

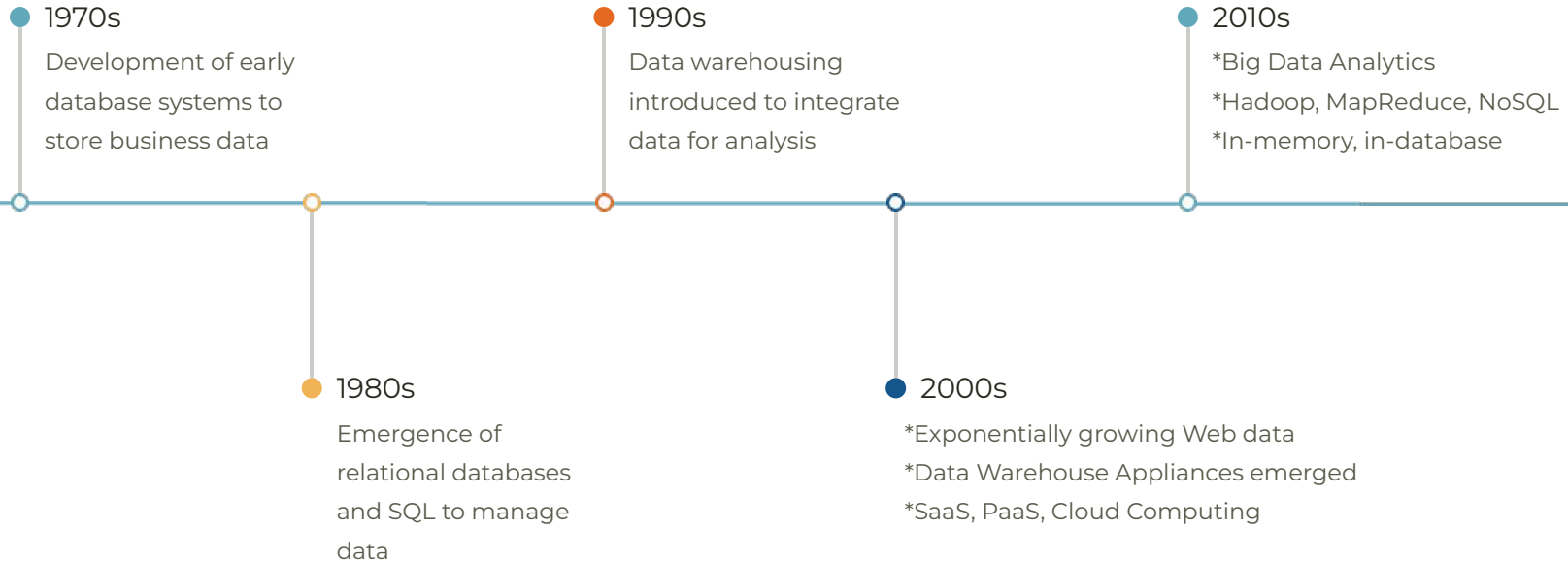


Data Warehousing Products

Avi Kapadia, Avinash Athota, Bill Xiao, Dylan Keskinian, Kyungmin Park



Historical Timeline

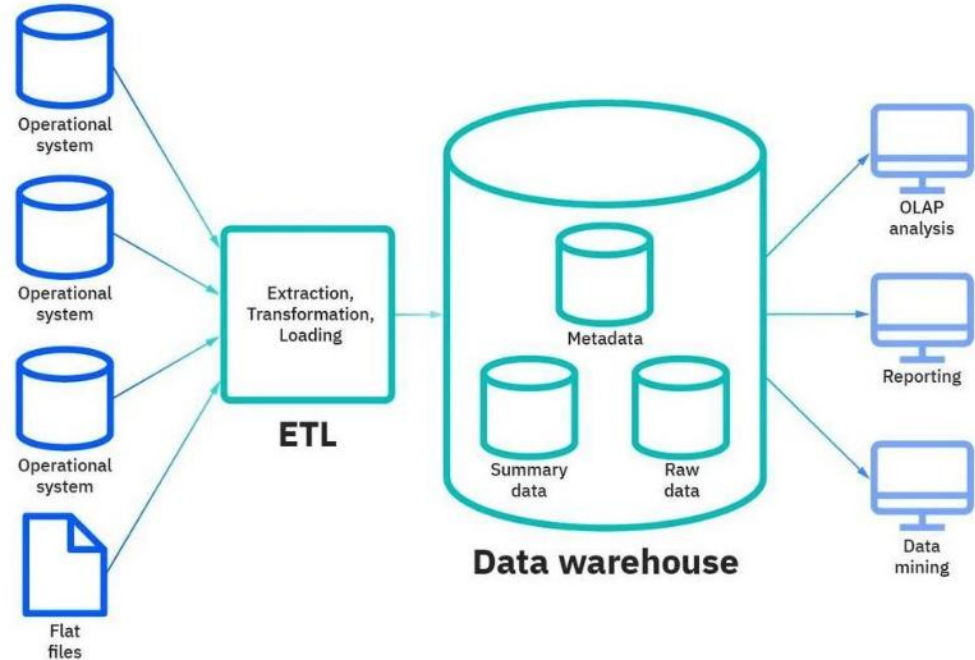


What is Data Warehouse?



Data warehouse

A central repository of integrated data from multiple sources, organized to support business intelligence activities



A data warehouse enables advanced business intelligence by consolidating data from multiple sources into a central repository optimized for analytics.

Data Warehouse Components



Staging area

Land raw data from multiple sources into a staging area before processing



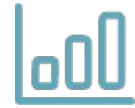
ETL process

Extract data from sources, transform and cleanse it, and load into the data warehouse



Database

Store integrated data in a relational database optimized for analytics



BI tools

Leverage BI tools to analyze data and create dashboards/reports

These core components work together to collect, store, and analyze data for business intelligence.

Business Use Cases



Business intelligence

Data warehouses enable advanced business intelligence analytics on large volumes of data, allowing businesses to gain insights.



Strategy Planning Team



Dashboards and visualizations

Data warehouses support interactive dashboards and data visualizations to help businesses identify trends, patterns, and insights.



Data Analytics Team



Ad-hoc querying

Data warehouses allow users to query large datasets on-demand for specific insights without relying on predefined reports.



Data Team



Predictive analytics

Data warehouses combined with statistical models and machine learning enable businesses to predict future outcomes and trends.



Sales, Product Development Team

Data warehouses are critical for businesses to store and analyze large amounts of data from multiple sources, powering a range of analytics and insights.

Business Value



Improved decision making

Data warehousing provides timely, accurate data for analysis to support better business decisions.



Increased operational efficiency

Data warehousing enables consolidation of data from multiple sources for unified reporting and analytics to uncover process improvements.



Enhanced customer service

