

Oracle Database 12c for Data Warehousing and Big Data

ORACLE WHITE PAPER | SEPTEMBER 2014







Table of Contents

Introduction	1
Big Data: The evolution of data warehousing	2
Oracle Database 12c and Oracle Exadata:	
A Data Warehouse as a Foundation for Big Data	3
Exadata	3
Oracle Database In-Memory	4
Query Performance	4
Data Management	5
Partitioning	5
Compression	6
Read Consistency and Online operations	6
Analytics	7
SQL Extensions for Analytics	7
Advanced Analytics	7
OLAP	8
Conclusion	8



Introduction

What is a data warehouse? Quite simply, a data warehouse is a database built for the purposes of analysis. “Data warehouses” encompass a huge range of applications today, from large-scale advanced analytical data stores supporting dozens of sophisticated analysts to pre-built business intelligence applications with tens of thousands of users, and from enterprise-wide data warehouses to departmental data marts. Data warehouses are now a mainstay of the IT infrastructure, enabling both long-term strategic planning and agile responses to new market conditions.

However, data warehousing is undergoing a major transition. The benefits of data warehouses are currently being realized in most organizations, partially if not wholly. The best practices for data warehouses are well-established, and the technology supporting data warehouses is becoming more and more mature. Today’s leading-edge organizations, seeking to further differentiate themselves through analytics, are expanding their data warehouses with “big data”.

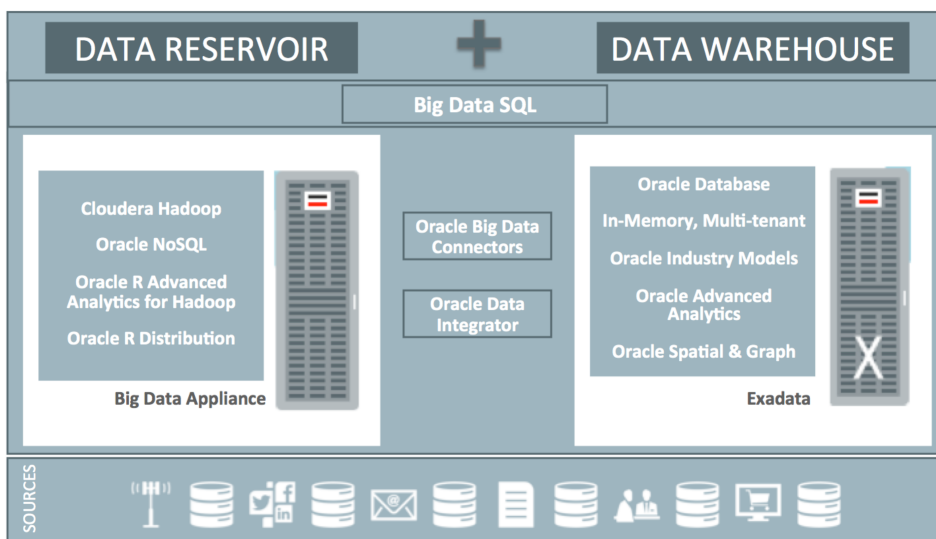
This paper focuses on Oracle Database 12c’s capabilities for data warehousing and big data. The first section of this paper describes an overall architecture for big data illustrating the role of Oracle Database 12c in a big data architecture. Subsequent sections highlight the capabilities in Oracle Database 12c to support data warehouses and big data, focusing on hardware integration, performance, scalability and analytics.

Big Data: The evolution of data warehousing


Big data and data warehousing share the same basic goals: to deliver business value through the analysis of data. However, big data and data warehousing differ in the scope of their data. While the promise of enterprise data warehousing has always been to analyze all of organization's data, in practice data warehouses have traditionally sourced data solely from other databases; that is, an enterprise's data warehouse contains data from its enterprise financials systems, its customer marketing systems, its billing systems, its point-of-sales systems, and so on. Organizations have recently recognized that there is an increasing amount of data in today's rapidly digitizing world which is not captured in operational databases: clickstream logs, sensor data, location data from mobile devices, customer support emails and chat transcripts, and surveillance videos, just to name a few. Big data systems harness these new sources of data, and allow enterprises to analyze and extract business value from these immense data sets.

