Cloud Data Warehouse Guide: Choose Relying on First-Hand Expertise

About ScienceSoft

ScienceSoft is a global IT consulting and software development company headquartered in McKinney, TX, US. Having 20+ years of experience in **data warehouse consulting and development services**, we advise companies on building cost-efficient and scalable cloud data warehouses and take on responsibility for end-to-end solution implementation. We also help businesses migrate their legacy DWH to the cloud to achieve considerable cost savings and performance gains.

- Advising on, designing and implementing DWH and BI solutions since **2005.**
- Data analytics expertise since **1989.**
- Big data consulting since **2013.**
- A **dedicated team** of industry-focused data analytics consultants, solution architects, data engineers, and data scientists to help bridge business and technology.
- Easy access to certified IT expertise and advanced technologies due to ScienceSoft's strategic partnership with **Microsoft, AWS, Oracle, and IBM.**

Cloud Data Warehouse Guide

Based on ScienceSoft's experience in designing and delivering data warehousing solutions, we've compiled a guide to help you evaluate potential cloud data warehouse vendors. The guide showcases the factors to consider when assessing the products, practical tips on how to deal with them, recommendations and precautions regarding particular products.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
Data volume	Choose a data warehouse based on your data needs in 5 years to avoid the need for migration in the near future. To define your target data volume, you may use the following formula: *Target data volume = Current data volume x (Annual growth) in power of 5* For example, your current data volume is 6 TB, the annual growth rate is 30 % (1,30 TB). Then your target data volume would be: *6 TB * 1,30 in power of 5 = 22 TB* ***NB:** The higher the data volume, the more critical the data storage cost is.*	Target data volume is < 8 TB: Azure SQL Database. Target data volume is 8 TB – 1PB (1,000 TB): **Azure Synapse Analytics, Amazon Redshift, Google BigQuery, Snowflake.** Target data volume is >1PB (1,000 TB): • **Amazon Redshift** runs analytic queries against up to 16 petabytes of data on a cluster. It queries exabytes of structured, semi-structured and unstructured data from a data lake (Amazon S3) for analyzing without loading and transformation. • **Google BigQuery** charges separately for storing cold and hot data, offering rather competitive prices. It accommodates petabyte-scale data warehouses with a low querying load (queries are charged separately).
Data sources	Examine the sets of pre-built connectors suggested by different cloud data warehouses to check if a particular data warehouse integrates with your existing data sources. Such ready-made connectors offer accelerated data loading due to optimized data transfer paths, significant cost savings, and zero maintenance.	**Azure Synapse Analytics** offers more than 95 native connectors, including connectors for Magento, Jira, PayPal, Salesforce, etc. – all backed up with Azure Data Factory. **Google BigQuery** offers native data integration with 150+ data sources via Cloud Fusion. **Amazon Redshift** console provides rapid and seamless integration with select AWS partners to bring data from such apps as Salesforce, Google Analytics, Facebook Ads, Slack, Jira, Splunk, etc. Note that using a partner solution incurs additional costs.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
Data types	Ensure that a cloud data warehouse offers the best infrastructure for storing and querying relevant data types: • Structured. • Semi-structured (JSON, Avro, ORC, Parquet, XML, etc.). • Unstructured (graph, image, invoices, IoT, weather data, etc.). To store heterogeneous data, consider data warehouse solutions, which offer: • Native support for needed data types (most cloud data warehouses support structured and semi-structured data). • Native integration with a data lake (enables storing all types of data in the data lake and SQL querying without loading data in the DWH). • Native integration with an operational database (enables storing operational data in the operational data store and its analytical querying without loading into the DWH).	Consider Amazon Redshift and Snowflake in case you need native support for semi-structured data. Amazon Redshift provides native integration with a data lake (Amazon S3) to enable querying exabytes of structured, semi-structured, and unstructured data for analysis without loading and transformation. Azure Synapse Analytics employs Azure Synapse Link to enable SQL querying on real-time operational data without data loading and transformation (cloud-native HTAP).
Data transformation and cleaning needs	A big portion of implementation and support costs lies in the ETL/ELT. The ETL/ELT costs vary because of the complexity of data aggregation and cleansing activities as well as requirements for data freshness. To enable cost savings, consider cloud data warehouse solutions that allow you to: • Decouple storage and compute resources, as well	For integrating data across AWS services into **Amazon Redshift,** the vendor offers AWS Data Pipeline, in other cases, go for AWS Data Migration Services. For batch and real-time data transformation, use AWS Glue and AWS Kinesis. In **Azure Synapse Analytics**, the data integration capabilities such as Synapse pipelines and data flows are based upon those of Azure Data Factory, which serves to move data from both on-premises and cloud source data stores.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
	as pause and resume compute resources. • Query data with no ETL (via a data lake, operational database integration). • Integrate with open-source ETL software (usually such tools offer a wide range of pre-built integrations with both source databases and a target data warehouse). Integrate with ETL software, which uses a fixed pricing model (the price doesn't depend on the workload) or charges a fixed monthly rate per connector.	
