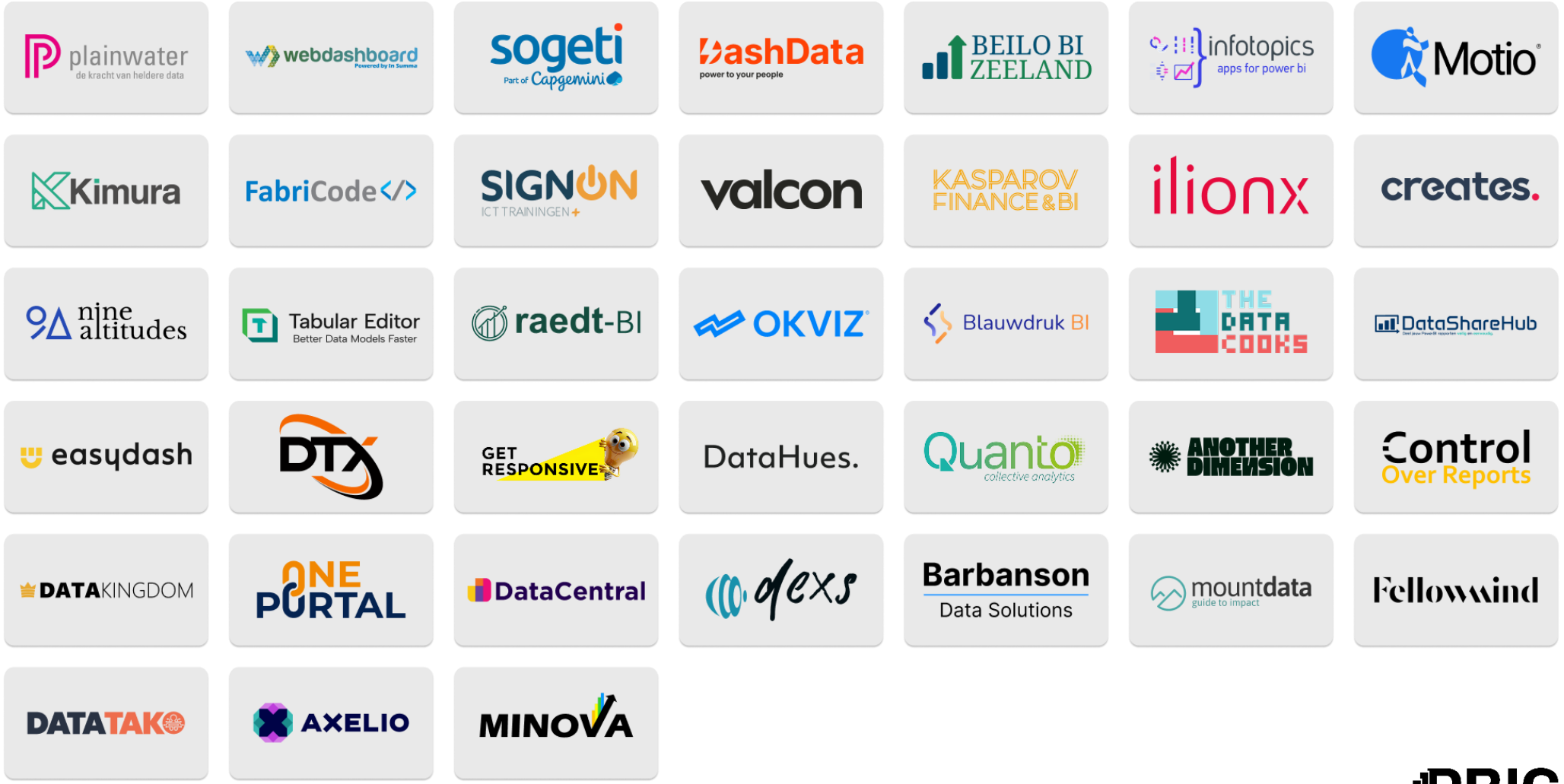
Demystifying the on-premises data gateway

Benni De Jagere

Slides

CAT

A big thank you to our partners



PBIG
POWER BI GEBRUIKERSGROEP

Microsoft

Benni De Jagere

Principal Program Manager | Azure Data Customer Advisory Team (Azure Data CAT)

Azure Data CAT

dataMinds .be Member

/bennidejagere

sessionize /bennidejagere

/bennidejagere

#SayNoToPieCharts

Before we go ..

Learning Goals, we'll cover:

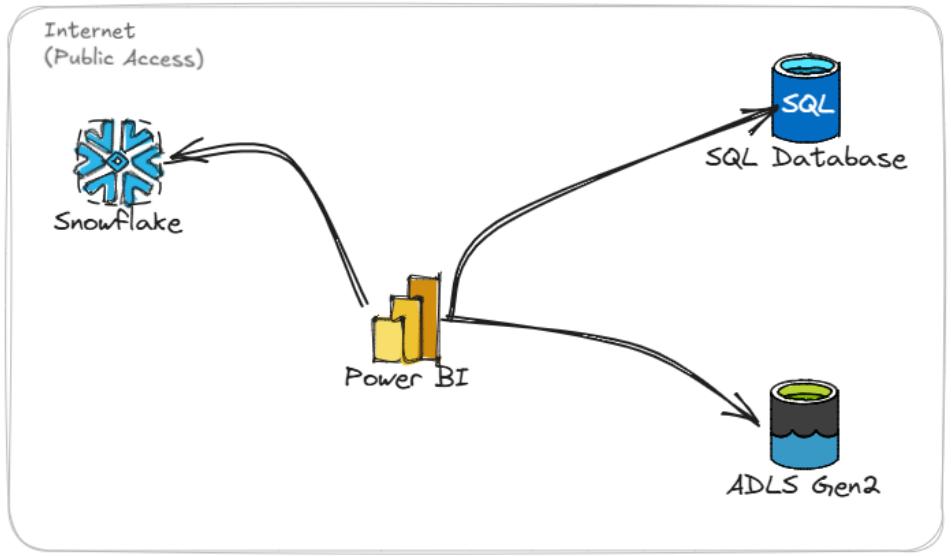
- where the on-premises data gateway can make a difference
- how the on-premises data gateway works, flow of data
- which metrics are relevant to monitor, and how
- how configuration options can have a big impact on stability, performance
- which questions need to be considered for planning decisions

