IDUG

IDUG

2024 NA **Db2** Tech Conference

Db2 Analytics Accelerator Product Update and Customer Experiences

Mehmet Cüneyt Göksu
Mehmet.Goksu@ibm.com

IBM

@IDUGDb2
#IDUG_NA24

Session Code: DMA1 | Platform: Cross Platform

- AGENDA

- Introduction
- Workload Assessment

- Infrastructure Enhancements
  - Transient Storage & NVMe
  - RoCE
  - Confined Nodes in multimode

- Software Enhancements
  - Performance
  - Functionality
  - Integrated Synchronization Enhancements
  - Monitoring Enhancements
  - Load Experiences…

- Accelerator Vnext.

Db2 Analytics Accelerator 7 on IBM Z

- Wide range of scalability (from very small to very large multi-node deployments)

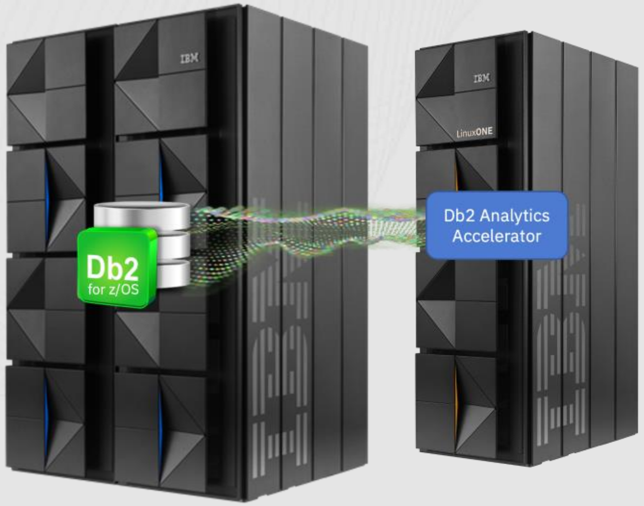
- Scalability for demanding workloads, optimized for large workloads with a flexible resource adjustments

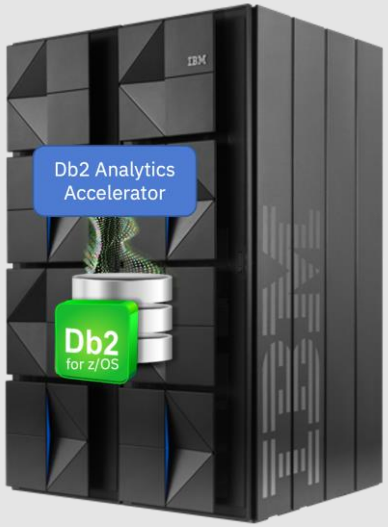
- Multi-node deployment that can grow from entry level (50 IFLs on IBM z16) to the largest size using all available IFLs on a system (200 IFLs on IBM z16) without reloading data

  - Multi-node accelerators can be deployed on any supported hardware for Db2 Analytics Accelerator on IBM Z (including IBM LinuxONE)

  - CKD or FB storage options can be used

  - Exploits new hardware features such as ROCE cards and NVMe Storage



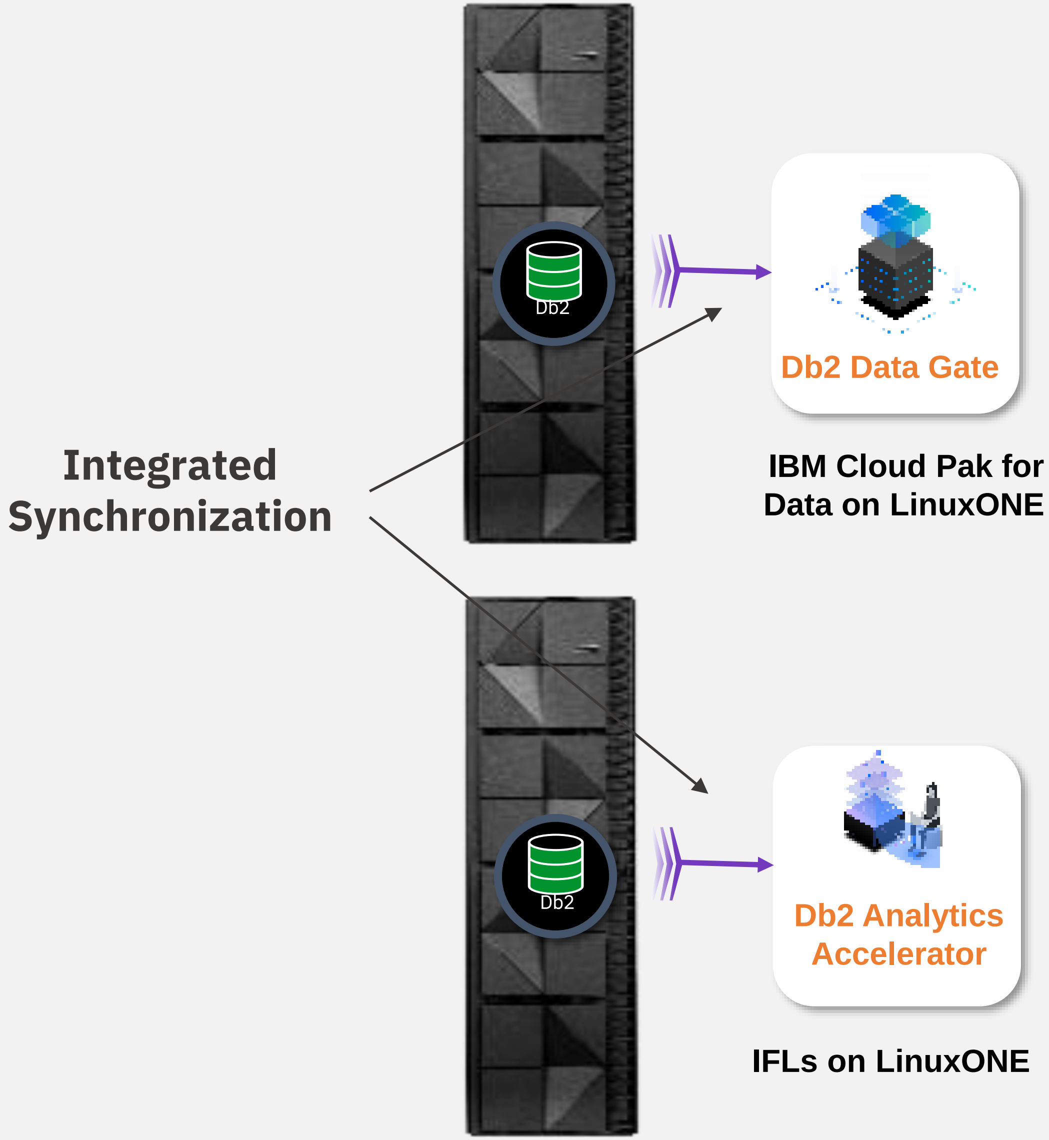


Integrated Synchronization

New replication protocol utilized in both IDAA and Data Gate

Purpose-built and optimized synchronization of data from Db2 to IDAA and Data Gate

- Fully integrated into Db2 for z/OS. Nothing to install on the mainframe. Easier to maintain
- Factors of improved overhead, throughput and latency compared to any other technology
  - Uses ½ the z/OS CPU of traditional replication
  - The remaining CPU is +98% zIIP eligible
- Allows for transactional consistency a.k.a HTAP, another competitive differentiator


Db2
Db2
Integrated Synchronization
Db2 Data Gate
IBM Cloud Pak for Data on LinuxONE
Db2 Analytics Accelerator
IFLs on LinuxONE

Db2 Analytics Accelerator assessment help you get more from Db2 for z/OS

- A workload assessment can help evaluate Db2 for z/OS workload and identify the benefits of routing a portion of the SQL execution to the Accelerator.
- A workload assessment could be useful in the following cases
  - There is no Db2 Analytics Accelerator, but want to analyze whether existing Db2 for z/OS workload would benefit from acceleration
  - There is Db2 Analytics Accelerator, but want to know whether new analytic queries or workloads would benefit from acceleration before you prepare accelerators for running these queries or workloads

Do I need Db2 Analytics Accelerator ?

Db2 Analytics Accelerator on IBM LinuxONE

Db2 Analytics Accelerator on IBM Z

Workload assessment – Overview

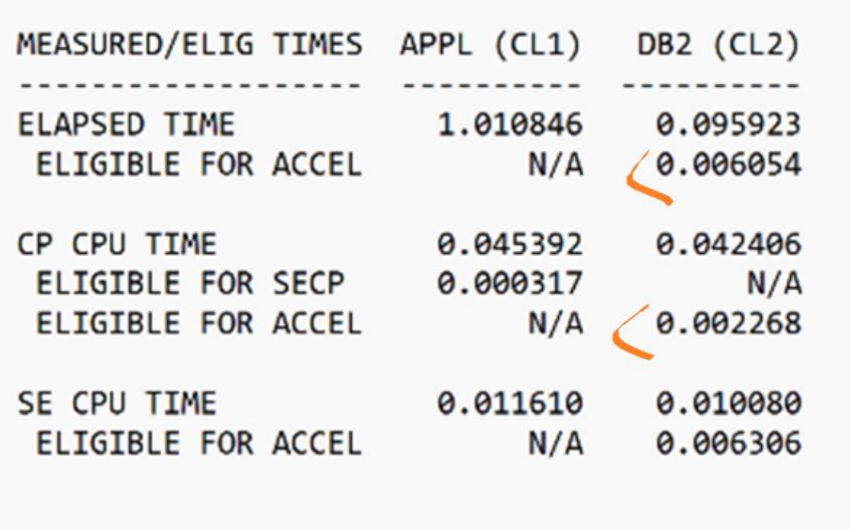
How to Analyze
-Collect Db2 for z/OS data from SMF and/or Dynamic Statement Cache
-Analyze Static and/or Dynamic SQL Workload
-Focus eligibility and savings

Final Report
-What is eligible / not-eligible
-Elapsed Time and CPU Savings for eligible workload
-Suggestions for non-eligible workload to make them eligible

References
-Details of the workload assessment are documented in this Technote
-IDUG Blog
- Determining whether queries can benefit from acceleration
-ACCELMODEL setting in Db2 for z/OS

