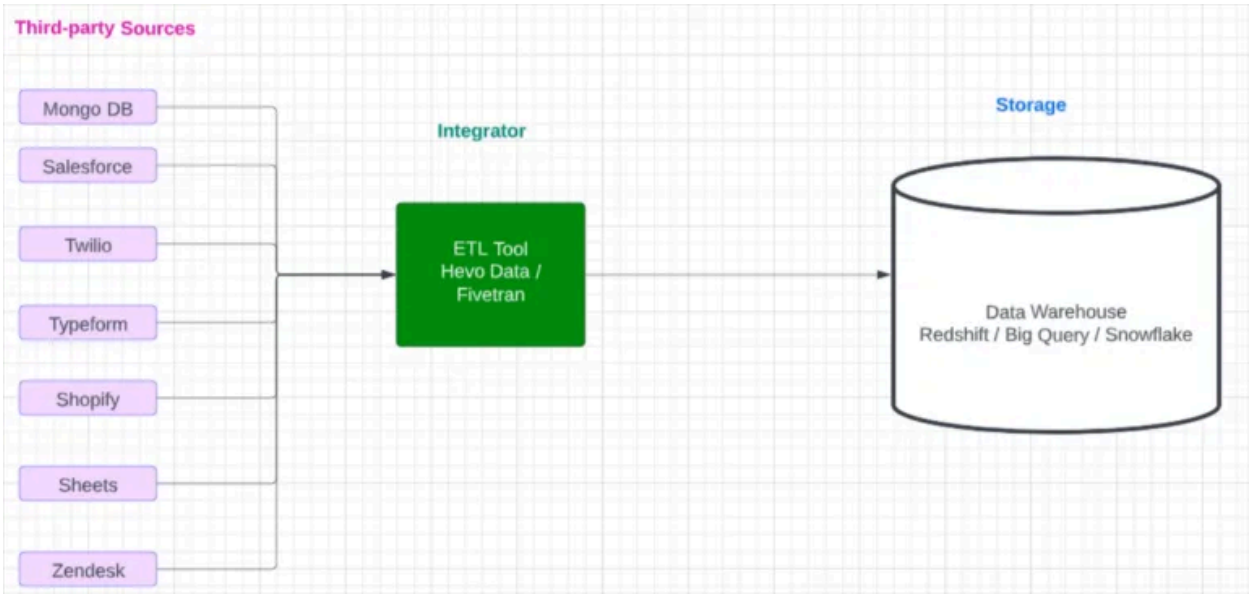
MONTE CARLO

DATA PLATFORMS Updated Feb 15 2023

Data Warehousing Guide: Fundamentals & Key Concepts



AUTHOR | Michael Segner

Since the inception of the cloud, there has been a massive push to store any and all data. On the surface, the promise of scaling storage and processing is readily available for databases hosted on AWS RDS, GCP cloud SQL and Azure to handle these new workloads.

The problem is that these databases belong to the OLTP (online transaction processing) category of databases, which are not built to handle billions of rows and take anywhere from 30 minutes to a few hours to return the resultset for one SQL query. The cost of running these queries is very high due to the lack of query optimization.

Cloud data warehouses solve these problems. Belonging to the category of OLAP (online analytical processing) databases, popular data warehouses like Snowflake, Redshift and Big

fastest growing in the world. Data analysts and engineers can run the queries they want to run when they want to run them without worrying about excessive load times or statement timeouts.

This article will define in simple terms what a data warehouse is, how it's different from a database, fundamentals of how they work, and an overview of today's most popular data warehouses.

What is a data warehouse?

A data warehouse is an online analytical processing system that stores vast amounts of data collected within a company's ecosystem and acts as a single source of truth to enable downstream data consumers to perform business intelligence tasks, machine learning modeling, and more.

What's the difference between a transactional database and a data warehouse?

It's possible to use a database meant for OLTP as a data warehouse, but as your data grows and the queries become more complex, operations start to slow down, ultimately resulting in deadlocks and missed data.

The distinction we're making here are tools, such as PostgreSQL that can be used as transactional databases, versus BigQuery, a data warehouse. A data warehouse tool should be optimized for analytical queries, with features such as columnar storage, that make it much faster to process common ad-hoc questions.

Let's imagine a scenario where you're collecting orders information. A transactional database would insert every new order as a new row, so you can imagine a row-based storage would fit this type of process.

If we imagine then analyzing one column, such as total order price, we'd want to add up the order prices across one column, making a columnar store data warehouse much more suitable for running that example analytical query.

