Data Warehousing on Microsoft/Azure SQL Platforms

Before cloud, Microsoft's offerings for SQL Server databases were straight forward to understand. You generally chose the most recent version available to you (i.e., 2005, 2008, 2012, 2016, etc...). All you needed to decide on was what edition you wanted (i.e., Standard or Enterprise) based on features and then license it for the number of cores you intended to use. In general, if you were embarking on a data warehousing project, Enterprise edition was the way to go as it had higher resource limits, better performance and offered Always On High Availability. As the years rolled on, this way of licensing remained consistent and a straightforward choice.

Fast forward to today and in the Azure cloud you have a multitude of SQL options:

- SQL Server running on a VM
- SQL Managed Instance
- SQL Server
    - General Purpose
    - Hyperscale
    - Business Critical
- Synapse Analytics
    - Dedicated SQL Pools
    - Serverless SQL Pools

Some organisations have data policies that prevent them from storing data on the cloud which does leave you with the on-premises offerings only. The intention of this article is to discuss which option to choose for a modern cloud-based data platform, so without further ado below are the ones we can automatically rule out for key reasons:

- **SQL Server running on a VM**: This option is almost identical to running SQL Server on-premises with the only difference being it's hosted on IaaS versus a physical host. Moving to the cloud is all about reducing management overhead and total cost of ownership and this option is not aligned.
- **SQL Server Managed Instance**: Whilst this is a PaaS offering, it's more targeted for lift and shift scenarios to get on-premises databases migrated onto the cloud with as little effort as possible. If you had an on-premises data warehouse that you wanted to migrate, that's the only reason I'd choose this option.
- **SQL Server – Business Critical**: Whilst this is a PaaS offering, it's more targeted for high transaction, low latency scenarios (i.e. 1ms-2ms). Data warehousing workloads are typically low transaction with standard latency requirements (5ms-10ms).

- **Synapse Analytics – Serverless SQL Pools**: In terms of traditional data warehousing, serverless SQL Pools aren't a good fit. They're a compute layer with external storage options only. They're a great option for querying, especially data lakes, and they certainly have a role to play in lakehouse architectures, but the focus here is on data warehouses which rules this out.

If you're embarking on a new modern data platform, hosted in the cloud, your options for data warehousing become a bit simpler, in fact they narrow down to three main candidates:

1. SQL Server – General Purpose
2. SQL Server – Hyperscale
3. Synapse Analytics – Dedicated SQL Pools

The table below details the key differences between the three options.

	Azure SQL Server		Synapse Analytics
	General Purpose	**Hyperscale**	**Dedicated SQL Pool**
Storage Size	Up to 4TB	Up to 100TB	Limitless
Processing	Symmetric Multi Processing (SMP)		Massively Parallel Processing (MPP)
Supports a large number of concurrent users	Yes		No
Supports Primary Keys and Foreign Keys	Yes		No
Supports Scaling	Scale up and out (for read-only replicas only)		Scale up and out
Availability (downtime per year)	99.99% - 99.995% (26.3 minutes – 52.6 minutes)		99.9% (8.77 hours)
Pricing	DTU vCore - serverless vCore – provisioned	vCore – provisioned	provisioned
Best for	Budget oriented balanced compute and storage options.	Most business workloads. Autoscaling storage size up to 100 TB, fast vertical and horizontal compute scaling, fast database restore.	Big data scenarios involving high volumes of data or highly complex queries.

One of the important differences to understand is the Symmetric Multi Processing (SMP) vs Massively Parallel Processing (MPP) options, as they relate directly to the ability to scale up and/or scale out. SMP relates to having a single "shared processor, bus and memory" system. Subsequently, these systems can only be scaled up by provisioning additional CPU's or memory up to the limit of that single system. Scale out can be achieved for read-only workloads, however it's important to understand that the workload will still only be handled by one of the replicas in the scale out cluster, not all of them. MPP relates to having separate machines, each with their own CPU and memory, to handle a request. Subsequently MPP systems can be scaled up and out. This is very different to SMP

because a request submitted to an MPP system can potentially involve all nodes in the scale out cluster. Therefore big data and complex queries are better suited to MPP systems.

Now that we have an understanding of processing architecture, we can focus on the factors that will impact your decision to choose one over the others. All three are suitable options, but which one is the best? … Well, like most things in life … it depends. What is your definition of best? Cheapest, fastest, most features, easiest to develop against, etc. At an absolute minimum, the following questions should be considered as they'll either automatically eliminate an option or strongly favour one, making it a more obvious choice.

- **What is my expected database size now and in a few years time?** >100 TB's? Synapse is your only option. <4 TB? then all three options remain on the table.
- **What is the expected complexity of queries on the data?** Highly complex queries on big data are best suited to Synapse due to its MPP architecture.
- **How much downtime am I willing to accept?** If 8.7 hours of downtime a year is too much, then Azure SQL Database is a better option.
- **What is the expected concurrency of requests?** If you expect a high concurrency of requests, the Azure SQL options are better suited. Synapse Analytics could still be an option if you employ data marts to handle concurrency, however this may introduce timing and data currency issues in needing to sync your Synapse Analytics instance to your data marts hosted in Azure SQL Database.

If after asking yourself these questions, you're still left with 2 or more of the options or even worse, no option fits your use case, then I think it's time to contact a Fujitsu Data & AI specialist, to help you choose the best option for your specific needs.

As if the above wasn't enough to consider, data Lakehouse architecture is quickly becoming the new standard approach to warehousing, which is essentially doing away with all the above, instead opting to represent your data warehouse in the data lake, hence the name data lakehouse. The **Synapse Analytics – Serverless SQL Pools** option which I excluded due to lack of its own storage is perfectly suited to a lakehouse architecture. So why did I even mention data lakehouse then? … I think I know what the topic for my next article will be, stay tuned.

Contact
Fujitsu Data & AI
+61 3 9924 3000