database engine

**Next Generation In-Memory**
In-memory columnar processing with
dynamic movement of data from storage

**CPU Acceleration**
Multi-core and SIMD parallelism
(Single Instruction Multiple Data)

**Analyze Compressed Data**
Patented compression technique that preserves
order so data can be used without decompressing

**Data Skipping**
Skips unnecessary processing of irrelevant data
Why switching to a Container-based dashDB Acceleration Engine?

New Query Engine (IBM dashDB), common across all Platforms

In-memory computing

Columnar (and row-based) data store

Vector processing

Faster ingest for incremental updates

Higher degree of concurrent queries and users

Better SQL compatibility with DB2

Better performance

Rich in-database analytics capabilities

Spark integration

Compatible to existing DB2 Analytics Accelerator installations (co-existence)
IBM z Systems

Hardware specification of DB2 Analytics Accelerator on Cloud

CPU: 2 x 12 cores, Intel Xeon E5-2690 v3 (Haswell) @ 2.6 Ghz
Memory: 256GB

Disk
- 10x 800GB SSD, RAID-10 (mirrored + striped, 2 spares), for data
- 2x 1TB HDD, RAID-1 (mirrored), for OS

Network
- 2x 10GbE adapters

Redundant Power Supply

Good for ~4 TB of user data
*assuming 4x compression
Supported features for DB2 Analytics Accelerator on Cloud

- Adding and loading tables
  - Encodings: EBCDIC, UNICODE
  - Same datatypes as in DB2 Analytics Accelerator on premises

- Query Acceleration for static and dynamic SQL
  - More SQL native support than in DB2 Analytics Accelerator on premises feature

- Query History (only first ~100 bytes of SQL statement available)

- EXPLAIN to check whether a query can be accelerated
  - Visual Explain showing the access path on the Accelerator currently not available

- Monitoring
  - DB2 Analytics Accelerator Studio, SMF data for e.g. Omegamon, DISPLAY command
    - some counters will be 0
  - DSNX881I messages in SYSLOG
    - Disk x% full, DBMS state change, Long running SQL

- Workload balancing within generation (on Cloud <-> on Cloud, on Premises <-> on Premises)
DB2 Analytics Accelerator on Cloud eliminates some SQL limitations of an Accelerator on premises

- Route all types of correlated subqueries

- More SQL native support
  
  - EBCDIC MBCS, DBCS, GRAPHIC (converted to UTF-8 on Accelerator on-premises)
    
    - QUERY_ACCEL_OPTION 1 not required any more
  
  - "FOR BIT DATA" subtype (only supported for EBCDIC SBCS on Accelerator on-premises)
  
  - TIMESTAMP value 24:00:00 (mapped to 23:59:59 on Accelerator on-premises)
  
  - TIMESTAMP precision 12 (truncated to precision 6 on Accelerator on-premises)
    
    - QUERY_ACCEL_OPTION 5 and 6 not required any more
  
  - Byte-based processing of built-in functions, e.g. CHAR(multi-byte string) (Accelerator on-premises supports only character based processing)
    
    - QUERY_ACCEL_OPTION 3 not required any more

- Improved support for mixed encodings: Can add EBCDIC tables when UNICODE tables already present
  
  - Joining of EBCDIC and UNICODE tables still not supported
IBM z Systems

Accelerator Studio

Acelelerator: WOWDEMO @ DWDB13

- **Acceleration:** Started Stop
- **Status:** Online Trace: DEFAULT / OFF
- **Used space:** 51.7 GB of 2.05 TB

**Monitoring**

- **Queries:** Successful: N/A Failed: 0
- **Queue:** Max length: N/A Avg. wait time: 0 ms Max. wait time: N/A
- **CPU cost:** Query execution: 22.4 s Data maintenance: 0 ms Replication: 0 ms