Data warehousing gives a 360-degree customer view to deliver personalized service and identify new opportunities.

In summary, data warehousing delivers significant business value through improved analytics and decision making, leading to greater efficiency, better customer service and more informed strategy.

Data Warehouse Tools

ORACLE



teradata.



Introduction to Data Warehousing Features

- Scalability
- Data Integration
- Data Storage
- Query and Analysis
- Data Security
- Performance Optimization
- Data Governance
- Cloud Deployment





Scalability

Scalability in data warehousing ensures that the system can efficiently grow to handle increasing data volumes and user demands without sacrificing performance.



Data Integration

Data Integration is the process of combining data from multiple sources, such as databases and applications, into a unified data warehouse, enabling comprehensive analysis and insights.



Data Storage

Data Storage encompasses the robust storage capabilities of a data warehousing system, ensuring efficient organization and secure storage of structured and unstructured data.



Query and Analysis

Query and Analysis capabilities enable users to run complex queries and perform in-depth analytical operations on large datasets, empowering organizations to derive valuable insights and make informed decisions.



Data Security

In data warehousing, ensuring robust data security is paramount to safeguarding sensitive information stored within the system and maintaining compliance with regulatory requirements.



Performance Optimization

Performance optimization in data warehousing involves implementing strategies and techniques to enhance the speed and efficiency of data processing and retrieval, ensuring optimal system performance for timely analysis and decision-making.



Data Governance

Data Governance in data warehousing involves establishing policies, procedures, and controls to ensure data quality, integrity, and compliance, fostering trust and confidence in the data.



Cloud Deployment

Cloud Deployment in data warehousing offers organizations the flexibility, scalability, and cost-effectiveness of deploying and managing their data infrastructure in a cloud environment, enabling rapid scalability and reducing operational overhead.

