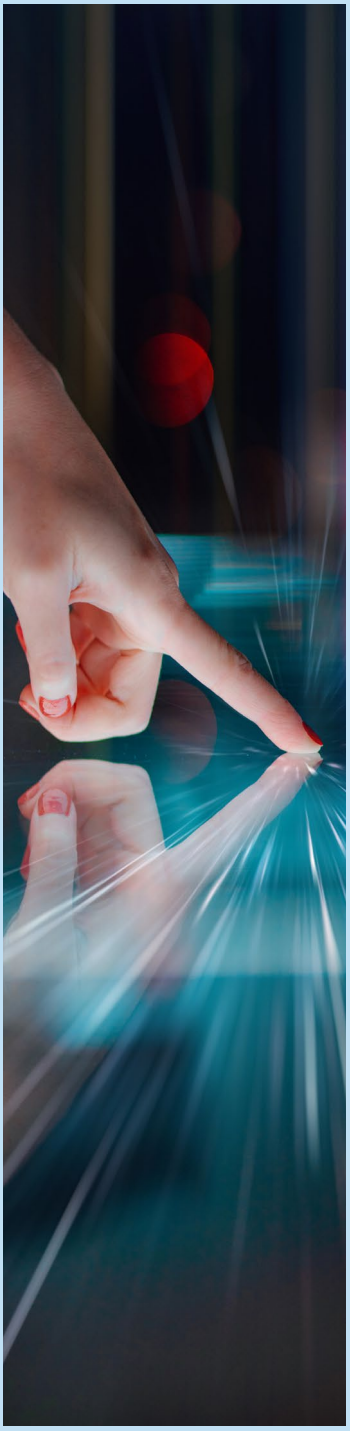
Infosys topaz

VIEW POINT



# POWER BI PREMIUM CAPACITY MANAGEMENT


## Abstract

There is an ever-growing demand for enterprise grade Business Intelligence solutions. Therefore, it is imperative that platform teams procure and manage enough capacity to cater to these use-cases. With Power BI, Microsoft offers two kinds of capacities to handle reporting workloads viz., standard shared capacity and premium capacity. As the name suggests, premium capacities are tailor-made for enterprise-grade or business-critical reporting systems.

However, premium capacities come at a heavy cost, and the cumulative expenditure for an enterprise could be significant if individual teams procure them on their own. In this article, we cover our recommendations on how enterprises should centralize, procure, and manage premium capacities to meet business demands, while keeping the costs down as much as possible.

We also explain our strategy for categorizing different reporting use-cases, and which capacities they should be allocated to. We also provide pointers on how to track the load on premium capacities (and when to buy more), and our recommendations & lessons we've learnt from our experience of managing shared premium capacities for one of the largest Power BI tenants in the world.

