Fixing A Flawed Data Warehouse Model Snowflake will not grow into its valuation by simply stealing share from the on-prem data warehouse vendors. Even if it got 100% of the data warehouse business it wouldn’t come close to justifying its market cap. Rather Snowflake must create an entirely new market based on completely changing the [...]
Fixing A Flawed Data Warehouse Model

Snowflake will not grow into its valuation by simply stealing share from the on-prem data warehouse vendors. Even if it got 100% of the data warehouse business it wouldn’t come close to justifying its market cap. Rather Snowflake must create an entirely new market based on completely changing the way organizations think about monetizing data.

Every organization we talk to says it wants to be – or already is – data driven. Why wouldn’t you aspire to that goal? There’s probably nothing more strategic than leveraging data to power your digital business and creating competitive advantage. But many businesses are failing or – will fail – to create a true data-driven culture, because they’re relying on a flawed architectural model hardened by decades of building centralized data platforms.

In this week’s Wikibon CUBE Insights, Powered by ETR, we make the case that the centralized warehouse/big data platform model is structurally ill-suited to support multi-faceted digital transformations. Rather we believe a new approach is emerging where business owners with domain expertise will become the key figures in a distributed data model that will transform the way organizations approach data monetization.

The Data Cloud Summit

On November 17th, theCUBE 365 is hosting the Snowflake Data Cloud Summit (*Disclosure below). Snowflake’s ascendency and its blockbuster IPO has been widely covered by us and many others. Since Snowflake went public, we’ve been inundated with outreach from investors, customers and competitors that wanted to either better understand the opportunities, or explain why their approach is better or different. And in this segment, ahead of Snowflake’s big event, we want to share some of what we’ve learned and how we see it evolving.

A Flawed Data Model

The problem is complex, no debate. Organizations must integrate data platforms with existing operational systems, many of which were developed decades ago. And there is a culture and sets of processes that have been built around these systems and hardened over the years.
Breaking Analysis: How Snowflake Plans to Change a Flawed Data Warehouse Model

The chart below tries to depict the progression of the monolithic data source, which for many of us began in the 1980’s when decision support systems (DSS) promised to bring actionable insights to organizations. The data warehouse became popular and data marts sprang up all over, which created proprietary stovepipes with data locked inside. The Enron collapse led to Sarbanes Oxley, which tightened up reporting requirements and breathed new life into the warehouse model. But it remained expensive and cumbersome.

The 2010’s ushered in the Big Data movement and data lakes emerged. With Hadoop we saw the idea of no schema on write where you put structured and unstructured data into a repository and apply structure to the data on read. What emerged was a fairly complex data pipeline and associated roles that involved ingesting, cleaning, processing, analyzing, preparing and ultimately serving data to the lines of business. This is where we are today with hyper-specialized roles around data engineering, data quality, data science with a fair amount of batch processing. Spark emerged to address the complexity associated with Map Reduce and definitely helped improve the situation.

And we’re seeing attempts to blend in real time stream processing with the emergence of tools like Kafka. But we’ll argue that in a strange way, these innovations compound the problem we want to discuss because they heighten the need for more specialization and more fragmentation in the data lifecycle.

The reality, and it pains us to say it – is the outcome of the big data movement, as we sit here in 2020 – is that we have created thousands of complicated science projects that have once again failed to live up to the promise of rapid, cost effective time to data insights.

**2020’s: More of the Same or a Radical Change?**

What will the 2020’s bring? What’s the next silver bullet? You hear terms like the Lakehouse, which Databricks is trying to popularize and the data mesh, which we’ll discuss later in this post. Efforts to modernize data lakes and merge the best of data warehouse and second generation data systems will unify batch and streaming frameworks. And while this definitely addresses some of the gaps, in our view it still suffers from some of the underlying problems of previous generation data architectures.
In other words, if the “NextGen” data architecture is incremental, centralized, rigid and focuses primarily on making the technology to get data in and out of the pipeline work faster...it will fail to live up to expectations...again.

Rather, we’re envisioning an architecture based on the principles of distributed data where serving domain knowledge workers is the primary target citizen and data is not seen as a byproduct (i.e. the exhaust) of an operational system, but rather a service that can be delivered in multiple forms and use cases across an ecosystem. This is why we say data is not the new oil. A specific quart of oil can either be used to fuel our home or lubricate our car engine but it can't do both. Data does not follow the laws of scarcity like natural resources.