Fundamentals

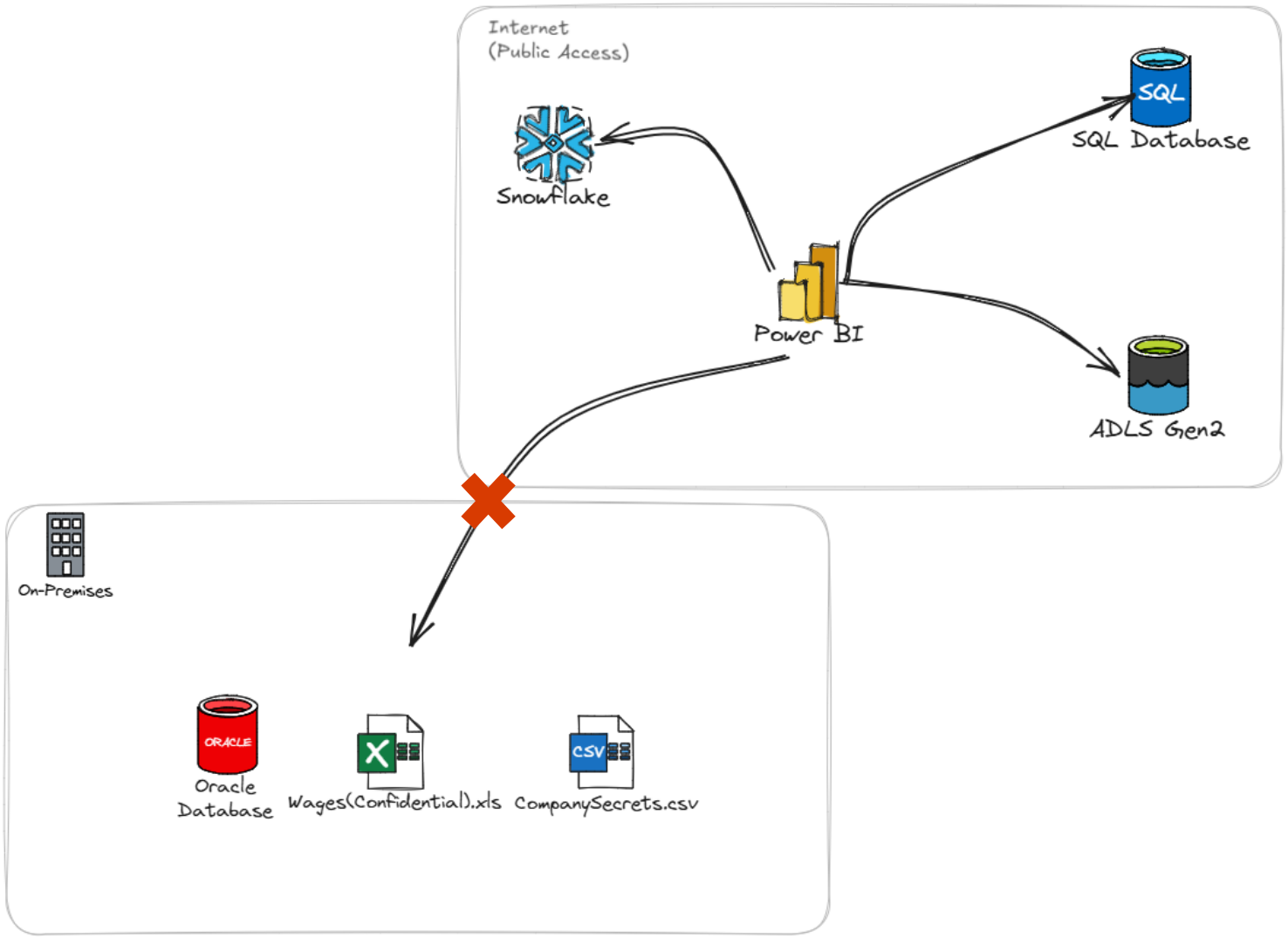
1

Problem Area

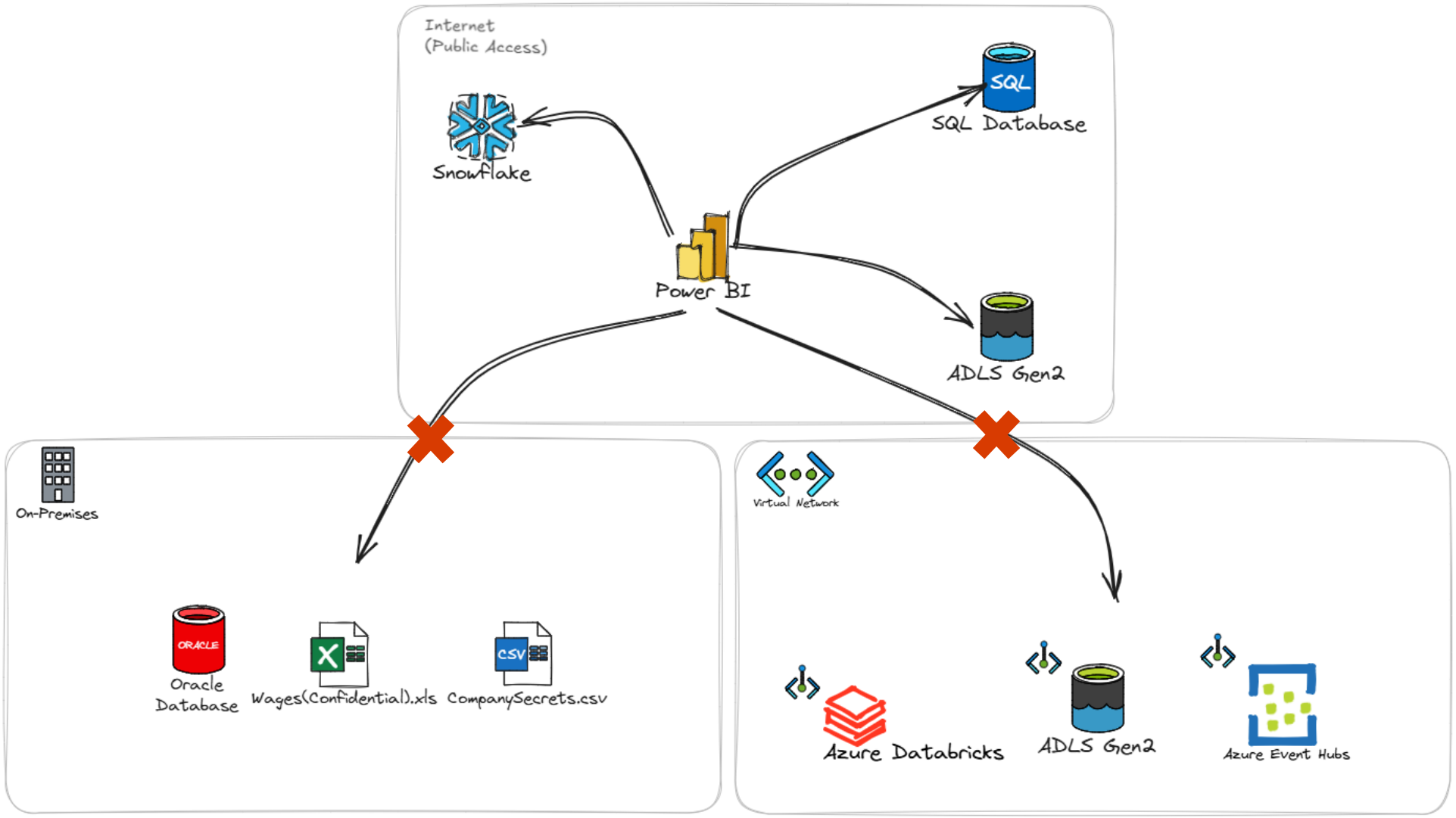
Why?


Internet
(Public Access)
SQL
SQL Database
Snowflake
Power BI
ADLS Gen2

Why?


Internet
(Public Access)
SQL
SQL Database
Snowflake
Power BI
ADLS Gen2
On-Premises
ORACLE
Oracle
Database
Wages(Confidential).xls
CSV
CompanySecrets.csv

Why?


Internet
(Public Access)
Snowflake
SQL
SQL Database
Power BI
ADLS Gen2
On-Premises
Oracle
Database
Wages(Confidential).xls
CompanySecrets.csv
Virtual Network
Azure Databricks
ADLS Gen2
Azure Event Hubs

Connecting to private resources (on-premises / cloud)

- Whitelisting Power BI IPs
  - ~600 IP address ranges, exceeding practical limits of Azure items
  - IP Ranges change weekly (multiple times)
- Azure Service Tags (docs)
  - Significantly reduces overhead, but still complex
  - Not all firewalls can handle these
- Azure Private Link Service Direct Connect (docs)
  - Public Preview, no clear GA yet
  - Requires set up, configuration, ..
- Homegrown solution, pushing data into Fabric
- **Gateways**
  - **On-premises data gateway:** Customer Managed Infra (IaaS), gives most flexibility
  - VNET data gateway: Microsoft Managed infra (SaaS)

Understanding how Power BI processes data (and where the on-premises data gateway fits in)

