Scalable Analytics: IBM System z Approach
Disclaimer

© Copyright IBM Corporation 2012. All rights reserved.
U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM, the IBM logo, ibm.com, DB2, and DB2 for z/OS are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (© or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml
Other company, product, or service names may be trademarks or service marks of others.
Enterprise Information Is on System z

66 out of the top 66 banks in the world
9 of the top 10 global life/health insurance providers
24 of the top 25 US retailers

Up to 4.8M connected threads

Zero unplanned downtime for over 10 years

Large U.S. Retailer
World’s largest SAP system

World’s largest known peak RDBMS workload
1.1 Billion SQL statements per hour

620M accounts, 120M online users

World’s largest banking benchmark result, a record 9,445 transactions per second

World’s largest known transaction processing RDBMS >40 TB
Traditional Systems Landscape

Applications

- operational
- analytical

OLTP → Staging Area → ODS → EDW → Data Marts

Historical reasons:
- Different access patterns
  - impact on performance
- EDW as the data integration hub
  - again, impact on performance
- Different life-cycle characteristics
  - and again, impact on performance
- Different Service Level Agreements (SLA)
  - Lack of broadly available workload management capabilities
  - Choice of lower cost-of-acquisition offerings

Negative ramifications:
- Complexity
  - both in systems management and in applications
- Difficulties in supporting real time analytics
- Inability to match ever more demanding SLA requirements
- High total cost of ownership
Road to Visionary Systems Landscape

Applications
- operational
- analytical

OLTP — Staging Area — ODS — EDW — Data Marts

Benefits
- Uniform policies and procedures for security, HA, DR, monitoring, same tools, same skills, ...
- Efficient data movement within the system, often not involving network (ELT vs. ETL)
- Uniform access to any data for types of applications
- Opportunity to remove, i.e. consolidate some of the layers, ultimately leading to a single database
**Universal DBMS**

**Applications**
- operational
- analytical

**Data**

**Benefits**
- Uniform policies and procedures for security, HA, DR, monitoring, same tools, same skills, ...
- Efficient data movement within the system, often not involving network (ELT vs. ETL)
- Uniform access to any data for types of applications
- **Opportunity to remove, i.e. consolidate some of the layers, ultimately leading to a single database**

**Challenges**
- Mixed workload management capabilities
- Ensuring continuous availability, security and reliability
- Providing seamless scale-up and scale-out
- Providing universal processing capabilities to deliver best performance for both transactional and analytical workloads without the need for excessive tuning

**Approaches**
- Columnar stores
- In-memory databases
- Hardware acceleration, special purpose processors
- Appliances

**Building on proven technology base**
- System z Data Sharing and Parallel Sysplex technology provides all the needed characteristics except one:
  - Special purpose processing for analytical workloads to minimize the need for manual tuning
DB2 Components

Applications
Application Interfaces
(standard SQL dialects)

DBA Tools, z/OS Console, ...
Operation Interfaces
(e.g. DB2 Commands)

DB2

Data Manager
Buffer Manager
IRLM
Log Manager
IBM DB2 Analytics Accelerator as a Virtual DB2 Component

<table>
<thead>
<tr>
<th>Applications</th>
<th>DBA Tools, z/OS Console, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Interfaces (standard SQL dialects)</td>
<td>Operation Interfaces (e.g. DB2 Commands)</td>
</tr>
</tbody>
</table>

DB2

- Data Manager
- Buffer Manager
- IRLM
- Log Manager
- Accelerator
DB2 Becomes a Hybrid Database Management System

Uniform and transparent access for transactional and analytical applications.

Applications
- Application Interfaces (standard SQL dialects)

DBA Tools, z/OS Console, ...
- Operation Interfaces (e.g. DB2 Commands)

DB2

- Data Manager
- Buffer Manager
- IRLM
- Log Manager
- IBM DB2 Analytics Accelerator

System z
Superior availability, reliability, security, workload management, OLTP performance ...