Big data is in many ways an evolution of data warehousing. To be sure, there are new technologies used for big data, such as Hadoop and NoSQL databases. And the business benefits of big data are potentially revolutionary. However, at its essence, big data requires an architecture that acquires data from multiple data sources, organizes and stores that data in a suitable format for analysis, enables users to efficiently analyze the data and ultimately helps to drive business decisions. These are the exact same principles that IT organizations have been following for data warehouses for years.

The new information architecture that enterprises will pursue in the big data era is an extension of their previous data warehouse architectures. The data warehouse, built upon a relational database, will continue to be the primary analytic database for storing much of a company's core transactional data: financial records, customer data, point-of-sale data and so forth. The data warehouse will be augmented by a big-data system, which functions as a 'data reservoir'. This will be the repository for the new sources of large volumes of data: machine-generated log files, social-media data, and videos and images -- as well as a repository for more granular transactional data or older transactional data which is not stored in the data warehouse. Data flows between the big data system and the data warehouse to create a unified foundation for analytics:



Big Data architecture using Oracle Engineered Systems



The majority of business users will access the data in this information architecture from the data warehouse, using SQL-based environments. With Oracle Big Data SQL, a feature of the Oracle Big Data Appliance, Oracle offers unified SQL access across both the database and the big data environment – further extending the applicability of SQL. Meanwhile, a smaller number of analysts and developers will interact directly with the data in the big data environment, using more specialized languages such as MapReduce or R.

While Big Data SQL is outside of the direct scope of this white-paper, it is a critical capability for integrating the data warehouse and the data reservoir. Interested readers should see “Oracle Big Data SQL Solutions Brief” for more information.

Oracle Database 12c and Oracle Exadata: A Data Warehouse as a Foundation for Big Data

Even as new big data architectures emerge and mature, business users will continue to analyze data by directly leveraging and accessing data warehouses. The rest of this paper describes how Oracle Database 12c provides a comprehensive platform for data warehousing that combines industry-leading scalability and performance, deeply-integrated analytics, and advanced workload management – all in a single platform running on an optimized hardware configuration.

Exadata

The bedrock of a solid data warehouse solution is a scalable, high-performance hardware infrastructure. One of the long-standing challenges for data warehouses has been to deliver the IO bandwidth necessary for large-scale queries, especially as data volumes and user workloads have continued to increase. While the Oracle Exadata Database Machine is designed to provide the optimal database environment for every enterprise database, the Exadata architecture also provides a uniquely optimized storage solution for data warehousing that delivers order-of-magnitude performance gains for large-scale data warehouse queries and along with very efficient data storage. A few of the key features of Exadata that are particularly valuable to data warehousing are:

- » **Exadata Smarts Scans.** With traditional storage, all database intelligence resides on the database servers. However, Exadata has database intelligence built into the storage servers. This allows database operations, and specifically SQL processing, to leverage both the storage servers and database servers to vastly improve performance. The key feature is “Smart Scans”, the technology of offloading some of the data-intensive SQL processing into the Exadata Storage Server: specifically, row-filtering (the evaluation of where-clause predicates) and column-filtering (the evaluation of the select-list) are executed on Exadata storage server, and a much smaller set of filtered data is returned to the database servers. “Smart scans” can improve the query performance of large queries by an order of magnitude, and in conjunction with the vastly superior IO bandwidth of Exadata’s architecture delivers industry-leading performance for large-scale queries.
- » **Exadata Storage Indexes.** Completely automatic and transparent, Exadata Storage Indexes maintain each column’s minimum and maximum values of tables residing in the storage server. With this information, Exadata can easily filter out unnecessary data to accelerate query performance.
- » **Hybrid Columnar Compression.** Data can be compressed within the Exadata Storage Server into a highly efficient columnar format that provides up to a 10 times compression ratio, without any loss of query performance. And, for pure historical data, a new archival level of hybrid columnar compression can be used that provides up to 40 times compression ratios.