What we’re envisioning is a re-thinking of the data pipeline and associated cultures to put the needs of the domain owner at the core and provide automated governed and secure access to data-as-a-service; at scale.

How will the Data Pipeline Change?

Let’s unpack the data lifecycle and look deeper into the situation to see how what we’re proposing will be different.

The picture above has been discussed in detail over the past decade and it depicts the data flow in a typical scenario. Data comes from inside and outside the enterprise. It gets processed, cleansed and augmented so that it can be trusted and made useful. And then we can add machine intelligence and do more analysis and finally deliver the data so that domain-specific consumers can essentially build data services. A report, a dashboard, a content service, an insurance policy, a financial product, a loan...a data “product” that is packaged and made available for someone on which to make decisions or a transaction. And all the metadata and associated information is packaged along with that data set as part of the service.

Organizations have broken down these steps into atomic components so that each can be optimized and made as efficient as possible. And below the data flow, you have these happy stick figures – sometimes
they’re happy – but they’re highly specialized and they each do their job to make sure that the data gets in, gets processed and delivered in a timely manner.

Here’s why This Model is Faulty

While these individual components are seemingly independent and can be optimized, they are all encompassed within the centralized big data platform. And by design, this platform is domain agnostic. Meaning the platform is the data owner, not the domain-specific experts. Knowledge workers are subservient to the platform and the processes surrounding it.

There are a number of problems with this model. First, while it’s fine for organizations with a small number of domains. But organizations with a large number of data sources and complex structures, struggle to create a common lingua franca and data culture. Another problem is that as the number of data sources grows, organizing and synchronizing them in a centralized platform becomes increasingly difficult because domain context gets lost. Moreover, as ecosystems grow and add more data, the processes associated with the centralized platform tend to further water down domain-specific context.

There’s more. While in theory, organizations are optimizing on the piece parts of the pipeline, the reality is that as a domain requires a change – for example a new data source or an ecosystem partnership requires a change in access or process that can benefit a data consumer – the reality is the change is subservient to the dependencies and the need to harmonize across the discreet parts of the pipeline. In other words, the monolithic data platform itself is really the most granular part of the system.

When we complain about this faulty structure, some folks tell us this problem is solved. That there are services that allow new data sources to be easily added. An example of this is Databricks Ingest, which is an auto loader that simplifies ingestion into the company’s Delta Lake offering. Rather than centralizing in a data warehouse, which struggles to efficiently allow access to machine learning frameworks, this feature allows you to place all the data into a centralized data lake – or so the argument goes.

The problem is, while this approach does admittedly minimize the complexities of adding new data sources, it still relies on a linear, end-to-end process that slows down the introduction of data sources from the domain consumer end of the pipeline. In other words the domain expert has to elbow her way to the front of the line – or pipeline – to get stuff done.

Finally, the way we are organizing our teams is a point of contention and we believe will cause more problems down the road. Specifically, we’ve again optimized on technology expertise where, for example, data engineers, while very good at what they do, are often removed from the operations of the business. Essentially we’ve created more silos and organized around technical expertise, versus domain knowledge. As an example, a data team has to work with data that is delivered with very little domain specificity and serve a variety of highly specialized consumption use cases. Unless they’re part of the business line, they don’t have the domain context.

Understandably, this service desk-like structure is established because the specialized skills are not abundant and sharing resources is more efficient. But the future in our view is to reduce the need for specialization by changing the organizational structure to empower domain leaders.

Snowflake is not the Perfect Data Warehouse

We want to step aside for a minute and share some of the problems people bring up with Snowflake. As we said earlier, we’ve been inundated by dozens and dozens of data points, opinions and criticisms of the company. We’ll share a few but here’s a deeper technical analysis from a software engineer that we found to be fairly balanced.
There are five Snowflake criticisms that we’ll highlight.

**Price transparency.** We’ve had more than a few customers tell us they chose an alternative database because of the unpredictable nature of Snowflake’s pricing model. Snowflake prices based on consumption – just like AWS and other public cloud providers (unlike SaaS vendors by the way). So just like AWS for example, the bill at the end of the month is sometimes unpredictable. Is this a problem? Yes but like we say about AWS...kill us with that problem. If users are creating value by using Snowflake then that’s good for the business. But this clearly is a sore point for some users, especially for procurement and finance which don’t like unpredictability. And Snowflake needs to do a better job communicating and managing this issue with tooling that can predict and manage costs.