Powered by Netezza
True appliance, Industry leading ease of performance
Query Execution Process Flow

Application Interface

Optimizer

IDAA DRDA Requestor

Heartbeat

SPU
CPU
FPGA
Memory

CPU
FPGA
Memory

CPU
FPGA
Memory

CPU
FPGA
Memory

DB2 for z/OS

Queries executed without IDAA
Queries executed with IDAA
Heartbeat (IDAA availability and performance indicators)
## Synchronization Options

<table>
<thead>
<tr>
<th>Synchronization options</th>
<th>Use cases, characteristics and requirements</th>
</tr>
</thead>
</table>
| **Full table refresh**      | ▪ Existing ETL process replaces entire table  
                              ▪ Multiple sources or complex transformations  
                              ▪ Smaller, un-partitioned tables  
                              ▪ Reporting based on consistent snapshot |
| **Table partition refresh** | ▪ Optimization for partitioned warehouse tables, typically appending changes “at the end”  
                              ▪ More efficient than full table refresh for larger tables  
                              ▪ Reporting based on consistent snapshot |
| **Incremental update**      | ▪ Scattered updates after “bulk” load  
                              ▪ Reporting on continuously updated data (e.g., an ODS), considering most recent changes  
                              ▪ More efficient for smaller updates than full table refresh |

Full table refresh
The entire content of a database table is refreshed for accelerator processing

Table partition refresh
For a partitioned database table, selected partitions can be refreshed for accelerator processing

Incremental update
Log-based capturing of changes and propagation to the accelerator with low latency (typically few minutes)
Initial Data Load and Refresh

DB2 for z/OS

Table A
- Part 1
- Part 2
- Part m

Table B
- Part 1
- Part 2
- Part 3

Table C
- Part 1
- Part 2
- Part 3

Table D
- Part 1
- Part 2
- Part 3

Explicit Invocation (e.g. by Job Scheduler)

Administrative Stored Procedures

UNLOAD 'lite'

USS Pipe

IBM DB2 Analytics Accelerator

SPU
- CPU
- FPGA
- Memory

SPU
- CPU
- FPGA
- Memory

SPU
- CPU
- FPGA
- Memory

SPU
- CPU
- FPGA
- Memory

Explicit Invocation (e.g. by Job Scheduler)
Incremental Updated Architecture

DB2 for z/OS

- Accelerator Stored Procedures
  - ACCEL_SET_TABLES_REPLICATION
  - ACCEL_CONTROL_ACCELERATOR
- JCL
- Capture Agent DB2 z/OS (Log reading)

NPS host

- Accelerator SERVER
  - Controller
  - Catalog information

Netezza Database

Accelerator SERVER

- Automation code
  - (creates data sources, subscriptions, etc.)
- Apply Agent on NPS host
  - (Receives log events)

Accelerator Studio

(insert, delete, update)

(private network)
Connectivity Options

Multiple DB2 systems can connect to a single IDAA

A single DB2 system can connect to multiple IDAAs

Multiple DB2 systems can connect to multiple IDAAs

Better utilization of IDAA resources
Scalability
High availability

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
Disaster Recovery Configuration Example

Prior to disaster

Site A
- System z
  - DB2
    - Member 1
  - CF
- Accelerator 1
  - Tab 1
  - Tab 2
  - Tab 3

Site B
- System z
  - DB2
    - Member 2
  - CF
- Accelerator 2
  - Tab 1
  - Tab 2
  - Tab 3

Synchronous replication between Site A and Site B.
Disaster Recovery Configuration Example

Disaster happens

Site A → Site B

Synchronous replication

System z

DB2

Member 1

CF

Accelerator 1

Tab 1

Tab 2

Tab 3

Tab 4

Tab 5

DB2

Member 2

CF

Accelerator 2

Tab 1

Tab 2

Tab 3

Tab 4

Tab 5
Disaster Recovery Configuration Example

After disaster

Site A
- System z
- Tab 1
- Tab 2
- Tab 3
- Tab 5
- Tab 4
- Accelerator 1
  - Tab 1
  - Tab 2
  - Tab 3

Site B
- System z
- Tab 1
- Tab 2
- Tab 3
- Tab 5
- Tab 4
- DB2
  - Member 2
- CF
- Accelerator 2
  - Tab 1
  - Tab 2
  - Tab 3
High Performance Storage Saver

