CHENEGA® CHENEGA PROFESSIONAL & TECHNICAL SERVICES, LLC

Cloud Data Warehouses

Analytical Database-as-a-Service

A White Paper by Chenega Professional & Technical Services, LLC

Adam Getz
Solution Architect, Data Analytics
June 2019

Chenega Professional & Technical Services

Headquarters:

609 Independence Parkway
Suite 210
Chesapeake, VA 23320

Phone:

757-549-5700

Web:

www.chenegapts.com

LinkedIn:

www.linkedin.com/company/chenega-professional-services-business-unit/

Evolution of Data Warehouse Technologies

Traditionally, data warehouses are data repositories optimized for retrieval and analysis of data from transactional, legacy, or external systems, applications, and sources that provide managers, executives, and other decision makers the ability to conduct data analysis. The traditional data warehouse provides an environment separate from operational systems and is designed for decision-support, business intelligence, data analysis, and ad-hoc queries. To take advantage of this new paradigm in information technology (IT), to solve the need to be more scalable, to lower administrative burden of IT staff members, and to satisfy the growing need of rapidly implementing full implementations of data warehousing environments, a new type of solution has entered the IT marketplace: **The Cloud Data Warehouse.**

Data warehouse technologies have existed in the past and have been succcesfully implemented within on-premise data centers. Initially, data warehouses were implemented within conventional relational database management systems (DBMS) including Oracle, MS SQL Server, DB2, Sybase, & Informix. However, these database technologies were designed to handle transactional systems with a focus on integrity and inserts of data rather than performance of large read queries. Early data warehouses implemented in relational database technologies were limited in size and system performance was a common issue.

Subsequently a wave of data warehouse specific databases came to the market that included new algorthms and technniques to increase system performance of data warehouses. This wave of technologies was identified by being software-only solutions that could be implemented on any type of commodity hardware and storage device. Technologies in this first wave included Sybase IQ, Greenplum, ParAccel, Vertica, & Kognitio, & Exasol.

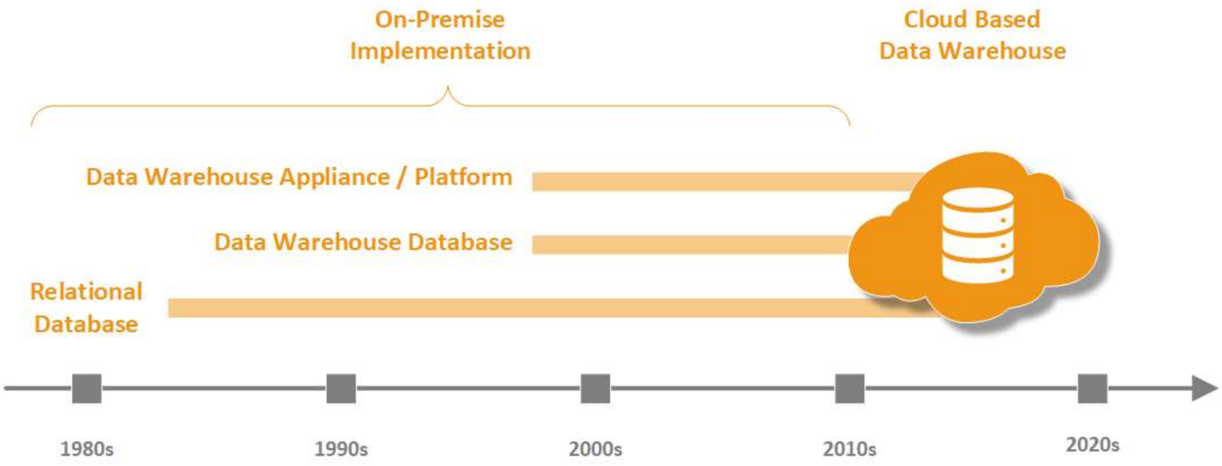
Around the same time, another wave of data warehouse technology came to the market that was known as either a datawarehouse platform or a data warehouse appliance. This complete solution included all of the software, operating system, hardware, and storage bundled together within one fully-integrated unit. Moreover, this wave of data warehouse technology was constructed entirely on proprietary hardware and storage devices. The data warehouse platform was not another application that was implemented within an on-premise data center. But rather, it was a complete solution that encompassed a whole technology stack inlcuding both hardware and software provided in one solution by the same vendor. Technologies included in this wave of data warehouse technology include the products: Netezza, Terradata, Oracle Exadata, IBM Infosphere, & DATAllegro.

Forrester Research, Inc. defines A cloud data warehouse as:

"An on-demand, secure, and scalable self-service data warehouse that automates the provisioning, administration, tuning, backup, and recovery to accelerate analytics and actionable insights while minimizing administration requirements."