AWS Redshift (Modes of Operation and Data Security)

Loading Data

- Supports built-in loading from:
 - AWS-S3
 - AWS-GlueCatalog
 - AWS-EMR
 - AWS-EC2
 - Other ETL-enabled tools
- COPY command

Querying Data

- Massively Parallel Processing (MPP) architecture

Data Security

- Encryption
 - AES-256
 - SSL
- Access Control
 - AWS Identity and Access Management (IAM) integration
 - Define roles for access into Redshift's cluster
- Network Isolation
 - Deploy into Amazon Virtual Private Cloud (VPC)



Comparison of Warehousing Technologies

Feature/Aspect	AWS Redshift	Snowflake	IBM Db2	Oracle	SAP	Teradata	GCP BigQuery	DynamoDB
Scalability	Y	Y	Y	Y	Y	Y	Y	Y
Data Integration	Y	Y	Y	Y	Y	Y	Y	Y
Data Storage	Y	Y	Y	Y	Y	Y	Y	Y
Query and Analysis	Y	Y	Y	Y	Y	Y	Y	Y
Data Security	Y	Y	Y	Y	Y	Y	Y	Y
Performance Optimization	Y	Y	Y	Y	Y	Y	Y	Y
Data Governance	Y	Y	Y	Y	Y	Y	Y	Y
Cloud Deployment	Y	Y	Y	Y	Y	Y	Y	Y
Unique Feature/Aspect	Complex queries + AWS integration	Auto-scaling compute resources	Deep integration with IBM	Extensive enterprise features	Advanced analytics & business	Industry-specific solutions	Seamless integration with GCP	Serverless NoSQL with auto-scaling

Technical Specifications

Amazon Redshift

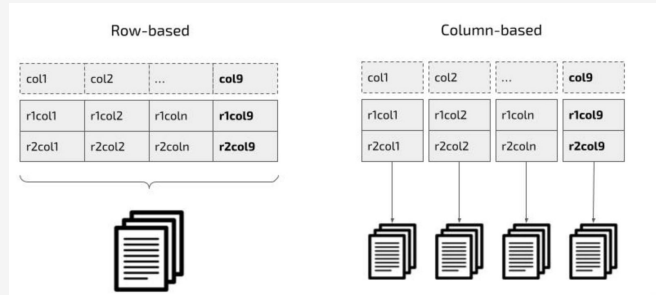


Amazon
Redshift

AWS Redshift (Architecture)

Columnar-Based

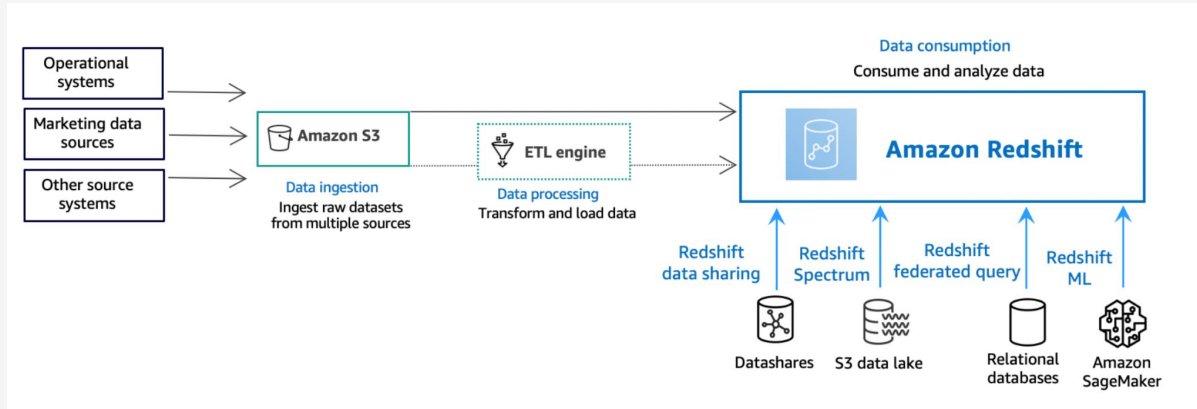
- Very fast read access
- Common in Data Warehousing Technologies
- Quickly reads aggregate data across columns



Cluster System

- Core component of Redshift Architecture
- Cluster: collection of 1+ nodes
 - Contains 1+ databases
- Leader Node
 - Receives queries from clients
 - Parses queries
 - Develops execution plans
 - Aggregates results from compute nodes
 - Returns result to client
- Compute node
 - Execute queries and computations
 - Has its own CPU, memory, disk storage