Streaming data ingestion	To enable near real-time workloads, consider the solutions, which either have a built-in streaming functionality or ingest real-time data via an add-on service, which doesn't require manual configuration for setup or running. ***NB:*** *Handling streaming should not cause high costs or much maintenance effort.*	To enable streaming data ingestion, **Azure Synapse Analytics** uses the built-in Apache Spark streaming functionality or Azure Stream Analytics event-processing engine. **Google BigQuery** suggests ingesting streaming data with the BigQuery Storage Write API or the legacy streaming API. **Snowflake** offers Snowpipe as an add-on service to enable real-time ingestion. **Amazon Redshift** does not offer a built-in capability for ingesting streaming data. However, for the near-real-time data load, you may use a native integration with Amazon Kinesis Firehose.
Scalability and concurrency	Prioritize a cloud data warehouse that decouples storage and compute resources and allows for automatic workload management. It helps accommodate and dynamically allocate different resources for different workloads (complex analytical querying, user concurrency, varying data volumes, etc.).	**Azure Synapse Analytics** decouples storage and compute and supports both manual and automatic workload management: serverless resources are scaled automatically, while the dedicated ones require manual configuration. **Snowflake** and **Google BigQuery** scale storage and compute independently, and the scaling is handled automatically. It is the optimal version for companies who want worry-free scaling flexibility.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
	Points to consider: • The number of users who will use the data stored in the data warehouse and their nature (C-level executives, analysts, managers, supervisors and IT specialists, etc.). • Share of users who will use the data from the data warehouse every day. • The need to run queries simultaneously (including from one account). • Most common data warehouse workloads (ETL, ongoing analytical querying, etc.)	**Amazon Redshift** decouples storage and compute only with the RA3 type nodes. To handle the extra load, an administrator has to add nodes to the cluster manually. If you need to add extra compute power to the cluster dynamically, you can use the concurrency scaling feature (incurs additional cost). Single databases in **Azure SQL Database** allow for automatic and manual scaling of compute resources, while Elastic pools can only be scaled manually. Azure SQL Database is a good fit for data warehousing scenarios with a large number of active users (concurrent requests can reach up to 6,400 with up to 30,000 concurrent sessions.
High availability and disaster recovery	Features you definitely need: • Automatic data backup and data replication. • Cloud data warehouse infrastructure health monitoring. • Recovering from a restore point. • Geo-redundancy.	**Google BigQuery** provides replicated storage in multiple locations charge-free by default. You can easily restore a table from any of table snapshots taken within the last 7 days. **Amazon Redshift** allows for manual and automated snapshots of a cluster. Then, the snapshots are replicated to S3 through an encrypted SSL connection in another region for disaster recovery. The snapshots are not deleted automatically – you can set the retention period for both automated and manual snapshots. Additionally, Amazon Redshift continuously monitors clusters and re-replicates data and replaces nodes in case of failures. **Azure Synapse Analytics** takes automatic snapshots of the data warehouse throughout the day to create restore points that are available for seven days (the retention period can't be changed). With the geo-backup capability, a data warehouse can be restored to a server in any other region in case the restore points in the primary region are not available.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
Data security and compliance	Major security capabilities to consider are: • Data access granularity (realized on row level, column level). • User authentication. • Data encryption (end-to-end, at rest only, by default or requiring deliberate configuration, etc.). • Encryption keys (managed by a service provider, by the customer, possibility for second-level encryption keys). • Network security (if the DWH can be deployed on a private cloud). • Compliance with regulatory requirements (whether the cloud data warehouse satisfies the compliance requirements for HIPAA, ISO 27001, PCI DSS, SOC 1 Type II, etc.). ***Note:*** *Cloud users are responsible for configuring the administrative and technical safeguards (permissions and system access, encryption standards, audit logging) for their sensitive data.*	Different cloud DWH vendors support the same security requirements differently. **Data access granularity** • **Azure Synapse Analytics** offers granular permissions on schemas, tables, views, individual columns, procedures, other objects. • **Google BigQuery** – on datasets, tables, views. • In **Amazon Redshift,** permissions apply to tables as a whole. **User authentication** • **Azure Synapse Analytics, Google BigQuery, Snowflake, Azure SQL Database** offer OAuth support to enable authorized account access without sharing or storing user login credentials; **Amazon Redshift** lacks this capability. All of the above-mentioned providers offer multi-factor authentication. **Data encryption** • Solutions enabling this functionality by default: **Google BigQuery, Azure SQL Database, Snowflake.** • For **Amazon Redshift** and **Azure Synapse Analytics,** data encryption must be manually enabled. **Compliance** • **Amazon Redshift, Azure Synapse Analytics, Snowflake, Azure SQL Database, Google BigQuery** satisfy compliance requirements for HIPAA, ISO 27001, PCI DSS, SOC 1 Type II, and SOC 2 Type II and provide support for network security.
Cost	To estimate the costs, you need to know the following inputs: • Current data volume. • Monthly data growth rate (or what growth of data volume	**Amazon Redshift** offers on-demand pricing varying on the type and number of nodes in the cluster. Reserved instance pricing offers a significant discount (up to 72%) compared to the on-demand option (3-year commitment). Concurrency scaling, managed