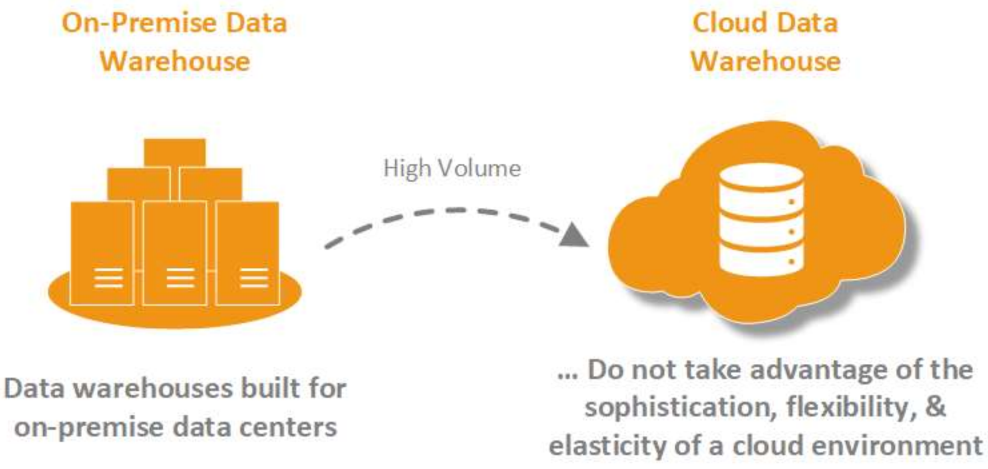
The first data warehouse technology built specifically for a cloud environment was made available with the release of Amazon Redshift in 2012. However, Amazon did more than just introduce a new database management system. Amazon changed the paradigm of data warehouse technolgies from being implemented entirely within on-premise data centers to being implemented entirely within a cloud environment. Amazon Redshift demonstrated the viability of the cloud data warehouse, and other cloud data warehouses quckly became available in the marketplace. Following the model of Amazon Redshift other database vendors either transformed their own database technologies to be cloud-enabled or developed completely new cloud-based technologies. And cloud data warehouses became available within other cloud environments other than Amazon Web Services (AWS). Now other cloud environments including Microsoft Azure, Google Cloud, IBM Cloud, SAP Cloud, Oracle Cloud, Alibaba Cloud, and Huawei Cloud have cloud data warehouse technologies within their portfolio of cloud solutions. And in addition to Amazon Redshift, other cloud data warehouse technologies that are available include: Snowflake Elastic Data Warehouse, Google BigQuery, Oracle Autonomous Data Warehouse, SAP HANA, Teradata IntelliCloud, IBM DB2 Warehouse, Microsoft Azure SQL Data Warehouse, Hortonworks Cloud, MarkLogic, Alibaba DataWorks, Pivotal Greenplum, Exasol, Micro Focus Vertica, and Huawei Data Warehouse Service. Now there are numeorous data warehouse technologies within several cloud environments that can be utlized for implementations of data warehouses.

Timeline of Data Warehouse Technologies


On-Premise Implementation
Cloud Based Data Warehouse
Data Warehouse Appliance / Platform
Data Warehouse Database
Relational Database
1980s
1990s
2000s
2010s
2020s

Benefits of a Cloud Data Warehouse

Many data warehouses deployed in the past and were built utilizing technologies designed for on-premise data centers typical of the time of deployment. As the cloud did not yet exist, data warehouses of the past were not able to take advantage of the sophisticated features that are now available and capabilities that can now be leveraged. This includes the capabilities of elastically scaling up, scaling down, and suspending as needed to meet continuously varying demands. On-premise data warehouse technologies did exist in the past but were limited compared to today's more advanced capabilities of cloud environments.


On-Premise Data Warehouse
Cloud Data Warehouse
High Volume
Data warehouses built for on-premise data centers
... Do not take advantage of the sophistication, flexibility, & elasticity of a cloud environment

The numerous **benefits** of implementing a **cloud data warehouse** include:

- **Query performance:** Cloud data warehouses are purpose-built for processing of complex queries that include large amounts of data., long table scans, large read blocks, and other advanced tasks required in large analytical environments.

- **Scalability for very large implementations:** Cloud data warehouses can be initially loaded with very large volumes of data and its architecture can support the processing of this data with very quick response time.

- **Rapid provisioning:** Cloud data warehouses can be built and deployed within a cloud environment in just few minutes.

- **Automatic scaling and elasticity:** Cloud data warehouses offer automation capabilities and the ability to scale up / scale down to any size data warehouse based on business needs and usage.

