The Modern Data Repository: Understanding Your Options

blog.netwrix.com/2022/12/09/the-modern-data-repository-understanding-your-options

Joe Dibley

Today, organizations have a variety of options for storing the data they generate, collect and use. Options for data repositories include:

- Relational database
- Data warehouse
- Data lake
- Data mart
- Operational data store

Choosing the best option for a given business situation depends on a variety of factors, including the needs of your user base, the skills of your DBAs and other database resources, the reporting and analysis requirements for business decisions, and whether you are storing structured or unstructured data.

The table below provides a brief summary of the characteristics of each type of data repository, and the rest of this blog post explains them in more detail.

	RDBMS	Data Warehouse	Data Lake	Data Mart	Operational Data Store
Purpose	OLTP	BI and reporting	Big data analytics and data discovery	Targeted business analytics	Consolidated OLTP
Type of Data	Structured transactional data	Structured data for OLAP	Structured & unstructured	Consolidated structured data from internal & external systems	Integrated & cleansed data from OLTP systems
Data Quality	Normalized & consistent	De-normalized & consistent	De-normalized or normalized & inconsistent	Normalized & partial subsets	Normalized & cleansed with some inconsistency

Relational Database

The relational database management system (RDBMS) dates back to 1970, and it remained the only option for most organizations until the late 1990s. Relational databases have traditionally been used to store structured transactional data from online transaction processing (OLTP) systems like CRM, ERP, HR, manufacturing and financial applications. The field of RDBMS solutions include numerous commercial and open-source options , such as Oracle, Sybase, IBM DB2, Microsoft SQL Server, PostgreSQL and MySQL.

The basic functions of any RDBMS system are to create, read, update and delete data (collectively referred to as CRUD). Data is stored in row-based tables using normalization, primary keys, foreign keys and constraints to ensure the reliability of the data. Structured Query Language (SQL) is used to find, access and manipulate the data.

The two key features of any RDBMS are:

- **Data normalization** — Normalizing data is the process of arranging related data in multiple tables in an intended and unambiguous manner that eliminates data redundancy.
- **Atomicity, consistency, isolation and durability (ACID)** — An RDBMS must preserve and guarantee the consistency of transactions at the database level. ACID enables developers to create enterprise software applications without having to consider the complexities of data integrity in the backend RDBMS.

Relational databases are well suited when security, accuracy, integrity and consistency of data is required.

Data Warehouse

In organizations that use relational databases, it is customary for each business application to have its own backend RDBMS. However, this segregation of data can impede business intelligence (BI) activities like decision support, enterprise reporting, just-in-time marketing and -hoc querying.

Enter the data warehouse, a term coined in the late 1980s. A data warehouse is a single repository that consolidates data from many different data sources, in de-normalized format. Usually, a data warehouse is a purpose-built relational database on premises or in the cloud:

- **On-premises** data warehouses include Teradata, Greenplum, IBM Netezza, Oracle DW Appliance and Oracle Exadata Server. These appliances are a combination of specialized hardware and software optimized for data warehousing workloads.
- **Cloud-based data warehouses** include Snowflake, Google BigQuery, Microsoft Azure SQL Data Warehouse and Amazon Redshift. Cloud-based data warehouses enable organizations to scale up on demand while saving them the cost of on-premises infrastructure and ongoing maintenance.

Data Marts

While data marts are frequently confused with data warehouses, they actually serve markedly different purposes. While data warehouses are used to make strategic decisions that might impact the entire organization, data marts are typically used to store data for making tactical decisions whose impact is limited to a specific business process or department. The data contained in a data mart is highly curated and might be normalized or de-normalized.

The most distinguishing characteristic of a data mart is the use of the *star schema* configuration, a framework that consists of one or more fact tables that reference many dimension tables, forming the shape of a star.

Data Lake

Unlike traditional databases, data lakes do not have a set structure because the data is stored in a raw or unrefined form, since its purpose is yet to be determined. Since they have no defined schema structure or taxonomic information and the raw data is not suitable for BI activities, data lakes are cheaper to set up and require little maintenance.

Data lakes are often built on top of a NoSQL database such as Apache Hadoop. Data is stored as a schema-less key-value pair, and the schema and data requirements are not defined until the data is actually queried. Data is split into shards across multiple nodes built using commodity hardware, which also provides fault-tolerance and redundancy. In a data lake built on the Hadoop platform, data is queried using MapReduce jobs. Open source or third-party add-ons such as Hive or HBase support SQL queries by converting SQL into MapReduce jobs.

Amazon, Microsoft, Oracle, Teradata, MongoDB and Cloudera all market data lake solutions with proprietary data management add-ons.

Operational Data Store (ODS)

Operational data stores are often confused with data warehouses because they are both used to consolidate large data volumes from multiple information systems. However, that is where the similarities end; they store data differently and serve different purposes.

An ODS is used to store detailed transactional data from different operational systems on a short-term basis. It can serve as an operational system or as an interim staging area before the data is cleansed, processed and submitted into a data warehouse. Operational data stores normally receive data on a continuous basis from other systems, either through real-time data replication or via batch extract-transform-load (ETL) processes. Data is normally stored in a denormalized format.

Operational data stores are ideal for querying small data sets to satisfy real-time or near-real-time reporting or ad-hoc querying needs. Operational data stores are also referred to as master data management (MDM) systems because they are used to create

appropriate sets of data required to conduct day-to-day business activities.

How can Netwrix help?

Netwrix offers data security tools that help you discover, classify and securing sensitive data, no matter where it resides — including data in your various data repositories. You might want to consider the following solutions:

- Netwrix StealthAUDIT helps you secure your sensitive data, simplify compliance and increase IT productivity by providing visibility into what's going on across your Oracle, SQL Server and Azure SQL environments, as well as your unstructured data repositories.
- Netwrix Auditor and Netwrix Data Classification together enhance the security of your most critical assets wherever they reside. They facilitate the processes of identifying sensitive data across your IT ecosystem, reviewing who has access to that data, monitoring activity around it and automatically moving it to a secure location if it is overexposed.

Joe Dibley
Security Researcher at Netwrix and member of the Netwrix Security Research Team. Joe is an expert in Active Directory, Windows, and a wide variety of enterprise software platforms and technologies, Joe researches new security risks, complex attack techniques, and associated mitigations and detections.