Workload assessment – Examples

https://www.ibm.com/docs/en/om-db2-pe/5.5.0?topic=blocks-measuredelig-times

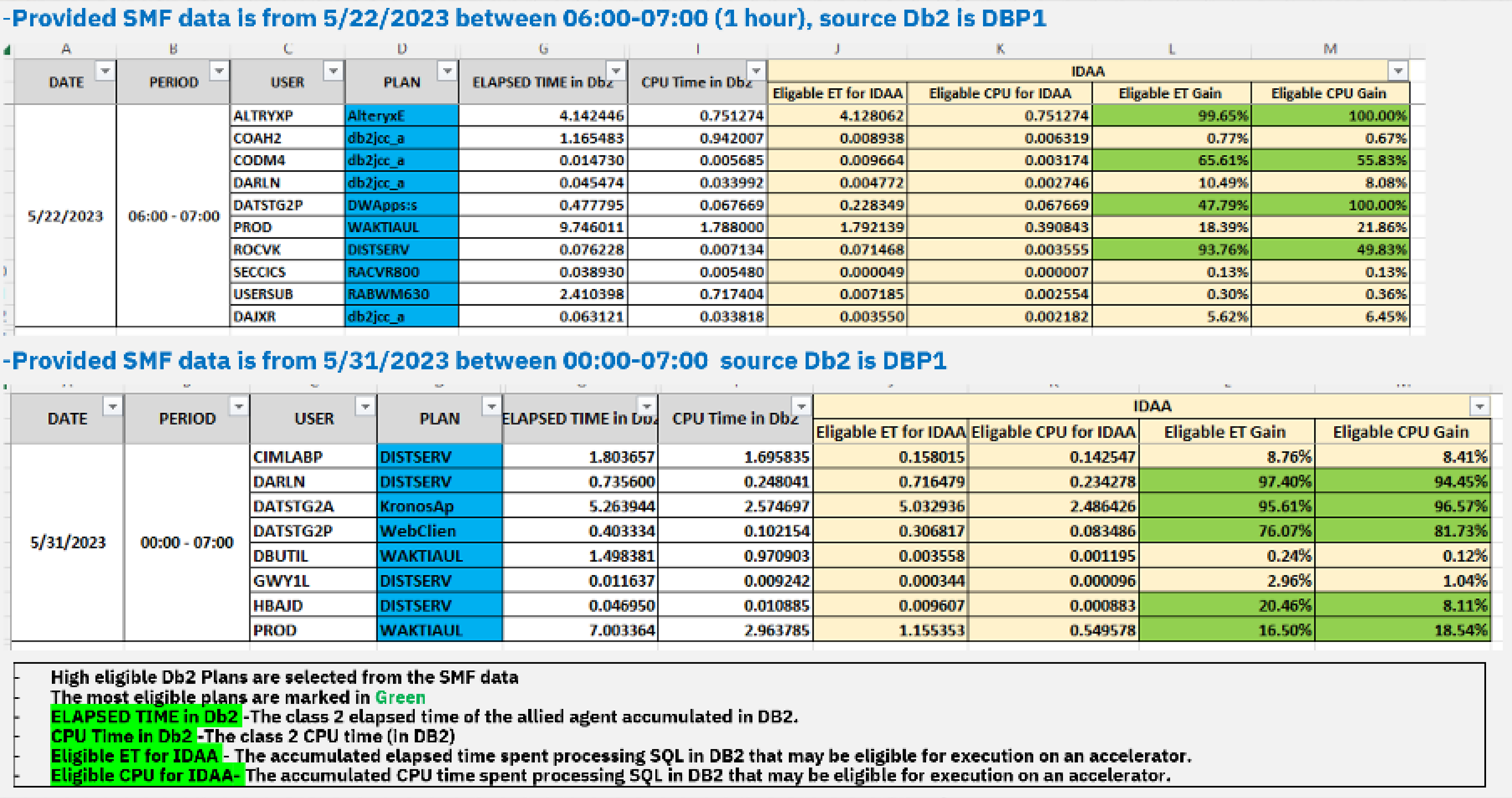


```
MEASURED/ELIG TIMES  APPL (CL1)   DB2 (CL2)
-------------------  ----------  ----------
ELAPSED TIME           1.010846    0.095923
 ELIGIBLE FOR ACCEL         N/A    0.006054

CP CPU TIME            0.045392    0.042406
 ELIGIBLE FOR SECP     0.000317         N/A
 ELIGIBLE FOR ACCEL         N/A    0.002268

SE CPU TIME            0.011610    0.010080
 ELIGIBLE FOR ACCEL         N/A    0.006306
```

DB2 (CL2) - ELAPSED TIME - ELIGIBLE FOR ACCEL: The accumulated elapsed time spent processing SQL in DB2 that may be eligible for execution on an accelerator.

DB2 (CL2) - CP CPU TIME - ELIGIBLE FOR ACCEL : The accumulated CPU time spent processing SQL in DB2 that may be eligible for execution on an accelerator.

Workload Assessment



-Provided SMF data is from 5/22/2023 between 06:00-07:00 (1 hour), source Db2 is DBP1

A	B	C	D	G	I	J	K	L	M
DATE	PERIOD	USER	PLAN	ELAPSED TIME in Db2	CPU Time in Db2	IDAA			
						Eligable ET for IDAA	Eligable CPU for IDAA	Eligable ET Gain	Eligable CPU Gain
5/22/2023	06:00 - 07:00	ALTRYXP	AlteryxE	4.142446	0.751274	4.128062	0.751274	99.65%	100.00%
		COAH2	db2jcc_a	1.165483	0.942007	0.008938	0.006319	0.77%	0.67%
		CODM4	db2jcc_a	0.014730	0.005685	0.009664	0.003174	65.61%	55.83%
		DARLN	db2jcc_a	0.045474	0.033992	0.004772	0.002746	10.49%	8.08%
		DATSTG2P	DWApps:s	0.477795	0.067669	0.228349	0.067669	47.79%	100.00%
		PROD	WAKTIAUL	9.746011	1.788000	1.792139	0.390843	18.39%	21.86%
		ROCVK	DISTSERV	0.076228	0.007134	0.071468	0.003555	93.76%	49.83%
		SECCICS	RACVR800	0.038930	0.005480	0.000049	0.000007	0.13%	0.13%
		USERSUB	RABWM630	2.410398	0.717404	0.007185	0.002554	0.30%	0.36%
		DAJXR	db2jcc_a	0.063121	0.033818	0.003550	0.002182	5.62%	6.45%

-Provided SMF data is from 5/31/2023 between 00:00-07:00 source Db2 is DBP1

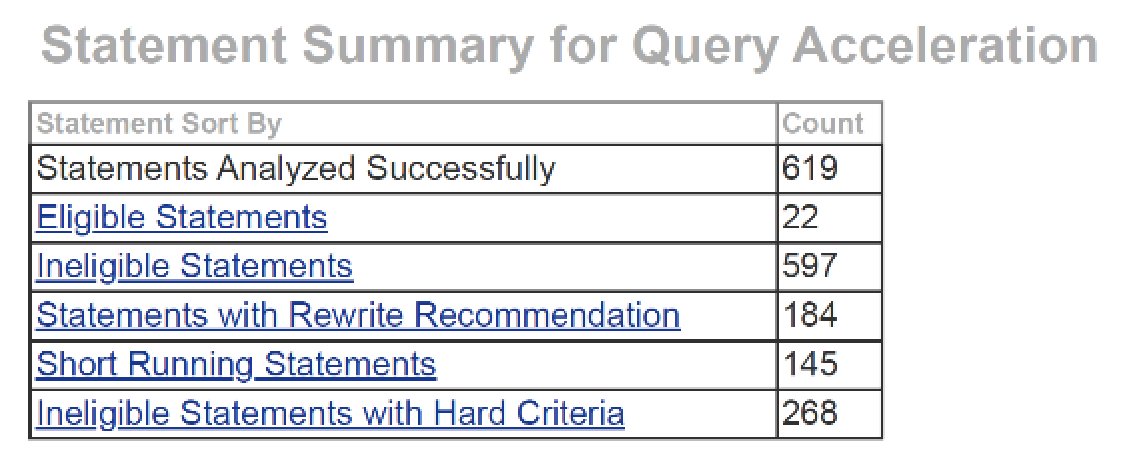
DATE	PERIOD	USER	PLAN	ELAPSED TIME in Db2	CPU Time in Db2	IDAA			
						Eligable ET for IDAA	Eligable CPU for IDAA	Eligable ET Gain	Eligable CPU Gain
5/31/2023	00:00 - 07:00	CIMLABP	DISTSERV	1.803657	1.695835	0.158015	0.142547	8.76%	8.41%
		DARLN	DISTSERV	0.735600	0.248041	0.716479	0.234278	97.40%	94.45%
		DATSTG2A	KronosAp	5.263944	2.574697	5.032936	2.486426	95.61%	96.57%
		DATSTG2P	WebClien	0.403334	0.102154	0.306817	0.083486	76.07%	81.73%
		DBUTIL	WAKTIAUL	1.498381	0.970903	0.003558	0.001195	0.24%	0.12%
		GWY1L	DISTSERV	0.011637	0.009242	0.000344	0.000096	2.96%	1.04%
		HBAJD	DISTSERV	0.046950	0.010885	0.009607	0.000883	20.46%	8.11%
		PROD	WAKTIAUL	7.003364	2.963785	1.155353	0.549578	16.50%	18.54%

- High eligible Db2 Plans are selected from the SMF data
- The most eligible plans are marked in Green
- ELAPSED TIME in Db2 -The class 2 elapsed time of the allied agent accumulated in DB2.
- CPU Time in Db2 -The class 2 CPU time (in DB2)
- Eligible ET for IDAA - The accumulated elapsed time spent processing SQL in DB2 that may be eligible for execution on an accelerator.
- Eligible CPU for IDAA- The accumulated CPU time spent processing SQL in DB2 that may be eligible for execution on an accelerator.

During the one-hour period, savings were
- Two threads from two plans were 100% eligible
- Two workloads reported more than 30% of CPU IDAA acceleration

During the seven-hour period, savings were
- Three plans reported more than 75% of CPU & ET savings with IDAA

Workload Assessment



Statement Summary for Query Acceleration

Statement Sort By	Count
Statements Analyzed Successfully	619
Eligible Statements	22
Ineligible Statements	597
Statements with Rewrite Recommendation	184
Short Running Statements	145
Ineligible Statements with Hard Criteria	268

- Estimated cost saving from query acceleration (sec): 11,527.56
- Estimated CPU saving from query acceleration (sec): 3,375.58
- For this workload, 58 tables were eligible for acceleration

Reasons for Ineligible statements

Reason	Remark
The query is an INSERT statement, but the Db2subsystem parameter QUERY_ACCEL_OPTIONS does not specify option 2 to enable its acceleration.	After zParm change, those queries could be eligible
The query contains an unsupported expression. The text of the expression is: TABLE	Unsuported function.
Db2 classified the query as a short-running query, or Db2 determined that sending the query to an accelerator server provided no performance advantage	Query performs better in db2 for z/OS
Query offloading is not supported for the followingtypes of statements: 1) DELETE and UPDATE statements; 2) INSERT statements that have nosubselect; 3) INSERT statements that have subselectswhen QUERY_ACCEL_OPTIONS is not set to 2.	
The query is not read-only.	Requires WITH UR

NVMe storage for Db2® Analytics Accelerator on IBM® LinuxONE

WHAT

NVMe (non-volatile memory express), is a protocol for highly parallel data transfer with reduced system overheads per input/output (I/O)

HOW

NVMe storage of IBM® LinuxONE system can be configured to be used as transient storage for temporary files

WHY

To provide a great stability and performance advantage when running workloads on IBM ® LinuxONE

WOW

Elapsed time improvement of a high double digit percentage value can be achieved

Benefits of the NVMe storage for Db2 Analytics Accelerator on IBM® LinuxONE

Improves load, query and replication stability

Protects the data pool

Protects the enterprise storage system

Improves query performance

7.5.12

NVMe storage for Db2 Analytics Accelerator on IBM® LinuxONE - Benefits

Improves load and replication stability	• Better isolation for loads, replication, and queries • Improved performance
Protects the data pool	• Large temporary files located on FCP or ECKD can be defined on NVMe
Protects the storage system	• Protects over utilization of I/O activity • Prevents bandwidth bottlenecks
Improves query performance	• Double digit improvements compared to using FCP attached Flash storage

NVMe storage for Db2 Analytics Accelerator on IBM® LinuxONE - Configuration

7.5.12

Local NVMe storage of a IBM® LinuxONE system can be configured to be used as transient storage for temporary files

New keyword in "runtime_environments" in JSON config file: "transient_storage": "NVMe"

Introduced in Accelerator V7.5.8 as technical preview and fully supported with maintenance level 7.5.12 or later.

- Remove any custom setting of the "temp_working_space" parameter because this parameter applies to transient data in the data pool only. As soon as transient data is processed on transient storage, the "temp_working_space" parameter becomes ineffective.
- Suggestion: Do not use temp_working_space and transient_storage at the same time. They are mutually exclusive

If specified, all available NVMe devices within a IBM® LinuxONE system (IBM® LinuxONE local storage) are used as temporary storage.