**Tables (131 of 131 loaded / 131 of 131 enabled for acceleration)**

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Acceleration</th>
<th>Last Load</th>
<th>Distribution Key</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPCH</td>
<td>22</td>
<td>8 of 8</td>
<td>8 of 8 tables</td>
<td>C_CUSTKEY</td>
<td>0.000</td>
</tr>
<tr>
<td>CUSTOMER</td>
<td>676</td>
<td>Enabled</td>
<td>10/17/16, 7:39 PM</td>
<td>L_ORDERKEY</td>
<td>0.000</td>
</tr>
<tr>
<td>LINEITEM</td>
<td>144</td>
<td>Enabled</td>
<td>10/17/16, 11:29 PM</td>
<td>N_NATIONKEY</td>
<td>0.000</td>
</tr>
<tr>
<td>NATION</td>
<td>1 MB</td>
<td>Enabled</td>
<td>10/17/16, 7:11 PM</td>
<td>O_ORDERKEY</td>
<td>0.000</td>
</tr>
<tr>
<td>ORDERS</td>
<td>403</td>
<td>Enabled</td>
<td>10/17/16, 8:44 PM</td>
<td>P_PARTKEY</td>
<td>0.000</td>
</tr>
<tr>
<td>PART</td>
<td>573</td>
<td>Enabled</td>
<td>10/17/16, 8:30 PM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Query Monitoring**

<table>
<thead>
<tr>
<th>SQL Text</th>
<th>User ID</th>
<th>Start Time</th>
<th>State</th>
<th>Wait Time</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
<tr>
<td>SELECT &quot;ProductLine&quot;:PRODUCT_LINE_CODE AS &quot;Product...</td>
<td>FNEUMAN</td>
<td>10/26/16, 5:18:3...</td>
<td>Successful</td>
<td>0 seconds</td>
<td></td>
</tr>
</tbody>
</table>

View: All Queries Show: 
Loading data from DB2 for z/OS tables to the Accelerator on Cloud heavily depends on:

- Internet connectivity
- Bandwidth available for the mainframe to upload data to cloud using the VPN connection
- Data load speed on the Accelerator

From Lab tests:

- 174 GB/h end-to-end load throughput for loading data from San Jose, CA to Dallas, TX
  - Internet connectivity: 2 x 1 Gb/s
  - 12 unload streams in parallel
Which features of DB2 Analytics Accelerator (on Premises) are currently not available for DB2 Analytics Accelerator on Cloud?

Incremental Update
High Performance Storage Saver (HPSS)
In-database transformation using Accelerator-only tables (AOTs)
In-database analytics using IBM Netezza Analytics (INZA) stored procedures or R
**DB2 Analytics Accelerator on Cloud – Example Usage Scenarios**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Analytics on departmental operational data on the mainframe</th>
<th>Reduce sprawl of distributed data marts</th>
<th>Analytics on IT operations data</th>
<th>Proof of Concept (PoC) and/or testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Employ analytics, reporting, drilldown and self-service reporting to LOB users on data that is not subject to privacy regulations</td>
<td>Reduce “data tourism”: Instead of moving data copies off the mainframe, leave them in DB2 and run the queries in the cloud</td>
<td>Enable, predictions, problem detection and advanced optimization on performance and diagnostic data from your IT environment</td>
<td>Develop and test applications for the concept of a DB2 accelerator</td>
</tr>
<tr>
<td><strong>Benefits</strong></td>
<td>Rapidly and affordably deployed</td>
<td>Simplify and consolidate complex infrastructures in affordable way</td>
<td>Support the company’s cloud strategy</td>
<td>Performance improvements and cost reduction while retaining z Systems security</td>
</tr>
<tr>
<td></td>
<td>With secure data movement and access</td>
<td>With secure data movement and access</td>
<td>Reduce IT cost</td>
<td>Compatible offload behavior, rapidly and affordably deployed</td>
</tr>
<tr>
<td></td>
<td>no additional mainframe resources needed</td>
<td></td>
<td>Increase IT QoS</td>
<td></td>
</tr>
</tbody>
</table>
Where can Accelerator on Cloud be deployed?
VPN Connection Options for DB2 Analytics Accelerator on Cloud

**Option 1: Hardware (Router with VPN capability)**
- DB2 for z/OS
- Intranet (not encrypted)
- e.g. Cisco Router
- Internet (encrypted)

**Option 2: “Software Appliance” on Intel Server**
- DB2 for z/OS
- Intranet (not encrypted)
- Vyatta Client
  - VPN, GW
- Intel Server
- Internet (encrypted)

Accelerator

Vyatta Server

VPN, NAT
Connectivity options for Accelerators on Cloud

- Multiple DB2 systems can connect to a single Accelerator
- A single DB2 system can connect to multiple Accelerators
- Multiple DB2 systems can connect to multiple Accelerators

**Full flexibility for DB2 systems:**
- residing in the same LPAR
- residing in different LPARs
- residing in different CECs
- being independent (non-data sharing)
- belonging to the same data sharing group
- belonging to different data sharing groups

**Note:**
Multiple DB2 systems connected to one Accelerator on Cloud share the resources on the Accelerator. No resource allocation percentages can be defined per connected DB2 system as it can be done on Accelerators on-premises (Workload isolation feature).

