TechTarget | SearchDataManagement

PRO+ Content

Data Repository Cheat Sheet

Databases vs. data warehouses vs. data lakes vs. data marts

Distribution of market share among the major industry players
Projected sales of main products in 2013
BGY
FEW
TRG
RTG
RDW
WEF

In this e-guide

In this e-guide:

With so much data now streaming into your business, coming from various places and used for unique purposes, it's essential that you know where your data should go – and why.

For example, imagine if you need to analyze customer data, but that data is misallocated to a data lake and isn't clean for analysis. That's potentially a massive financial loss.

This cheat sheet will help you understand how each repository differs based on the type, purpose and quality of the data.

In addition, get an introduction to the features of data catalog software, a user-friendly metadata management system.

Beyond the RDBMS: Data warehouse vs. data lake vs. data mart

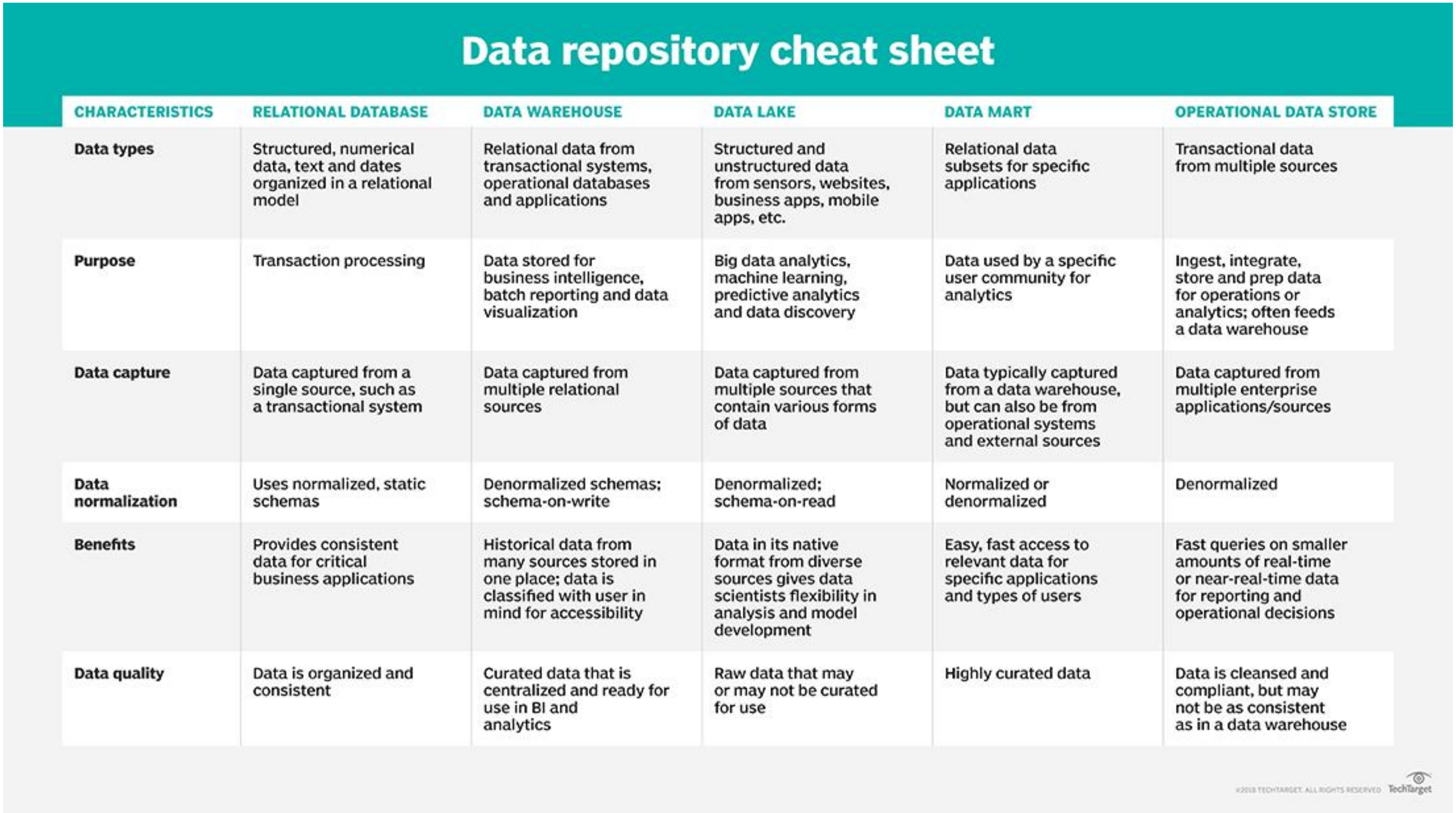
Bridget Botelho, Editorial Director

The vast amount of data organizations collect from various sources goes beyond what traditional relational databases can handle. This leads to the data warehouse vs. data lake question -- when to use which one and how each compares to data marts, operational data stores and relational databases.