AWS Redshift (Typical Working Scenario)



- Data is first loaded into system
- Then perform analysis with SQL
 - Create reports and dashboards based on analysis

AWS Redshift (Modes of Operation and Data Security)

Loading Data

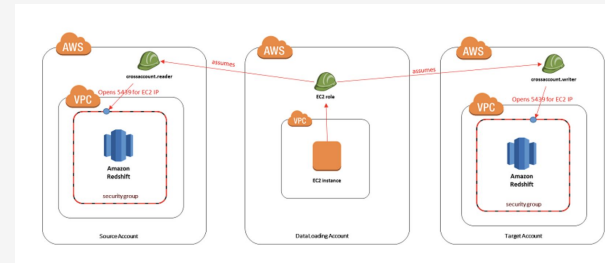
- Supports bulk loading from:
 - AWS S3
 - AWS DynamoDB
 - AWS ER
 - AWS Glue
 - Other SSH-enabled hosts
- COPY command

Querying Data

- Massively Parallel Processing (MPP) architecture

Data Security

- Encryption
 - AES-256
 - SSL
- Access Control
 - AWS Identity and Access Management (IAM) integration
 - Define roles for access into Redshift cluster
- Network Isolation
 - Deploy into Amazon Virtual Private Cloud (VPC)



Technical Specifications

Snowflake



Snowflake (Architecture)

Database Storage Layer

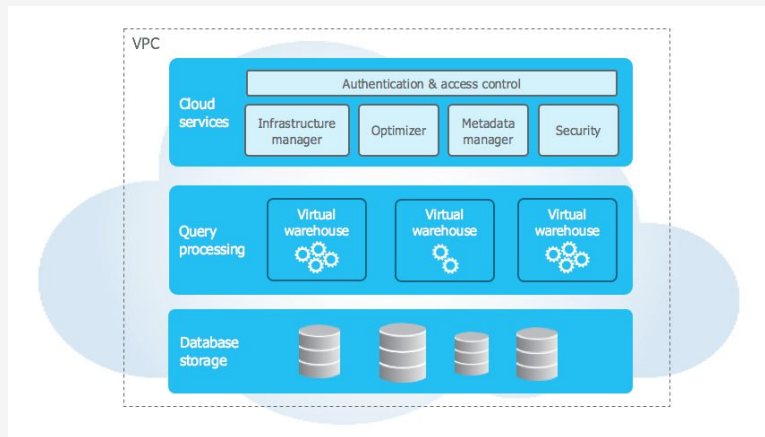
- Manages all aspects of how data is stored
 - Cloud storage
- Columnar Storage
 - Very fast read access

Cloud Services Layer

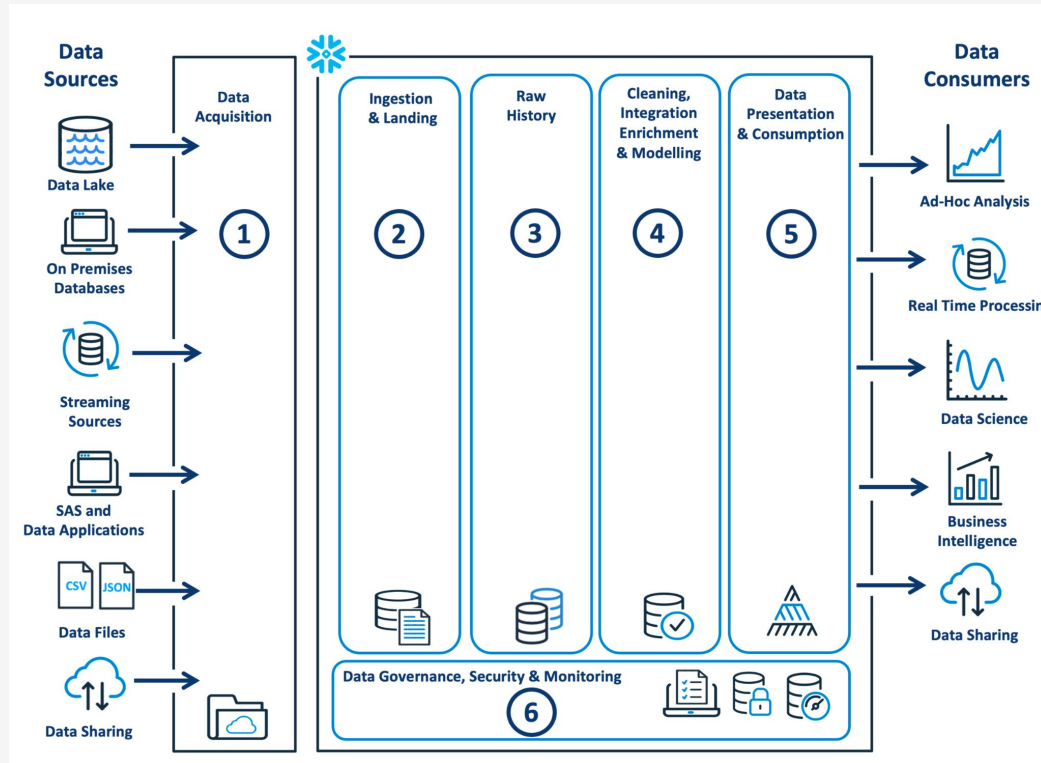
- Coordinates all activities across Snowflake
- Ensures that compute layer can scale independently from storage