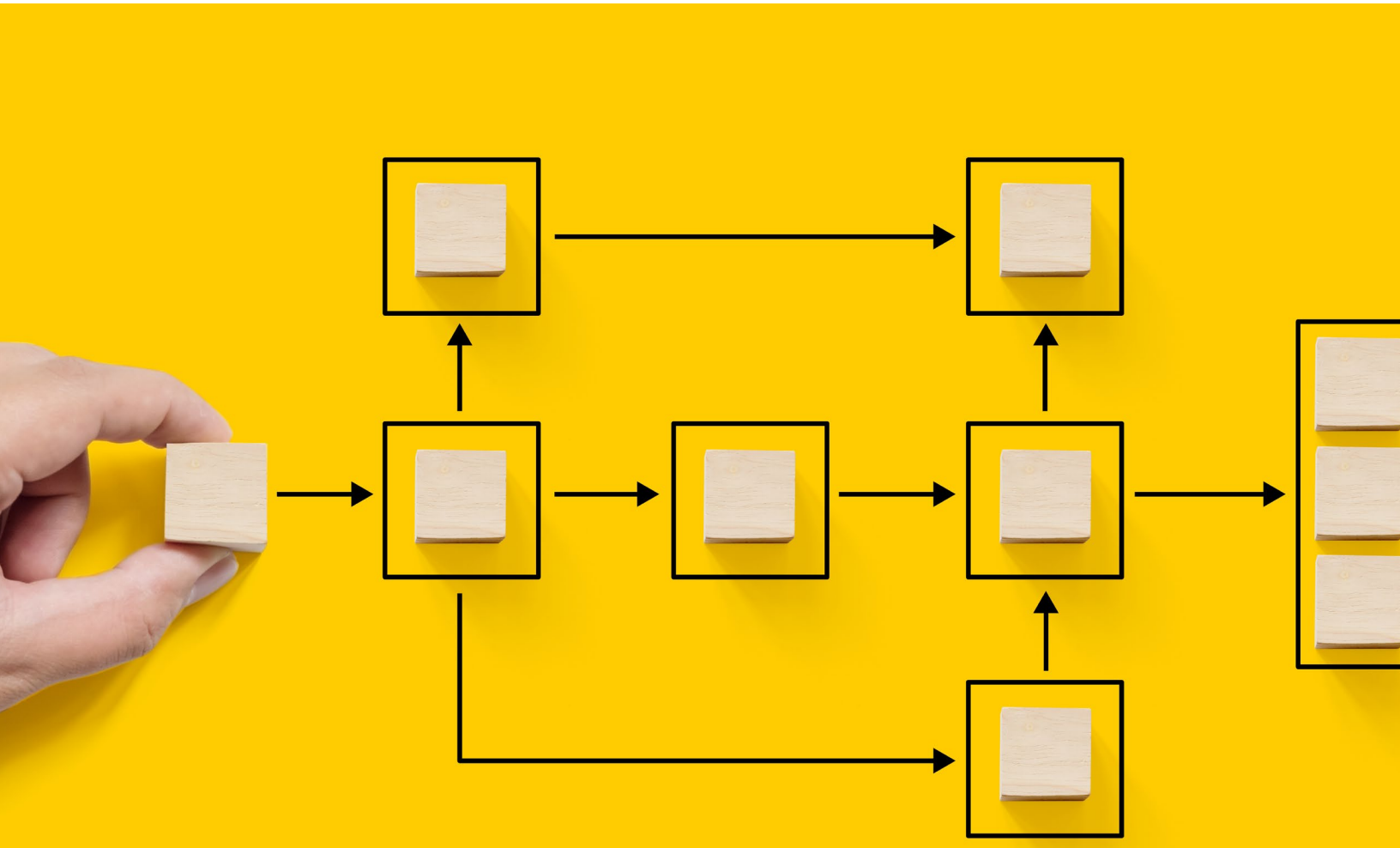

Infosys®
Navigate your next

# Table of Contents



## What is Power BI? and Power BI Premium?

- Power BI, as one of the components of Microsoft's Power Platform, is marketed as a self-service business intelligence & reporting solution. The underlying philosophy behind Power Platform is enabling business users and citizen developers to build reports, apps, automated flows, or simple chat bots on their own. The guiding principle is ease of use to begin with i.e., low-code, no-code design.
- This implies there is an assumption that Power Platform is meant to be used for building smaller, personal, team or departmental/domain level solutions. While you could build large scale apps, reports, chat bots or flows with Power Platform products, the performance would not be comparable to other products on the market which offer much more customizability in terms of infrastructure as well as design.
- While Power BI is meant for self-service BI reporting, it has also been widely adopted by various organizations as the enterprise-wide reporting tool of choice. Furthermore, we are seeing an increasing trend of initially self-serve reports eventually scaling up to serve the wider organization, and therefore there is an increasing need for central IT departments and teams to take up the ownership of the underlying artifacts so that they meet the performance and security considerations expected of an enterprise-wide reporting & data solution.
- Although, Microsoft offers other products that specifically cater to enterprise-wide, performant semantic data models (same technology that powers Power BI Datasets) i.e., Azure Analysis Services; going forward, Microsoft is going to make Power BI Premium the platform of choice for this purpose. This would help consolidate the self-service as well as enterprise BI reporting within Power BI environment itself.
- Power BI Premium offers the full range of capabilities and functionalities within Power BI, starting with increased thresholds for memory & CPU consumption of the reporting solutions. It also offers possibility of up to 48 scheduled refreshes of the dataset per day, enhanced compute functionalities for dataflows, support for datamarts, unlimited distribution of artifacts within and outside the tenants to Power BI Free license users, and more.