Oracle Database In-Memory

While Exadata tackles one major requirement for high-performance data warehousing (high-bandwidth IO), Oracle Database In-Memory tackles another requirement: interactive, real-time queries. Reading data from memory can be orders of magnitude faster than reading from disk, but that is only part of the performance benefits of In-Memory: Oracle additionally increases in-memory query performance through innovative memory-optimized performance techniques such as vector processing and a new in-memory aggregation algorithm. Key features include:

- » **In-memory Column Store.** Data is stored in a compressed columnar format when using Oracle Database In-Memory. A columnar format is ideal for analytics, as it allows for faster data retrieval when only a few columns are selected from a table(s), especially when the query accesses a large portion of the rows from those table(s). Compression is a fundamental component of In-Memory, since enables more data to be stored in memory. Columnar data is very amenable to efficient compression; data is typically compressed 2-20x, often with better performance than non-compressed columnar data.
- » **SIMD Vector Processing.** When scanning data stored in the IM column store, Database In-Memory uses SIMD vector processing (Single Instruction processing Multiple Data values). Instead of evaluating each entry in the column one at a time, SIMD vector processing allows a set of column values to be evaluated together in a single CPU instruction, for example in applying a where-clause predicates. In this way, SIMD vector processing enables the Oracle Database In-Memory to scan and filter billion of rows per second.
- » **In-Memory Aggregation.** Analytic queries require more than just simple filters and joins. They require complex aggregations and summaries. A new aggregation algorithm, specifically optimized for the join-and-aggregate operations found in typical star queries, has been introduced with Oracle Database 12.1.0.2. This algorithm allows dimension tables to be joined to the fact table, and the resulting data set aggregated, all in a single in-memory pass of the fact table.


Oracle Database In-Memory is useful for every data-warehousing environment. Oracle Database In-Memory is entirely transparent to applications and tools, so that it is simple to implement. Unlike a pure in-memory database, not all of the objects in an Oracle database need to be populated in the IM column store. The IM column store should be populated with the most performance-critical data, while less performance-critical data can reside on lower cost flash or disk. Thus, even the largest data warehouse can see considerable performance benefits from In-Memory.

Query Performance

Oracle provides performance optimizations for every type of data warehouse environment. Data warehouse workloads are often complex, with different users running vastly different operations, with similarly different expectations and requirements for query performance. Exadata and In-Memory address many performance challenges, but many other fundamental performance capabilities are necessary for enterprise-wide data warehouse performance.

Oracle meets the demands of data warehouse performance by providing a broad set of optimization techniques for every type of query and workload:

- » **Advanced indexing and aggregation techniques** for sub-second response times for reporting and dashboard queries. Oracle's bitmap and b-tree indexes and materialized views provide the developer and DBA's with tools to make pre-defined reports and dashboards execute with fast performance and minimal resource requirements.
- » **Star query optimizations** for dimensional queries. Most business intelligence tools have been optimized for star-schema data models. The Oracle Database is highly optimized for these environments; Oracle Database In-Memory provides fast star-query performance leverage its in-memory aggregation capabilities. For other database environments, Oracle's "star transformation" leverages bitmap indexes on the fact table to efficiently join multiple dimension tables in a single processing step. Meanwhile, Oracle OLAP is a complete multidimensional



analytic engine embedded in the Oracle Database, storing data within multidimensional cubes inside the database accessible via SQL. The OLAP environment provides very fast access to aggregate data in a dimensional environment, in addition to sophisticated calculation capabilities (the latter is discussed in a subsequent section of this paper).

- » **Scalable parallelized query access methods.** Parallel execution is one of the fundamental database technologies that enable users to query any amount of data volumes. It is the ability to apply multiple CPU and IO resources to the execution of a single database operation. Oracle's parallel architecture allows any query to be parallelized, and Oracle dynamically chooses the optimal degree of parallelism for every query based on the characteristics of the query, the current workload on the system and the priority of requesting user.
- » **Partition pruning and partition-wise joins.** Partition pruning is perhaps one of the simplest query-optimization techniques, but also one of the most beneficial. Partition pruning enables a query to only access the necessary partitions, rather than accessing an entire table – frequently, partition-pruning alone can speed up a query by two orders of magnitude. Partition-wise joins provide similar performance benefits when joining tables that are partitioned by the same key. Together these partitioning optimizations are fundamental for accelerating performance for queries on very large database objects.

The query performance techniques described here operate in a concerted fashion, and provide multiplicative performance gains. For example, a single query may be improved by 10x performance via partition-pruning, by 5x via parallelism, by 20x via star query optimization, and by 10x via Exadata smart scans – a net improvement of 10,000x compared to a naïve SQL engine.