Compute Layer

- Contains independent virtual warehouses
 - MPP compute clusters with multiple compute instances
- Warehouses can be scaled up/down without affecting storage layer or other warehouses



Snowflake (Typical Working Scenario)



Snowflake (Modes of Operation)

Data Loading






- COPY command usage to load data
- Use of Snowpipe for ingestion
 - Loading of real-time data as it arrives

Querying and Analysis

- Analysis through SQL queries, data analysis, and DML operations
- Performance can be scaled through adjusting size of warehouse

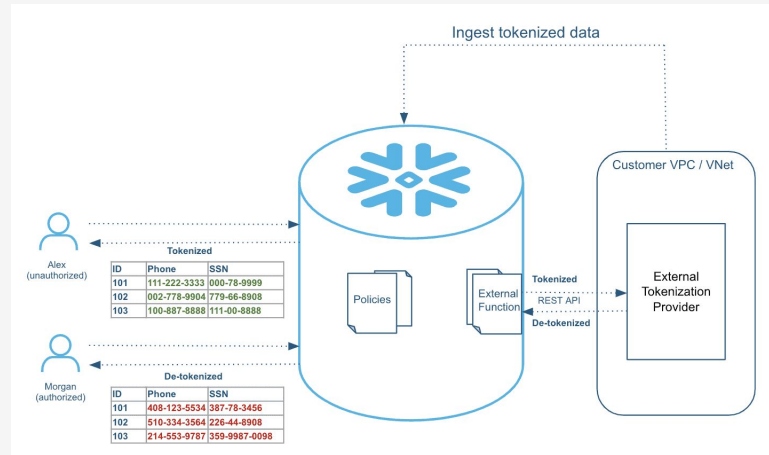
Data Sharing and Exchange

- Secure data sharing solutions with clients
- Share subsets of their data with clients or partners without copying or moving the data

STANDARD	PREMIER	ENTERPRISE	ENTERPRISE FOR SENSITIVE DATA	VIRTUAL PRIVATE SNOWFLAKE (VPS)
				
Complete SQL Data Warehouse Data Sharing Business hour support M-F 1 day of time travel Always-on enterprise grade encryption in transit and at rest Customer dedicated virtual warehouses	Standard + Premier Support 24 x 365 Faster response time SLA with refund for outage	Premier + Multi-Cluster warehouse Up to 90 days of time travel Annual rekey of all encrypted data Materialized Views	Enterprise + HIPAA support PCI compliance Data encryption everywhere Tri-Secret Secure using customer-managed keys AWS PrivateLink support Enhanced security policy	Enterprise for Sensitive Data + Customer dedicated virtual servers wherever the encryption key is in memory Customer dedicated metadata store Additional operational visibility
\$2.00 compute cost per credit	\$2.25 compute cost per credit	\$3.00 compute cost per credit	\$4.00 compute cost per credit	

Snowflake (Data Security)

- Encryption
 - All data encrypted using AES-256 bit
- Access Control
 - Role-based access control (RBAC)
 - Define user roles and assign privileges
- Network security
 - Can configure VPCs or virtual networks to control access to their Snowflake environment
- Security can also be controlled via third-party tokenization