- Most of the data in an ODS or EDW is static
  - The large tables are partitioned by time
  - Older partitions are never changed
  - The most recent partition is frequently changed
- Many DBMS vendors provide multi-temperature (MT) data solutions
  - MT concept is not the same as archiving, but they have lots in common
  - Level of sophistication in implementing MT varies
    - the industry leading solutions are so called 'near-line storage servers'
    - 'near-line' means 'near-online'
- MT's value proposition is twofold:
  - Move less frequently accessed data to cheaper storage
  - Improve performance for both queries and administrative operations accessing more recent data
- The drawback is degraded performance of queries that access old data
- Better solution is needed if the query access pattern includes both
  - Transactional, i.e. accessing limited amount of data, predominantly from the most recent partition, and
  - Analytical, i.e. accessing large amount of data across all the partitions
- IBM DB2 Analytics Accelerator can offer such a solution
  - High Performance Storage Saver
  - 'Online Storage Server' as opposed to 'Near-line Storage Server'
  - Netezza provides very large disk capacity at a fraction of cost of the System z disk subsystem
  - IBM DB2 Analytics Accelerator technology provides the basis for access to data irrespective of where they reside (on DB2 disks or Netezza disks)
Use Cases:

- **Workload Isolation**
  Ensure that the workload of one DB2 subsystem does not monopolize the resources of a shared accelerator. A development subsystem, attached to the same accelerator as a production subsystem, should not be able to drain all accelerator resources.

- **Query Prioritization**
  More important queries should be executed before and faster than less important queries that are sent from the same DB2 subsystem against the accelerator.
Routing Control Knobs

Special register CURRENT QUERY ACCELERATION

- Can be set implicitly by inheriting the value of the system parameter, or
- Explicitly by SET CURRENT QUERY ACCELERATION

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONE</td>
<td>No query is routed to the accelerator</td>
</tr>
<tr>
<td>ENABLE</td>
<td>A query is routed to the accelerator if it satisfies the acceleration criteria including the cost and heuristics criteria. Otherwise, it is executed in DB2. If there is an accelerator failure while running the query, or the accelerator returns an error, DB2 will return a negative SQL code to the application.</td>
</tr>
<tr>
<td>ENABLE WITH FAILBACK</td>
<td>A query is routed to the accelerator if it satisfies the acceleration criteria including the cost and heuristics criteria. Otherwise, it is executed in DB2. Under certain conditions the query will run on DB2 after it fails in the accelerator. In particular, any negative SQLCODE will cause a failback to DB2 during PREPARE or first OPEN. No failback is possible after a successful OPEN of a query.</td>
</tr>
<tr>
<td>ELIGIBLE</td>
<td>A query is routed to the accelerator if it satisfies the acceleration criteria irrespective of the cost and heuristics criteria. Otherwise, it is executed in DB2.</td>
</tr>
<tr>
<td>ALL</td>
<td>A query is routed to the accelerator. If it cannot be executed there, the query fails and a negative return code is passed back to the application.</td>
</tr>
</tbody>
</table>
Disk Enclosures

Slice of User Data
Swap and Mirror partitions
High speed data streaming
High compression rate
EXP3000 JBOD Enclosures
12 x 3.5" 1TB, 7200RPM, SAS (3Gb/s)
max 116MB/s (200-500MB/s compressed data)
e.g. in model 1000-12:
8 enclosures → 96 HDDs
32TB uncompressed user data (→ 128TB)

IDAA Server
SQL Compiler, Query Plan, Optimize Administration
2 front/end hosts, IBM 3650M3 or 3850X5 clustered active-passive
2 Nehalem-EP Quad-core 2.4GHz per host

Processor & streaming DB logic
High-performance database engine streaming joins, aggregations, sorts, etc.
e.g. in model 1000-12: 12 back/end SPUs
(more details on following charts)

SMP Hosts

Snippet Blades™ (S-Blades, SPUs)

Powered by Netezza
### Scaling Out

<table>
<thead>
<tr>
<th>Model</th>
<th>1000-3</th>
<th>1000-6</th>
<th>1000-12</th>
<th>1000-24</th>
<th>1000-36</th>
<th>1000-48</th>
<th>1000-72</th>
<th>1000-96</th>
<th>1000-120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cabinets</td>
<td>1/4</td>
<td>1/2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Processing Units</td>
<td>24</td>
<td>48</td>
<td>96</td>
<td>192</td>
<td>288</td>
<td>384</td>
<td>576</td>
<td>768</td>
<td>960</td>
</tr>
<tr>
<td>Capacity (TB)</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>96</td>
<td>128</td>
<td>192</td>
<td>256</td>
<td>320</td>
</tr>
<tr>
<td>Effective Capacity (TB)</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>384</td>
<td>512</td>
<td>768</td>
<td>1024</td>
<td>1280</td>
</tr>
</tbody>
</table>

**Capacity** = User data space

**Effective Capacity** = User data space with compression (4x compression assumed)