## Which reporting use-cases warrant a premium capacity?

Reporting use-cases can come in all sizes and shapes. It is a key decision in the report design and governance process to ensure that there are enough resources/capacity allocated to handle the kind of workload the reporting use-cases warrant. In our view, we can broadly categorize the reporting use-cases into 3 categories, namely bronze, silver, and gold (explained below).

Please note, these categorizations are meant to be used as guiding principles only. There are no specific metrics which can be considered as decision boundaries between categories, and we recommend that when a reporting use-case straddles between two categories – go for the higher category to ensure best possible user-experience.

Bronze: These tend to be the overwhelming majority of reporting use-cases within a tenant. They are generally developed by citizen developers across the organization. These use-cases could be connecting to and importing small to moderate data models (say less than 1 GB Power BI file size on disc); for team/department level reporting, with limited concurrent usage requirements.

- In this case, the default standard shared capacity which you get with standard app workspaces (which you can create if you have a Power BI Pro license) is the way to go. There are no additional costs. Premium capacity is not necessary for such reporting use-cases.

Silver: These could be use-cases that have moderate to large data models, with tens of millions of rows (>1GB Power BI Dataset size on disc), for multi-department/multi-geo or enterprise-wide reporting use cases, with high concurrent usage requirements but with acceptable levels of occasional slowdowns in interactive performance (not business critical). It could also be that the use-cases aren't CPU intensive per-say but could be leveraging any Power BI Premium only functionality (datamarts etc.)

- In this case, you may opt to allocate the use-cases to a Shared Premium Capacity: wherein you get to leverage premium features for a fraction of the cost of purchasing a dedicated premium capacity. With shared capacity, you do have to deal with noisy-neighbour problems, which result in brief periods of slowdowns due to occasional overloading of the capacity – this should be acceptable to the team whose workspace(s) you as a capacity admin are allocating to a shared premium capacity.
- But overall, shared premium capacity offer significantly better performance than standard shared capacity, and you get significant burst capacity (as shared premium capacities are generally higher SKU capacities – which offer higher throughput thresholds for CPU usage per evaluation window).

Gold: These could be large enterprise-wide or business critical reporting use-cases. They could have large data models, with tens or hundreds of millions, to billions of rows, very large dataset size, for enterprise reporting requirements for business-critical decision making, with high levels of expected concurrent usage, or critical reports that have very high-performance requirements (no level of throttling is tolerable).

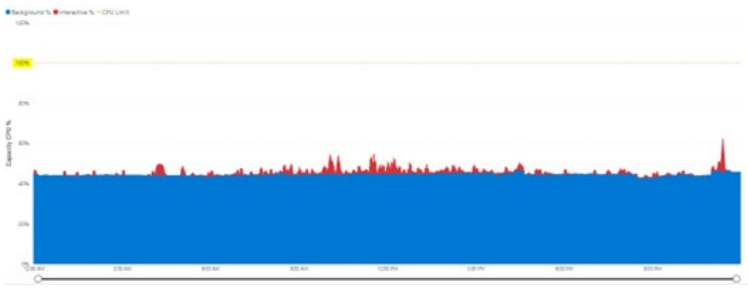
- In this case, it is best to isolate these use-cases to a dedicated premium capacity, to eliminate any risks of noisy-neighbour problems. You get the full capacity to run your business-critical reporting.
- However, full ownership of the costs also goes to the requesting team, and depending on the capacity SKU, these costs could be significant. They can also miss out on the benefits of generally higher SKU shared premium capacities which have higher throughput thresholds for CPU usage than the generally smaller SKU dedicated premium capacities.