Current sizing recommendation

- num_of_carriers=num_NVMe cares, (one carrier+NVMe)/LPAR, up to two carriers/LPAR
- It is suggested 2 carriers/LPAR
- If LPAR memory is up to 4TB then use one carrier with 15TB per LPAR. For larger LPARs use 2 carriers with 15TB each

7.5.12

NVMe storage for Db2 Analytics Accelerator on IBM® LinuxONE HA Considerations with Accelerator release 7.5.12 (1/2)

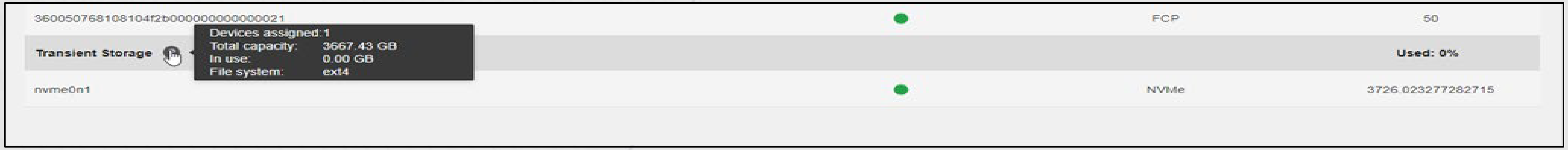
NVMe failure During Initialize	• The accelerator will hang in the system state "starting". After several hours after a timeout is reached, the system state becomes "failed ". • In the Admin UI storage panel, no storage information is displayed for the LPAR with the failed NVMe card or the storage panel is not shown at all. For other LPARs (in case of a multi-node deployment) the storage and NVMe card information is displayed correctly. • In the Admin UI logs panel no log messages are shown for the LPAR with failed NVMe card after messages related to network connection. For other LPARs (in case of a multi-node deployment) further log messages are written.
NVMe failure During Operation	• Executing queries that write transient data are hanging, but they do not fail. Internally they are waiting for the transient storage to become available, therefore they do not fail. • In the Admin UI no storage information is displayed for the LPAR with the failed NVMe card. The display hangs. For other LPARs (in case of a multi-node deployment) the storage and NVMe card information is displayed correctly.
Single-Node customers	• It is possible to connect multiple NVMe to the LPAR. The upper limit still exists as 16/CEC

7.5.12

NVMe storage for Db2 Analytics Accelerator on IBM® LinuxONE HA Considerations with Accelerator release 7.5.12 (2/2)

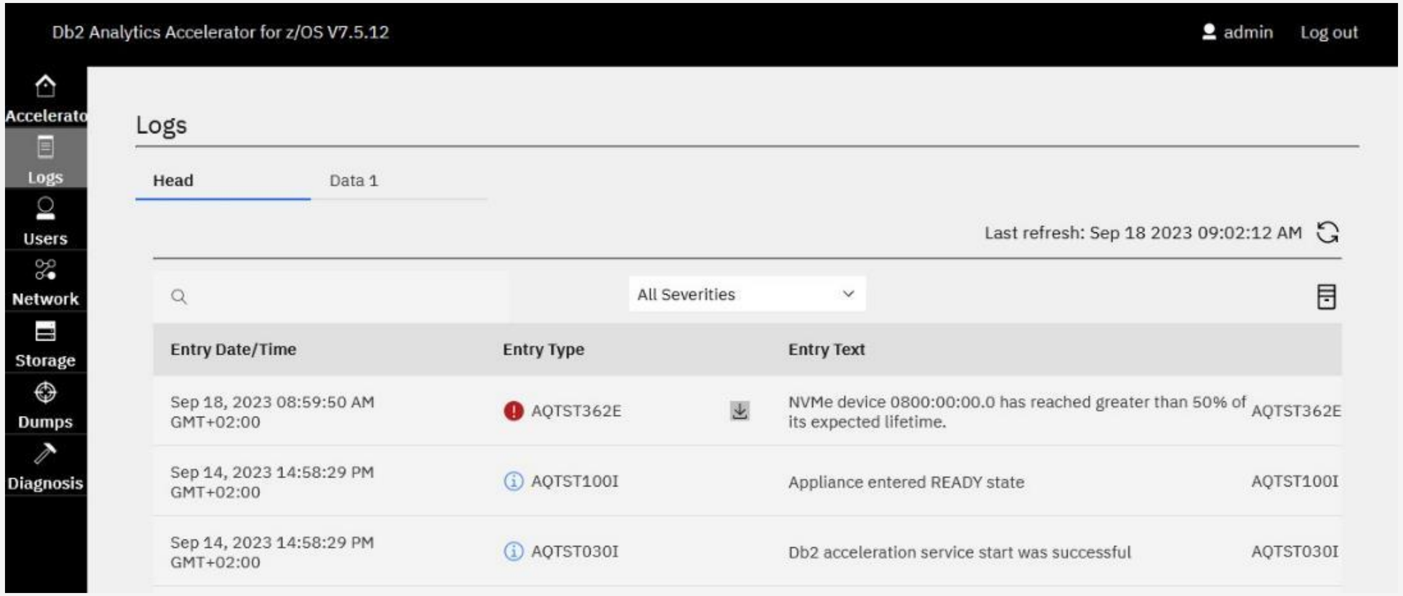
Accelerator	If NVMe fails while Accelerator is running, running queries do not fail but 'hang'. Once customer realizes hanging queries, customer can check the storage tab in the Admin UI.
Startup Code	A reset is enough in such a case, the startup code recognizes the failed NVMe and takes it out. It is an accelerator outage approximately 15min depends on the accelerator setup. It is always recommended to have 2 NVMe/LPAR. If one of the NVMe fails in such scenario, after Accelerator reboot, it keeps on running
Future releases	In future releases of the Accelerator, there will be messaging enhancements for NVMe. So that customers can monitor and automate necessary actions if the card fails. There are planned enhancements about HA as well. It means with 2xNVME/LPAR, even one NVMe fails, Accelerator could continue running the workload without immediate outage and customer can plan the replacement in future with a planned outage.

For the node with the failed NVMe, no storage information is shown, the display 'hangs' as well. For the other nodes storage and NVMe information is shown.



```
360050768108104f2b000000000000021        ●        FCP        50
Transient Storage                                            Used: 0%
    Devices assigned:1
    Total capacity:    3667.43 GB
    In use:            0.00 GB
    File system:       ext4
nvme0n1                                  ●        NVMe       3726.023277282715
```

NVMe Monitoring



Logs

Last refresh: Sep 18 2023 09:02:12 AM

Entry Date/Time	Entry Type	Entry Text	
Sep 18, 2023 08:59:50 AM GMT+02:00	AQTST362E	NVMe device 0800:00:00.0 has reached greater than 50% of its expected lifetime.	AQTST362E
Sep 14, 2023 14:58:29 PM GMT+02:00	AQTST100I	Appliance entered READY state	AQTST100I
Sep 14, 2023 14:58:29 PM GMT+02:00	AQTST030I	Db2 acceleration service start was successful	AQTST030I

7.5.12

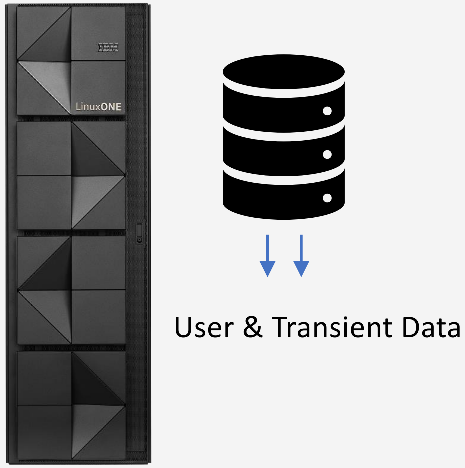
- **Accelerator Web UI**
- In 7.5.12, there is a new message in Accelerator Web UI as follows. It is a warning message. Once it is received, customer can start planning to replace the card.
- In a future release, there will be an enhancement to DSNX881I console message so that customer does not need to check Accelerator Web UI.

Troubleshooting

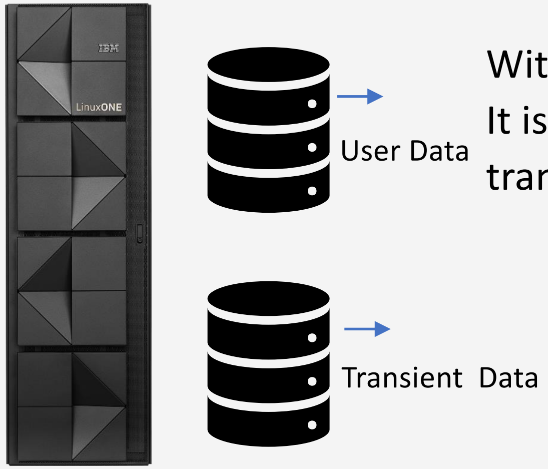
First level of support will be the Accelerator team, since Accelerator team is detecting the errors. The TLS team is following up on how to involve the multi vendor support team

Transient (Temporary) Storage

7.5.12


User & Transient Data

Prior to 7.5.12
Transient storage was part of Data Storage Pool which includes both user & transient data


User Data

Transient Data

With 7.5.12+
It is possible to define dedicated storage for transient data

- Could be a dedicated external storage with the same or different storage technology with user data
- Could be NVMe (LinuxONE Only)

What is Transient Storage

- Data is not persistent, will be deleted immediately after the task has finished.

What uses Transient Storage

- *Temporary results of extensive sort operations* during query processing that cannot be executed exclusively in the system memory.
- *Replication spill queues* when a replication-enabled table is loaded to the accelerator and at the same time new data changes on this table are replicated.
- *Query results* that tend to arrive faster than they can be picked up by the receiving client.

7.5.12

How to define Transient Storage

Define a dedicated transient storage pool on external storage

1. The new optional keyword “transient_devices” in the “storage_environment” section of the JSON configuration file
2. A separate storage pool for transient data can be defined on FCP- or FICON-attached storage.

For instance, the data pool could be deployed on FICON-attached storage and the transient pool on FCP-attached storage.

```
"transient_devices": {
    "type": "dasd",
    "devices": [
     "0.0.9b12"
    ]
  }
```

Use local NVMe (non-volatile memory express) storage

- Adding of the keyword “transient_storage”:”NVME” to the “runtime_environments” section of the JSON configuration file
- This option
  - is only available for LinuxONE deployments
  - is highly recommended for multi-node deployments.

```
"runtime_environments": [
      {
         "cpc_name": "Z16_4",
         "lpar_name": "LPAR1",
         "transient_storage": "NVMe",
         "network_interfaces": [
                 .
                 .
```

7.5.12

Heads up!

- In a multi-node environment, define “transient_devices” for all LPARs. You must assign the same amount of storage to the transient pool on each LPAR.
- If “transient_devices” (dedicated external storage) and “transient_storage” (NVMe) are defined in the JSON configuration file, the setting of “transient_storage” takes precedence and “transient_devices” is ignored.
- Remove any custom setting of the "temp_working_space" parameter because this parameter applies to transient data in the data pool only. As soon as transient data is processed on transient storage, the "temp_working_space" parameter becomes ineffective.
- Only additional storage can be used to increase the transient storage pool. It is not possible to move storage from the data pool to the transient storage pool
- After the changes are completed in the JSON configuration file, upload it to the accelerator as described in this chapter: https://www.ibm.com/docs/en/daafz/7.5?topic=z-updating-existing-configuration
- After the JSON configuration file has been successfully uploaded and applied, accelerator uses the new transient storage pool for writing transient data.
- More details are in this tech note : https://www.ibm.com/support/pages/node/7138629

7.5.11

Support of RoCE cards for inter-node communication in multi-node installations (1/2)

- The use of 'RDMA over Converged Ethernet' (RoCE) Express cards for the cluster network in Db2 Analytics Accelerator on Z multi-node installations is supported and highly recommended.
- Very high workload in multi-node installations of Db2 Analytics Accelerator on IBM Z can lead to inefficient query processing, resource contention, and potential system instability.
- Using RoCE cards will reduce IFL resource contention, ensure predictable query runtimes, and enhance operational stability. Therefore, RoCE cards must be used for the inter-node communication (internal cluster network) between LPARs instead of HiperSockets.
- RoCE cards are supported with Accelerator maintenance level 7.5.10.1 or later.