Snowflake Sample Applications

E-commerce Optimization

- Virtual Warehouses enable on-demand, scalable resources which supports dynamic pricing models
- Storing and querying of semi-structured data allows for easy segmentation of customer data effectively
- Native support for ML tools and external ML services helps in creating highly targeted marketing campaigns

Content Personalization and Delivery

- Secure Data Sharing allows live data to be shared across business units or external partners without copying or moving data
- Native support for ML to analyze customer behavior and predict trends
- Better processing and analyzing of large data volume, helping for personalization of content recommendations and optimization for CDNs

Major Companies Using Snowflake

Instacart

- Data Sharing allowed for quicker access between large datasets across departments without duplication
- Automating scaling of Virtual Warehouses helped in meeting fluctuating demands of data analytics



Doordash

- Snowpipe allowed for loading and analyzing of data from app and website in real time for immediate insights
- Geospatial Data Analysis capabilities to enhance efficiency of delivery routes and times



Netflix

- Handling of semi-structured data like JSON for deep analysis of viewer interactions
- Scalable data processing, real-time analytics, cost-effective data management, and data sharing



Netflix Using Snowflake



Scalable Data Processing

Snowflake's scalable architecture handles large volumes of user data which Netflix needs in order to process and analyze data from millions of users seamlessly



Real-time Analytics

Snowflake helps Netflix personalized recommendations and content delivery by helping in performing real-time analytics on viewer behavior, preferences, and content consumption patterns.



Data Sharing

Netflix securely shares live data with revocable access rights across different business units and partners using Snowflake.



Cost-Effective Data Management

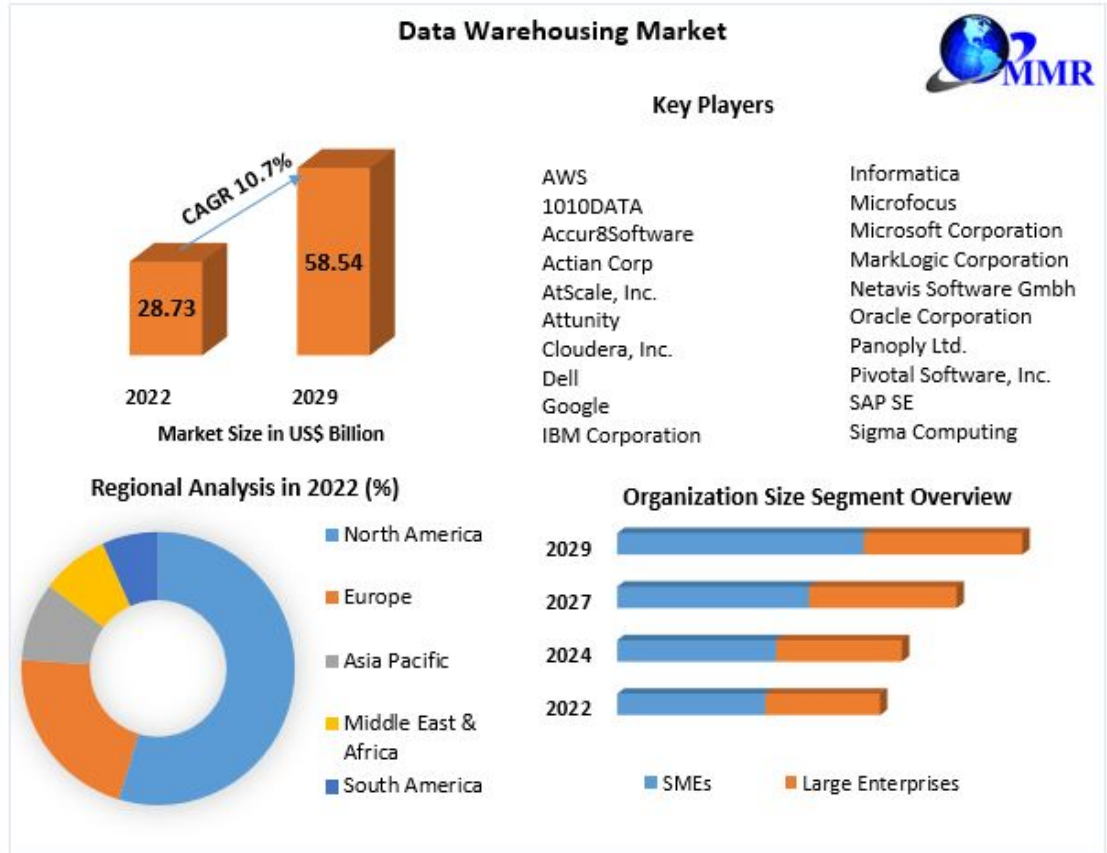
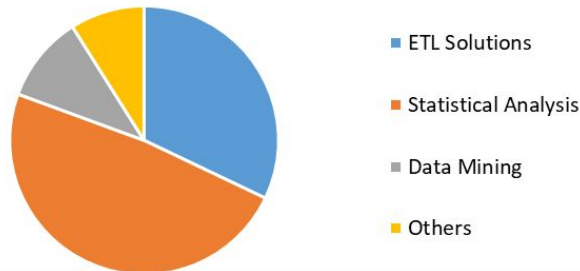
Snowflake's pay-as-you-go model allows Netflix to manage costs effectively based on their processing needs.

