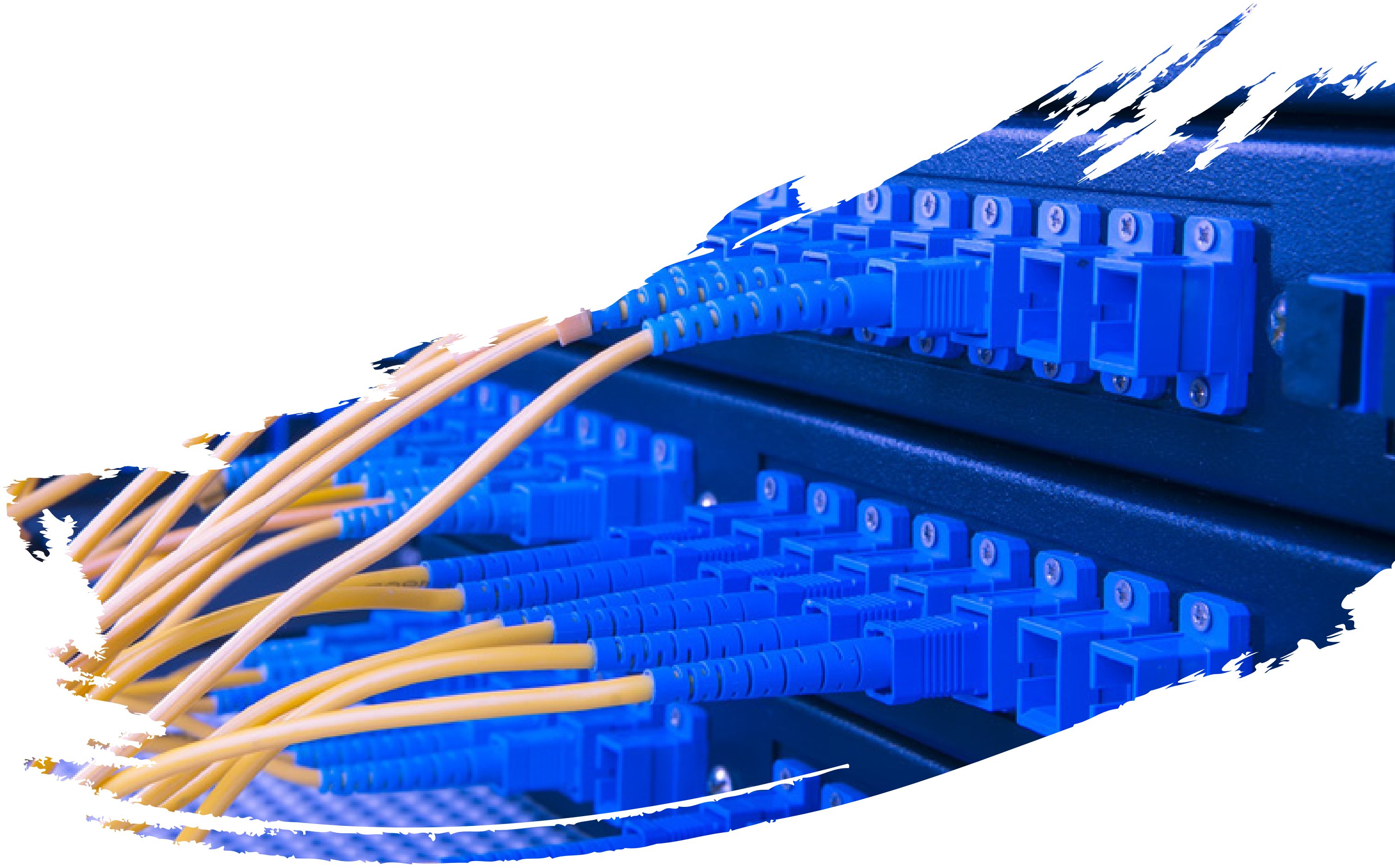
Whitepaper

Enabling Superior Data-driven Decisions with Performance Engineering on Snowflake

By C.Kalyana Chakravarthy

LTIMindtree

Table of Contents



The Eleventh Hour

Data and Analytics are core differentators for today's enterprise success. From being a just an enabler for operations, data analytics dictates the future of the organization. Data is no more considered a by-product, but an asset which facilitates modern day prescriptive and predictive analysis. Despite being the sole source of decision support systems, data has always challenged human, current infrastructure, processes, and pipelines with its enormous volume, ever-growing velocity, and distinguished variety, causing the veracity and value to take the test of time.

