ORACLE

Why Exadata is the Best Platform for Database In-Memory

August 2021 | Version 5.0

PURPOSE STATEMENT

This document provides an overview of features and enhancements when using Database In-Memory with Exadata. It is intended solely to help you assess the business benefits of using Database In-Memory with Exadata and to plan your I.T. projects.

DISCLAIMER

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

INTRODUCTION

Oracle Database In-Memory (Database In-Memory) transparently accelerates analytic queries by orders of magnitude, enabling real-time business decisions. Database In-Memory uses a "dual-format" architecture that enables data to be maintained in both row format and a pure in-memory columnar format. This columnar format allows data to be scanned much faster than row formatted data. Database In-Memory is able to further speed up scan performance by taking advantage of SIMD (Single Instruction, Multiple Data) vector processing and In-Memory Storage Indexes. With Database In-Memory it is possible to scan billions of rows per processor core per second purely in-memory, and this now makes it feasible for businesses to run real-time analytics on their critical business data without impacting the performance of their existing systems.

With the benefits of Database In-Memory, does it matter what platform you run your database on? Yes, the Oracle Exadata Database Machine (Exadata) has been the preferred platform for running Oracle Database since its release in 2008, and it provides distinct advantages for running Database In-Memory as well. The following are key advantages that Exadata uses with Database In-Memory:

- Exadata efficiently scales Database In-Memory
- Exadata provides a very fast interconnect with special protocols to speed up Database In-Memory scale-out
- Exadata provides high storage bandwidth to quickly populate the Database In-Memory column store
- In-Memory columnar formats in flash cache
- Exceed DRAM limits and transparently scale across Memory, Flash and Disk
- Exadata is Oracle's Database In-Memory development platform
- Exadata is a database consolidation platform and Database In-Memory further enables consolidation opportunities
- In-Memory fault tolerance
- In-Memory Aggregation optimization can be offloaded to Exadata storage cells
- Database In-Memory support for Active Data guard only on Exadata
- Column-level decryption and decompression greatly reduces storage CPU and IO

In this paper we will examine each of these points and explain in detail why Exadata is the best platform for running Database In-Memory.

EXADATA EFFICIENTLY SCALES DATABASE IN-MEMORY

Exadata uses a scale-out architecture for both database servers and storage servers. The Exadata configuration carefully balances CPU, I/O and network throughput to avoid bottlenecks. As an Exadata system grows, database CPUs, storage, and networking are added in a balanced fashion ensuring scalability without bottlenecks. This scale-out architecture can accommodate any size workload and allows seamless expansion from small to extremely large configurations while avoiding performance bottlenecks and single points of failure.

In a Real Application Clusters (RAC) environment, objects with the INMEMORY attribute specified can be distributed across the cluster by rowid range, by partition or by subpartition. Exadata is architected to accommodate the increased parallelism and interconnect messaging when the In-Memory column store (IM column store) is distributed across multiple RAC nodes.

EXADATA PROVIDES A VERY FAST INTERCONNECT WITH SPECIAL PROTOCOLS TO SPEED UP DATABASE IN-MEMORY SCALE OUT

Exadata uses an InfiniBand interconnect between the database servers and storage servers. Each InfiniBand link provides 40 Gb/second of bandwidth – many times higher than traditional storage or server networks. Further, Oracle's interconnect protocol uses direct data placement (DMA – direct memory access) to ensure very low CPU overhead by directly moving data from the wire to database buffers with no extra data copies. The InfiniBand network has the flexibility of a LAN network, with the efficiency of a SAN. By using an InfiniBand network, Exadata ensures that the network will not bottleneck performance. The same InfiniBand network also provides a high-performance cluster interconnect for RAC nodes. When scaling out Database In-Memory on Exadata this high-speed transfer and large bandwidth for messaging between IM column stores keeps the IM column stores transactionally consistent and in sync with each other. This enhances scale out for distributed objects as well as objects that have been duplicated.

EXADATA PROVIDES HIGH STORAGE BANDWIDTH TO QUICKLY POPULATE THE DATABASE IN-MEMORY COLUMN STORE

When data is initially populated into the IM column store it is read directly from disk in its row format, converted to a columnar format and then compressed. The faster you can read the data, the faster you can complete the population process. Exadata storage offers outstanding IO performance ensuring the data population process is not I/O bound.

The population process is conducted by a set of background worker processes. These worker processes can operate in parallel to populate the IM column store as fast as data can be read off disk and CPUs can process that data. This is where the high I/O performance and CPU resources of Exadata come into play to make the population of the IM column store as fast as possible. The number of background worker processes can also be controlled to take further advantage of Exadata's scalability.