- **Increased flexibility:** Cloud-based services can instantly meet increases demand of a data warehouse, rather than undergoing a complex and expensive update to information technology infrastructure.

- **Disaster recovery:** Cloud providers provide quick data recovery for all kinds of emergency scenarios, from natural disasters to power outages.

- **Savings on equipment:** Cloud computing eliminates the capital expense of buying hardware and software and setting up and running on-site datacenters before deploying a data warehouse.

- **Pay for computing usage:** Cloud computing provides the ability to bill customers on actual resource consumption rather than expected consumption.

- **Guarantee of uptime:** Cloud providers have service level agreements (SLAs) with their customers that provide assurance that the application will be operational for a certain amount of time (i.e. 99.9% of the time).

- **Low barrier of entry:** Organizations can start utilizing cloud data warehouses without building an expensive and time-consuming on-premise data center.

- **Economies of scale:** Cloud data warehouses are installed for large numbers of customers rather than just one customer, meaning that infrastructure costs are shared among lots of organizations.

- **Automatic software updates:** Cloud providers install database software releases and patches as soon as they are readily available from the vendor. This means that there is no longer a need for organizations IT staff to keep-up-to-date and install database software.

- **Increased data security:** A primary focus of cloud providers is to carefully monitor security. This is significantly more efficient than a conventional on-premise system, where an organization must divide its efforts between a myriad of information technology concerns, with security being only just one of those concerns.

- **Simplified Management:** Cloud data warehouses include intuitive tools and controls to conduct systems administration that require minimal knowledge to maintain and have a low learning curve to conduct basic administrative tasks.

- **Reduced Administrative Burden:** Cloud data warehouses provides enhanced and simplified database administrator (DBA) operations and maintenance capabilities. Further, they include few unnecessary features that typically consume much of a DBA's time and efforts.

- **Competitive Edge:** Organizations using a cloud data warehouse can implement more rapidly than competitors who must devote information technology resources to managing infrastructure.

- **Accelerated Return on Investment (ROI):** Cloud data warehouse allow for rapid installation and implementation schedules as all the technical components of the data warehouse are included. The customer has no need to spend valuable time collecting and installing all the individual hardware and software components and has little need to conduct regression and integration testing of both hardware and software.

- **Utilization of latest technologies:** Cloud data warehouses are leading-edge technologies and are constantly being updated to take advantage of advances in technology.

- **Use of best practices:** Cloud data warehouses utilize techniques that have been generally accepted as superior to alternative techniques. And cloud data warehouses produce results that are superior to those achieved by other means.

Techniques of Cloud Data Warehouse Optimization

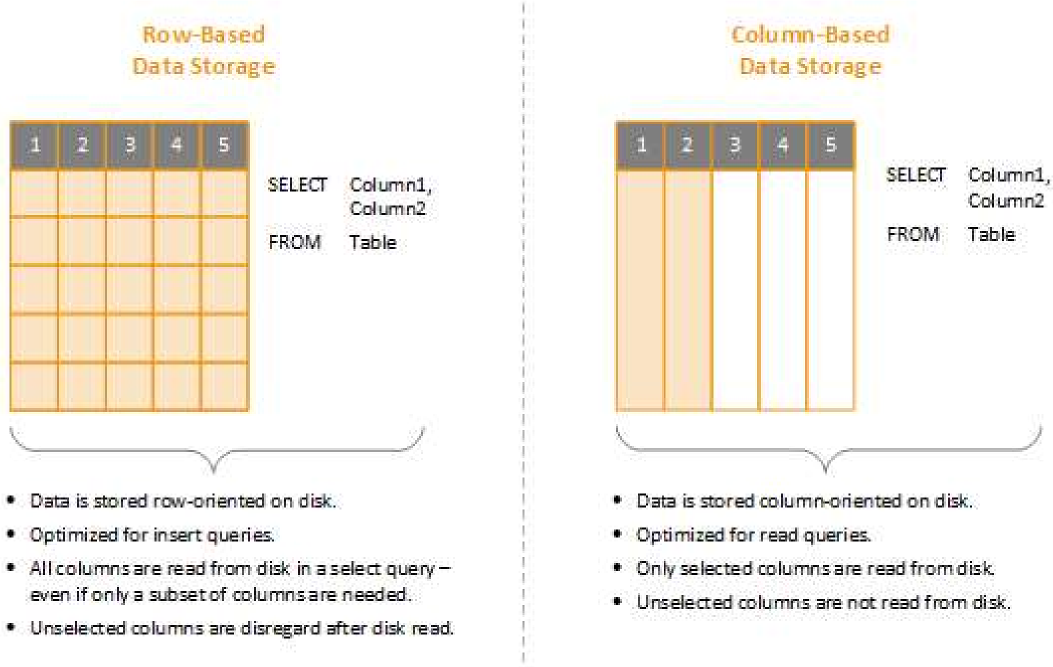
As the main purpose of the cloud data warehouse is the enablement of rapid querying of large and complex sets of data for data analysis, the technology needs to be designed in a way to handle this purpose. Thus, **four basic techniques** are included in cloud data warehouse technologies to enable extremely fast processing of read queries.