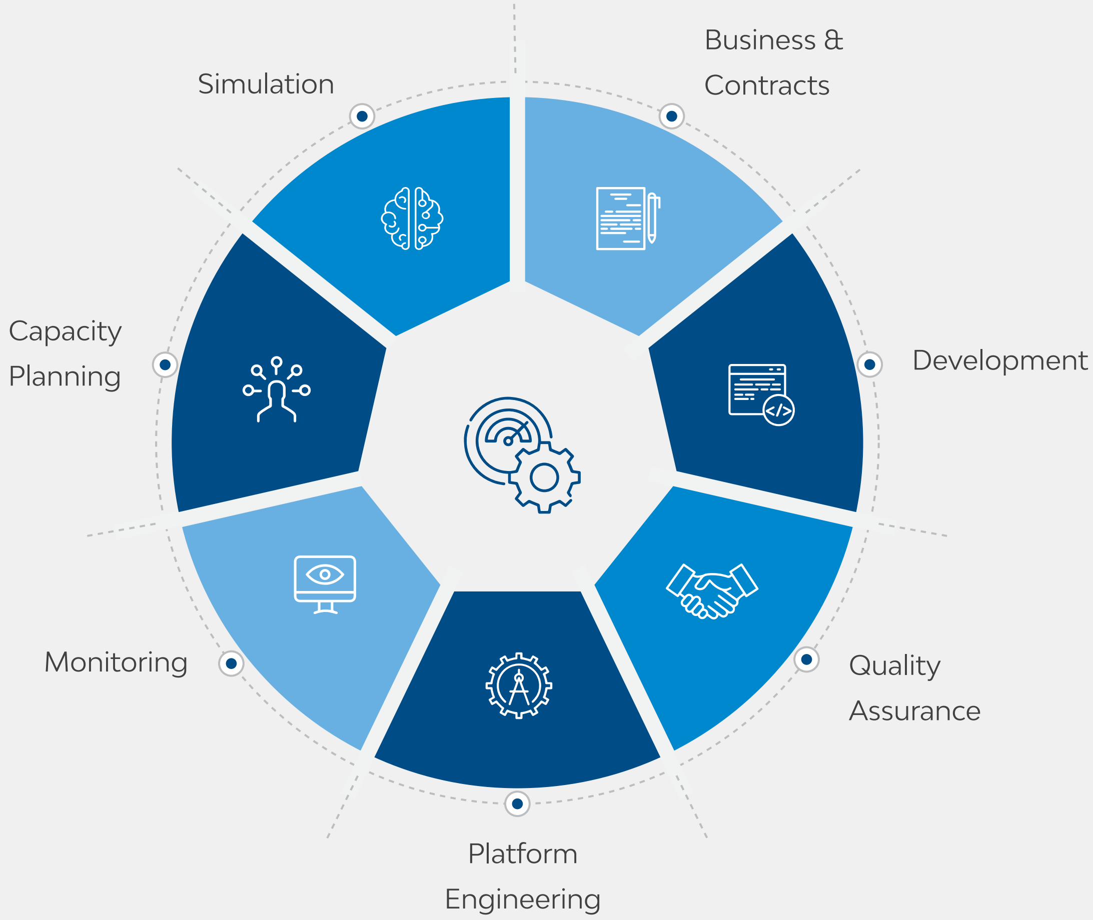
With the advancement in methodology (Symmetric processing, Massive parallel processing, Hadoop) that has happened in the past, nothing could give a holistic approach or process to completely extract value out of data. These improvements were myopic at best, not solving the holistic problem. We could not eliminate data silos paving the way for Single Source of truth. Semi-structured and unstructured data pipelines were extremely complex, demanding dedicated hardware, software, and highly skilled professionals.

Amidst this complexity, cloud-based solutions have addressed the traditional challenges of infra-structure, diversified variety of data with unconventional architecture by decoupling storage with computing.

With its simple yet elegant adherence to ANSI SQL, pay-on-the-go model, effortless data sharing features, and many more, Snowflake positioned itself as a pioneer in the world of data warehousing, taking a complete transformation to the Data Cloud platform.

Snowflake is becoming ubiquitous with its unprecedented innovation, path-breaking features, and ease of use. In this context, we aim to peep through the Snowflake ecosystem with a Performance Engineering lens.