Database In-Memory will also repopulate IMCUs when the number of stale entries in an IMCU reaches a staleness threshold. Again, with Exadata's high I/O performance this can occur in the background with no noticeable effect on application performance.

IN-MEMORY COLUMNAR FORMATS IN FLASH CACHE

On Oracle Exadata, the Oracle Database In-Memory feature automatically enables data to be encoded in the Smart Flash Cache of the Storage Servers using the same in-memory columnar formats as those used in the database tier. This flash-based columnar storage increases the total effective columnar capacity of the system to **100s of Terabytes.** This feature is a huge advantage over both traditional in-memory databases (in which all data must be loaded into costly DRAM) and over commodity flash storage arrays (that lack database processing algorithms).

All SQL operations offloaded to storage automatically benefit from this flash-resident in-memory columnar format using *Single Instruction, Multi-Data* (SIMD) vector instructions - the ability to process multiple values in one instruction. In fact, the very same algorithms and optimizations used by the Database In-Memory feature in the database servers are used by the storage servers for flash-resident in-memory column formatted data.

In addition to accelerating scans by **3-5x**, the in-memory columnar format on storage also accelerates many aggregation operations such as GROUP BY, MIN, MAX, SUM, etc., which are also offloaded to the storage tier. These operations are processed via an in-memory aggregation algorithm (known as **vector group by**), leading to an additional performance gain of up to **5x** compared to performing the aggregation in the database tier.

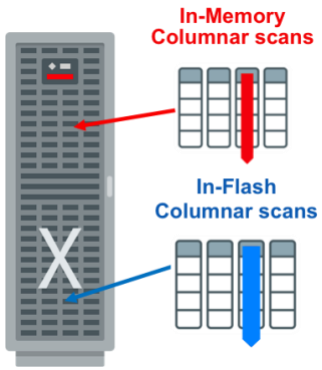

Figure 3. All of the benefit of In-Memory columnar now available on Exadata Flash

EXCEED DRAM LIMITS AND TRANSPARENTLY SCALE ACROSS MEMORY, FLASH AND DISK

With Exadata, your application can make use of all storage tiers (memory, flash, & disk) without having to be aware of where the data resides or suffer suboptimal performance when not all of the data resides in-memory in the IM column store. On Exadata data can reside in the IM column store, in the database buffer cache, in flash storage in columnar or row format, or on disk storage and your application never needs to be aware of data location because Oracle Database can seamlessly access that data.

When data resides on Exadata storage servers, Exadata's Smart Flash Cache feature or In-Memory Column Cache can dramatically accelerate Oracle Database processing by speeding I/O operations. Exadata Smart Flash Cache provides intelligent caching of database objects to avoid physical disk I/O. Exadata storage also provides an advanced compression technology, Hybrid Columnar Compression (HCC), that typically provides 10x level of data compression and boosts the effective data transfer by an order of magnitude.

This means that all data access, and not just data that has been populated into the IM column store in DRAM, will be as efficient as possible.

Exadata also includes Smart Scan, a unique technology that offloads data-intensive SQL operations into the Oracle Exadata storage servers. This is similar to and complements Database In-Memory processing by pushing SQL processing into the Exadata storage servers when data is not in the IM column store. Data filtering and processing occurs immediately and in parallel across all storage servers as data is read from disk or flash. Exadata Smart Scan reduces database server CPU consumption and greatly reduces the amount of data moved between storage and database servers. This enables scaling and efficient SQL processing across all storage tiers whether data resides in the IM column store, on flash storage or on disk storage.

EXADATA IS ORACLE'S DATABASE IN-MEMORY DEVELOPMENT PLATFORM

Exadata is the development platform for Database In-Memory. Thus, Database In-Memory issues are discovered and fixed on Exadata first. Exadata is also the primary platform for Oracle Database testing, HA best practices validation, integration and support. The same reasons it is the best platform for Oracle Database apply to Database In-Memory.

EXADATA IS A DATABASE CONSOLIDATION PLATFORM AND DATABASE IN-MEMORY FURTHER ENABLES CONSOLIDATION OPPORTUNITIES

Database consolidation is one of the major strategies that organizations use to achieve greater efficiencies in their operations. Increasing the utilization of hardware resources while reducing administrative costs are primary goals of consolidation projects. Exadata is optimized for Data Warehouse and OLTP database workloads, and its balanced database server and storage grid infrastructure make it an ideal platform for database consolidation. Exadata is a modern architecture featuring scale-out industry-standard database servers, scale-out intelligent storage servers and a high-bandwidth low-latency InfiniBand network that connects all servers and storage. In many ways Database In-Memory "completes" Exadata by applying in-memory performance techniques that are similar to those that are used by Exadata on flash and disk. Exadata allows customers to simultaneously optimize performance and cost for analytic workloads by using Database In-Memory columnar formats in-memory and flash in conjunction with existing Exadata features to increase consolidation capacity for all data. The result is a solution that gives the speed of DRAM, the IOPs of flash, and the cost effectiveness of disk.