- **Columnar Data Storage**
- **Database Compression**
- **Massive Parallel Processing (MPP)**
- **In-Memory Processing**

Columnar Data Storage: While traditional databases store data in rows or records to limit the number of operations on data, cloud data warehouse often store data in columns or fields to increase read query performance. By storing data in columns, the database can more precisely access the data it needs to answer a query rather than scanning and discarding unwanted data in rows. Columnar data storage for database tables drastically reduces the overall number of input/output (I/O) operations to disk storage and reduces the amount of data needed to be loaded from physical disk.

Columnar data storage data warehouse technologies ignore all the unnecessary data contained in database table rows that doesn't apply within a read query, and only retrieves data from the selected columns of query. With columnar data storage, a database query only reads the values of columns required for processing a given query and avoids bringing into memory irrelevant fields that have not been selected within the query.

Columnar Data Storage



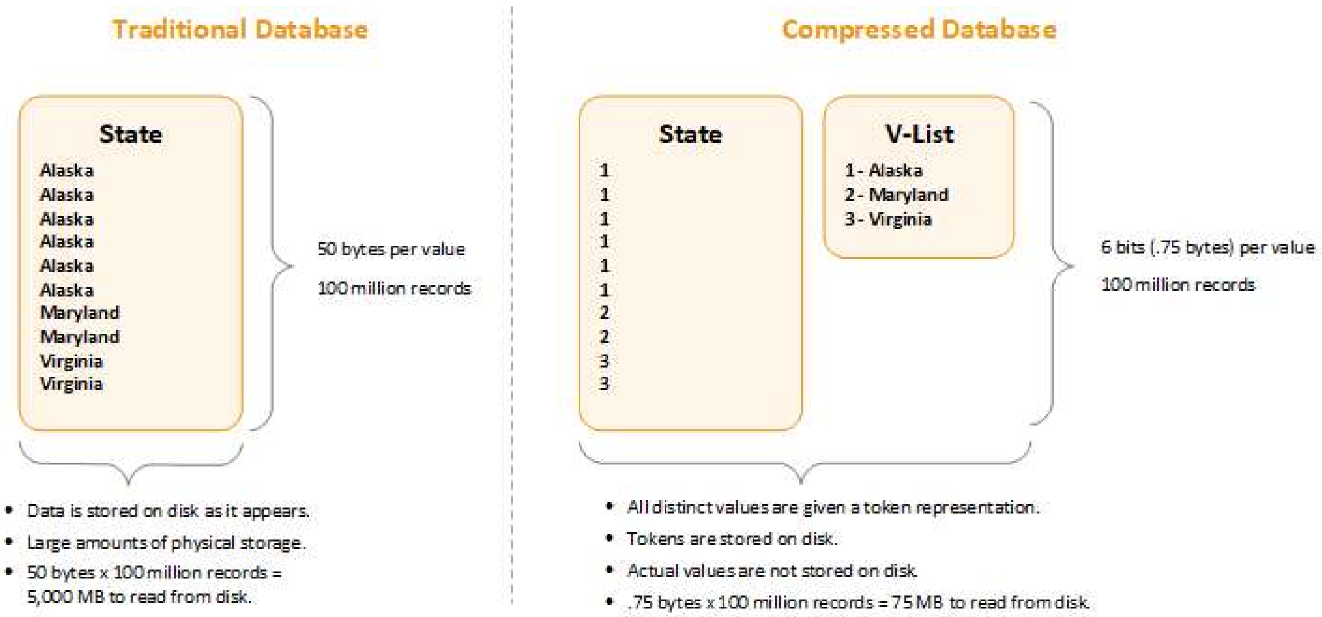
- Data is stored row-oriented on disk.
- Optimized for insert queries.
- All columns are read from disk in a select query – even if only a subset of columns are needed.
- Unselected columns are disregard after disk read.

- Data is stored column-oriented on disk.
- Optimized for read queries.
- Only selected columns are read from disk.
- Unselected columns are not read from disk.