All of these data repositories have a similar core function: housing data for business reporting and analysis. But their purpose, the types of data they store, where it comes from and who has access to it differs.

In general, data comes into these repositories from systems that generate data -- CRM, ERP, HR, financial applications and other sources. The data records created from those systems are applied against business rules and then sent to a data warehouse, data lake or other data storage area.

Once all the data from the disparate business applications is collated onto one data platform, it can be used in business analytics tools to identify trends or deliver insights to help make business decisions.



Data repository cheat sheet

CHARACTERISTICS	RELATIONAL DATABASE	DATA WAREHOUSE	DATA LAKE	DATA MART	OPERATIONAL DATA STORE
Data types	Structured, numerical data, text and dates organized in a relational model	Relational data from transactional systems, operational databases and applications	Structured and unstructured data from sensors, websites, business apps, mobile apps, etc.	Relational data subsets for specific applications	Transactional data from multiple sources
Purpose	Transaction processing	Data stored for business intelligence, batch reporting and data visualization	Big data analytics, machine learning, predictive analytics and data discovery	Data used by a specific user community for analytics	Ingest, integrate, store and prep data for operations or analytics; often feeds a data warehouse
Data capture	Data captured from a single source, such as a transactional system	Data captured from multiple relational sources	Data captured from multiple sources that contain various forms of data	Data typically captured from a data warehouse, but can also be from operational systems and external sources	Data captured from multiple enterprise applications/sources
Data normalization	Uses normalized, static schemas	Denormalized schemas; schema-on-write	Denormalized; schema-on-read	Normalized or denormalized	Denormalized
Benefits	Provides consistent data for critical business applications	Historical data from many sources stored in one place; data is classified with user in mind for accessibility	Data in its native format from diverse sources gives data scientists flexibility in analysis and model development	Easy, fast access to relevant data for specific applications and types of users	Fast queries on smaller amounts of real-time or near-real-time data for reporting and operational decisions
Data quality	Data is organized and consistent	Curated data that is centralized and ready for use in BI and analytics	Raw data that may or may not be curated for use	Highly curated data	Data is cleansed and compliant, but may not be as consistent as in a data warehouse

Data warehouse vs. data lake

Organizations typically opt for a data warehouse vs. a data lake when they have a massive amount of data from operational systems that needs to be readily available for analysis. Data warehouses often serve as the single source of truth because these platforms store historical data that has been cleansed and categorized.

While data warehouses retain massive amounts of data from operational systems, a data lake stores data from more sources. It is essentially a collection of various raw data assets that come from an organization's operational systems and other sources.

Because the data within data lakes may be uncurated and can originate from sources outside of the company's operational systems, it isn't a good fit for the average business analytics user; rather, data lakes are the playground of data scientists and other data analysis experts.

To remember how a data warehouse vs. data lake differ, picture actual warehouses and lakes: warehouses store curated goods from specific sources, where a lake is fed from rivers, streams and other sources, and the content is raw.

Data warehouse vendors include AWS, Cloudera, IBM, Google, Microsoft, Oracle, Teradata, SAP, SnapLogic and Snowflake, to name some of the many options. Data lakes are available from AWS, Google, Informatica, Microsoft, Teradata and other data management providers.

Data warehouse vs. data mart

Data marts are often confused with data warehouses, but the two serve markedly different purposes.

A data mart is typically a subset of a data warehouse; the data within it often comes from a data warehouse -- though it can come from another source. The data sent to a data mart is highly curated for a specific community of users -- such as a sales team -- to allow them to find the data they need quickly. The data is held there for specific uses, such as financial analytics.

Data marts are also much smaller than data warehouses -- they hold tens of gigabytes vs. the hundreds of gigabytes to petabytes of data that can be held in a data warehouse.

Data marts can be built from an existing data warehouse or other data source system by designing and constructing the database table, populating it with relevant data and deciding who can access it.