Performance **Engineering**



Performance Engineering is a holistic approach to ensure system the delivers optimum performance at a given price point. Performance should be incorporated as a culture across teams and it must become an integral part of teams activities.

Pillars of **Performance Engineering**

In this paper, we aim to detail certain critical elements for a solution's success.

- **Development**
- **Finance and Cost Management**
- **Operations**

Let's look at each element with a broader and deeper understanding to ensure optimal performance across the system.
When all the dots are connected – data acquisition, ingestion, and transformation will be interwoven beautifully with intact and robust data pipelines paving the way for new-age analytics.

First pillar- **Development**

We will do a sampling of few key attributes in Snowflake Ecosystem.

Capacity planning

01

Capacity Planning was a prerequisite for any project in the times past, where team plans, estimates, and avails resources for peak time usage.

02

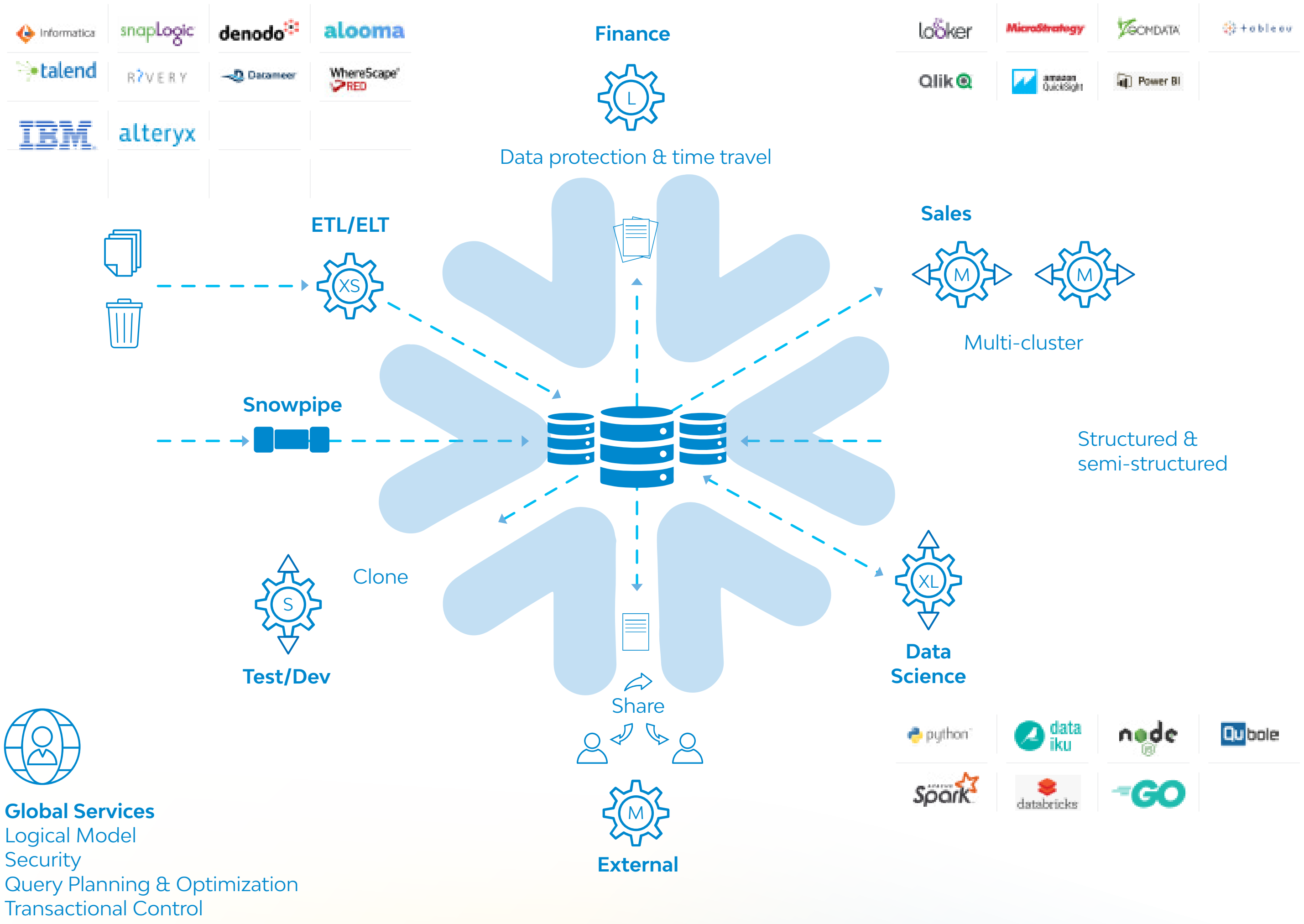
The very onset of cloud-based solutions and SAAS has eliminated this inundated availing of resources well ahead of time to be prepared for peak time usage.