Database Compression: Cloud data warehouses use compression to save disk storage space by using fewer database pages to store data. Fundamentally, database compression includes a reduction in the number of bits needed to represent data. Because more data can be stored per database page, fewer database pages must be read to access the same amount of data. Therefore, queries on a compressed table need fewer disk input-output (I/O) operations to access the same amount of data and performance of read queries is increased.

In general, data warehouses have large data volumes and large amounts of data redundancy with lots of repeating values. This allows database compression to be very effective and enables very high data compression ratios (i.e. the ratio between the uncompressed size and compressed size). Generally, the higher the database compression ratio, the higher the performance gains in read queries that can be achieved. With **database compression** being utilized, the database can store more records per database page and fewer database pages must be read to access the same amount of data. And more data will be stored within the system buffer pool (i.e. the area of system memory that has been allocated by the database to cache table and index data). And since more data resides within the buffer pool, the likelihood that needed records are located within the buffer pool increases significantly. When a database record needs to be accessed, the related buffer is first searched, and a disk access operation can be avoided if the record is found within the buffer. By compressing records, the effective capacity of the buffer increases and consequently its hit ratio (the probability that a record will be found in buffer) increases. Thus, database compression improves query performance through improved buffer pool hit ratios.

Database Compression

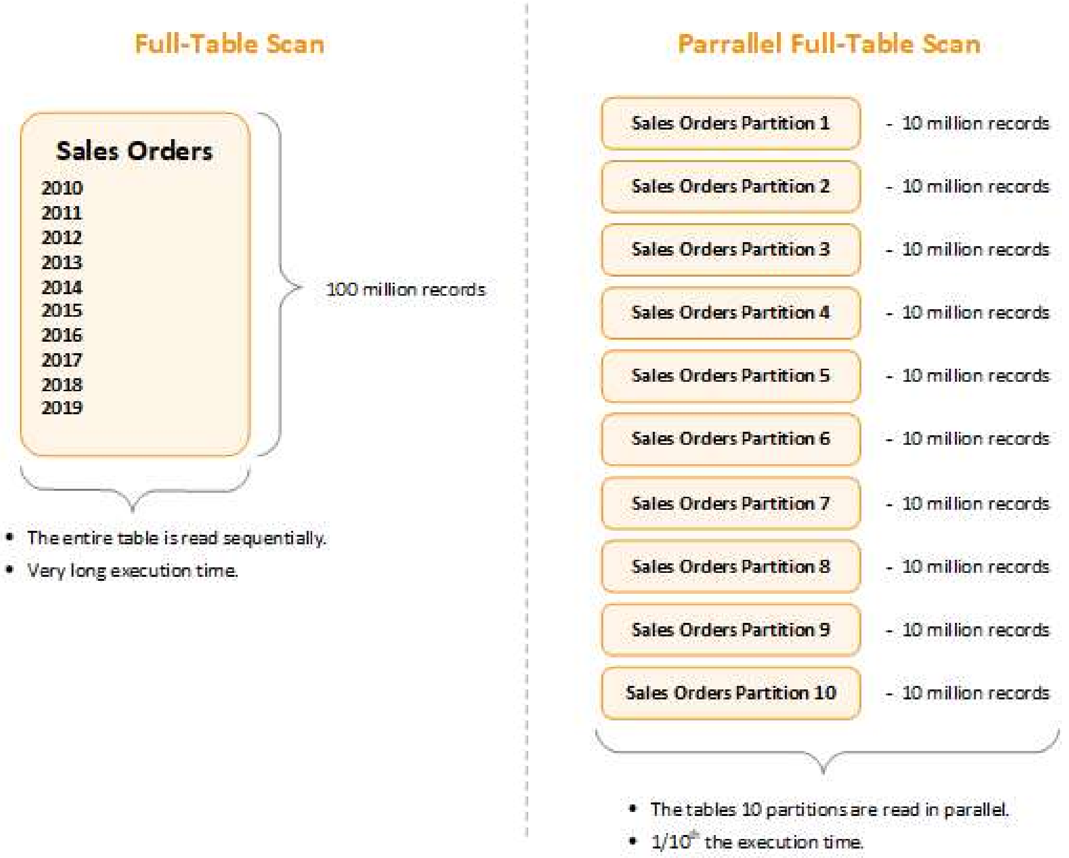


- Data is stored on disk as it appears.
- Large amounts of physical storage.
- 50 bytes x 100 million records = 5,000 MB to read from disk.

- All distinct values are given a token representation.
- Tokens are stored on disk.
- Actual values are not stored on disk.
- .75 bytes x 100 million records = 75 MB to read from disk.