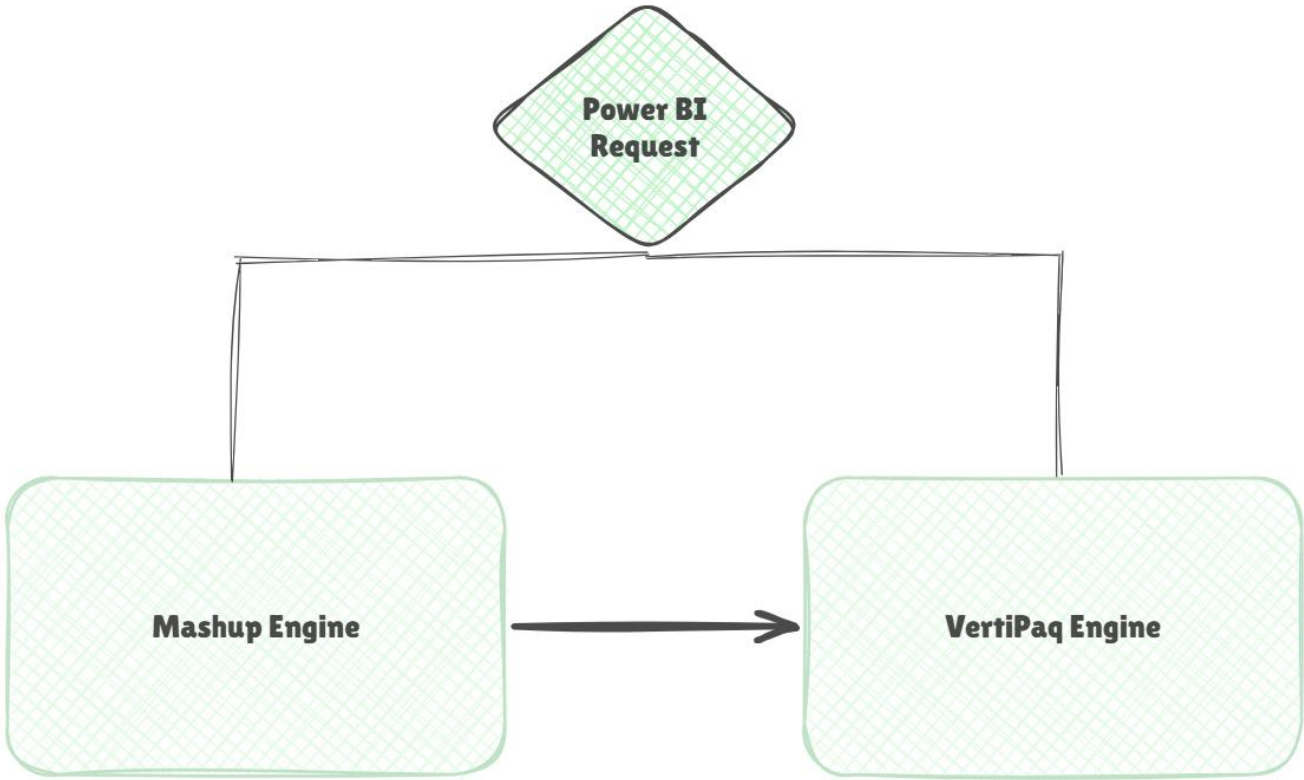
How Power BI requests are handled (simplified)

Power BI requests use two engines

Mashup (PQ)
Data transformation, data movement

VertiPaq (DAX - Storage & Formula):
Handles compression, simple transformations, relationships, storage, ..

Request Scenarios:



How Power BI requests are handled (simplified)

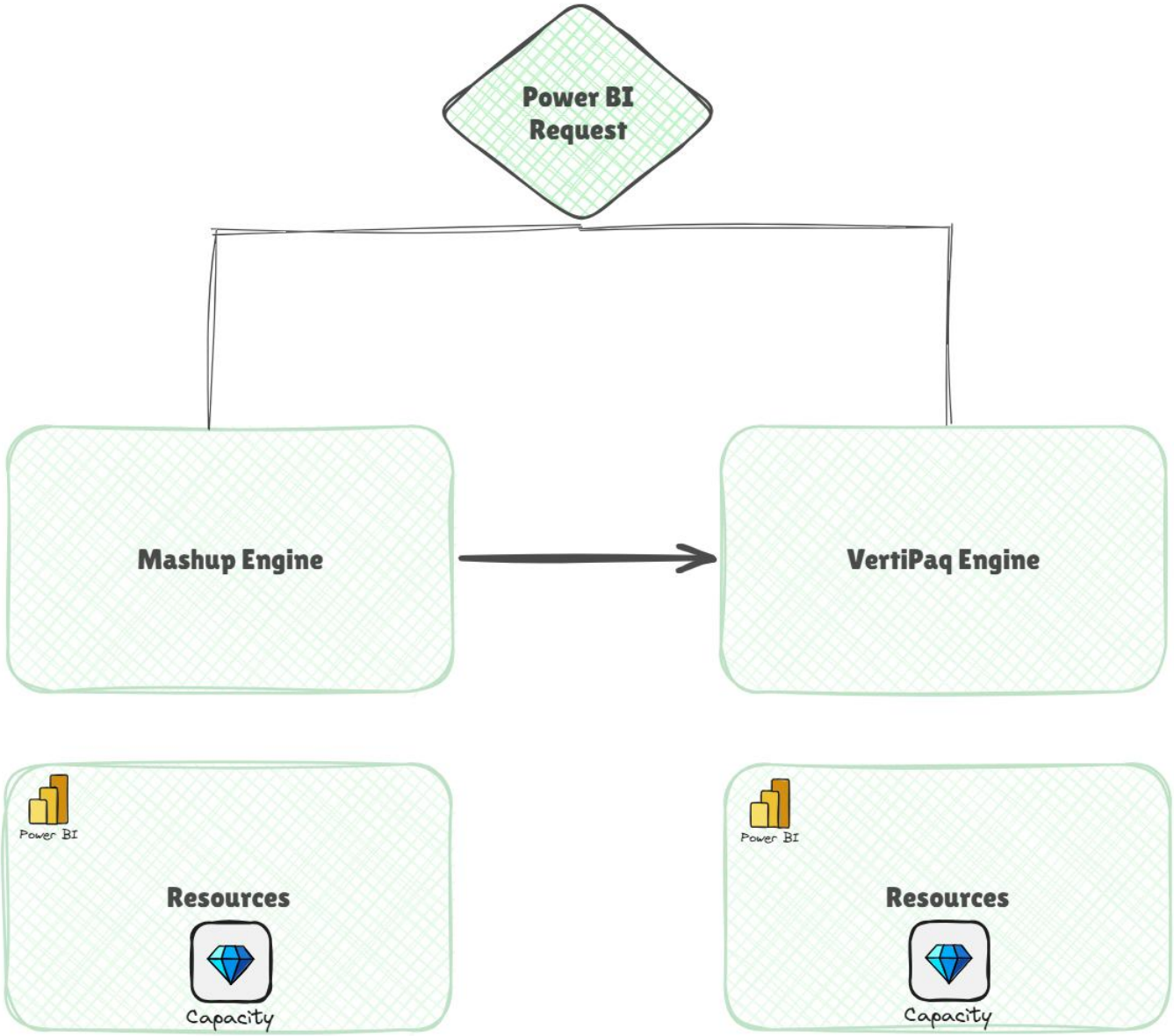
Power BI requests use two engines

Mashup (PQ)
Data transformation, data movement

VertiPaq (DAX - Storage & Formula):
Handles compression, simple transformations, relationships, storage, ..