03

Decoupling of Storage and Compute, on-demand procurement of resources has disrupted the age-old theory of advanced buying of resources almost making the <INITIAL> Capex minimal.

Effective workload modelling

- Snowflake provides solitary compute layer which spins up resources dynamically basing on the load. With its mutually exclusive warehouse concept, effective workload segregation is possible.
- Based on the use case, the nature of queries executing warehouse size and clusters can be configured.



Landscape Analysis

Inventory analysis of existing system will help in setting the stage for migration.

This will assist in deciding cost affecting parameters:

For storage : considerations while choosing storage features.

Comprehensive Data Protection (CDP) Matrix: based on our experience.

Table Type	Permanent	Permanent	Permanent	Permanent	Permanent
Table size	Low-Medium	Low-Medium	Huge	Huge	Huge
Churning	Low/Medium/High	Low/Medium/High	Moderate	Moderate	Moderate
Criticality	Low/High	Low/High	Low-Medium	High	High
Recovery of data or performing Full load @ source	Easy	Complicated	Easy	Moderate	Complex
CDP Retention period from our experience	3-4 weeks	3-4 weeks	3-4 weeks	4-6 weeks	4-8 weeks

For Compute : considerations while choosing compute resources.

Compute Matrix :

Complexity of queries	Query Tuning	Scale up (Long running queries)	Scale out	Auto Suspend
Simple	Yes	NA	When queuing happens	ETL - 1- 3 min Reporting 5-10 min
Medium	Yes	Consider if query is well tuned	When queuing happens	ETL - 1- 3 min Reporting 5-10 min
Complex	Yes	Consider if query is well tuned	When queuing happens	ETL - 1- 3 min Reporting 5-10 min

Design Strategy

Stitching best practices into the system in every progression of the life cycle.

Migration Challenges

With each requirement, something that works for one system may not hold good for other. This will be an inevitable reality to face when migrating from traditional/legacy systems to new age cloud solutions. If it is a mere movement of code and data onto the cloud, system will suffer over a period in some cases.

Second Pillar- **Finance**

Once the solution is built and operationalized – then arises the question of optimal run.

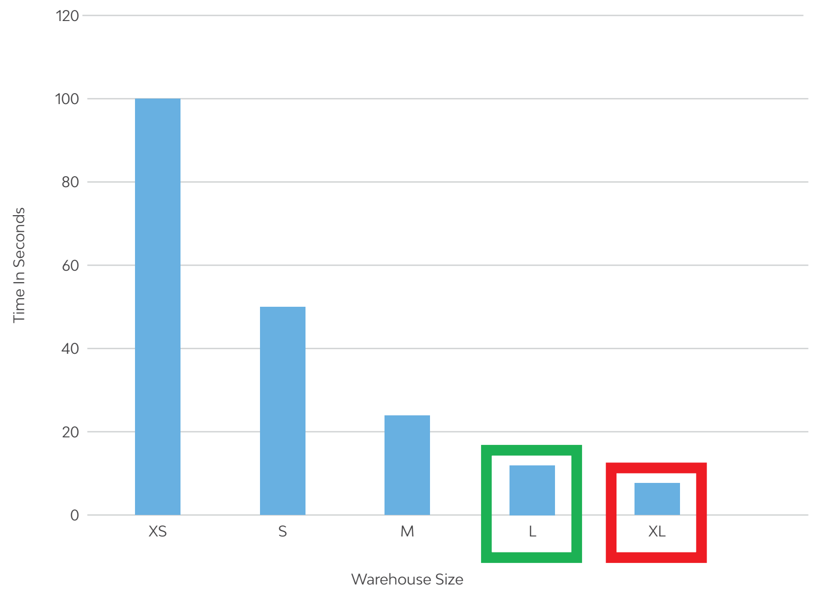
Now, let us take a dive at how better we can optimize the costs.

Finance and Management

Let us zoom into the compute costs arena.