Massive Parallel Processing (MPP): Massive parallel processing typically consists of independent processors, servers, or nodes that all execute in parallel and allow for optimal query performance and platform scalability. Also known as a “shared nothing architecture”, this type of data warehouse performance enhancement technique is characterized by a design in which every embedded processor or node is self-sufficient and controls its own memory and disk operations. Read queries executed against the database are broken into smaller components, and all components are worked upon both independently and simultaneously to deliver a single combined result set. Additionally, this divide-and-conquer approach allows for massive parallel processing databases to scale linearly as new processors are added. Massively parallel processing refers to the fact that when a query is issued, every node works simultaneously to process the data that resides within that node.

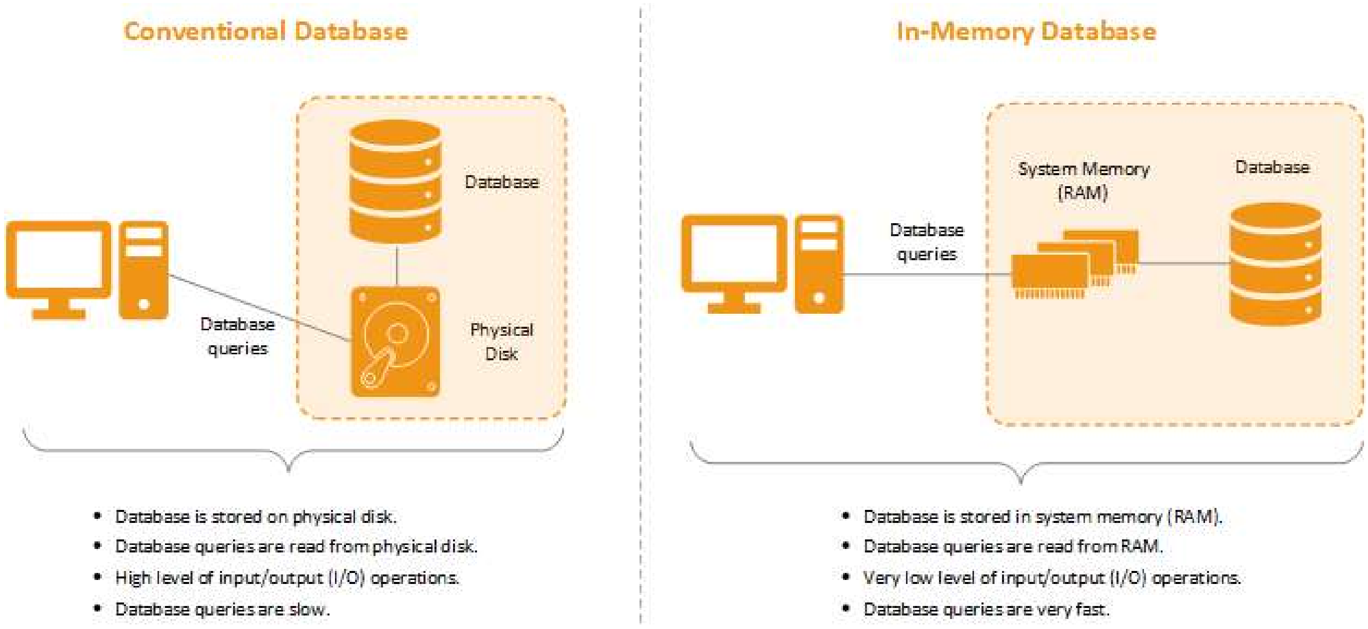
Massive Parallel Processing (MPP)



In-Memory Processing: Data processed in the cloud data warehouse is stored within system random-access memory (RAM) rather than conventional database management system that processes data stored on physical disks. This allows the processing of queries from reads of system memory rather than reads from disk devices. Accessing data in memory eliminates both seek time and input/output (I/O) operations when querying data. Seek time is the time taken for a disk drive to locate the area on the disk where is stored. This provides faster and more predictable performance than queries conducted from data on residing on disk since retrieving data from disk storage is the slowest part of data processing. The less data that needs to be retrieved from disk, the faster the data retrieval process.

The concept behind **in-memory processing** is relatively simple. Traditionally, data is placed physically within a storage device known as a disk. Then, only when needed, the data is accessed from disk, moved to system memory, and acted upon within system memory. Transferring data from disk to system memory results in a bottleneck that reduces query performance. Without the bottleneck of having to access data in storage, in-memory databases can swiftly process data and convert the data into information in a highly effective way.