- For inter-node communication it is recommended to bond two 25 GbE RoCE cards. The primary purpose of bonding is to deliver a performance increase to 50 Gb. A second purpose is to provide high availability.

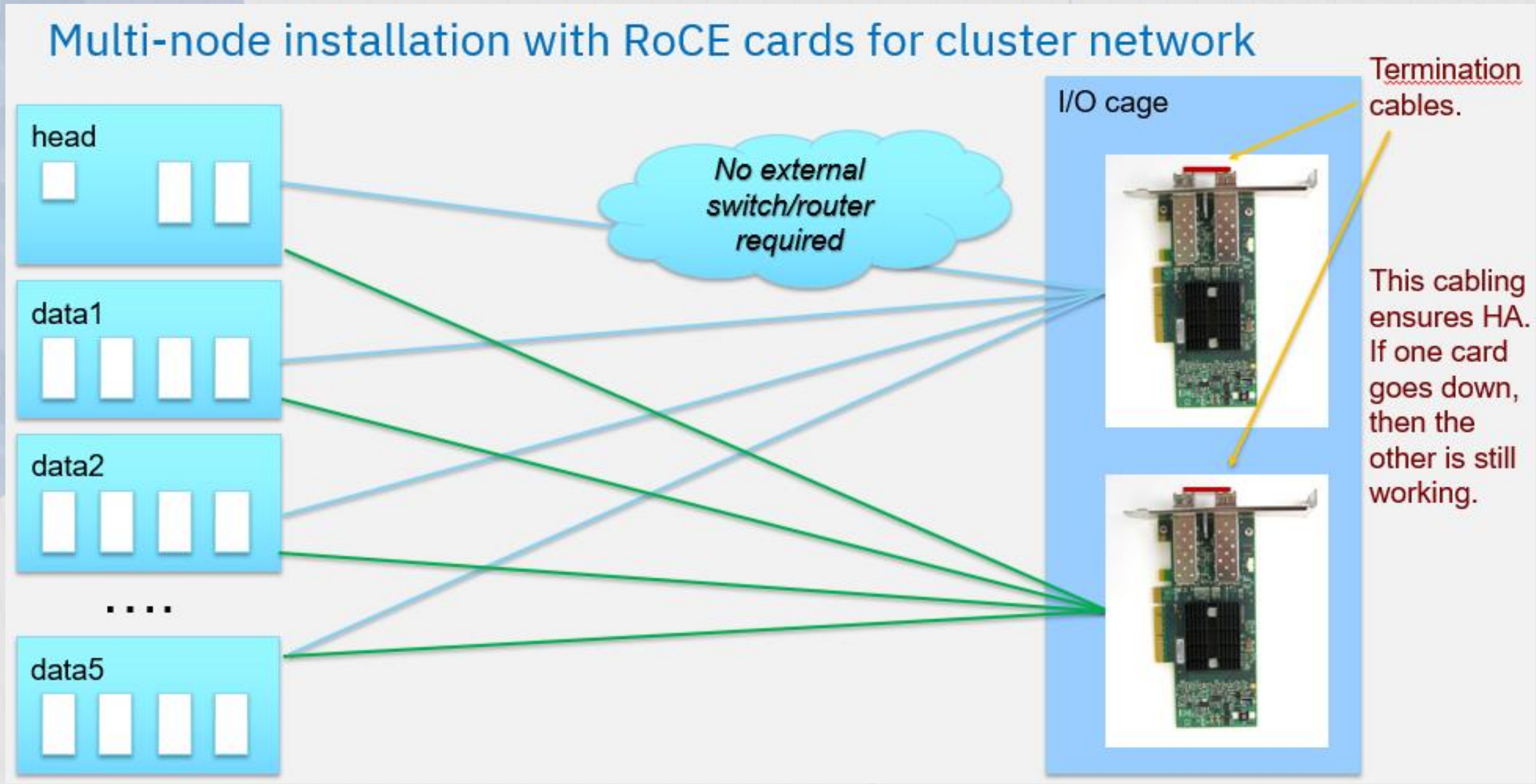
Support of RoCE cards for inter-node communication in multi-node installations (2/2)

7.5.11

Benefits of using RoCE cards for inter-node communication

- RoCE cards do not consume CPU resources (IFLs)
  - All configured IFLs are used for workload execution (e.g. queries, load, replication)
- Using RoCE cards reduces IFL resource contention. Available CPU capacity for workload execution is increased by roughly 10% compared to HiperSockets
  - For example, if a multi-node installation is configured with 70 IFLs then
    - With HiperSockets ~63 IFLs are used for workload execution and 7 IFLs for internode-communication
    - With RoCE cards all 70 IFLs are used for workload execution
- Using RoCE cards enhances operational system stability in high workload situations and ensure predictable query runtimes

- Each RoCE Express card has two ports. Both ports must be connected by using a termination cable to ensure high availability.
- Multiple accelerators in a single CEC can share RoCE Express cards.
- Within the CEC, the RoCE cards are represented by Function IDs (FIDs), which are assigned to each port of the RoCE card during the hardware configuration. They can be obtained from the I/O Definition File (IODF), the Hardware Management Console (HMC), or the Dynamic Partition Manager (DPM) for each LPAR.
- FIDs are different for each Db2 Analytics Accelerator LPAR.
- More details are in this tech note : https://www.ibm.com/support/pages/node/7031391


Multi-node installation with RoCE cards for cluster network
head
data1
data2
....
data5
No external switch/router required
I/O cage
Termination cables.
This cabling ensures HA. If one card goes down, then the other is still working.

Confined head node – new multi-node installation for z16

For multi-node installations on z16, it is now supported to set it up with 3 or 4 LPARs, depends on the number of drawers on the CEC. This is a deployment option for high-end installations

Benefit: Improved load and workload performance because cross-drawer traffic is avoided, no PR/SM interaction is involved
New JSON config parameter to define the MLN distribution for such a setup: “mln_distribution” : “4:4”

For IBM z16 systems with 2 or 4 drawers, 4 LPARs are recommended.

For IBM z16 with 3 drawers, it is recommended to install 3 LPARs

If starting with a system of 2 drawers and there is need to grow system capacity, adding IFLs to the existing drawers until you reach 50 IFLs per drawer should lead to very good scalability of performance. Same is true if adding 2 additional drawers

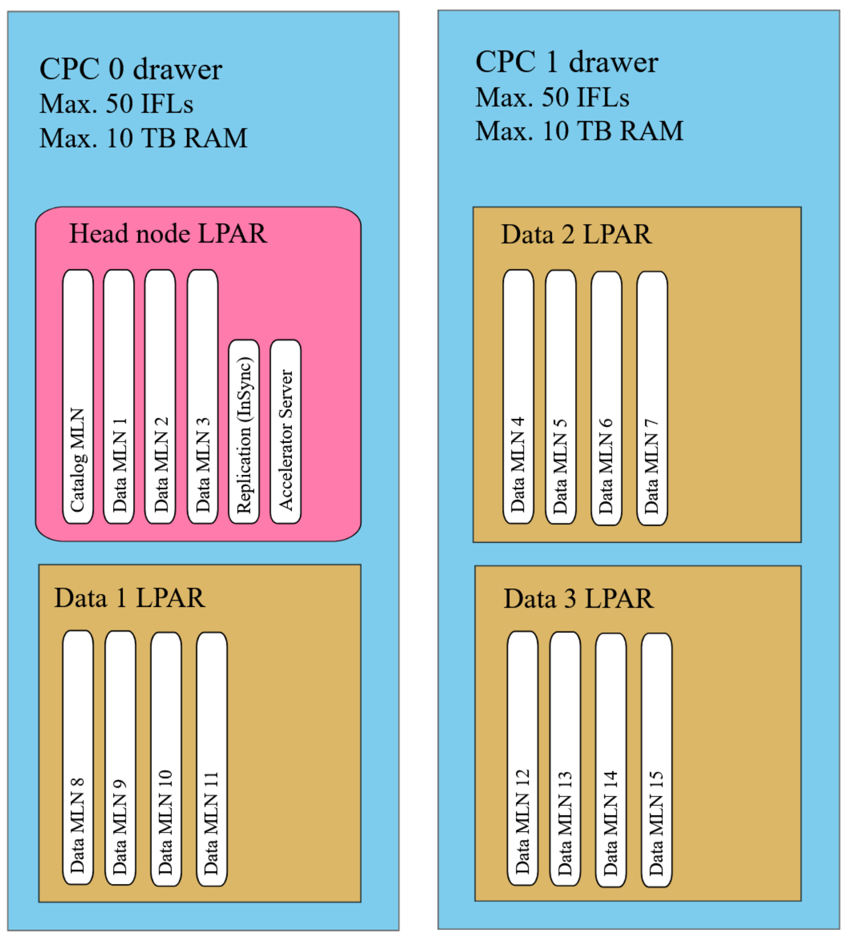
If starting with a 2-drawer system (4 LPAR) and then growing to a 4-drawer system (4 LPAR), normal migration works - so instead of sharing one drawer with 2 LPARs, every LPAR gets its own LPAR.

To change the number of LPARs, a complete re-load of the Accelerator and a fresh install is required. In other words, if starting with a 3-drawer system (3 LPARs) as recommended and then adding a fourth drawer, it is highly recommended to re-install the cluster with 4 drawers.