Orchestrating the query capabilities of the Oracle database are several foundational technologies. Every query running in a data warehouse benefit from:


- » A **query optimizer** that determines the best strategy for executing each query, from among all of the available execution techniques available to Oracle. Oracle's query optimizer provides advanced query-transformation capabilities, and, in Oracle Database 12c, the query optimizer adds Adaptive Query Optimization, which enables the optimizer to make run-time adjustments to execution plans.
- » A **sophisticated resource manager** for ensuring performance even in databases with complex, heterogeneous workloads. The Database Resource Manager allows end-users to be grouped into 'resource consumer groups', and for each group, the database administrator can set policies to govern the amount of CPU and IO resources that can be utilized, as well as specify policies for proactive query governing, and for query queuing. With the Database Resource Manager, Oracle provides the capabilities to ensure that data warehouse can address the requirements of multiple concurrent workloads, so that a single data warehouse platform can, for example, simultaneously service hundreds on online business analysts doing ad hoc analysis in a business intelligence tool, thousands of business users viewing a dashboard, and dozens of data scientists doing deep data exploration.
- » **Management Packs** to automate the ongoing performance tuning of a data warehouse. Based upon the ongoing performance and query workload, management packs provide recommendations for all aspects of performance, including indexes and partitioning.

Data Management

Data warehouses, being the largest database in an IT organization, can present different data management challenges than typical OLTP database. Oracle provides unique advantages for running Oracle data warehouses online, with all data always available.

Partitioning

Oracle Partitioning is essential for managing large databases. It enables a "divide and conquer" technique for managing the large tables in the database, especially as those tables grow.



Although your database may have twice as much data next year as it does today, your end-users are not going to tolerate their application running twice as slow, your database is not going to be given twice as much time to complete maintenance and batch processing, and your IT managers are not going to double the hardware budget for the data warehouse. Partitioning is the feature that allows a database to scale for very large datasets while maintaining consistent performance, without unduly increasing administrative or hardware resources. Partitioning divides large tables up into smaller pieces, and thus allows many types of typical maintenance operations and end-user queries to be maintained at constant performance level even as the data grows.

Oracle leads the industry with the most comprehensive set of partitioning technologies, with numerous methods for partitioning tables, along with the capability for DBA's to define custom partitioning schemes; a rich set of administrative commands for partitioned tables; and a partition adviser to guide administrators on how best to implement partitioning.

Partitioning also enables Oracle12c's Advanced Data Optimization capabilities. A single table, when partitioned, can be distributed across multiple storage tiers. Older, less-frequently accessed data, corresponding to older partitions, can be compressed and/or stored on less expensive storage tiers, while newer data is stored on faster storage tiers. Advanced Data Optimization automates the process of modify data's storage characteristics based on the usage of that data.

Compression


Compression capabilities are used within every large data warehouse. As customers look to store larger and larger volumes of data, compression is a natural solution. Oracle provides industry-leading compression technology, with a breadth of compression techniques that enables every table in a data warehouse to be compressed:

- » "OLTP" Compression: Part of Oracle Advanced Compression, this technique delivers a typical compression ratio of 3:1 for data warehouses, with virtually no negative impact on query performance. This compression technique enables efficient updates to support compression for even data warehouse tables which are 'trickle-fed' or otherwise updated frequently.
- » Query Compression: Based upon Exadata Hybrid Columnar Compression, this technique delivers a typical compression ratio of up to 10:1 for data warehouses, with virtually no negative impact on query performance.
- » Archive Compression: Based upon Exadata Hybrid Columnar Compression, this technique delivers a typical compression ratio of up to 40:1 for data warehouses, but does entail trade-offs in query performance.

Read Consistency and Online operations

Oracle provides unique, patented read-consistency model to ensure that data loads never impact query performance. Oracle solves the challenges of concurrent access through a technology called multi-version read consistency; this unique technology has been the foundation of Oracle's concurrency model for over 15 years. Multi-version read consistency guarantees that a user always sees a consistent view of the data requested. If another user changes the underlying data during the query execution (such a trickle-feed update of a large data warehouse table), Oracle maintains a version of the data as it existed at the time the query began. The data returned to the query always reflects the state of the database (including all committed transactions) at the point in time at which the query was submitted regardless of what other updates may be occurring while the query is running. With this technology, Oracle is uniquely positioned to handle near real-time data loads within data warehouse environments, since queries and updates can occur simultaneously without blocking each other.

Oracle also provides online operations. Even for occasional data-maintenance operations, taking a data warehouse table offline is not viable. Thus, Oracle allows maintenance operations, such as moving a partition or table to be performed online, without impacting ongoing query or DML operations.