**Workload management.** Or lack thereof. If you want to isolate higher performance workloads with Snowflake, you just spin up a separate virtual warehouse. It works generally but will add expense. We’re reminded of the philosophy of Pure Storage and its approach to storage management. The engineers at Pure always designed for simplicity and this is the approach Snowflake is taking. The difference between Pure and Snowflake, as we’ll discuss in a moment, is Pure’s ascendency was based largely on stealing share from legacy EMC systems. Snowflake in our view has a much larger market opportunity than simply shifting share from legacy data warehouses.

**Caching architecture.** At the end of the day, Snowflake is based on a caching architecture. And a caching architecture has to be working for a time to optimize performance. Caches work well when the size of the working set is small. Caches work less well when the working set size is very large. In general, transactional databases have pretty small data sets. And in general, analytics data sets are potentially much larger. Isn’t Snowflake targeting analytics workloads you ask? Yes. But the good thing that Snowflake has done is enable data sharing and its caching architecture serves its customers well because it allows *domain experts* to isolate and analyze problems/opportunities based on tactical needs. Quite often these data sets are relatively smaller and are served well by Snowflake’s approach.

However, very big queries across the whole data set (or badly written queries that scan the entire data set) are not the sweet spot for Snowflake. Another good example is if you’re doing a large audit and need to analyze a huge data set, Snowflake is probably not the best solution.
Breaking Analysis: How Snowflake Plans to Change a Flawed Data Warehouse Model

Complex joins. The working sets of complex joins by definition are larger. See above explanation.

Read only. Snowflake is pretty much optimized for read only data. Stateless data is perhaps a better way of thinking about this. Heavily write-intensive workloads are not the wheelhouse of Snowflake. So where this is maybe an issue is real time decision making and AI inferencing. Now over time, Snowflake may be able to develop products or acquire technology to address this opportunity.

We would be more concerned about these issues if Snowflake aspired to be a data warehouse vendor. If that were the case, this company would hit a wall, just like the MPP vendors that preceded them and were essentially all acquired because they ran out of market runway. They built better mousetraps but their use cases were relatively small compared to what we see for Snowflake’s opportunity.

Our premise in this Breaking Analysis is that the future of data architectures will be to move away from a large, centralized warehouse or data lake model to a highly distributed data sharing system that puts power in the hands of domain experts in the line of business.

Snowflake is less computationally efficient and less optimized for classic data warehouse work. But it’s designed to serve the domain user much more effectively. Our belief is that Snowflake is optimizing for business effectiveness. And as we said before, the company can probably do a better job keeping passionate end users from breaking the bank...but as long as these end users are making money for their companies, we don’t think this will be a problem.

What are the Attributes of a new Data Architecture?

In the chart above we depict the a total flip to the centralized and monolithic big data systems we’ve known for decades. In this new architecture, data is owned by domain-specific business leaders, not technologists. Today, it’s not much different in most organizations than it was twenty years ago. If we want to create something of value that requires data, we need to cajole, beg or bribe the technology and data teams to accommodate. The data consumers are subservient to the data pipeline whereas in the future we see the pipeline as the second class citizen where the domain expert is elevated. In other words getting the technology and the components of the pipeline to be more efficient is not the key objective. Rather the
time it takes to envision, create and monetize a data service is the primary measure. The data teams are cross functional and live inside the domain versus today’s structure where the data team is largely disconnected from the domain consumer.

Data in this model is not the exhaust coming out of an operational system or external source that is treated as generic, rather it’s a key ingredient of a service that is domain-driven.

And the target system is not a warehouse or a lake, it’s a collection of connected domain-specific data sets that live in a global mesh.

What Does a Domain-Centric Architecture Look Like?

A domain-centric approach is based on a global data mesh. What is a global data mesh? It is a decentralized architecture that is domain-aware. The data sets in this system are purposefully designed to support a data service or data product if you prefer. The ownership of the data resides with the domain experts because they have the most detailed knowledge of the data requirements and its end use. Data in this global mesh is governed and secured and every user in the mesh can have access to any data set as long as it is governed according to the edicts of the organization.

In this model, the domain expert has access to a self-service and abstracted infrastructure layer that is supported by a cross functional technology team. Again, the primary measure of success is the time it takes to conceive and deliver a data service that can be monetized. By monetized we mean a data service that cuts costs, drives revenue, saves lives or whatever the mission is of the organization.