https://www.ibm.com/docs/en/daafz/7.5?topic=setup-multi-node-z16

Confined head node – new multi-node installation for z16

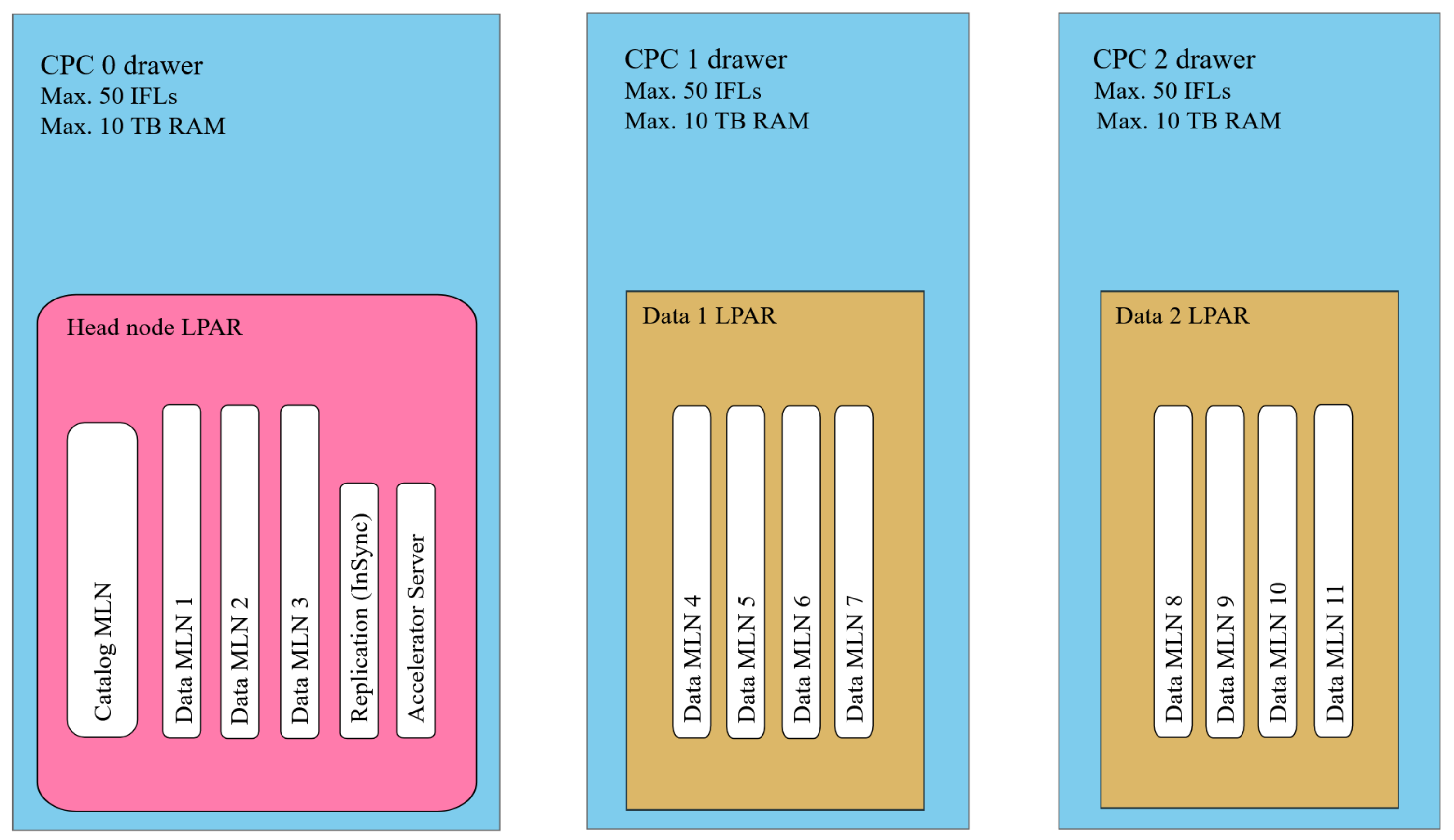
2 drawers: For IBM z16 systems with two drawers, 4 LPARs are recommended. This translates to 2 LPARs per drawer.



If you want to increase the system capacity, you can add IFLs to the existing drawers until you reach 50 IFLs per drawer. This leads to a very good scalability of the performance. The same is true for adding two additional drawers.

Confined head node – new multi-node installation for z16

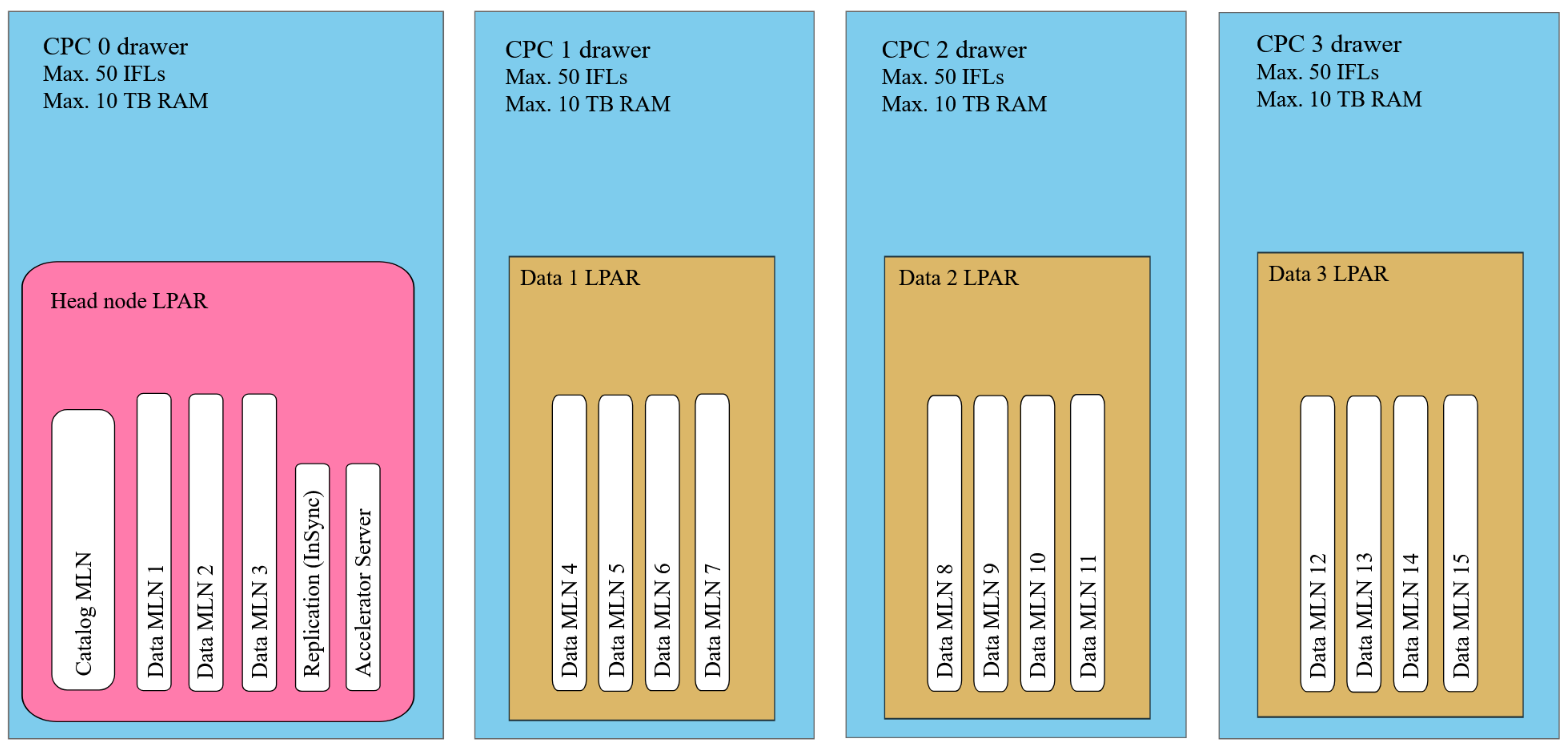
3 drawers: For an IBM z16 or higher with three drawers, install 3 LPARs, that is, one LPAR per drawer



If you ran four LPARs on a system with only three physical drawers, or used a system with an unbalanced number of IFLs and an unbalanced amount of memory, the overhead caused by the resulting cross-drawer memory access could be so significant that a three-drawer system would be barely faster than a two-drawer system. This is why three LPARs are recommended for a three-drawer system (one LPAR per drawer).

Confined head node – new multi-node installation for z16

4 drawers: For IBM z16 systems with four drawers, four LPARs are recommended. That means one LPARs per drawer



The term MLN refers to a database limit that defines the maximum number of "multiple logical nodes" per processing node. In this case, you have one logical node for the catalog and up to three logical nodes for data processing in the head node LPAR, plus two or four logical nodes for data processing in each data LPAR.

7.5.12

REORG REBALANCE toleration on replicated PBR table

- REORG REBALANCE to redistribute data across partitions which might cause change in partition IDs in DB2 z/OS. Internal Synchronization holds the partition ID which helps to apply some utilities that impact partitions like REORG DISCARD, LOAD DUMMY.
- Before 7.5.12, REORG REBALANCE is not supported, moves the tables in the subjected tablespace to ERROR state which requires reload.
- With 7.5.12+, If a REORG REBALANCE operation is executed on a replication-enabled table in PBR tablespace then replication now continues instead of suspending the table from replication
- Since a REORG REBALANCE, operation changes the limit keys of a table in PBR tablespace, a subsequent execution of REORG DISCARD or LOAD DD DUMMY will suspend the table from replication. In this case a full reload or reload with 'detectChanges' enabled will bring back the table to active replication state.
- The REORG REBALANCE toleration mode is not enabled by default. Open a support case to have IBM Support turn on the toleration in a remote maintenance session for each applicable connected Db2 subsystem. The activation requires a restart of the Integrated Synchronization processes.
- If the table is replicated across multiple accelerators it is required to have the same level of code in each accelerator to support tolerance for REORG REBALANCE.

7.5.11

New non-routing query option for correlated subqueries

- Accelerator V7 in general supports the routing of correlated subqueries

- However, executing a correlated subquery on the Accelerator could result in suboptimal query performance compared to running the query in Db2 for z/OS especially if the correlation can not be decorrelated by the Db2 Warehouse optimizer automatically

- Db2 for z/OS APAR PH52030 for Db2 12 and Db2 13 delivers an option to block the routing of specific correlated subquery patterns to the Accelerator and execute on Db2 for z/OS instead

- See the APAR description for query examples. To set the option contact IBM Support by opening a support case

- Note, that Accelerator maintenance level 7.5.11 is **not** required as prerequisite for using this new option

7.5.12.2

Availability of timestamp of when a table was last used by a query

- The stored procedure ACCEL_GET_TABLES_INFO now returns the following 2 new values per table:
  - lastAccessTimestamp: Timestamp when the table was last used by a query
  - accessCount: number of queries which used the table since the table was added to the accelerator

- Allows smarter table management on the accelerator, e.g. these values could be used to identi tables that are rarely used or not used at all and potentially could be removed from the accelerator
- Note, that the values are not exposed to a graphical user interface such as Accelerator Studio

7.5.12.2

Availability of timestamp of when a table was last used by a query

- What is noticed, ACCESS COUNT on an accelerator shadow table has been increasing but there was no query running in the system!

- IDAA code updates the Last Access Timestamp and the access count whenever a table is accessed when user issues a query, replication or when IDAA internally, access the table.

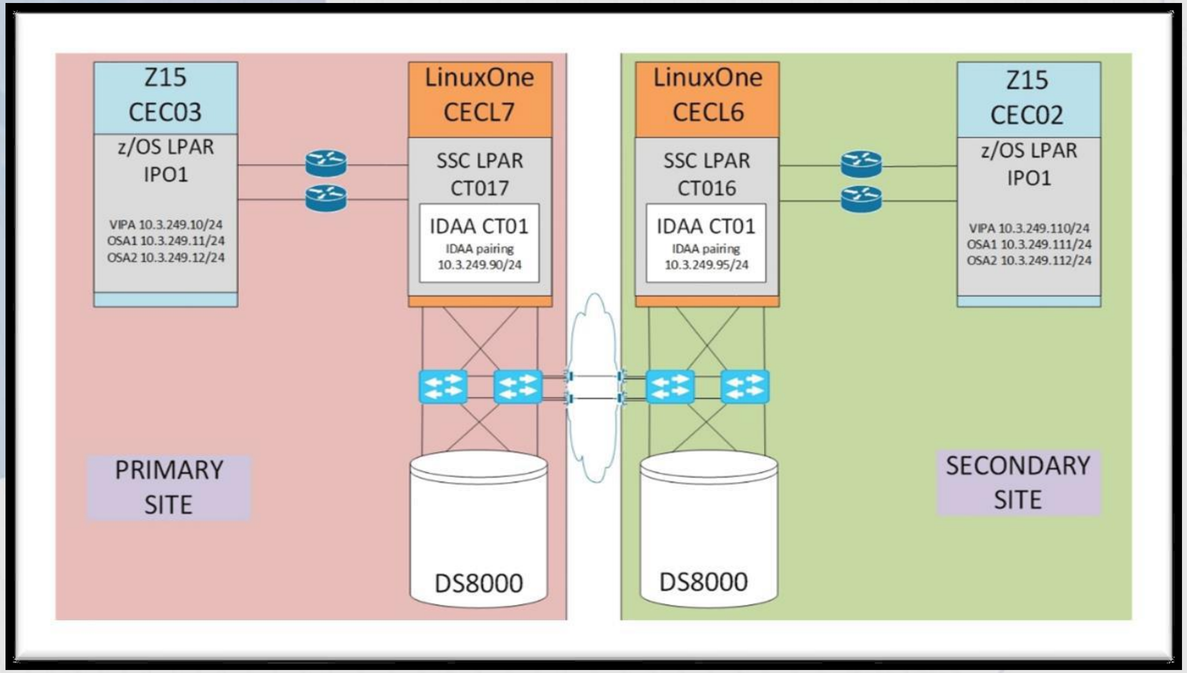
- AccessCount and LastAccessTimeStamp are updated if replication is touching the table or internally touched such as REORG. However, documentation says that only queries will cause the counter to increase. https://www.ibm.com/docs/en/daafz/7.5?topic=procedures-sysprocaccel-get-tables-info

- the documentation is correct, but the implementation was incorrect. The implementation is corrected with 7.5.12.2

Automate failover: Move Db2 pairing IP address into runtime environment i JSON file

Environment

- IDAA LPAR CT017 on CECL7 and IDAA LPAR CT016 on CECL6
- In a normal situation, only the primary site is active, the secondary site is inactive.
- The secondary site is activated in 2 situations
  1. DR Situation: In case of failure on the primary site, the secondary site is activated while the primary site is stopped.
  2. **Application Test Situation**: App test is designed to test a copy of the production configuration on the secondary site. In this case, both primary and secondary sites are active at the same time.