In-Memory Processing


Conventional Database
Database
Physical Disk
Database queries
• Database is stored on physical disk.
• Database queries are read from physical disk.
• High level of input/output (I/O) operations.
• Database queries are slow.
In-Memory Database
System Memory (RAM)
Database
Database queries
• Database is stored in system memory (RAM).
• Database queries are read from RAM.
• Very low level of input/output (I/O) operations.
• Database queries are very fast.

Conventional databases already have a concept known as buffer pools that caches data by storing some data within system memory. But with buffer pools, the database only caches commonly used data within system memory. **In-memory processing** extends the concept of caching of data way beyond the use of the buffer pools. In-memory processing caches either entire tables, indexes, or databases within system memory. And in-memory processing databases are not limited by the size of the buffer pool. In the past, **in-memory processing** was limited as the cost of RAM was far higher than the cost of space on disk drives. And in the past, operating systems had 32-bit architectures that could only process 4 GBs of data stored in system memory. But now the cost of RAM has come way down and operating systems now have 64-bit architectures. The means that the cost of RAM is about the same as the cost of hard disk space. And operating systems can now support 16 exabytes of system memory. The result is that large volumes of data can be stored and processed within system memory in a cost-effective manner resulting in extremely fast query execution times.

Cloud Data Warehouse Products and Vendors

Each of the cloud data warehouse technologies uses a different combination of the four performance enhancement techniques, with most of the cloud data warehouses utilizing columnar data storage, database compression, and parallel processing. Of the cloud data warehouses that utilize in-memory processing, SAP HANA is the only one built entirely on a 100% in-memory architecture which places the entire database within system memory. Oracle Autonomous Data Warehouse includes an optional in-memory column store that allows for specific columns or fields to be stored within system memory. Microsoft Azure SQL Data Warehouse includes memory-optimized tables and memory-optimized indexes which allow for specific tables and indexes to be stored within system memory. Additionally, Exasol and MemSQL have options to store parts of the database within system memory.

And the most common cloud providers for implementations of cloud data warehouses include:

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud
- Oracle Cloud
- IBM Cloud
- SAP Cloud

Cloud data warehouses can be provided in one of two service models: Platform-as-a-Service (PAAS) or Software-as-a-Service (SAAS). PAAS means that a cloud provider offers access to a cloud-based environment in which user organizations can build applications from. With PAAS, the user organization does some of the system administration of the software product, but the cloud provider maintains the operating systems, servers, infrastructure, network, and storage. SAAS means that a cloud provider offers user organizations access to a full cloud-based application, and the cloud provider conducts all the system administration of both the software product and environment. With SAAS, the cloud provider offers a web user interface to use the product, open database connectivity / java database connectivity (ODBC/JDBC) to access data stored within the product, and application programming interfaces (API) to conduct application integrations with the product.

Currently Snowflake Elastic Data Warehouse, Panoply Smart Data Warehouse, 1010Data are the only cloud data warehouse products offered in a SAAS service model. All other cloud data warehouses are provided in a PAAS service model. The SAAS cloud data warehouses provide a web user interface to use the product, while the PAAS cloud data warehouses provide the ability to conduct system administration of the product.

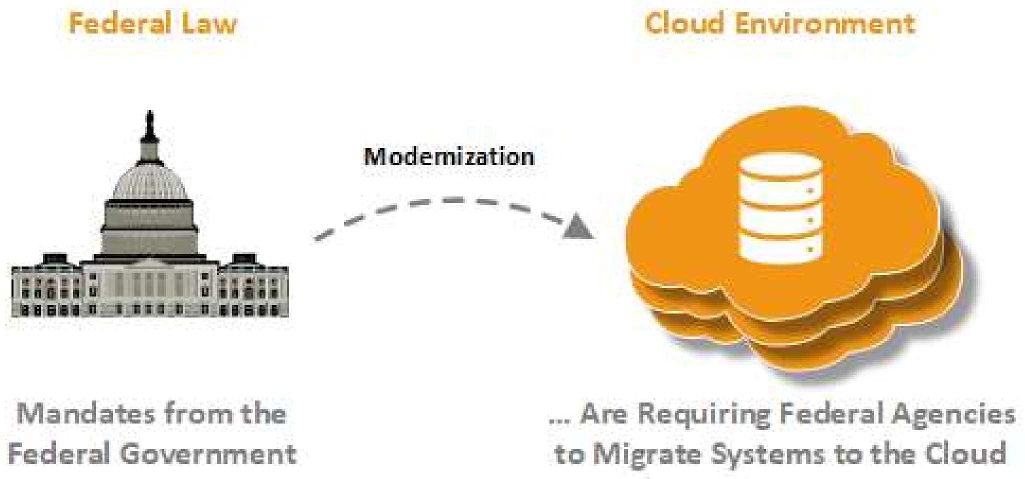
Cloud Data Warehouse	Performance Techniques	Cloud Providers
Amazon Redshift	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS)
Snowflake Elastic Data Warehouse	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure
Google BigQuery	Columnar Data Storage Database Compression Parallel Processing	Google Cloud
Oracle Autonomous Data Warehouse	Columnar Data Storage Database Compression Parallel Processing In-Memory	Oracle Cloud
SAP HANA	Database Compression In-Memory	SAP Cloud Platform
Teradata IntelliCloud	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure
IBM DB2 Warehouse	Columnar Data Storage Database Compression Parallel Processing	IBM Cloud
Microsoft Azure SQL Data Warehouse	Columnar Data Storage Database Compression Parallel Processing In-Memory	Microsoft Azure
Hortonworks Cloud	Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Google Cloud IBM Cloud
MarkLogic	Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Google Cloud
Alibaba DataWorks	Columnar Data Storage Database Compression Parallel Processing	Alibaba Cloud