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
	is expected during the next 12, 24, 36 months). • A number of everyday DWH data users and their nature. • A number of queries per day and per user. • Average data usage per query. • Data transformation and cleaning needs. ***NB:*** *Cloud data warehouse vendors calculate costs differently, so the final cost will also depend on these differences.* ***Note****: Most cloud data warehouse vendors offer free trial plans.* • Cost optimization capabilities to look for: • Automatic data compression. • The pause and resume feature to save on compute resources. • Self-tuning capability to pay only for additional capacity during large workload spikes. • Materialized views support. • Query results caching. • Flexible indexing options, etc.	storage, etc., are billed separately. No charge for the amount of data processed. **Azure Synapse Analytics** bills for compute and storage resources separately. On-demand pricing or pre-purchased reserved storage at a discount. No charge for the amount of data processed. **Google BigQuery** charges for data storage (separate rates for cold and hot data storage) and the number of bytes processed by each query. Discounted flat-rate pricing is available. Streaming inserts are billed separately. BigQuery is an optimal option for storing voluminous data less frequently queried. **Snowflake** offers on-demand pricing or pre-purchasing storage capacity at a discount. Compute time is billed separately.
Ease of maintenance and administration	Choose the level of self-optimization based on your company's size and data needs. Smaller companies may look for automated data warehouse management if they don't want to invest in the hiring of corresponding expertise. Larger companies may benefit more from maintaining the data	**Snowflake** requires simple provisioning – you need to choose a cloud provider and the size of the virtual warehouse. **Google BigQuery** is completely serverless – all provisioning is automatic. **Amazon Redshift** requires regular monitoring and configuration: you have to choose the instance size, scale nodes manually, etc.

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
	warehouse manually as it allows for greater flexibility and control. Maintenance activities may be fulfilled either by a consulting vendor or a cloud vendor.	**Azure Synapse Analytics** offers both serverless and dedicated resource provisioning options.
Cloud independence	Cloud neutrality allows running a data warehouse across a variety of cloud vendors, avoiding vendor lock-ins, ensuring unprecedented fault tolerance and considerable cost savings. ***NB:*** *The multi-cloud support is viable only in case of true necessity as it substantially narrows down the vendor choice – there are few cloud data warehouses ready for the task.*	If you are looking for a truly multi-cloud experience, go for **Snowflake**, which is available on AWS, Azure, and GCP. Multi-cloud analytics support is also provided by **Google BigQuery** (Omni) – it allows querying data across AWS and Azure (coming soon) without data copying.
Current cloud usage	None of the cloud vendors charge for transferring data into their clouds, but all charge egress fees for moving it out. ***NB:*** *If you use Azure for your ERP or CRM, in case of choosing Amazon Redshift for the DWH, you will be charged for every data transfer from a data source to the data warehouse.* Data egress fees vary considerably based on: • A cloud provider (it is highly recommended to use clouds with higher egress fees only for the workloads that require the capabilities of that specific cloud). • The volume of data to move (use data deduplication and compression where applicable). • Data transfer destination (within or across availability zones, clouds).	Egress fees on: • **Azure:** from $0.01 per GB to $0.181 per GB. • **AWS:** from $0.01 per GB to $0.09 per GB. • **Google Cloud Platform:** from $0.01 per GB to $0.23 per GB. A typical 10TB data transfer would roughly cost $800 for **Azure**, $850 for **AWS**, and $1,100 for **Google Cloud Platform.**