Snowflake helps Netflix streamline their data analytics processes, improve viewer experience through personalized recommendations, and optimize content delivery based on real-time insights.

Data Warehousing Market Overview

- \$28.73 billion market cap globally, projected to reach \$58.54 billion by 2029
- Equally dominated by large and small/medium enterprises, but SME expected to grow
- Majority of market use centered around statistical analysis but ETL solutions is rising in %

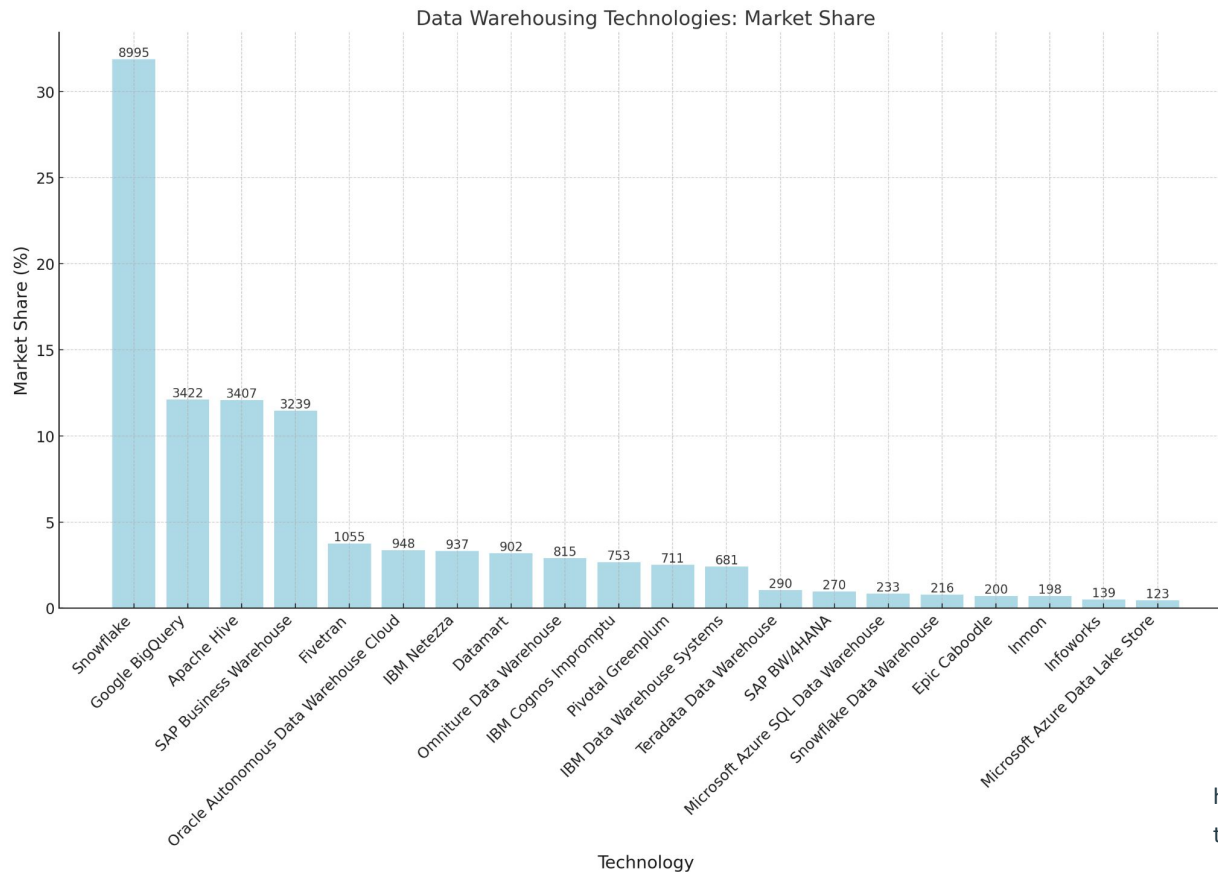
Data Warehousing Market, by Offering Type 2022 (%)



