Cloud Data Transformation RoadMap

By: Krunal Shah

Index

Introduction 3
Performance at Scale 3
Elasticity 4
Cost Efficiency 5
Supports Structured and Semi-Structured Data 6
Data Granularity 7
Deployment Options 8
Cloud Ecosystem Integrations 8
Conclusion & Food for Thought: 9

Introduction

This whitepaper talks about some of the key criteria one must think while designing Data Warehousing Solutions on any Enterprise Cloud platform.

It is always said and known that it is less complex to upload thousands of GBs of RAW data to AWS or any public cloud considering the wide range of Cloud native services available like Azure Migrate, AWS Server Migration, AWS Storage Gateway etc.

It is rather complicated task to transform such a high data from RAW to transformed data which can be used for Business and Analytics purpose.

To achieve Hugh data transformations on Public Cloud one must build data lakes which does these complex data transformations.

This whitepaper talks more about what are the different parameters one should be considering while handling and transforming data on Cloud.

Performance at Scale

The rapid growth in diversity and size of data can make querying and ETL/ELT processes tedious and frustrating. When evaluating the performance of cloud data warehouses, consider your use-cases. If you're running customer facing analytics, performance can really make it or break it. But even in internal BI use-cases, the ability to run more queries instead of waiting for data, and to have fresh data quickly accessible for querying can be crucial for success.
There are two main bottlenecks that hold performance back:

- Storage bottleneck: data lakes are built for infinite storage but are terribly slow when large amounts of data need to be scanned and moved to the compute layer for querying.

- Compute bottleneck: century-old techniques for processing data are not efficient enough for today's data sets. Queries that

cannot be accelerated through scale-out completely diminish the end-user experience.

In many cases, for higher performance you'll have to consume more compute resources resulting in heavy costs.

Modern cloud data warehouses enable users to analyse large amounts of data at high granularity, and with near real-time query response time, by combining a range of tailored techniques including compression, low-footprint indexing, sparse indexing, cost-based optimizers.

Key Questions to be added as Evaluation Criteria.

- How does the technology scale in terms of query latency and concurrency?
- What does the vendor's compute scaling model look like?
- What is the performance profile of a vendor's storage architecture?
- Does the technology support data compression, indexing and pruning?
  Is the compute utilization in line with your current use case?
- Will compute resources meet future scaling needs?
- What add-ons are required to provide support for sub-second response times?
- Does the technology support join, indexing and automated materialized views?

Elasticity

In traditional data warehouses, as usage or data scales up, users will note that performance is no longer meeting their business needs. The notion of decoupling storage and compute enables seamless scaling up or down to support any workload, amount of data and concurrent users. The flexibility to seamlessly resize nodes without expensive and time-consuming re-clustering, vacuuming, fragmentation, and other heavy lifting tasks, is crucial for handling ever changing resource requirements on demand.

By isolating resources, different teams can start/stop compute engines for different workloads/tasks like ETL, heavy querying and exploration while getting the performance that they need.

Key Questions to be added as Evaluation Criteria.

- How does the technology provide the ability to start / stop compute resources?
- Can compute and storage scale independently for cpu, memory, capacity, and performance?
- How does the scaling model provide for workload isolation?
- Does the technology provide APIs to automate environment spin-up and spin-down?
- Does the technology provide elastic scaling capabilities?

Cost Efficiency

The process of understanding cloud data warehouse pricing models is not straightforward, as they are dependent on different parameters like speed, scale and usage. Common pricing models include:

Pay per TB scanned (Athena, BigQuery):

The 'Pay Per TB Scanned' model includes storage and executed query costs. This means that pricing heavily depends on usage and the size of your workload. While this pricing model works well for small-mid-sized data sets, it starts to become pricey when dealing with big-data use cases, where a lot of data needs to be scanned.

Pay for consumed cloud resources (Redshift, Snowflake, Firebolt):
The cost of this model depends on how much you use the platform, performance requirements and dataset sizes, and users are typically charged per hour or second.

Relevant use-cases:

- Medium-large data set size

- High performance
- Unlimited amount of users
- Frequent usage

Key Questions to be added as Evaluation Criteria.

- What is your data set size and performance requirement today and what will it be in 1-2 years?
- Will the pricing and scaling model support your future growth?
- Does the vendor charge for data set capacity or consumed-compressed capacity?
- What types of storage does the vendor recommend (block, filesystem, object storage etc)?