The power of this model is that it accelerates the creation of value by putting authority in the hands of those individuals who are closest to the customer and have the most intimate knowledge of how to monetize data. It reduces the diseconomies at scale of having a centralized or monolithic data architecture and scales much better than legacy approaches because the atomic unit is a data service that is controlled by the domain, not a monolithic warehouse or lake.

### Requirements of a Domain-centric Architecture

- Secure and governed
- Trusted
- Easily discovered
- Accessible
- Self-service
- Interoperable

Zhamak Dehghani is a software engineer who is attempting to popularize the concept of a global mesh. Her
work is outstanding and has strengthened our belief that practitioners see this the same way we do. To paraphrase her view (see above graphic), a domain centric system must be secure and governed with standard policies across domains. It has to be trusted. Discoverable via a data catalog with rich metadata. The data sets must be self-describing and designed for self-service. Accessibility for all users is crucial as is interoperability, without which, distributed systems fail.

What Does this Have to do With Snowflake?

Snowflake is not Just a Data Warehouse. In our view, Snowflake has always had the potential to be more than a data warehouse. Our assessment is that attacking the data warehouse use case gave Snowflake a straightforward, easy to understand narrative that allowed it to get a foothold on the market.

Data warehouses are notoriously expensive, cumbersome and resource intensive. But they’re critical to reporting and analytics. So it was logical for Snowflake to target on-premises legacy warehouses and their smaller cousins, the data lakes, as early use cases. By putting forth (and demonstrating) a simple data warehouse alternative that could be spun up quickly, Snowflake was able to gain traction, demonstrate repeatability and attract the capital necessary to scale to its vision.

The chart below shows the three layers of Snowflake’s architecture that have been well documented. The separation of compute from storage and the outer layer of cloud services. But we want to call your attention to the bottom part of the chart – the so-called Cloud Agnostic Layer that Snowflake introduced in 2018.

![Snowflake Region Architecture](image)

This layer is somewhat misunderstood. Not only did Snowflake make its cloud native database compatible to run on AWS, then Azure and in 2020 Google Compute Platform. What Snowflake has done is abstract cloud infrastructure complexity and create what it calls the Data Cloud.

What is the Data Cloud?

We don’t believe that the Data Cloud is just a marketing term with little substance. Just as SaaS simplified application software and IaaS made it possible to eliminate the value drain associated with provisioning
infrastructure, a data cloud, in concept, can simplify data access, break down data fragmentation and enable shared access to data, globally.

Snowflake has a first mover advantage in this space. We see five fundamental aspects that comprise a data cloud:

- Massive scale with virtually unlimited compute and storage resource, enabled by the public cloud;
- A data/database architecture that is built to take advantage of native public cloud services;
- A cloud abstraction layer that hides the complexity of infrastructure;
- A governed and secure shared access system where any user in the system, if allowed, can access any data in the cloud;
- The creation of a global data mesh

Earlier this year, Snowflake introduced a global data mesh.

Over the course of its recent history, Snowflake has been building out its data cloud by creating data regions strategically tapping key locations of AWS regions and then adding in Azure and GCP regions (see above graphic). The complexity of the underlying cloud infrastructure has been stripped away to enable self-service and any Snowflake user becomes part of this global mesh, independent of which cloud they're on.

So let’s go back to what we were discussing earlier. Users in this mesh can be...will be...are... domain owners. They’re building monetizable services and products around data. They are most likely dealing with relatively small, read-only data sets. They can ingest data from any source very easily and quickly set up security and governance to enable data sharing across different parts of an organization or an ecosystem.

Access control and governance is automated. The data sets are addressable, the data owners have clearly defined missions and they own the data through the lifecycle. Data that is specific and purpose-shaped for their missions.
What About the Infrastructure and Specialized Roles?

By now you’re probably asking...What happens to the technical team and the underlying infrastructure and the clusters...and how do I get the compute close to the data and what about data sovereignty and the physical storage layer and the costs? All good questions. The answer is these are details that are pushed to a self-service layer managed by a group of engineers that serve the data owners. And as long as the domain expert / data owner is driving monetization, this piece of the puzzle becomes self-funding. The engineers by design become more domain aware and incentive structures are put in place to connect them more closely to the business.

As we said before, Snowflake has to help these users optimize their spend with predictive tooling that aligns spend with value and shows ROI. While there may not be a strong motivation for Snowflake to do this, our belief is that they’d better get good at it or someone else will do it for them and steal their ideas.