[illegible]

<https://www.g2.com/categories/data-warehouse>

Data Warehousing Market Usage Comparison



The Evolution & Future of Data Warehousing

- Hybrid and Multi-Cloud Deployments
- Advanced Analytics and AI Integration
- Data Privacy and Governance
- Serverless Data Warehousing
- Data Lakes Integration



1. Hybrid and Multi-Cloud Deployments

- Organizations are increasingly adopting hybrid and multi-cloud strategies to leverage a combination of on-premises and cloud-based infrastructure.
- The adoption of hybrid and multi-cloud approaches offers benefits such as improved data accessibility and scalability.
- Hybrid and multi-cloud deployments help mitigate the risks associated with vendor lock-in by allowing organizations to diversify their cloud providers.



2. Advanced Analytics & AI Integration

- The integration of predictive analytics and machine learning enables organizations to uncover valuable insights from their data.
- Advanced analytics and AI empower organizations to make informed decisions based on data-driven insights.
- Utilizing real-time data processing allows organizations to extract actionable insights and respond promptly to changing conditions.



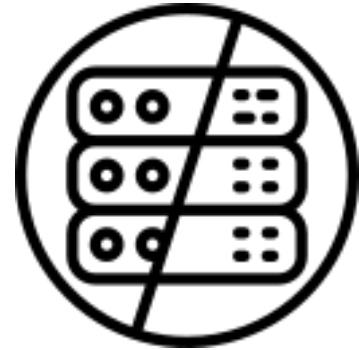
3. Data Privacy and Governance

- There is a focus on enhancing data security and compliance within organizations.
- Robust encryption methods are being implemented to safeguard sensitive data from unauthorized access.
- Automation of compliance management processes streamlines regulatory compliance efforts and ensures adherence to data protection laws.



4. Serverless Data Warehousing

- Serverless data warehousing solutions are experiencing a rise in popularity among organizations.
- These solutions offer benefits such as ease of use, scalability, and cost-effectiveness.
- Serverless data warehousing abstracts the underlying infrastructure management, allowing organizations to focus on data analytics without worrying about infrastructure provisioning and maintenance.



5. Data Lakes Integration

- There is an increasing importance placed on integration with data lakes in data warehousing strategies.
- Data lakes integration enables seamless data movement and analytics across both structured and unstructured data.
- Organizations can derive insights from diverse datasets more effectively by integrating with data lakes.



Problems

- Security and Privacy
- Data Integration
- Maintaining Data Quality
- Keeping Data Relevant

<https://www.linkedin.com/advice/1/what-potential-pitfalls-data-warehousing-architecture-qbzxe>

Research paper 1

- *A generic, flexible smart city platform focused on citizen security and privacy* by Stamatiou, Y., Halkiopoulos, C., & Antonopoulou, H. (2023).
- E-governance: structures and policies that allow city governments to deploy and use ICT's and data warehouses in order to promote interaction and cooperation between citizens and their local government
- Current warehouses' infrastructure include serious security risks:
 - Invasion of privacy
 - Sensitive data leaks
- This paper proposes and describes the basic operations of a new adaptable smart city infrastructure that considers securing people's data its top priority
- The goal is to foster trust between a city's people and government so that their prepared to face future challenges

Research paper 2

- *Cloud-based hybrid simulation model for optimizing warehouse yard operations* by Farhan, M., Ngoko, P., Halawa, F., & Mohammed, R. (2023)
- Efficiency of Amazon fulfillment center rely on their ability to convert digital orders into yard actions
- This article describes how data houses can be used to reduce yard congestion
- Data can be taken directly from Amazon Redshift to run simulations in order to find the most optimal way to handle orders