Fundamentals of Data Warehouses

Three types of common data transformations within adata warehouse are:

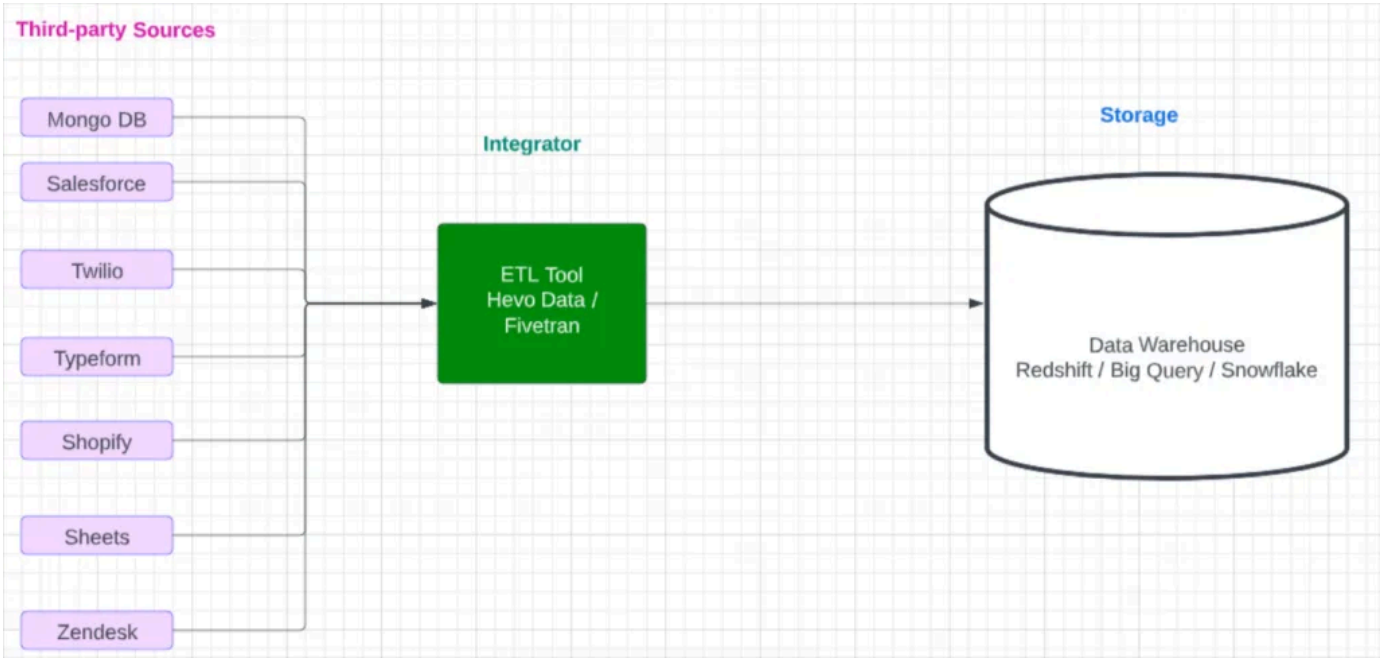
- **tegration** – It acts as an endpoint for data from various sources such as APIs and databases

valuable information

Integration

A data warehouse comes to life when massive troves of data get ingested into it. In order to ingest a huge volume of data, a data warehouse needs to be tightly integrated with all popular third-party data-generating tools. Most data warehouses either have native connectors with third-party tools or rely on ETL players to extract data from these sources and insert them into the data warehouse.

The following diagram explains how integrations work.



A company's production data, third-party ads data, click stream data, CRM data, and other data are hosted on various systems.

An ETL tool or API-based batch processing/streaming is used to pump all of this data into a data warehouse. Once all of the siloed data is cleaned and consolidated inside a data warehouse, insights can be derived from them.

ɹning

Generally, raw data can be unclean and enter into the data warehouse, but once inside data engineers can run data quality checks using *DISTINCT*, *ROW_NUMBER()*, etc. keywords. The *IS NULL* keyword is used to check for missing values. The *MIN() and MAX()* functions are used to check the range of values. If there are any outliers they are replaced with the appropriate value. More commonly, data engineers will build unit tests within the pipeline itself, for example using the Airflow ShortCircuitOperator to prevent bad data from landing in the warehouse.

There are a few challenges with this approach. First, it is impossible to build tests for all the ways data can break let alone scale those tests across all of your constantly evolving pipelines. Second, to reduce your time-to-detection you need to be end-to-end across your entire data system which may include warehouses or lakes from other vendors or other components of the modern data stack. Finally, where and how the data pipeline broke isn't always obvious.

Monte Carlo solves these problems with our our data observability platform that uses machine learning to help detect, resolve and prevent bad data.