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
	Data transfer method (whether you use a public cloud or a private network).	
Analytics ecosystem	If you already use any products from a particular cloud vendor, consider deploying a data warehouse on the same cloud to get considerable discounts, quicker deployment, easier and cheaper integrations, etc.	**Amazon Redshift** handles analytics workloads with the native integration with the AWS analytics services: • AWS Lake Formation for data lake setup. • AWS Glue for ETL. • Amazon Kinesis Data Firehorse for streaming analytics. • Amazon EMR for big data processing. • Amazon QuickSight for BI and data visualization. • Amazon SageMaker for ML management, etc. **Azure Synapse Analytics** native integrations include: • Azure Data Lake Storage for data lake setup. • Azure Data Factory for ETL. • Azure Stream Analytics for streaming analytics. • Spark engine for big data processing. • Power BI for BI and data visualization. • Azure Machine Learning, Azure Cognitive Services, and Power BI for ML management. **Google BigQuery** facilitates business intelligence with Looker and integrates with the whole Google Cloud Analytics ecosystem. **Snowflake's** analytics capabilities are supported with the Snowflake platform and Snowflake's technology partners (enabled by the Snowflake Partner Connect feature).

SELECTION FACTOR	TIPS	PRODUCT RECOMMENDATIONS
Existing data warehouse	In case there is a data warehouse in place, the technology it runs on may influence the choice of the cloud data warehouse vendor. Most leading cloud data warehouse solutions provide free migration services/tools to simplify the migration from popular cloud and on-premises data warehouse platforms.	Migration to **Azure Synapse Analytics** is simplified with Azure Synapse Pathway – a code translation tool that supports code conversion of a database, schemas, and tables for Amazon Redshift, Google BigQuery, IBM Netezza, Microsoft SQL Server, Snowflake, and Teradata. **AWS** provides the Schema Conversion Tool to streamline the migration to Amazon Redshift from Azure SQL Database, Microsoft SQL Server, Azure Synapse Analytics, Netezza, Oracle, Teradata, Vertica, Snowflake, etc. **Google Big Query** offers the BigQuery Migration Service to enable migration from Teradata. To facilitate the efficient migration from a particular data warehouse ecosystem, **Snowflake** partners with various tech solutions.

Consider Advisory Services for Your Cloud DWH

ScienceSoft's team can help you choose the optimal data warehouse technology to decrease your cloud data warehouse implementation and maintenance costs and ensure high ROI. For tailored recommendations based on your particular data storage and processing needs, use our cloud DWH configurator or request a PoC to ensure a particular DWH platform can handle your workload.

Contact Us

The United States

McKinney, Texas
5900 S. Lake Forest Drive, Suite 300
McKinney, Dallas area, TX-7507
+1 214 306 68 37
contact@scnsoft.com

Atlanta, Georgia
3372 Peachtree Rd., Suite 115
Atlanta, GA-30326
+1 214 306 68 37
contact@scnsoft.com

Europe

Riga, Latvia
Aspazijas bulvāris 20
Rīga, LV-1050
+371 2569 2767
eu@scnsoft.com

Vantaa, Finland
Rajatorpantie 8
Vantaa, FI-01600
+358 92 316 30 70
nordics@scnsoft.com

Gulf Cooperation Council

Fujairah, United Arab Emirates
Fujairah - Creative Tower
Fujairah, POB 4422
+971 585 73 84 33
gulf@scnsoft.com