IN-MEMORY FAULT TOLERANCE

The Oracle Database In-Memory feature optimizes data and accesses for column-oriented real-time analytic operations and processing. When Oracle Database In-Memory is deployed on Exadata, the Oracle Database leverages the high speed, high bandwidth RDMA over Converged Ethernet (RoCE) Network Fabric to duplicate objects in the Database In-Memory Column Store across multiple database servers. This means that each In-Memory Compression Unit (IMCU) populated into the IM column store will have a mirrored copy placed on one of the other nodes in the RAC cluster. Mirroring the IMCUs provides in-memory fault tolerance as it ensures data is still accessible via the IM column store even if a node goes down. It also improves performance, as queries can access both the primary and the backup copy of the IMCU at any time. In addition to improving star-schema table join operations, duplicated Column Store objects provide workload fault tolerance for planned and unplanned downtime.

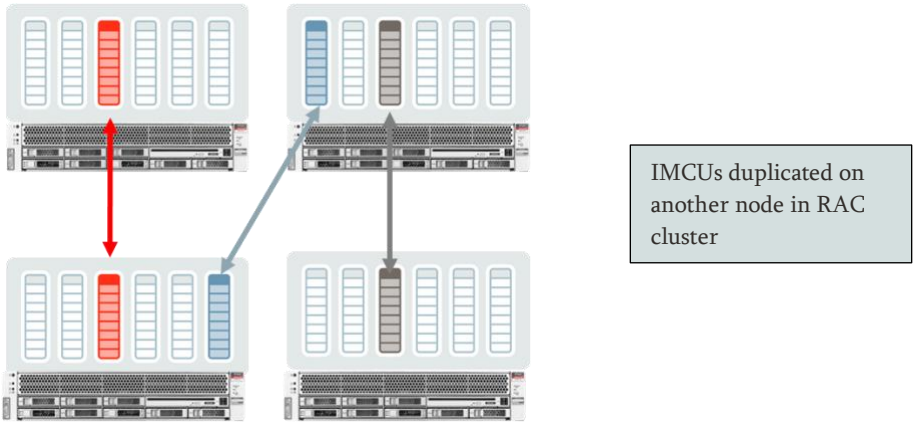

Figure 1. Objects in the IM column store on an Exadata Database Machine can be mirrored to improve fault tolerance

Should a RAC node go down and remain down for some time, the only impact will be the re-mirroring of the primary IMCUs located on that node. Only if a second node were to go down and remain down for some time would the data have to be redistributed.

If additional fault tolerance is desired, it is possible to populate an object into the IM column store on each node in the cluster by specifying the DUPLICATE ALL sub-clause. This will provide the highest level of redundancy and provide linear scalability, as queries will be able to execute completely within a single node.

The DUPLICATE ALL option may also be useful to co-locate joins between large, distributed fact tables and smaller dimension tables. By specifying the DUPLICATE ALL option on the smaller dimension tables a full copy of these tables will be populated into the IM column store on each node. In the example in Figure 2, when a query joins a partition of the sales table to one or more of the dimension tables all the data required for the join will be in the local node, avoiding having to fetch data across nodes to complete the join.

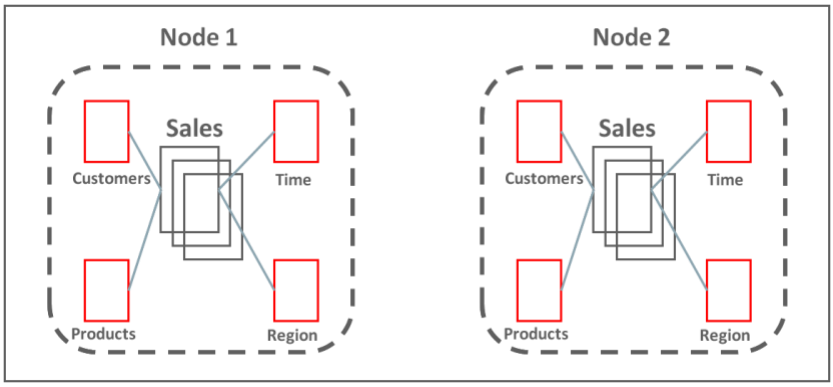

Figure 2. Distributed fact table with duplicated dimension tables

IN-MEMORY AGGREGATION OPTIMIZATION CAN BE OFFLOADED TO EXADATA STORAGE CELLS

With the introduction of Database In-Memory comes the new In-Memory Aggregation optimization, or vector group by feature. In-Memory Aggregation (IMA) provides new SQL execution operations that accelerate the performance of a wide range of analytic queries against star and similar schemas. These include the KEY VECTOR USE and VECTOR GROUP BY operations which enable the use of a vector transformation plan that minimizes the amount of data that must flow through the execution plan. This minimizes the amount of CPU used as compared to alternative plans.