Supports Structured and Semi-Structured Data

Data no longer arrives only in predictable and structured formats; it arrives in different types and from different sources. Semi-structured data can enrich your analytics, but most traditional data warehouses aren't equipped to handle such data. Users waste time and costs on inefficient flattening/unnesting/exploding, which multiplies the number of rows with the number of cells in the arrays. As a result, you end up with a much bigger table, more unnecessary costs and painfully slower performance.

A modern data warehouse will enable you to query semi-structured data with standard SQL and without complicated ETL processes which flatten and blow-up data set sizes and costs. This can be achieved with native array manipulation functions, and without compromising speed and efficiency.

The right way to handle semi-structured data:

1. Load semi-structured data without transformation: JSON manipulation functions are used to seamlessly cleanse, fix and organize semi-structured data as it's ingested

2. Automatically convert semi-structured data to make it ready for querying: Semi-structured data is automatically converted and stored as arrays of primitive types. Users query the data extremely fast with native array manipulation functions while using standard SQL

Key Questions to be added as Evaluation Criteria.

- What data sources are primarily driving your use cases?
- Does the technology support semi-structured data?
- What are some recommended strategies for managing semi-structured data?
- Does the technology provide the ability to store raw semi-structured data as well as flattened data structures?

Data Granularity

The issue with choosing the right levels of granularity is that detailed data can be too voluminous. That's why granularity is another reason not to compromise on the first factor - choose a platform that supports high performance at scale, without a heavy cost trade-off.

For example, technologies like sparse indexes enable users to only pull rows that are relevant for the query at the most granular level. This is crucial for performance in data lake environments, as fetching unnecessary data from the low storage layer has a huge performance penalty. Bottom line, think twice before you agree to compromise granularity for performance, a good platform will provide both.

Key Questions to be added as Evaluation Criteria.

- What level of granularity do your workloads need? (e.g.: partitions, micro-partitions, files, block level etc)
- What is the granularity at which the platform operates?
- Does the platform require add-ons to address granularity?

Deployment Options

One must be able to deploy the data warehouse you select on the cloud you're using. Some data warehouses are exclusively deployed specifically on AWS, GCP or Azure while others offer multi-cloud deployments.

Most leading solutions are available on AWS, due to its dominance in the public cloud market. AWS has a huge array of services, as well as the most comprehensive network of worldwide data centres.

However, the ideal approach is multi-cloud, which enables companies to avoid vendor lock-ins and provides flexibility to negotiate rates and capitalize on services offered by different cloud providers.

Key Questions to be added as Evaluation Criteria.

- Is your current cloud platform lacking features that you can find in other cloud platforms?
- What do the security and operational models look like as you go across multiple clouds?
- What ecosystem integrations does the cloud provider offer?

Cloud Ecosystem Integrations

Healthy ecosystem partners are important for a smooth integration with the tools you already use. Typically, data warehouses provided by cloud vendors will have the most extensive integrations with the other tools the vendor offers. Seamless integration with your BI tools,

ingestion frameworks and data lake will substantially shorten your time to market.

Key Questions to be added as Evaluation Criteria.

- Which tools are included in your stack?
- How easily do they integrate with the cloud data warehouse?
- Does the platform provide API and SQL access?
- Does the platform support JDBC, ODBC drivers?

Conclusion & Food for Thought

Evaluating the data warehouse platform needs to be done holistically, by understanding not only feature-functionality differences but by focusing on downstream impact of must-have capabilities and nice-to-have features. Downstream impact can be long lasting in terms of customer satisfaction, cost, and operational impact. These could be features built around machine learning, data integration, data sharing, data catalog etc. Start the evaluation with the core platform and what the downstream impacts are in mind.

About Sogeti

Sogeti is a leading provider of technology and engineering services. Sogeti delivers solutions that enable digital transformation and offers cutting-edge expertise in Cloud, Cybersecurity, Digital Manufacturing, Digital Assurance & Testing, and emerging technologies. Sogeti combines agility and speed of implementation with strong technology supplier partnerships, world class methodologies and its global delivery model, Rightshore®. Sogeti brings together more than 25,000 professionals in 15 countries, based in over 100 locations in Europe, USA and India. Sogeti is a wholly-owned subsidiary of Capgemini SE, listed on the Paris Stock Exchange.

Learn more about us at

www.sogeti.com