Using correct-sized warehouse :

- To know the best fit size for the warehouse – snowflake recommends running homogenous queries on a warehouse.
- Initial scale up will progress linearly – which means the time taken will be reduced by half as the warehouse size is doubled.
- At a point in time, time taken would not reduce by half which means that the query will not get benefit by doing a scale up.
- That is when we need to scale down to previous size as that will get an optimal performance for the cost incurred.
- In the example below: linear scalability stopped at warehouse – XL. So, optimal size for the workload will be Large.



Picture depicting linear scalability

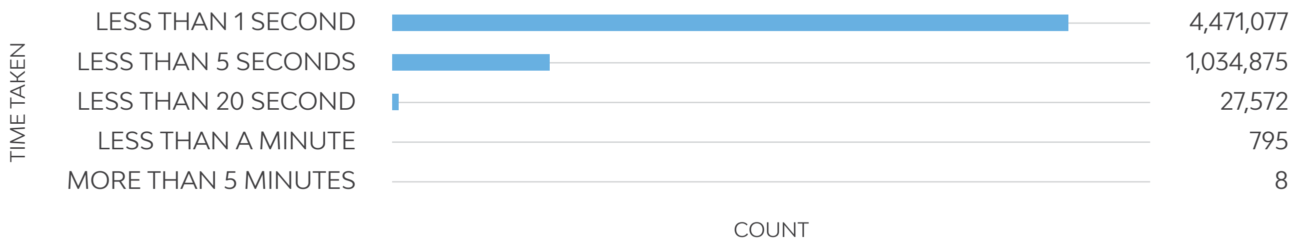
Knowing when to go for a scale up/down :

The best way to decide on scale up or down is to analyse the workloads/queries executing on the warehouse. If major proportion of queries are completing in less than a second – then a Large/XL warehouse will not be a right fit.

Exemplifying:

Row Labels	Sum of NUMBER OF QUERIES	Percentage
LESS THAN 1 SECOND	2530157	80.00
LESS THAN 20 SECONDS	18102	0.57
LESS THAN 5 SECOND	613970	19.41
LESS THAN A MINUTE	351	0.01
MORE THAN 5 MINUTES	8	0.00
Grand Total	**3162588**	**100**

Performance - Latest



Warehouse query load patterns

- Picture depicts that 99% of queries are completing in less than 5 seconds.
- This will enable us to conclude that most of the queries executing on this warehouse are simple and does not need a large warehouse.
- In case if a large warehouse is used, scaling down will help in saving compute costs. If there is a concern about concurrency, clusters size can be increased to handle it as and when the load goes high.
- In case of a hybrid situation where simple and complex queries are in equal proportion, routing the short running queries to a smaller warehouse can be considered if the segregation of workloads is feasible.

Third Pillar- **Development and Security**

First and Second pillars we have discussed above, will give deeper insights into Development and Security.

- Operations is a continuous and crucial life cycle activity for success of any application/system.
- Given right take off conditions with required thrust and force, system will auto pilot itself once it hits off the ground to clear skies.
- Seamless and effective operations will fall in place if care is taken in every preceding activity of design, coding, testing, and operationalization.

Conclusion

Performance Engineering is an underlying discipline which decides the outcome of the solution. This is an incremental and a continuous activity which demands its catalytic presence in every detail of the solution. Right from the onset of requirements gathering to the code migration, every phase will have more than one option to implement. Picking up the optimal choice at each phase will ensure no leaks in the system.

Cognizant and cautious care should be taken while developing an application. Functional implementation remains the prime focus in the development phase.Adding to this, long term performance impacts should also be given a lion share of thought. This enables system to tackle any performance issues which may arise later.

When the development phase is well thought through and implemented, system leaks will be minimal which ensures no cost spillage on process gaps or undesirable code components.

Once the solution is deployed, operations will be effective and productive as all the known measures would have been taken to ensure an optimal performance.

Data value will be optimal with a prompt time-to-market value and accuracy of data analysis with robust and future proof data pile lines. This will enable business decision support system to forecast and predict better outcomes by focussing on business need instead of infrastructure requirements.

About the Authur

C.Kalyana Chakravarthy is working as a Data Engineer in Snowflake Cloud Data platform. He is a data enthusiast with over a decade of experience in Data Warehousing. He worked on ETL tools like Abinitio, Informatica, Reporting tools - SAP BO, currently working in Cloud Data platform. He aspires to be an architect in Snowflake - designing systems which serves the client to be faster, cheaper, and better. He is certified in Snowpro core and Snowpro Advanced Architect. He enjoys reading books, travelling, and listening to music at leisure.