snowflake

How to Craft Your Data Warehouse POC

APPLES OR APPLE PIE? WHICH DO YOU PREFER?

CHAMPION GUIDES

snowflake

What's inside:

Rethink an old paradigm

It's time to replace your data warehouse. Waiting days or even weeks to analyze your data is no longer acceptable. Your executive team wants real-time insights to match the pace of business, and your data scientists are frustrated by limitations placed on their queries and the inability to load, transform and integrate semi-structured data.

However, it's no small feat to research and conduct a full assessment of data warehousing options. How do you evaluate systems? What's the best method for figuring out what you need? How do you select a solution that is both innovative and can significantly outpace your incumbent system based on performance, scalability, elasticity, flexibility and affordability?

Modern cloud data warehousing offers a new way to think about what's possible when you move from on-premises to the cloud. In this eBook, you'll learn why the standard apples-to-apples comparison falls short when you're looking to upgrade your data warehouse. Rather then let the features of your existing system set limits on your evaluation of a new solution, consider a more comprehensive approach to exploring all the possibilities your new data warehouse can deliver today and in the years to come.

snowflake

Apples to apples: a half-baked concept

It's second nature to compare apples to apples. The analogy works well when you have two things that look, act or function more or less alike but differ enough to form a basis for comparison. For example, mobile phones from Apple and Google are worthy candidates for an apples-to-apples comparison. But the logic falls flat when it comes to data warehousing options.

Let's say your incumbent system is a legacy, on-premises data warehouse or a "cloud-washed" version of one of these systems. Chances are, your team has optimized the data warehouse to leverage it fully and run highly-tuned queries. Congratulations, you have a polished apple.

The temptation exists to take a modern data warehouse and see how it compares to this shiny incumbent apple. However, this strategy has some obvious flaws. For starters, any contending system during a trial run will not be optimized in the same manner as the incumbent. You'll have to strip down the new solution to its basics in order to evaluate performance and compare it against what you have today.

The comparison is even more unbalanced when you put an incumbent system up against a data warehouse built for the cloud. Shared-disk architectures retain simple storage management by centralizing data but with the tradeoff of a performance bottleneck between storage and compute. Shared-nothing architectures avoid the bottleneck between compute and storage but still carry the tradeoff of complicated storage management. For example, resizing the system requires redistributing and re-replicating data. Only a modern warehouse built for the cloud, one that truly separates compute from storage, can effectively capitalize on everything cloud architecture can enable, including secure data sharing. For an applicable comparison, you'd have to disable all the bells and whistles that represent the modern cloud architecture.

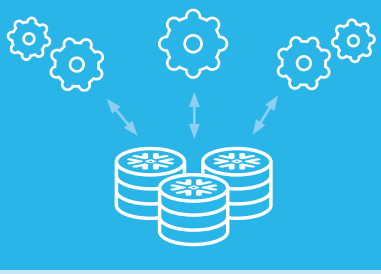
TRADITIONAL VS. CLOUD-BUILT ARCHITECTURE

Traditional Architectures

SHARED DISK
SHARED STORAGE
SINGLE CLUSTER

SHARED NOTHING
DECENTRALIZED LOCAL STORAGE
SINGLE CLUSTER

Built-for-the-cloud Architecture



MULTI-CLUSTER, SHARED DATA
CENTRALIZED SCALE-OUT STORAGE
MULTIPLE, INDEPENDENT COMPUTE CLUSTER

snowflake

To illustrate, consider the following steps often taken in an apples-to-apples comparison that can sideline the best a modern cloud data warehouse has to offer:

Choose a suboptimal warehouse size.

In order to match the incumbent system's number of cores or amount of RAM used, you must choose a suboptimal warehouse size for the cloud-built contender. This decision is predicated on how much it would cost to run the system 24/7. While the decision may seem fair, the rationale is faulty when comparing an incumbent system to a cloud-built solution with instant elasticity, which comprises:

- Instant provisioning
- On-demand performance
- On-demand concurrency
- Per-second pricing

Start with cold cache warehouses.

To justify the preconception that shared-nothing architectures are the fastest, try to factor out any local caching of data being retrieved from your cloud storage provider. Therefore, the test starts with a cold-cache warehouse—a scenario you'd never perform in real life.