Cloud Data Warehouse	Performance Techniques	Cloud Providers
Pivotal Greenplum	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Google Cloud
Exasol	Columnar Data Storage Database Compression Parallel Processing In-Memory	Amazon Web Services (AWS) Microsoft Azure
Micro Focus Vertica	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Google Cloud
Huawei Data Warehouse Service	Columnar Data Storage Database Compression Parallel Processing	Huawei Cloud
Panoply Smart Data Warehouse	Database Compression Parallel Processing	Amazon Web Services (AWS)
Qubole	Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Oracle Cloud
MemSQL	Columnar Data Storage Database Compression Parallel Processing In-Memory	Amazon Web Services (AWS) Microsoft Azure
1010Data	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS)
Cloudera	Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure
Treasure Data Customer Data Platform	Columnar Data Storage Database Compression Parallel Processing	Amazon Web Services (AWS) Microsoft Azure Google Cloud

Why Implement a Cloud Data Warehouse in the Federal Government?

The Federal Managing Government Technology (MGT) Act was signed into law on December 12, 2017 and apportions $500 million to the Technology Management Fund (TMF). This fund applies to federal agencies with either large IT systems or federal agencies that have a need to make their data stored within their numerous systems more transparent to constituents. The MGT Act is important as it includes a mandate to modernize federal systems along with appropriating funding for federal agencies to modernize. Although the MGT Act does not explicitly mention data warehousing systems, it does focus on migrating government systems to cloud environments. And many federal agencies are also now being required to provide constituents with more interactive access to their agency's data. Implementations of cloud data warehouses is consistent with the mandate to make more federal data available, to enable constituents with the ability to interact with federal data, and to modernize federal IT systems by migrating them to cloud environments. Although the MGT Act focuses only on federal agencies and federal IT systems, US states and large municipalities also have system modernization initiatives like the federal initiatives.

Even before the MGT Act has been signed into law, **modernization of data warehouses within the federal government** has already begun. More and more, there has been a need for federal agencies to centralize data within one data repository, to support big data initiatives, to make data available to constituents, and to provide the capability to analyze both structured and unstructured data. Federal agencies have started understanding that legacy technologies and analysis techniques will not satisfy the increasing demand for data and will not satisfy the mandate to be more transparent. And federal agencies have started understanding that implementations of data warehouses within on-premise data centers has limited federal agencies to not be able to truly modernize.


Federal Law
Cloud Environment
Modernization
Mandates from the Federal Government
... Are Requiring Federal Agencies to Migrate Systems to the Cloud

About Chenega Professional & Technical Services

Chenega Professional & Technical Services, LLC (CPTS) is a leading information technology services firm providing consulting and advisory services to clients within the federal government. CPTS specializes in building and implementing solutions that provide both rapid and long-term value in the areas of:

- Data Analytics
- Data Warehousing
- Business Intelligence

Our consultants are both experienced and knowledgeable in the business and technology aspects of IT systems implementations. Plus, we have high degree of subject matter expertise in the areas that are critical to many federal agencies. CPTS is the right organization to deploy the enabling technologies that help federal agencies make more informed decisions through smart use of their data.

CPTS is also a wholly-owned subsidiary of the Chenega Corporation. CPTS utilizes shared services provided by the corporation, which affords us the stability and support of a large business while maintaining the flexibility and rates of a small business. Chenega Corporation has the dual mission of succeeding in business to assist its shareholder, descendants and family members in their journey to economic and social self-determination and self-sufficiency, and to create and support comprehensive cultural, societal and community activities.

For more information visit us at either www.chenegapts.com or www.linkedin.com/company/chenega-professional-services-business-unit.