Consolidation

The main goal of ingesting data into a data warehouse is to perform SQL or Python-based analysis using business intelligence tools/notebooks or to prepare the data for downstream machine learning models. In both of these cases, the data needs to be consolidated.

Given that a data warehouse stores data from multiple sources, SQL queries are written to consolidate data from the multiple sources. Transformation tools like dbt are also very popular low-code, no-code alternatives to build data models and consolidate data so that it is ready to be consumed.

For example, during the consolidation process:

- clickstream data is combined with financial data to gauge the spending of users versus interaction with the app
- Product analytics data is combined with ads data to gauge the retention of users

Masking of personally identifiable information and transformation of data is done during the lidation process as well.

MONTE CARLO

Data Loading

This is one of the key functions of any data warehouse. Data can be loaded using a loading wizard, cloud storage like S3, programmatically via REST API, third-party integrators like Hevo, Fivetran, etc. Data can be loaded in batches or can be streamed in near real-time. Structured, semi-structured, and unstructured data can be loaded.

> **Can a data warehouse store unstructured data?**
> Yes, data warehouses can store unstructured data as a blob datatype. Later this blob can be used for analysis, text mining, natural language processing, etc.

Data Transformation

Raw data ingested into a data warehouse may not be suitable for analysis. They need to be transformed. Data engineers use SQL, or tools like dbt, to transform data within the data warehouse.

Data Security

Data Warehouses achieve security in multiple ways. For example, some data warehouses:

- Can only be accessed using a private cloud.
- Can only be accessed using a specific machine or location.
- Can only be accessed during a certain time of the day
- Can only be accessed using multi-factor authentication.

Access to a data warehouse does not guarantee access to all of its databases, schema, tables, or views.

Every user is tied to a role and every role only has needs-based access. Sometimes access can go nular as column level masking.

Synapse Analytics.

Snowflake

Snowflake was the first company to split compute and storage inside a data warehouse. Snowflake is the pioneer in cloud data warehousing. They have a marketplace where third-party data can be bought and leveraged for a variety of use cases. They recently announced Snowflake Data Sharing, a critical new methodology where one users can share data with another entity similar to sharing a file. [Read: How To Migrate To Snowflake Like A Boss]

AWS Redshift

Redshift is the official data warehouse of Amazon cloud services. Since Redshift belongs to the AWS ecosystem it is easy to move data from AWS RDS, S3 bucket, etc into Redshift. Redshift offers a serverless tier where one can forget about infrastructure management.

Azure Synapse

Azure Synapse studio is powered by Microsoft and comes with a lot of features like Ingest, Explore, Analyze, and Visualize. It also natively integrates with Apache Spark. It has its own notebooks, dataflow integrations, and spark job definitions.

Google BigQuery

BigQuery is famous for giving users access to public health datasets and geospatial data. It has connectors to retrieve data from Google Analytics and all other Google platforms. It natively powers Google Data Studio and supports operationalizing machine learning models that can be exported for downstream analysis. They recently enabled cloud Pub/Sub to write directly into BigQuery making data streaming a simpler and less time consuming endeavor.

No matter which you choose, all modern data warehouses:

- e encrypted at rest
- eparate storage from compute

- Store data in a columnar format for faster query processing
- Can query semi-structured data

Summary – Data Warehouses Cheat Sheet

Prefer to learn in bullet point form? Here's our cheat sheet with everything you need to know about data warehouses.

Fundamentals

1. All popular data warehouses use SQL as their primary querying language.
2. Every data warehouse is different therefore their SQL flavors are also different
    1. For instance, Snowflake has the *median & ratio_to_report* functions which are not available in other data warehouses.
3. All popular modern data warehouses run on the cloud
    1. This means there is no need to install and deploy any kind of hardware
    2. Maintenance is taken care of by providers themselves
4. Data warehouses are split into a few layers
    1. Storage
    2. Processing
    3. Infrastructure management
5. One can connect to data warehouses using
    1. Web clients
    2. Command line
    3. BI tools like Tableau & Mode analytics
    4. Programmatic languages like Python
    5. Third-party ETL connectors like Hevo

Key Concepts

1. There is no need to provision hardware or software.
2. There is no need to update or maintain modern data warehouses
    1. Data warehouse providers take care of it
3. Based on need, the hardware can be changed by clicking a few buttons

2. Performance at scale
    1. Data warehouses are built to query billions of rows (structured & semi-structured)
    2. Data can be streamed in and out of warehouses multiple times within one second
    3. More than 15,000 rows can be inserted in one go

3. Usage-based pricing
    1. Pay only when you query, load, or unload.
    2. Pay only for the resources used at that point in time