The result of this is that IMA can transform joins to KEY VECTOR filters on the fact table and aggregate data in a single pass while lowering CPU use. This is extremely fast when the entire table resides in the IM column store, but what if the entire table doesn't fit into the IM column store? On Exadata when tables are accessed, and they have not been populated in the IM column store, IMA is enhanced by the ability to offload the KEY VECTOR USE operation to Exadata storage servers. This might occur when the table is partitioned and only the most recent partitions are loaded into the IM column store and the other partitions are on disk. The offload capability distributes key vector processing across Exadata storage servers and minimizes the volume of data that must be returned to the database nodes.

DATABASE IN-MEMORY SUPPORT FOR ACTIVE DATA GUARD ONLY ON EXADATA

Oracle Active Data Guard is the most comprehensive solution available to eliminate single points of failure for mission critical Oracle databases. It prevents data loss and downtime in the simplest and most economical manner by maintaining a synchronized physical replica of a production database at a remote location. If the production database is unavailable for any reason, client connections can quickly, and in some configurations transparently, failover to the synchronized replica to restore service. It also eliminates the high cost of idle redundancy by allowing reporting applications, ad-hoc queries, and data extracts to be offloaded to read-only copies of the production database.

When either the primary or the Active Data Guard standby database is on an Exadata Database machine, Database In-Memory can be used to further accelerate reporting queries on the Active Data Guard standby. As redo is applied to update the Active Data Guard standby, the in-memory column store on the database servers of the standby is automatically maintained consistently, and analytic queries on the standby can benefit from the same in-memory optimizations as they benefit from on the primary database, with a 10-100x performance advantage. Moreover, a different set of tables can be kept in-memory on the standby from the primary database increasing the total effective in-memory columnar capacity of the combined system. Of course, the in-memory format in the smart flash cache is always available on both primary and the standby databases when the Database In-Memory option is enabled.

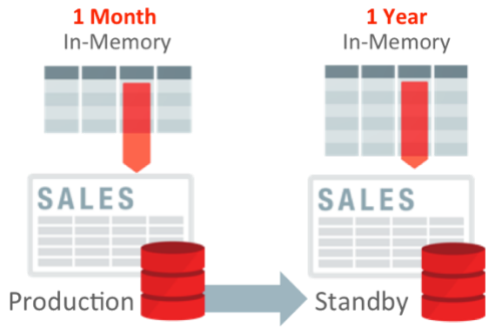

Figure 4. Example of how the IM column store on the standby database can have very different content to the primary

COLUMN-LEVEL DECRYPTION AND DECOMPRESSION GREATLY REDUCES STORAGE CPU AND IO

Rows are stored in columnar representation with both Database In-Memory Column Format and Exadata Hybrid Columnar Compression storage formats. Each column is separately compressed and (optionally) encrypted. Unlike a traditional storage array with encryption/compression support, only the columns referenced by the query need to be decrypted and decompressed, resulting in significant storage CPU savings and I/O savings as much as **10x** or more (depending on the selectivity of the operation). Exadata smart storage software can additionally pipeline the decryption and decompression operations together to get additional performance benefits. None of these benefits can be realized using third-party infrastructure.

CONCLUSION

Oracle Database and Exadata together represent a unique combination of class-leading enterprise hardware and state-of-the-art database software running both on database servers and storage servers, all engineered and optimized together. It is therefore not possible to recreate the combination of capabilities that the Exadata platform provides, using generic servers, networking, and storage.

This makes Exadata the best platform for running Oracle Database and Database In-Memory. Database In-Memory takes full advantage of Exadata's unique hardware features, enabling better performance than any other hardware platform. These features include a very fast interconnect enabling IM fault tolerance and scale-out, high storage bandwidth and IOPs enabling fast IM column store population, seamless access to all storage tiers and the running of mixed workload environments. All of this along with Oracle's commitment to ensuring that all hardware and software components are pre-configured, pre-tuned and pre-tested to work seamlessly together for the best possible performance and reliability in the industry make Exadata the best platform for running Oracle Database and Database In-Memory.

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com facebook.com/oracle twitter.com/oracle

Why Exadata is the Best Platform for Database In-Memory
August, 2021
Author: Andy Rivenes