Disable ResultSet cache.

Because the current system needs to run every query, the ResultSet cache of the contender is disabled. Again, you're turning off features that illustrate the vast difference between your incumbent system and what a cloud-built experience enables.

Test with single queries, not concurrent workloads.

Because your incumbent system can't handle concurrency well, or at all, isolated queries are measured instead of unlimited concurrent workloads from a cloud-built solution.

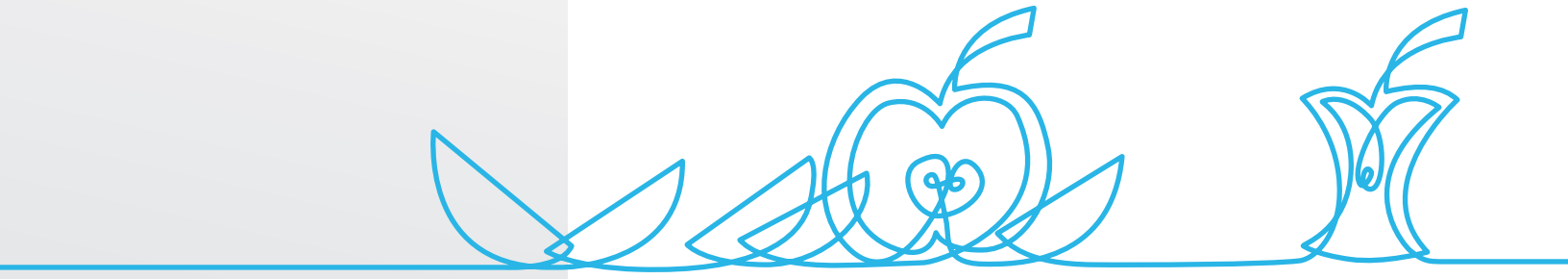
Run short queries on modest datasets.

To keep things "fair," you might choose to run short queries on modest datasets instead of big queries on big data. For a cloud-built data warehouse, the bigger the challenge, the brighter the results. Short, low-latency queries reveal the fixed overhead built into the architecture for networking, compilation and data loading while big queries wash out those costs.

Don't test semi-structured data.

Because the incumbent system can't handle JSON, Avro, Parquet or XML data, this semi-structured data is ignored in the comparison.

The result? You end up comparing your incumbent apple to nothing more than an apple core.



snowflake

Apple pie makes everybody happy

You'll always get a subpar outcome if you allow the features of an existing data warehouse to control your search for a modern, cloud-built solution. An apples-to-apples comparison between a data warehouse built for the cloud and your incumbent solution will always be incomplete.

Instead of settling for such a comparison with a system you want to replace, consider a full-featured, "apple-pie" comparison, which boasts:

SCALABILITY

- Independent, open-ended scalability of compute and storage.
- Multiple independently scalable compute clusters that share data but eliminate contention between workloads.
- Auto-scaling when concurrency surges

AFFORDABILITY

- Pay for per-second usage, not for what you might need on the busiest day of the year, every day of the year.
- Incrementally or automatically increase and decrease resources for new use cases, and for monthly, quarterly or seasonal data surges.
- Reduce capacity planning exercises from multi-year to simply assessing what you need by the month, week, day or right now.

SEMI-STRUCTURED DATA

- Full support for JSON, Avro, Parquet and XML data.

MODERN DATA SHARING

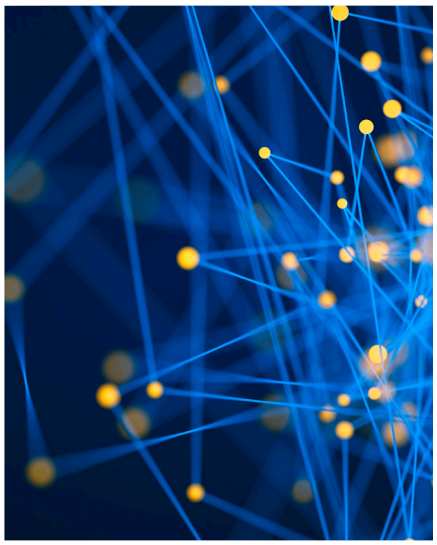
- Zero-copy cloning for test and dev, research or processes independent from continuous ELT or ETL.
- Live enterprise data sharing between data providers and data consumers for richer analytics.
- Cross-region data replication for data providers with worldwide data consumers and multi-region data resiliency.
- Cross-cloud data replication for data providers with worldwide, multi-cloud data consumers for extreme reliability and data resiliency.