Request Scenarios:

No Gateway
all components use capacity resources



How Power BI requests are handled (simplified)

Power BI requests uses two engines

Mashup (PQ)
Data transformation, data movement

VertiPaq (DAX - Storage & Formula):
Handles compression, simple transformations, relationships, storage, ..

Request Scenarios:

No Gateway
all components use capacity resources

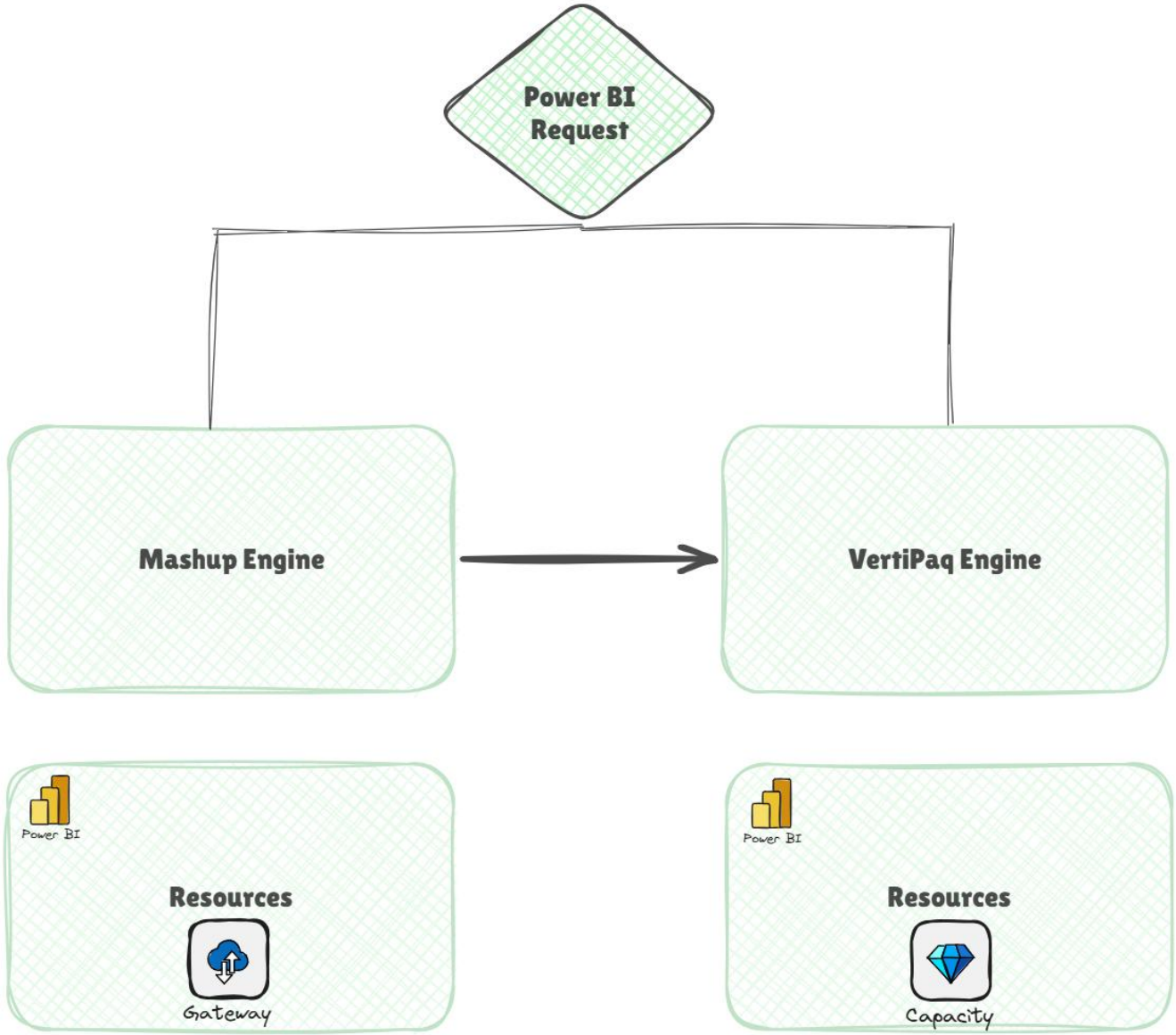
Gateway

Mashup Containers use Gateway resources

Cost Optimization Decision

Gateways tend to be cheaper than capacities
Balance cost with performance

Factor in management overhead



Understanding the Architecture and Process Flow

Architecture

Gateway Cloud Service

Azure Hosted Service that routes data requests between Power BI and gateway clusters