What Does the Spending Data Say About Snowflake?

Let’s end with some ETR data to see how Snowflake is getting a foothold on the market.

Followers of Breaking Analysis know that ETR uses a consistent methodology to go to its practitioner base each quarter and ask them a series of questions. They focus on the areas that the technology buyer is most familiar with and ask questions to determine the spending momentum around a company within a specific domain.

The chart below shows one of our favorite examples. It depicts data from the October ETR survey of 1,438 respondents and isolates on the data warehouse and database sector. Yes, we just got through telling you that the world is going to change and Snowflake is not just a data warehouse vendor but there’s no construct today in the ETR data set to cut the data on a data cloud or a globally distributed data mesh.

What this chart shows is Net Score on the Y-axis. That is a measure of spend velocity. It’s calculated by asking customers are you spending more or less on a platform and then subtracting the lesses from the mores. It’s more granular than that but that’s the basic concept.
On the X-Axis is Market Share which is ETRs measure of pervasiveness in the survey. You can see superimposed in the upper right hand corner, a table that shows the Net Score and the Shared N for each company. Shared N is the number of mentions in the dataset within, in this case, the data warehousing sector.

Snowflake, once again, leads all players with a 75% Net Score. This is a very elevated number and is higher than that of all other players, including the big cloud companies. We’ve been tracking this for a while and Snowflake is holding firm on both dimensions. When Snowflake first hit the dataset, it was in the single digits along the horizontal axis. And it continues to creep to the right as it adds more customers.

How Snowflake Customers are Spending

Below is a “wheel chart” that breaks down the components of Snowflake’s Net Score.

The lime green is adoption, the forest green is customers spending more than 5% the gray is flat spend, the pink is declining by more than 5% and the bright red is retiring the platform. So you can see the trend. It’s all momentum for this company.

And what Snowflake has done is grabbed a hold of the market by simplifying data warehouse but the strategic aspect of that is that it enables the data cloud leveraging the global mesh concept. And the company has introduced a data marketplace to facilitate data sharing.

We envision domain experts and their developers collaborating across an ecosystem to build new data-oriented applications, products and services leveraging this global mesh.

Metcalfes’s Law Applied to Data Sharing

This is all about network effects. In the mid to late 1990’s as the Internet was being built out, this author worked at IDG with Bob Metcalfe who was the publisher of Infoworld at the time. During that period we would go on speaking tours all over the world and we all listened carefully as Bob applied Metcalfe’s Law to the Internet. The law states that the value of the network is proportional to the square of the number of
connected nodes or users on the system (see below). Said another way, while the cost of adding new nodes to a network scales linearly, the consequent value scales exponentially.

Now apply this powerful concept to a data cloud. The marginal cost of adding a user is negligible...practically zero. But the value of being able to access any data set in the cloud...well let’s just say that there’s no limitation to the magnitude of the market. And our prediction is that this idea of a global mesh will completely change the way leading companies structure their business. Putting data at the core. Not by creating a centralized data repository, but rather creating a structure where the technologists serve domain specialists...as it should be.

What do you think?

Ways to get in touch: Email david.vellante@siliconangle.com | DM @dvellante on Twitter | Comment on our LinkedIn posts.

Remember these episodes are all available as podcasts wherever you listen.

(*Disclosure: theCUBE is a paid media partner for the Snowflake Data Cloud Summit. Neither Snowflake, the sponsor for theCUBE’s event coverage, nor any other sponsors have editorial control over content on Wikibon, theCUBE or SiliconANGLE.)

Image credit: kei907

Watch the full video analysis:
David Vellante is co-CEO of SiliconANGLE Media, as well as co-founder and Chief Analyst at The Wikibon Project, the world’s leading open source technology research community. Dave is a long-time tech industry analyst, entrepreneur, writer and speaker. As co-host of theCUBE – “The ESPN of Tech,” Vellante has interviewed over 5,000 experts since 2010. He is also a co-founder of CrowdChat, an angel funded startup based in Palo Alto using big data techniques to extract business value from social data. Prior to these exploits, Dave founded a CIO consultancy and spent a decade growing and managing IDC’s largest business unit. He lives in Massachusetts with his wife and four children where he is active in town activities including serving as the president of his town’s local “Kiddie Sports” association. Dave holds a B.S. in Applied Mathematics from Union College.

David Vellante
@dvellante
david.vellante@wikibon.org