4. Central repository (Single source of truth)
    1. Data from multiple sources are migrated into data warehouses
    2. They are cleaned, transformed, and aggregated to derive insights
    3. Business Intelligence tools like Sisense data and Tableau are plugged on top of aggregated data
    4. Aggregated data are also fed into machine learning models

5. Highly secure
    1. Data warehouses are often HIPPA and GDPR compliant
    2. Data can be masked as needed
    3. Supports MFA (multi-factor authentication)
    4. Data is encrypted during transport & rest
    5. Access controls can go as granular as possible

6. Data Marketplace
    1. Data warehouses come with a marketplace which gives us access to
        1. Geospatial data
            1. This data can be joined with production data to perform demography-based analysis

MONTE CARLO

3. Public sector financial data
    1. This can be used to monitor the economy

Interested in improving the data quality within your data warehouse? Set up a time to talk to us in the form below?

Your work email*

First name* Last name*

Talk To Us!

Our promise: we will show you the product.

Frequently Asked Questions

What is a real-life example of data warehousing?

A real-life example of data warehousing is using it to integrate, clean, and consolidate data from various sources within a company to perform business intelligence tasks, such as analyzing customer behavior or combining clickstream data with financial data to gauge spending and interactions.

What is a major feature of a data warehouse?

A major feature of a data warehouse is its ability to store vast amounts of data in a columnar format for faster query processing, enabling it to handle complex analytical queries efficiently. Data warehouses also offer features like data integration, transformation, and high security.

Search...

MONTE CARLO

You're ready to experience the power of data observability. Awesome! So, what's next?

Get the Guide

An enterprise data platform, often referred to as a 'modern data stack,' is the central processing hub for an organization's data ecosystem. The data platform manages the collection, normalization, transformation, and application of data for a given data product—from business insights and dashboards to ML and AI engineering.

Read the Blog

Freshly's Journey to Building Their 5-Layer Data Platform Architecture

For Freshly, food isn't the only thing that needs to be delivered fresh and fast; our data also needs to be reliable, timely, and most importantly, accurate. To achieve this, we invested in a cloud-based data platform architecture (often referred to as a 'modern data platform') that enables us to deliver insights quickly to data consumers across the company.

Read the Case Study

Freshly's Journey to Building Their 5-Layer Data Platform Architecture

For Freshly, food isn't the only thing that needs to be delivered fresh and fast; our data also needs to be reliable, timely, and most importantly, accurate. To achieve this, we invested in a cloud-based data platform architecture (often referred to as a 'modern data platform') that enables us to deliver insights quickly to data consumers across the company.

Read the Case Study

Building the Next Generation Data Platform

As companies increasingly rely on data to drive decision making and power digital products, the need to build

Monte Carlo Recognized as the #1 Data Observability Platform by G2 for 6th Consecutive Quarter

MONTE CARLO

Watch On Demand

is especially meaningful to our team because G2 relies on feedback and ratings from real customers — individuals who use these tools daily to accomplish their tasks and create more value for their business.

Learn More About Monte Carlo

Read more posts.

5 Steps To A Successful Data Warehouse Migration

Read more ▶

MONTE CARLO

The Hidden Threats in Your Data Warehouse Layers (And How to Fix Them)

Read more ▶

.he Future of Data Management: 8 Fast Growing Trends

MONTE CARLO

The Future of Data Warehousing

Read more ▶

MONTE CARLO

The 6 Data Quality Dimensions with Examples

Read more ▶

Eliminate data downtime.

Request a Demo

Product

What is data observability?

:tect anomalies

Triage incidents

Company

About us

Customers

Pricing

MONTE CARLO

Integrations

Solutions

Verticals

Financial services

Healthcare and life sciences

Technology

Advertising, media, and entertainment

Retail/CPG

Startups

Use cases

Data quality monitoring and testing

Data mesh and self serve

Report and dashboard integrity

Customer-facing data products

Cloud migrations

Infrastructure and cost management

Resources

Blog

Case Studies

Docs

eBooks, Reports, & Guides

Webinars

Commercial Terms and Conditions

Data Processing Addendum

Website Terms of Use

Privacy Policy

Cookie Policy

Report Issues

Security

From the blog

Monitoring Unstructured Data with Monte Carlo

Learn how to monitor unstructured data in Monte Carlo to ensure the reliability of your production AI.

Read now ▶

The Future of Reliable Data + AI—Observing the Data, System, Code, and Model

derstanding what stands in the way of reliable AI applications – and what you can do about it.

MONTE CARLO

Managing a data catalog manually is error-prone and time-consuming. These 20 popular tools can help your team maintain consistent data definitions.

Read now