Azure Relay

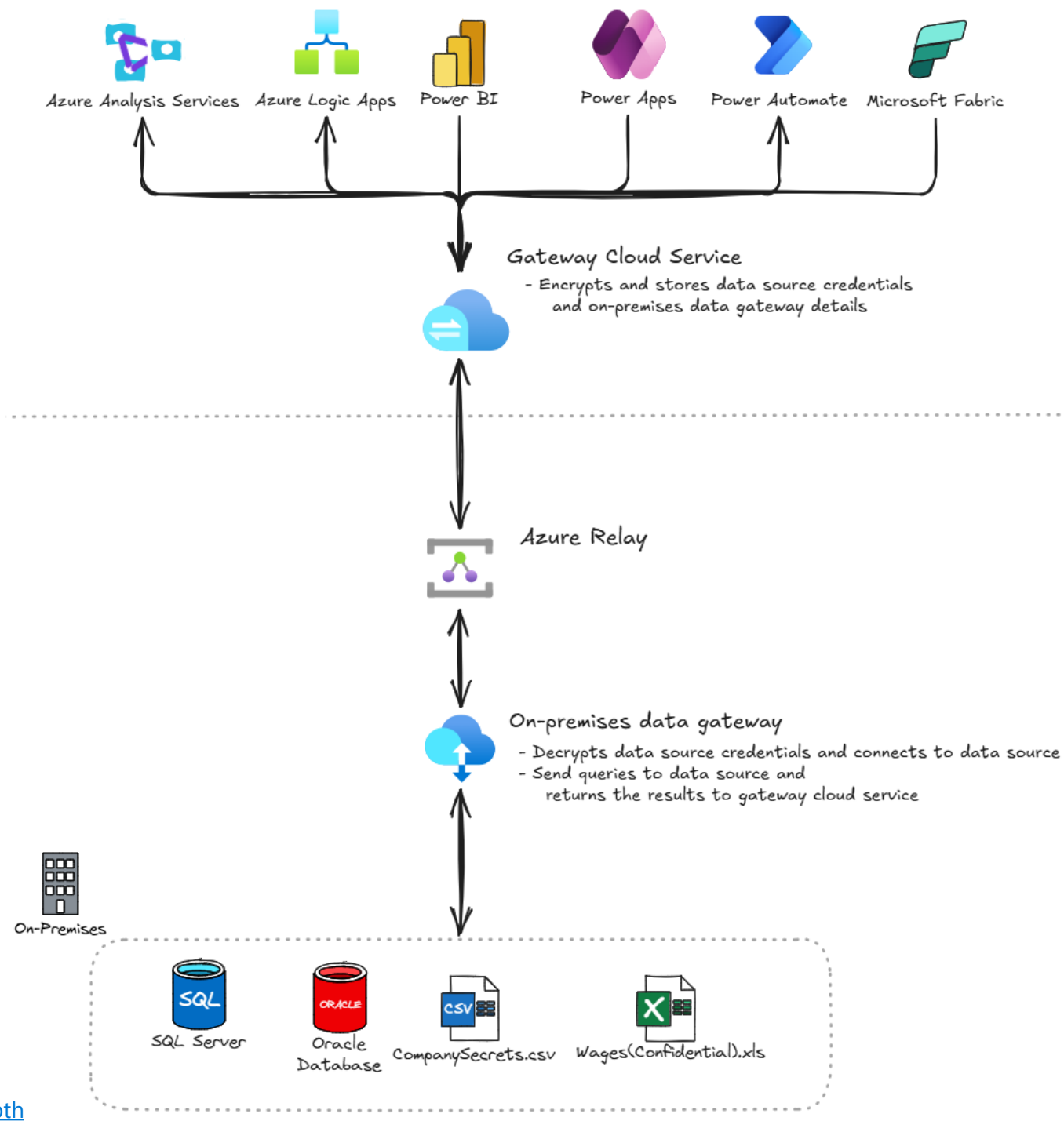
Azure Service that enables cloud services to exchange traffic with on-premises network, only through outbound firewall ports

Gateway Cluster

Group of gateways (nodes) working together to ensure high availability, and/or load balancing

On-premises data gateway

Software installed on corporate machines, enabling data transfer between isolated sources and Power BI Service



https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-onprem-indepth

Process Flow (Single node)

1. The cloud service creates a query and the encrypted credentials for the on-premises data source. The query and credentials are sent to the gateway queue for processing when the gateway polls the service periodically.