## How does the Shared Premium Capacity Model work?

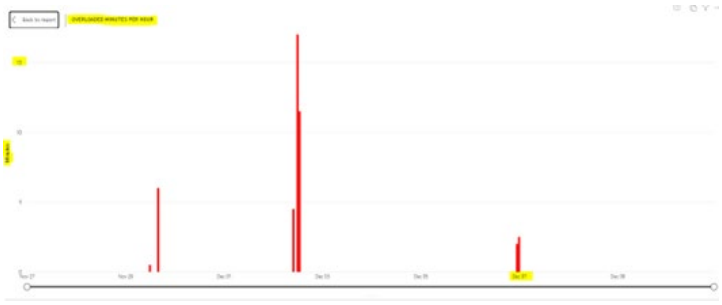
- Instead of letting individual teams within the organization directly purchase smaller premium capacities from Microsoft, the Power BI Administrators can purchase larger SKU premium capacities, and then let the internal teams subscribe to these on an internal monthly pricing model for a fraction of the cost, therefore reducing the overall cost to the organization significantly.
- In our case, we also purchased a smaller "development" premium capacity, where we let first time users and teams try out premium for free for a trial period of 90 days. This shared capacity is meant to support teams building POCs on premium prior to opting for a subscription on the larger "production" shared premium capacities.
- Shared premium capacities are subject to noisy-neighbour problems. A sub-optimal artifact from some other team/person can push the entire premium capacity into overload, which triggers throttling on interactive utilization in the subsequent evaluation window (artificial delays injected by Microsoft on the running queries), leading to bad user-experience. Administrators need to continuously track the overall utilization on the shared capacities, and upscale as and when the level of throttling becomes more than an accepted threshold (given budgetary constraints).
- However, the throttling, or the artificial delays in interactive queries, is applied only to the subsequent 30 second evaluation interval, and if the capacity utilization falls below the threshold in the subsequent window, the throttling/delay is removed for the evaluation window after that.
- In our view, considering all the options, the pros of shared premium capacities outweigh the con(s). Majority of reporting use-cases don't warrant a dedicated premium capacity, and with significant savings for the organization by leveraging economies of scale, plus more burst capacity to cater to periods of intense utilization for premium artifacts makes going for a shared premium capacity model a sensible choice.

## Tracking capacity workloads and when to buy more?

- Microsoft provides the Capacity Metrics Analysis report for all your premium capacities for capacity admins to monitor & track the utilization trends. It pulls last 2 weeks' worth of detailed utilization data for all the artifacts hosted on a premium capacity.
- This app shows the split of background and interactive operations as Microsoft considers it. Long running refresh operations, be it scheduled, or on-demand are considered as background operations; whereas queries generated when users are interacting with artifacts hosted on premium (opening reports, slicing/dicing data etc.) are considered as interactive operations.
- The CPU load for background operations once they complete is spread evenly across each 30 second evaluation window for the next 24 hours i.e., once a data refresh operation ends, the total CPU time consumed for that is spread evenly across the next 2880 evaluation windows (24 hours * 60 minutes * 2 half-minute windows).
- The CPU load for interactive operations is considered in real time. When Gen2 premium capacity was initially rolled out, the full interactive CPU load for a query within an evaluation window was considered in evaluating the total CPU load within the said evaluation interval. However, Microsoft has since introduced some smoothing, wherein they distribute the interactive CPU load (after an interactive query ends) up to 10 minutes.
- At any evaluation interval (30 second periods), if the sum of smoothed background and interactive CPU load goes over the capacity threshold (ex: for a P1 capacity, the capacity threshold is 240 seconds per evaluation window), the capacity is considered to be in overload, which triggers some level of throttling in the subsequent evaluation window.
- The Capacity metrics analysis report shows the CPU load over time (shown below). In the screenshot below, the background operation, which is spread across each evaluation window is hovering around 40-45% of the capacity threshold. We also see occasional spikes for interactive utilization, mostly clustered around the working hours (as expected since that's when people would be interacting with the reports and datasets).



- The CPU over time visual presents the live picture of the levels of load on any given premium capacity. In our view, we recommend that the background CPU load should not exceed 50% of the capacity threshold. In other words, we should leave at least 50% of the capacity for interactive workloads.
- Another important visual is Overloaded minutes per hour, which shows the level of throttling on the capacity over the past 2 weeks. For example, in the snapshot below, we can see there were some instances of throttling on this capacity – the worst case being the capacity being throttled for ~18 minutes in a single hour on December 2



.

- The overloaded minutes per hour visual should inform the decision to purchase more capacity. If the throttling is occasional, and within certain accepted parameters buying more capacity is not necessary. However, if you're seeing throttling consistently for periods longer than 10-15 minutes per hour every day, we recommend buying more capacity, or rebalancing the load on your shared premium capacities (if you have another shared albeit less utilized premium capacity).