Data warehouses vs. ODS

An operational data store (ODS) is a type of database that serves as the interim holding area for all the data that's about to enter the warehouse. Think of it as the warehouse loading dock, where goods are delivered, examined and verified. While in the ODS, data can be scrubbed, checked for redundancy and checked for compliance with business rules before entering the warehouse.

Queries can be made against data in the ODS, but the data there is transient, so it only supplies information for queries about, say, the status of a customer order that's in progress.

An ODS is typically run on a relational database management system or on the Hadoop platform. Data is supplied to the ODS using data integration and data ingestion tools, such as Attunity Replicate or Hortonworks DataFlow.

Relational databases vs. data warehouses and data lakes

The main difference between a data warehouse vs. data lake vs. relational database system is that a relational database is used to store and organize structured data from a single source, such as a transactional system, while data warehouses are built to hold structured data from multiple sources. Data lakes differ from both in that they store unstructured, semi-structured and structured data.

Relational databases are relatively simple to create and can be used to store and organize transaction data. The downside of relational databases is that they don't support unstructured data or the vast amount of data being generated today. That brings us to the data warehouse vs. data lake decision. Still, many companies continue to rely on relational databases today for tasks such as operational data analysis or trend analysis.

Relational databases available on premises or in the cloud include Microsoft SQL Server, Oracle Database, MySQL and IBM DB2, as well as Amazon Relational Database Service, Google Cloud Spanner and others.

↘ Next Article

In this e-guide

What are the main features of data catalog software?

Bridget Botelho, Editorial Director

Employees who rely on self-service business analytics tools to make data-driven business decisions need access to a lot of data, but they can't be allowed to just pull raw data out of a data lake or other big data repositories; the data they use must be curated to ensure it is accurate and appropriate. That's where data catalog software comes in.

A data catalog is a type of metadata management system that is user-friendly enough for the average business user. Data catalogs are used to build portals in which users can find data that has been curated by data stewards or other data professionals. They classify the data in terms that business users understand and provide context around the data so it can be used in analytics applications.

This type of metadata management tool is in high demand as businesses struggle to inventory all the data they collect, as well as to comply with data privacy rules, such as the European Union's General Data Protection Regulation.

Analyst firm Gartner recommends the use of data catalog software to curate inventories of available data assets and to map information supply chains. These tools are an essential component of corporate data management strategies, according to the firm.

How data catalog software works

Sharon Graves, enterprise data evangelist and Tableau Server administrator at web hosting giant GoDaddy, implemented data catalog software from Alation Inc. in 2015 to reduce the time analytics users spend searching for the right data and to ensure the data they access has been vetted by data stewards.

"There is a problem where we have users who don't know anything about which data source to use or where to find the data. We needed to point users to a tool," she said. "We wanted our analysts to be spending their time doing analysis, and we wanted to support end users doing simple charting and crosstabs."

The data catalog pulls in metadata from various locations -- Hadoop, Amazon Redshift, Apache Hive, Tableau Server, Teradata and other sources -- and gathers it all in a portal where users can search for relevant data. It sorts the data based on a number of factors, including whether the data steward has endorsed the data for use in certain applications, and by the popularity of the data – which can be finagled by data experts to ensure the right data surfaces first, Graves said. Data teams can also build unified or packaged data sets that take care of data joins for users, she added.

Traditional metadata management capabilities are at the core of data catalog software, including business glossaries, data lineage and impact analysis, along with modern features, such as self-generating topic extraction, taxonomy generation, semantic discovery, machine-learning pattern mapping and knowledge graphing, according to Gartner. All in all, data catalogs enable companies to get the most value out of the data that sits in data lakes by making it easy to find and apply in business analysis.

In addition to Alation, other vendors offer data catalog software either as part of their metadata management tools or as stand-alone offerings, including Attivio, Cambridge Semantics, Collibra, Informatica, Microsoft, Oracle, SAP and Waterline Data.

↘ Next article

In this e-guide

About SearchDataManagement

SearchDataManagement.com is a knowledge guide for data management professionals and business intelligence leaders. We offer a rich collection of insights, tips and advice on how you can efficiently manage your data supply chain.

With the ever-growing volume of internal and external data flooding into businesses, educated data management strategies are essential.

Our site helps DBAs, developers, programmers, and data scientists to process relevant data efficiently, integrate traditional warehouses with modern tools like Hadoop and Spark, select a compliant governance strategy, ensure data quality/trustworthiness, and better streamline techniques and practices to enable informed, knowledgeable business decisions.

For further reading, visit us at:
SearchDataManagement.com

Images: Fotalia