2. The gateway cloud service analyzes the query and pushes the request to Azure Relay.

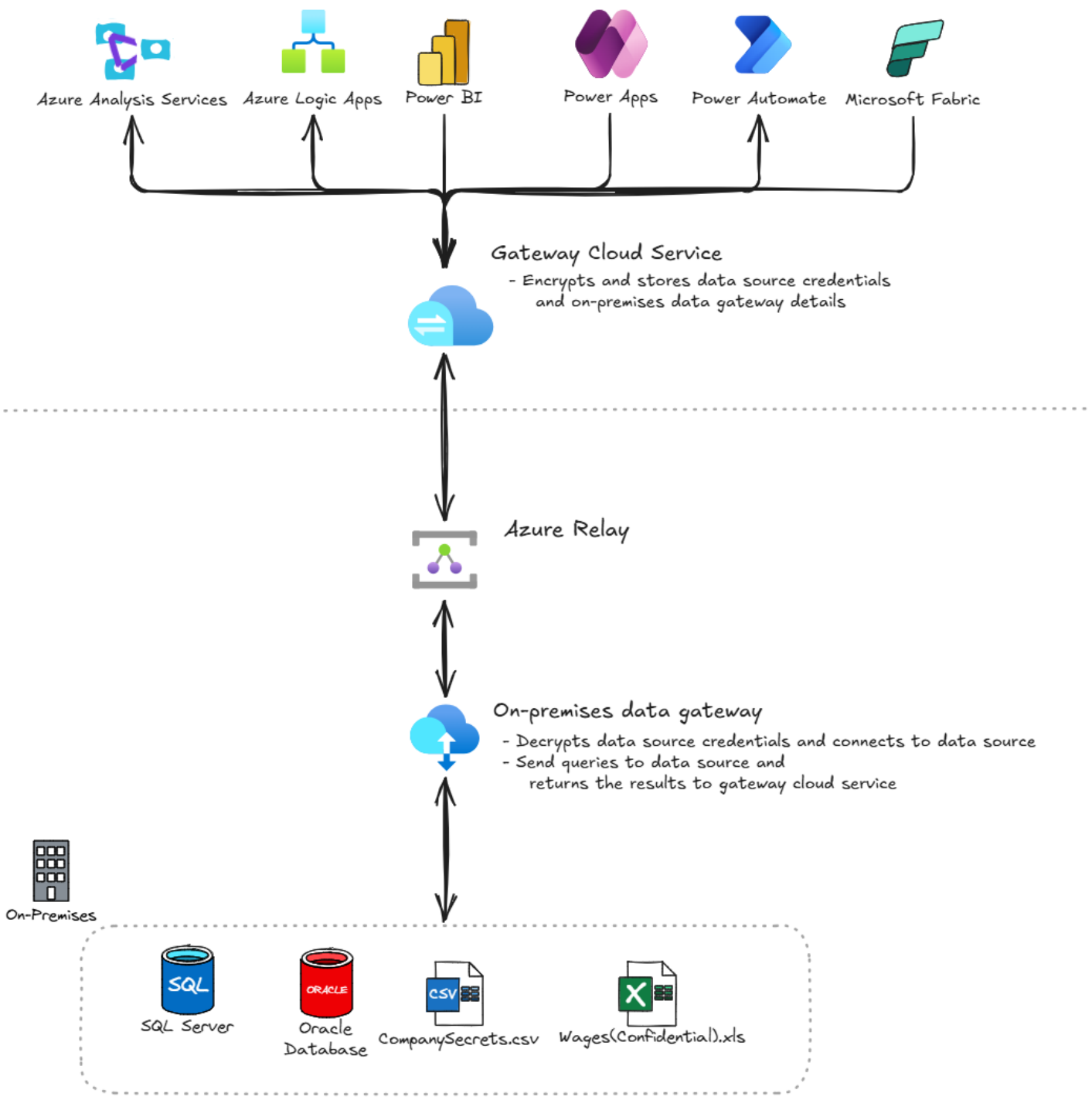
3. Azure Relay sends the pending requests to the gateway when it polls periodically. Both the gateway and Power BI service are implemented to only accept Transport Layer Security (TLS) 1.2 traffic.

4. Node selection (easy when it's single node ..)

5. The gateway gets the query, decrypts the credentials, and connects to one or more data sources with those credentials.

6. The gateway sends the query to the data source to be run.

7. The results are sent from the data source back to the gateway and then to the cloud service. The service then uses the results.

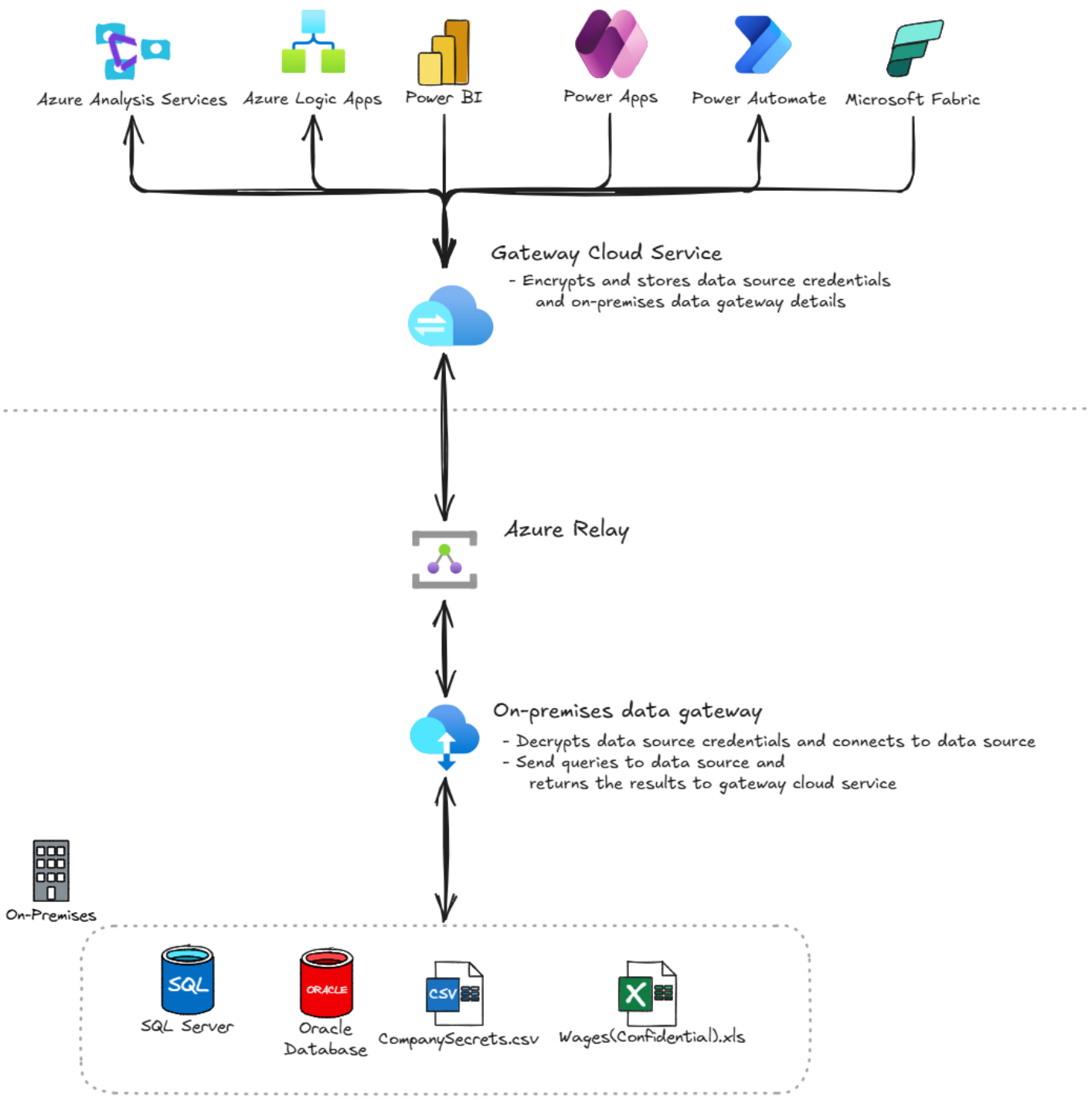


Process Flow (Cluster)

1. The cloud service creates a query and the encrypted credentials for the on-premises data source. The query and credentials are sent to the gateway queue for processing when the gateway polls the service periodically.