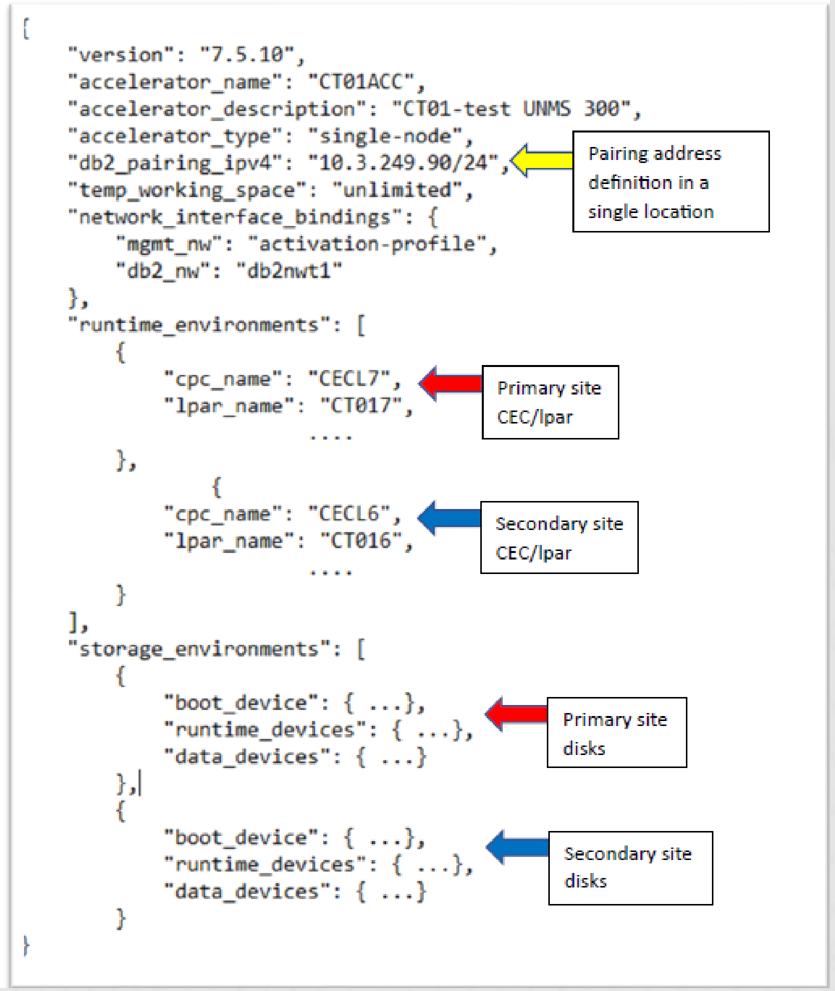
In this case, both primary and secondary sites are active at the same time.

- The disk replication is removed to have 2 sets of usable disks.
- A copy of production z/OS IPO1 is started on CEC02
- A copy of the CT01 IDAA is started in SSC lpar CT016.

Automate failover: Move Db2 pairing IP address into runtime environment in JSON file

Application Test Considerations

- Since both sites are active at the same time, the IP addresses must be changed on z/OS and on SSC.
    - Changing the z/OS IP address is easily done at startup.
    - Changing the IP address of the SSC (wall IP address or pairing address) is not possible at startup. The pairing IP address is specified in the JSON file as “DB2_pairing_ipv4” and cannot be dynamically changed or duplicated



```
[
    "version": "7.5.10",
    "accelerator_name": "CT01ACC",
    "accelerator_description": "CT01-test UNMS 300",
    "accelerator_type": "single-node",
    "db2_pairing_ipv4": "10.3.249.90/24",
    "temp_working_space": "unlimited",
    "network_interface_bindings": {
        "mgmt_nw": "activation-profile",
        "db2_nw": "db2nwt1"
    },
    "runtime_environments": [
        {
            "cpc_name": "CECL7",
            "lpar_name": "CT017",
                        ....
        },
                {
            "cpc_name": "CECL6",
            "lpar_name": "CT016",
                        ....
        }
    ],
    "storage_environments": [
        {
            "boot_device": { ...},
            "runtime_devices": { ...},
            "data_devices": { ...}
        },
        {
            "boot_device": { ...},
            "runtime_devices": { ...},
            "data_devices": { ...}
        }
}
```

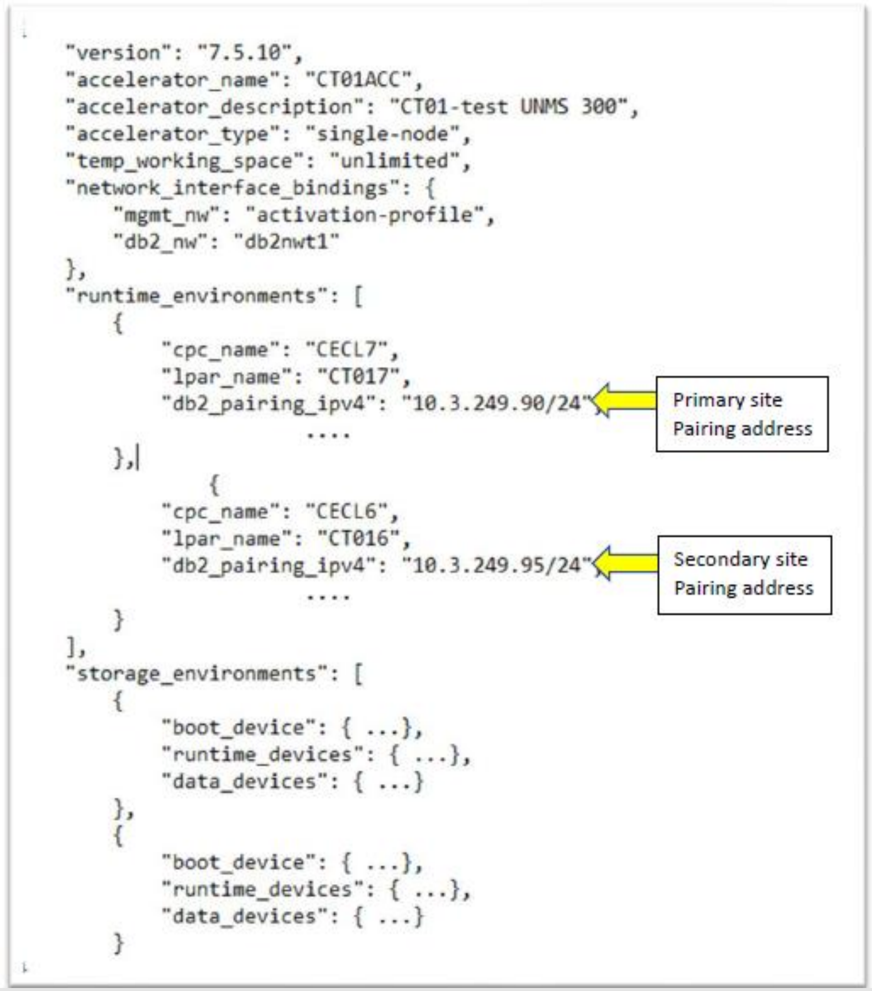
Pairing address definition in a single location

Primary site CEC/lpar

Secondary site CEC/lpar

Primary site disks

Secondary site disks

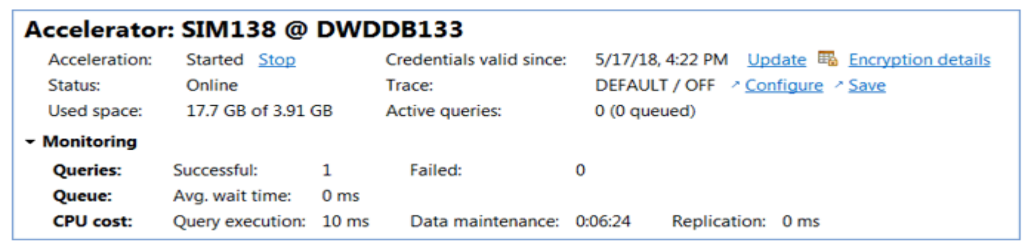


```
    "version": "7.5.10",
    "accelerator_name": "CT01ACC",
    "accelerator_description": "CT01-test UNMS 300",
    "accelerator_type": "single-node",
    "temp_working_space": "unlimited",
    "network_interface_bindings": {
        "mgmt_nw": "activation-profile",
        "db2_nw": "db2nwt1"
    },
    "runtime_environments": [
        {
            "cpc_name": "CECL7",
            "lpar_name": "CT017",
            "db2_pairing_ipv4": "10.3.249.90/24"
                        ....
        },
                {
            "cpc_name": "CECL6",
            "lpar_name": "CT016",
            "db2_pairing_ipv4": "10.3.249.95/24"
                        ....
        }
    ],
    "storage_environments": [
        {
            "boot_device": { ...},
            "runtime_devices": { ...},
            "data_devices": { ...}
        },
        {
            "boot_device": { ...},
            "runtime_devices": { ...},
            "data_devices": { ...}
        }
```

Primary site Pairing address

Secondary site Pairing address

Accelerator on Z Monitoring Enhancements - Health



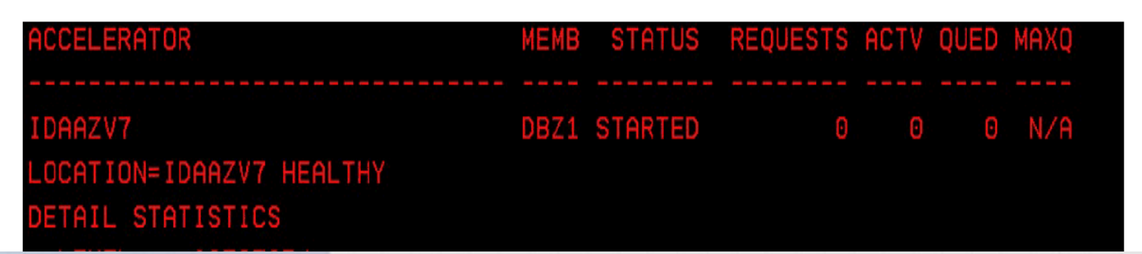
Accelerator: SIM138 @ DWDDB133

Acceleration:	Started Stop	Credentials valid since:	5/17/18, 4:22 PM Update Encryption details
Status:	Online	Trace:	DEFAULT / OFF Configure Save
Used space:	17.7 GB of 3.91 GB	Active queries:	0 (0 queued)