Data warehouses have become mission critical, and an area that is frequently overlooked is that they must now function as mission critical systems, by being fully available during load/update operations and during data-maintenance operations.

Analytics

Technologies such as OLAP, spatial analysis, statistical analysis, and predictive analytics are hardly new to data warehousing and business intelligence. However, OLAP products typically have their own calculation engine, statistics products have their own data engine, and predictive analytics products have their own mining engines. In short, an enterprise-wide business intelligence environment could maintain a half dozen different types of 'data engines', each requiring their own servers, their own copies of the data, their own management infrastructure, their own security administration, and their own high-availability infrastructure. Each engine has its own API's and its own set of developer tools and end-user tools. The complexity and cost of replicating entire stacks of BI technologies is significant.

Oracle Database provides a completely different approach by, first, continuing to extend the SQL language to perform more calculations within standard SQL and second by integrating analytics inside the database engine. Instead of moving data from a data warehouse to other analytic engines for further analysis, Oracle has instead brought the advanced analytic algorithms into its database, where the data resides.

Beyond the considerable advantages of consolidating the back-end data architecture of an enterprise business intelligence environment, the integration of analytics within the Oracle Database provides a host of advantages unavailable to stand-alone environments. For example, does your standalone OLAP server scale across large clusters of servers? How easily does your statistics engine integrate into your user authentication server? And can it transparently implement all of your data security policies? How easily can you integrate the results of your spatial analysis with your data warehouse data? Within Oracle Database, all of these issues are solved simply due to the deep integration of analytic capabilities in the database.

SQL Extensions for Analytics


The SQL language continues to evolve to enable more and more complex analysis. Moving averages, lag/lead, ranking, and ratio-to-report calculations are ubiquitously used in data warehouses today – and Oracle helped to pioneer these standardized SQL extensions.

With Oracle Database 12c, Oracle continues to extend SQL with its new Pattern Matching capabilities. SQL Pattern Matching introduces a new SQL syntax, along with optimized performance, for detecting patterns in a sequence of events stored in a database table. For example, one might want to look for trends in a stock price, or for suspicious behavior stored in activity log. With Oracle 12c, these queries can be expressed in a simple syntax, without requiring recursive joins or other complex SQL constructs.

Advanced Analytics

Oracle Advanced Analytics offers a combination of powerful in-database predictive-analytics algorithms and open source R algorithms, accessible via SQL and R languages.

These analytic capabilities include a dozen data-mining algorithms implemented in the Oracle Database (including algorithms for classification, clustering, regression, anomaly detection, and associations); SQL functions for basic statistical techniques; and tight server-side integration with open-source R to enable R programmers to realize the full performance and scalability of the Oracle database platform and also provide access to the entire functionality of the R ecosystem on data stored in the Oracle database.



By providing a range of GUI and IDE options, Oracle Advanced Analytics allows business users, statisticians, and data scientists to tap directly into the large data volumes and extensive processing capacity of the Oracle Database, using an environment appropriate to each end-user. SQL-savvy data scientists can directly use SQL, statisticians with experience in R can continue to write R programs and utilize R-based GUIs, and business users can leverage Oracle's Data Miner extension in SQL Developer.

OLAP

Oracle OLAP is a full-feature online analytical processing (OLAP) engine embedded in the Oracle Database. Oracle OLAP enhances data warehouses by improving query performance (as discussed in the performance section) and by adding enriched analytical content.

The core feature of Oracle OLAP is cubes. Managed within the Oracle database, cubes store data within a highly optimized multidimensional format. Cubes provide scalable and compressed storage of dimensional data, fast incremental update, fast query performance, and the ability to compute or store advanced analytical calculations.

Oracle's strategy with Oracle OLAP is to bring these core OLAP advantages into the data warehouse. This is achieved by exposing the key capabilities of Oracle OLAP via standard SQL, so that any business intelligence tools or other SQL-based application can leverage OLAP.

Conclusion





The Oracle Database is the market leader for data warehousing, built upon a solid foundation of scalability and performance, and augmented by innovative features such as Oracle's unique read-consistency model for near-real-time data warehouses and a flexible and powerful set of in-database analytic capabilities. The combination of the Oracle Database and an Oracle Exadata storage grid delivers the highest levels of performance for IO intensive workloads, and, with the Oracle Exadata Database Machine, Oracle delivers a complete hardware and software solution for data warehousing.



Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

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

Hardware and Software, Engineered to Work Together

Copyright © 2014, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0914