## Conclusions & Recommendations

- To make sure shared premium capacities are not overloaded beyond acceptable limits, capacity admins must continuously track the loads on these capacities using the Capacity Metrics Analysis app that Microsoft provides. This reporting is built on telemetry data captured by Microsoft for all the activities on the concerned premium capacity.
- Admins should also configure notifications for when the capacity hits the 95% or 100% of the capacity threshold for CPU seconds. Capacity admins can then look at the Capacity Usage Metrics app to see if the overloading is continuous.
- If the capacity is indeed overloaded beyond acceptable limits, then admins can identify the top contributing artifacts to the overload from the report and reach out to the concerned team/developers to explore avenues for optimization as a long-term fix. As a short-term fix, if you have multiple shared premium capacities, capacity admins can also rebalance the loads to ensure a judicious use of all the available resources.
- We recommend capacity administrators to disable/discourage XMLA read/write workloads on shared premium capacities. XMLA read operations are considered as interactive utilization, and based on our experience, are the single biggest contributors to high levels of continuous throttling. If there is a legitimate requirement for XMLA endpoint operations, these artifacts should be isolated to a dedicated premium capacity to ensure these operations don't cause long periods of slowdowns for other teams on a shared premium capacity.
- All things considered; shared premium capacities can help an organization avoid significant costs. Instead of a string of small dedicated premium capacities, you can consolidate and purchase larger premium capacities, which multiple teams can share, and benefit from the higher CPU throughput thresholds, along with higher thresholds for memory per artifact.
- However, if there are reporting use-cases which are heavy enough to warrant a larger capacity by itself, or if they are business critical, wherein any level of throttling or performance hit is intolerable, provisioning dedicated premium capacities for these teams and reporting use-cases is recommended.
- In any organization, the tenant & capacity administrators should continuously aim for a judicious mix of shared and dedicated premium capacities. Fundamentally, it's a fine-balancing act of adding capacity/resources to meet the ever-growing demand, while balancing the costs by making sure that the reporting use-cases are well designed and don't use more resources than they really need to.
- We recommend provisioning a trial/development premium capacity wherein you can let users in your organization try out the available functionalities and features and build their premium reporting use cases. With our internal pricing model, we could offer first time users and teams the option of hosting their workspaces in a shared "Development" capacity for free for 90 days.
- In our case, we created a Power App to serve as the one-stop solution for users/teams to request for premium capacity allocation in our tenant. We also automated the financial approval process by leveraging a Power Automate flow to ensure best possible user experience. We recommend adopting similar strategies as they can streamline the process of requesting and allocating workspaces to premium capacities, which can otherwise devolve into hundreds of emails floating around.
- In our view, capacity administrators need to be extremely careful with the auto-scaling feature. In principle, if you are only seeing rare instances of overloads (say once or twice a weak), and any level throttling is unacceptable, configuring few auto-scale cores can be helpful. Auto-scale cores are active for a period of 24 hours after they come live (after the load goes past the capacity threshold). But, if instances of overloading, and resulting unacceptable levels of throttling occur more often than not, we'd recommend purchasing additional capacity as that would be more economical.
- We also recommend periodically auditing/reviewing the heaviest utilizers on any shared premium capacity with an aim to either try and optimize the reporting solutions or deciding whether it is time to push it to a dedicated premium capacity. In our experience, some reporting use-cases started out small which we could host in our shared capacities but, grew over-time to such a degree that they warranted their own dedicated premium capacity.
- And lastly, no amount of premium capacity can compensate for bad design. Power BI Administrators should aim to enforce discipline as much as they can and ensure users aren't abusing premium capacities. Instead of throwing money at the problem (which may not be there in the first place), it is always better to continuously push for the adoption of best data preparation, data modeling, and data visualization practices within the organization.

References:


- What is Power BI Premium? - Power BI | Microsoft Learn
- Manage Microsoft Power BI Premium capacities - Power BI | Microsoft Learn
- Monitor Power BI Premium capacities with the Premium metrics app. - Power BI | Microsoft Learn


## About the Author :


**Atul Singh**
Consultant, SURE Practice – Infosys Consulting


Atul Singh is a consultant in SURE practice, specializing in Data Analytics, Engineering & Business Intelligence domain. He is a Microsoft certified Data Engineering and Data Analyst Associate, with technical expertise in Microsoft's cloud native data & BI stack. In his current role, he serves as one of the Power BI Administrators for one of the largest energy majors in the world, along with working with business as well as technical stakeholders and supporting their critical reporting solutions.


For more information, contact askus@infosys.com


Infosys®
Navigate your next