▾ **Monitoring**

Queries:	Successful:	1	Failed:	0			
Queue:	Avg. wait time:	0 ms					
CPU cost:	Query execution:	10 ms	Data maintenance:	0:06:24	Replication:	0 ms	

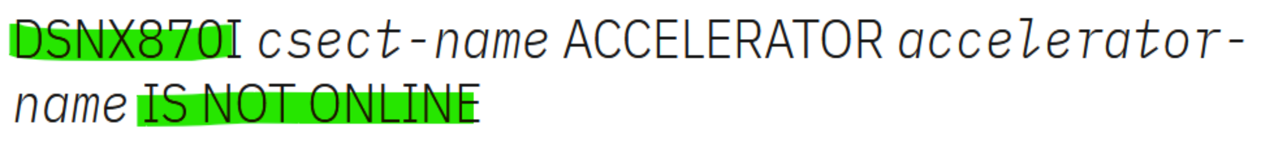
Data Studio - Status



```
ACCELERATOR                    MEMB  STATUS   REQUESTS ACTV QUED MAXQ
------------------------------ ---- -------- -------- ---- ---- ----
IDAAZV7                        DBZ1 STARTED         0    0    0  N/A
LOCATION=IDAAZV7 HEALTHY
DETAIL STATISTICS
```

The DISPLAY ACCEL command

https://www.ibm.com/docs/en/db2-for-zos/12?topic=accelerators-display-accel-db2



DSNX870I *csect-name* ACCELERATOR *accelerator-name* IS NOT ONLINE

Through DSNX870I https://www.ibm.com/docs/en/db2-for-zos/13?topic=messages-dsnx870i

Through DSNX881I https://www.ibm.com/support/pages/node/5694807

Problem & Enhancement: A Linux kernel crash causes on outage of all data nodes of a multi-node accelerator, with no corresponding DSNX881I or DSNX870I message occurring in DB2MSTR, the system indicates a HEALTHY status of the accelerator although there is a crash.

APAR PH59061 fixes for this issue with Accelerator maintenance level 7.5.12.2.

Accelerator on Z Monitoring Enhancements – Monitoring AOTs

Problem

Current monitoring features may not provide key information for specific issues. Adding additional monitoring in Accelerator could take long time and/or deliver

Monitoring Using AOTs

- Easy to deploy, not release bound
- Populates on an interval determined by the user, i.e. every 10 minutes.
- AOT is accessible to the user to view results and do cleanup.

ACCEL_ACTIVITY_TABLE AOT

- Provides utilized resources in Db2wh by all Accelerator operations: load, query, Accelerator server monitoring, etc.
- Includes: temp space, IO, CPU, and memory usage

ACCEL_ACTIVITY_TABLE

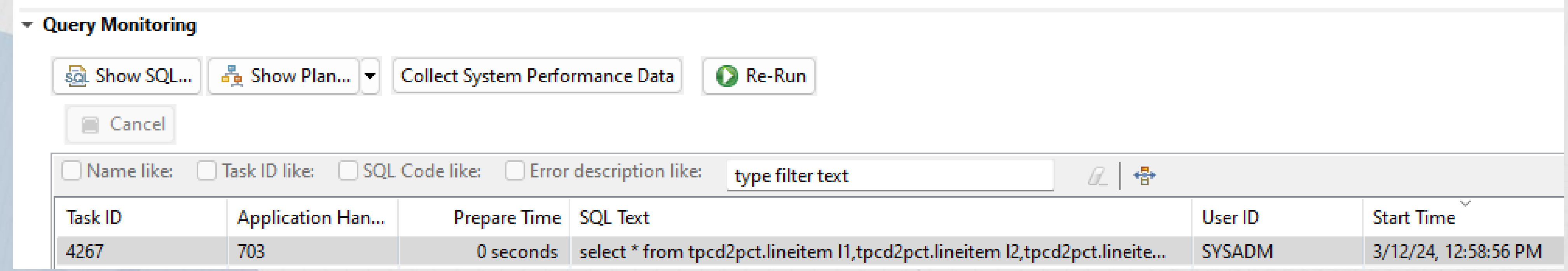
```
CREATE TABLE <creator>.ACCEL_ACTIVITY_TABLE (
ACTIVITY_TIMESTAMP TIMESTAMP NOT NULL,
APPLICATION_HANDLE BIGINT NOT NULL,
```

>> The application handle value in the AOT maps to Query Monitoring in DataStudio, Identify and possibly cancel queries consuming too much CPU, memory, etc.

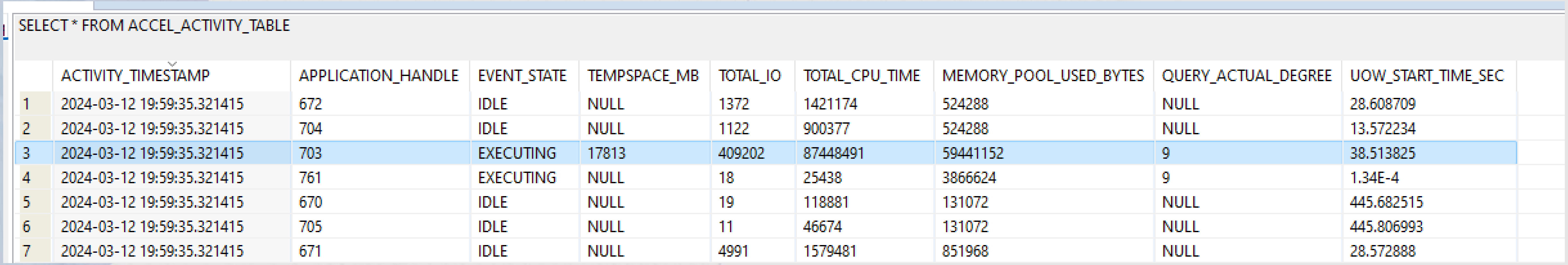
```
EVENT_STATE VARCHAR(32),
TEMPSPACE_MB BIGINT,
TOTAL_IO BIGINT,
TOTAL_CPU_TIME BIGINT,
MEMORY_POOL_USED_BYTES BIGINT,
QUERY_ACTUAL_DEGREE INTEGER,
UOW_START_TIME_SEC DOUBLE)
IN DATABASE "<dbname>"
CCSID UNICODE IN ACCELERATOR <accel>;
```

Accelerator on Z Monitoring Enhancements – Monitoring AOTs

Using the AOT



Task ID	Application Han...	Prepare Time	SQL Text	User ID	Start Time
4267	703	0 seconds	select * from tpcd2pct.lineitem l1,tpcd2pct.lineitem l2,tpcd2pct.lineite...	SYSADM	3/12/24, 12:58:56 PM



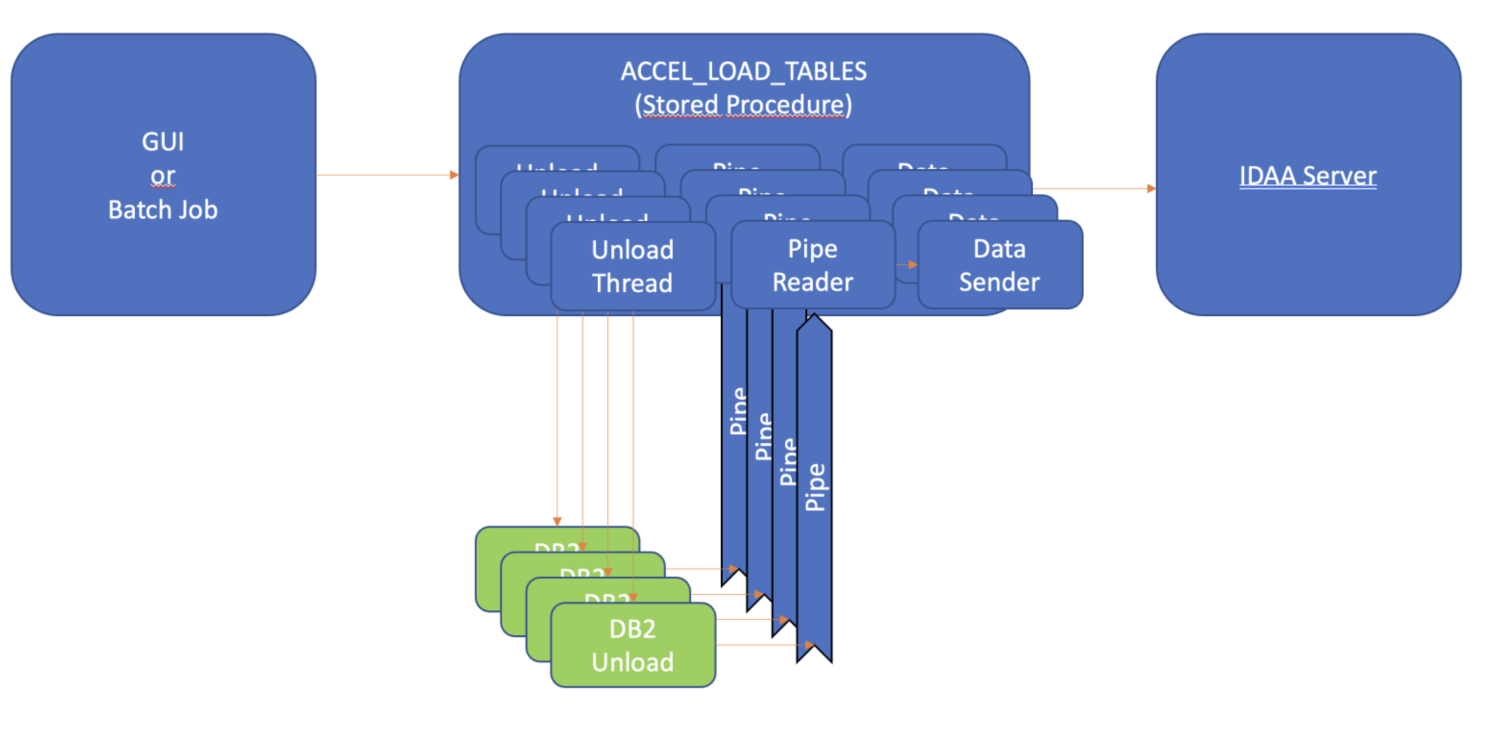
```
SELECT * FROM ACCEL_ACTIVITY_TABLE
```

	ACTIVITY_TIMESTAMP	APPLICATION_HANDLE	EVENT_STATE	TEMPSPACE_MB	TOTAL_IO	TOTAL_CPU_TIME	MEMORY_POOL_USED_BYTES	QUERY_ACTUAL_DEGREE	UOW_START_TIME_SEC
1	2024-03-12 19:59:35.321415	672	IDLE	NULL	1372	1421174	524288	NULL	28.608709
2	2024-03-12 19:59:35.321415	704	IDLE	NULL	1122	900377	524288	NULL	13.572234
3	2024-03-12 19:59:35.321415	703	EXECUTING	17813	409202	87448491	59441152	9	38.513825
4	2024-03-12 19:59:35.321415	761	EXECUTING	NULL	18	25438	3866624	9	1.34E-4
5	2024-03-12 19:59:35.321415	670	IDLE	NULL	19	118881	131072	NULL	445.682515
6	2024-03-12 19:59:35.321415	705	IDLE	NULL	11	46674	131072	NULL	445.806993
7	2024-03-12 19:59:35.321415	671	IDLE	NULL	4991	1579481	851968	NULL	28.572888

Accelerator on Z – Anatomy of Accelerator Load

How ACCEL_LOAD_TABLES works?