Additionally all queries run with the same priority. WLM importance level from DB2 for z/OS is currently not passed to the Accelerator.
Connectivity options for coexisting Accelerators on Cloud and on Premises

A single DB2 system can connect to multiple coexisting Accelerators

Multiple DB2 systems can connect to multiple coexisting Accelerators

Full flexibility for DB2 systems:
- residing in the same LPAR
- residing in different LPARs
- residing in different CECs
- being independent (non-data sharing)
- belonging to the same data sharing group
- belonging to different data sharing groups
Query routing considerations for coexisting Accelerators on Cloud and on Premises

DB2 routes a query to that Accelerator that has the required tables enabled for acceleration.

- Define in advance which workload should run on the Accelerator on Cloud and which one on the Accelerator on Premises and load and enable the tables appropriately

If an Accelerator on Cloud and an Accelerator on Premises has the tables required by a query enabled for acceleration then **DB2 routes the query first to the Accelerator on Cloud**

- If this fails for any reason then it routes it to the Accelerator on Premises

If multiple Accelerators on Cloud exist and have the tables required by a query enabled for acceleration then DB2 does **workload balancing** and routes the query to the less utilized Accelerator

If multiple Accelerators on Premises exist and have the tables required by a query enabled for acceleration then DB2 does **workload balancing** and routes the query to the less utilized Accelerator

- Same as in V5, no new behavior
Support of R functions
In-Database Analytics Support for DB2 for z/OS with DB2 Analytics Accelerator

Available for Accelerator on premise only

Requires at least DB2 Analytics Accelerator V5 PTF2

Based on the ibmdbR-package for dashDB

Support for DB2 for z/OS with DB2 Analytics Accelerator available with version 1.48.0 of ibmdbR

https://cran.r-project.org/package=ibmdbR

Requires update of INZA package on Accelerator with INZA-IDAA 3.2.2.1 (IBM Netezza Analytics 3.2.2.1 for System z)

- Available on FixCentral
  3.2.2.1-IM-Netezza-ANALYTICS-IDAA-fp114213
Enables a data scientist to perform in-database analytics without needing any SQL skills

- Uses an ODBC connection to a database
- Translates R-operations to SQL-statements
  - DB tables and views are mapped to R data frames
  - Operations on data frames like filter, projection and merge are mapped to the corresponding SQL constructs

**Example:**

- Create data frame for IRIS table
  ```r
  iris.df <- ida.data.frame("IRIS")
  ```
- Determine Species and ID values for rows with “SepalWidth > 3”:
  ```r
  iris.df[iris.df$SepalWidth > 3, c("ID", "Species")]
  ```
  ```sql
  SELECT "ID","Species" FROM IRIS WHERE ("SepalWidth">3)
  ```
Positioning
VERY complementary Solutions for a wide Variety of Analytics

A purpose-built appliance for the acceleration of SQL queries and select SPSS algorithms against an optimized internal data store, tightly integrated with DB2 application semantics and tools.

Primary use case: accelerated BA/BI and modeling of DB2 current and deep historical data. Other data sources, on platform and off, can be loaded into the accelerator in a warehouse or data lake topology.

A platform for the execution of a variety of common advanced analytics algorithms written in Scala, Python, R or Java that operates on data in memory, independent of the underlying data store.

Primary use case: advanced analysis of current data from a variety of data sources in-place, on platform and off, without first requiring data to be consolidated into a warehouse or data lake topology.

When both solutions are present, Spark SQL calls that qualify will be accelerated on the Accelerator transparently.

Source: Paul T. DiMarzio
IBM z Systems Analytics Marketing Diamond Team Lead
Worldwide Portfolio Marketing Manager, z Systems Cognitive & IoT
Thank You