SECURITY, MANAGEMENT AND DATA PROTECTION

- Out-of-the-box, always-on, secure data environment, including a virtual private cloud (VPC).
- Near-zero administration, with built-in performance tuning so there's no infrastructure to tweak, no knobs to turn and no tuning required.
- Zero-copy cloning for test and dev, research or other processes independent from continuous ELT or ETL, so there's no impact to performance.
- Eliminate the need for conventional backups with zero-copy clones and rollback features such as time travel.

snowflake

A POC plan that works



Now that we're comparing apple pie to apple pie, here are guidelines and best practices for setting up a proof of concept (POC) approach that will allow you to properly evaluate a modern cloud data warehouse against your incumbent solution.

These criteria represent the features and benefits that your business needs in order to compete in a data-driven world. With scalability, concurrency and access to all of your data, speedy and accurate decision-making becomes a piece of cake.

GUIDELINES FOR POC CRITERIA

Each of these criteria may be impossible to test with your incumbent system. But they represent the future for data-driven organizations and the gold standard for today's modern cloud data warehouse.

Concurrency testing

Create a workload framework that includes a mix of small, medium and large queries. Watch how the system handles surges in loads with auto-scaling.

Scalability

Include performance testing to experience linear scalability to see how a cloud-built data warehouse handles two, three and four times the warehouse size you have today. Be sure to note the cost of those configurations for your incumbent system versus a cloud-built data warehouse.

Software lifecycle

Clone to create Dev, QA, staging and research datasets before you spin up warehouse clusters for each function. Run only when needed to see how cost is connected to usage.

ETL and BI workloads

Spin up separate, concurrent warehouses for ETL and BI, without compute contention, so the ETL can provide fresh data with no impact to the performance of BI and other workloads.

Semi-structured data and queries

Explore the value to the business of loading, integrating and analyzing JSON, Avro, Parquet and XML data with your structured data in a single data warehouse.

snowflake

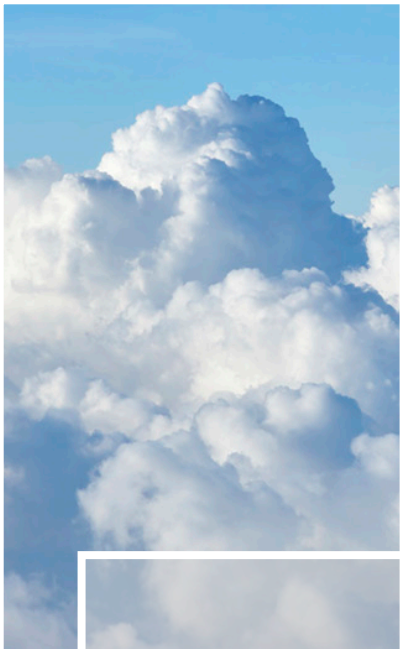
Performance testing best practices

We recognize that this POC evaluation criteria includes new ways of thinking about and interacting with your data warehouse. To help guide you through the evaluation process, here's a shortlist of best practices around performance testing for a data warehouse built for the cloud.

- Don't open a connection for each statement.
- Don't load single large files and expect linear scaling with bigger warehouses.
- Test with or without ResultSets.
- Test with multiple warehouse sizes and multi-cluster warehouses.
- Test multiple workloads in separate warehouses.
- Cluster the data, if appropriate.

Time for a bake-off!

Is your goal is to reveal all the differences between your incumbent data warehouse and one built for the cloud? We hope you will take the apple pie POC test and discover what it feels like to really get cookin'. If you need an extra pair of hands in the kitchen, please contact us here.



About Snowflake

Snowflake is the only data warehouse built for the cloud. Snowflake delivers the performance, concurrency and simplicity needed to store and analyze all data available to an organization in one location. Snowflake's technology combines the power of data warehousing, the flexibility of big data platforms, the elasticity of the cloud, and live data sharing at a fraction of the cost of traditional solutions. Snowflake: Your data, no limits. Find out more at www.snowflake.net.