Within the ACCEL_LOAD_TABLES stored procedure, there are three threads of interest per partition that is being loaded.
1. The UNLOAD Utility thread, which calls DSNUTILU (UNLOAD)
2. The Pipe Reader thread which reads the data from the UNLOAD utility (via pipe in USS) and puts the data into a buffer (to compensate for throughput changes).
3. The Data Sender thread, which reads the data from the buffer and sent it over the network to the accelerator.



- AQT_MAX_UNLOAD_IN_PARALLEL specifies how many partitions can be unloaded in parallel (4 is default).
- The total number of threads in an ACCEL_LOAD_TABLES invocation is about (AQT_MAX_UNLOAD_IN_PARALLEL *3)

Accelerator on Z – Anatomy of Accelerator Load

How ACCEL_LOAD_TABLES works?

- Data is loaded or refreshed using ACCEL_LOAD_TABLES stored procedure
- Non-partititioned tables and table partitions are loaded in parallel
- Max number of tables and partitions loaded in parallel determined by AQT_MAX_UNLOAD_IN_PARALLEL environment variable (load streams)

- AQT_MAX_UNLOAD_IN_PARALLEL (default 4) applies to each stored procedure call. It is always per ACCEL_LOAD_TABLES call (per stored procedure).

- Let’s assume, five ACCEL_LOAD_TABLES are running at any time. There can be up to 5*4 = 20 UNLOAD utilities running in parallel. If you’re using one table per SP call, even if you run 5 SPs in parallel, parallelism for UNLOAD will be max 4 (unless all input tables are non-partitioned)

- The Accelerator checks the number of all load requests coming in in parallel (from one subsystem or multiple connected subsystems) and dependent on resource consumption of load requests, the percentage of resource consumption per subsystem and the max number of parallel load streams per subsystem will be adjusted
  - All tables or partitions of tables specified in one ACCEL_LOAD_TABLES call are efficiently loaded in parallel. Load streams are efficiently used
  - Reduces the need for customers to manage load steams efficiently by calling ACCEL_LOAD_TABLES in parallel
  - Ensures that Accelerator manages the resources used for loading data proactively to prevent resource bottlenecks and overloading

Accelerator on Z – Anatomy of Accelerator Load

How to optimize Loading

- Let Accelerator optimize usage of parallel insert streams, for instance
  - Keep AQT_MAX_UNLOAD_IN_PARALLEL=4 setting
  - Reduce the number of parallel load jobs running, ideally to 5
    - During Performance test we got best number with about 20 parallel load streams across tables.
  - Group the tables to be loaded into buckets and provide list of tables per job. In each job, do one ACCEL_LOAD_TABLES call and give a list of tables.
  - Error Handling for loading multipe tables in a single SP call is documented here https://www.ibm.com/docs/en/daafz/7.5?topic=procedures-sysprocaccel-load-tables

US Customer example

- Total 92 Tables that would need to be loaded
- AQT_MAX_UNLOAD_IN_PARALLEL=8
- 2 Jobs with 1 SP call per job.
- 46 Tables per Job

ACCEL_LOAD_TABLES to load multiple tables. Like this, yes?

```
<?xml version="1.0" encoding="UTF-8" ?
>
<aqt:tableSetForLoad
xmlns:aqt=http://www.ibm.com/xmlns/prod/dwa/2011 version="1.1">
<table schema="SCHEMA1" name="TABLE1“ />
<table schema="SCHEMA1" name="TABLE2“ />
<table schema="SCHEMA2" name="TABLE1"/>
... <table schema="SCHEMA10" name="TABLE15"
/>
</aqt:tableSetForLoad>
```

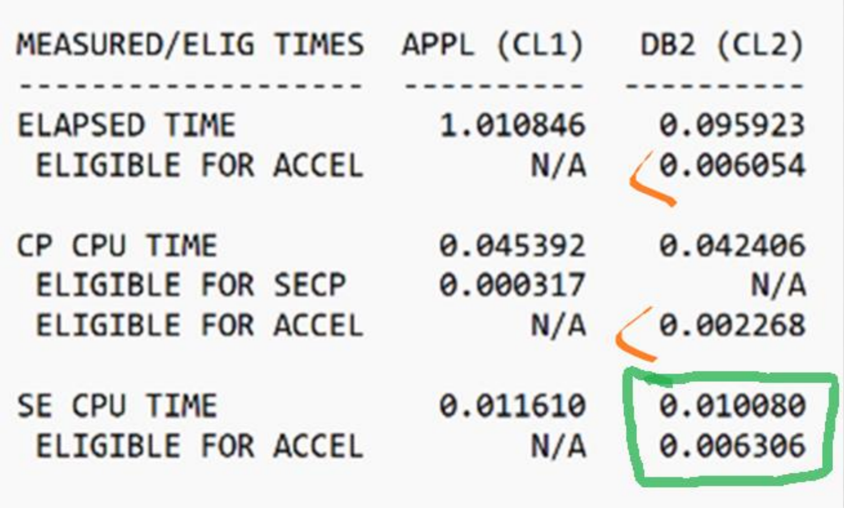
Our Suggestion to increase the parallelism

- Do not use one table per SP call approach.
- Add more tables per SP call and give a chance to ACCEL_LOAD_TABLES to enable parallel UNLOADs

Q&A’s from Webcast on April 30

IBM Webcast - April 30 @ 11:00 AM ET
A Db2 for z/OS must-have... IBM Db2 Analytics Accelerator
https://event.on24.com/wcc/r/4544573/4EC37E39BC0911EDEF0B2389E6049163

Q – Is zIIP offload looked at for assessment or just elapsed time & GP CPU usage?



```
MEASURED/ELIG TIMES  APPL (CL1)   DB2 (CL2)
-------------------  ----------  ----------
ELAPSED TIME           1.010846    0.095923
 ELIGIBLE FOR ACCEL         N/A    0.006054

CP CPU TIME            0.045392    0.042406
 ELIGIBLE FOR SECP     0.000317         N/A
 ELIGIBLE FOR ACCEL         N/A    0.002268

SE CPU TIME            0.011610    0.010080
 ELIGIBLE FOR ACCEL         N/A    0.006306
```

DB2 (CL2) - SE CPU TIME : The accumulated and consumed class 2 time on an IBM specialty engine (SE) that consists of times for non-nested, stored procedures, user-defined functions, triggers, and parallel tasks.

DB2 (CL2) - SE CPU TIME - ELIGIBLE FOR ACCEL : The accumulated CPU time consumed on an IBM specialty engine while processing SQL in DB2 that may be eligible for execution on an accelerator.

Q&A’s from Webcast on April 30

Q – Is the NVMe storage capable of being replicated for HA purposes? Or is it available on the local accelerator only?

A - NVMe storage of IBM® LinuxONE system is only used as transient storage for temporary files such as

- Writing temporary results of extensive sort operations during query processing that cannot be executed exclusively in the system memory.
- Writing replication spill queues when a replication-enabled table is loaded to the accelerator and at the same time new data changes on this table are replicated to the accelerator.
- Writing query results that tend to arrive faster than they can be picked up by the receiving client.

- Transient Storage does not keep persistent user data.
- There is no need to replicate NVMe (transient storage data) for HA purposes.

Q – Which Db2 for z/OS bufferpool is used in AOT retrieval?

A - There could be a small to zero BP hit during prepare of the query but it is negligible. For the actual fetching of results there is no hit to Db2 for z/OS Bufferpool, It's just like a network connection pulling data from Db2WH -> IDAA-> Db2 for z/OS DDF -> Application

Q&A’s from Webcast on April 30

Q – If we are using DRDA 3 part naming, how to route SQL to IDAA or should we use the IDAA federation

A - SELECT * FROM STLEC1.TPCD.T1, STLEC2.TPCD.T2; Does not work. Those SQLs are not routed.

However, Following will work
- on DB2LOCA

```
  CONNECT TO DB2LOCB;
  SET CURRENT QUERY ACCELERATION=ALL;
  SELECT * FROM T1;
```

Q – What is the best practices for large number of AoT backup

Option 1 - Use a DB2 “Staging” Table
- DB2 Cross-Load data from AOT into a normal DB2 Table. For recovery needed LOAD into AOT via IDAA Loader

Option 2 - DB2 Cross-Loader and IDAA Loader
- Cursor accesses AOT1 on IDAA1 and IDAA Loader loads AOT_Backup on IDAA2

Option 3 - Unload to sequential data set + IDAA Loader
- Backup with DSNTIAUL and use standard IDAA Loader for restore

Option 4 - Unload to sequential data set + IDAA Loader
- IDAA Loader Backup and Restore

Db2 Analytics Accelerator and Loader vNext in Q4 2024

🔵 → planned for Accelerator
🟢 → Planned for Loader

New workloads

LOB Support 🔵 🟢

vNext is planned to support adding LOB data into accelerator. The LOB date can be consumed with query acceleration.

Copy from Accl2Accl 🔵

vNext will provide the ability to copy tables between accelerators. In vNext it will be possible to copy AoTs between accelerators without running the AoT populate scripts in each accelerator .

Monitoring

AOT tables for monitoring 🔵

AOT tables can be used to store performance monitoring data which will improve the system status analysis process and give the ability to have a proactive actions. Having it in an AOT will ease the way of retrieving the needed information as a normal query.

Hardware Monitoring 🔵

Hardware monitoring such as storage, NVMes, etc. is considered to be one of the main focus elements to be measured and monitored.

Integration with UI 🔵 🟢

The monitoring journey is planned to have its effect on the UI to be retrieved and integrated in a convenient way with our UI.

External health monitoring 🟢

Monitor the load process of the external use cases: IMS, VSAM , SMF, etc. which provide an overview about the load status as well as giving the ability to have an automated proactive actions.

Db2 Analytics Accelerator and Loader vNext in Q4 2024

🔵 → planned for Accelerator
🟢 → Planned for Loader

User experience

Smooth user experience with Admin Foundation 🔵 🟢

Admin foundation is the strategic GUI for Db2 Analytics Accelerator vNext. The vNext is considered to deliver a smooth user experience with Admin foundation through the admin services that offers more than a functional parity with its former alternatives as Data Studio. We are targeting functional parity for the Loader with Admin foundation.

JSON Configuration file tracking 🔵

Changing the JSON configurations including the key baseline parameters and the software maintenance levels made the traceability too difficult on the user side. The vNext will provide access on older versions of the configuration's file.

Highly voted Ideas

Eligible with Failback 🔵

The Eligible option has been favored by different customers. Although it's effectiveness but users have voted to have a new option "Eligible with Failback"

Automate change Insync IP in batching process 🔵

As a better step in automation and reducing the manual steps, The vNext supposed to have an automation on changing the IP of Insync in the batching process without the need to change manually in the admin console.

Db2 Analytics Accelerator Focus Strategy

Where?

- IBM Z and Sailfish → **Focus on (IFLs) IBM Z and IBM LinuxONE**
- Smooth migration from version to version on IFLs
- Current v7 users will be entitled to move to vNext

How?

- Admin foundation and Data studio (EOS March 2025) → **Focus on Admin Foundation**
- Functional Parity between DS and AF for the Accelerator

Replication

Integrated Synchronization and CDC → **Focus on Integrated Synchronization** proved its performance advantage over CDC

IDUG

IDUG

2024 NA **Db2** Tech Conference

Db2 Analytics Accelerator Product Update and Customer Experiences

Mehmet Cüneyt Göksu
Mehmet.Goksu@ibm.com

IBM

@IDUGDb2
#IDUG_NA24

Session Code: DMA1 | Platform: Cross Platform