2. The gateway cloud service analyzes the query and pushes the request to Azure Relay.

3. Azure Relay sends the pending requests to the gateway cluster.



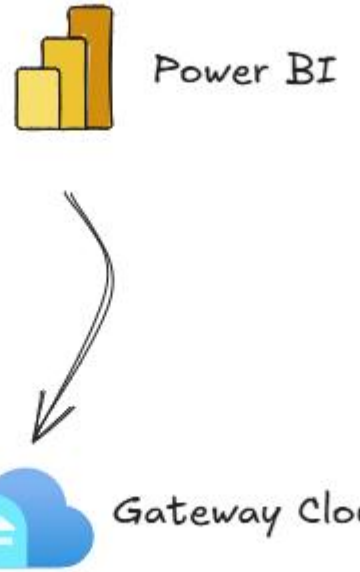
Process Flow (Cluster)

4. A cluster member is selected

Power BI Service sends a query to the Gateway Cloud Service

Based on the matching connection, cluster is identified

Depending on config, 3 options exist:



Process Flow (Cluster)

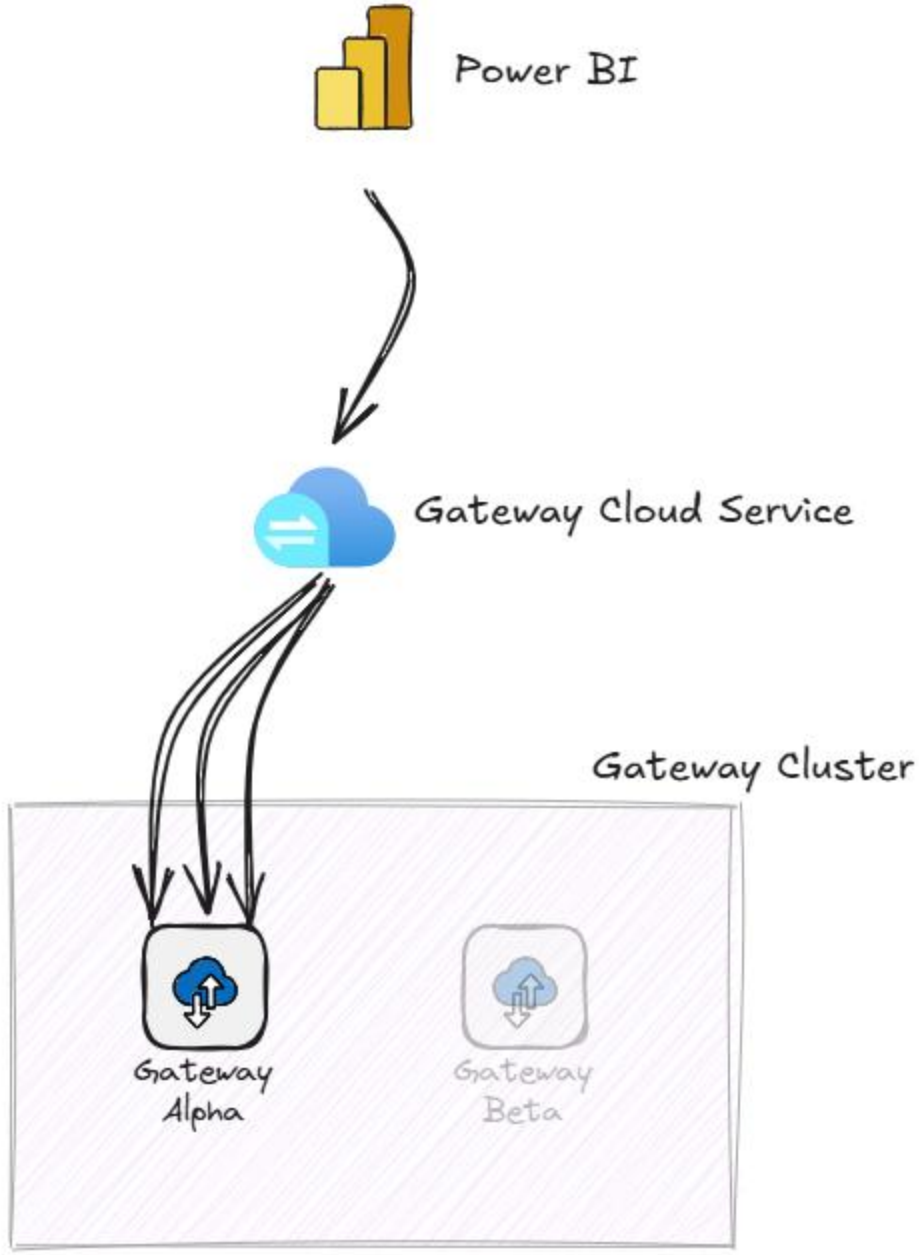
4. A cluster member is selected

Power BI Service sends a query to the Gateway Cloud Service

Based on the matching connection, cluster is identified

Depending on config, 3 options exist:

Send to Alpha, unless unavailable



Process Flow (Cluster)

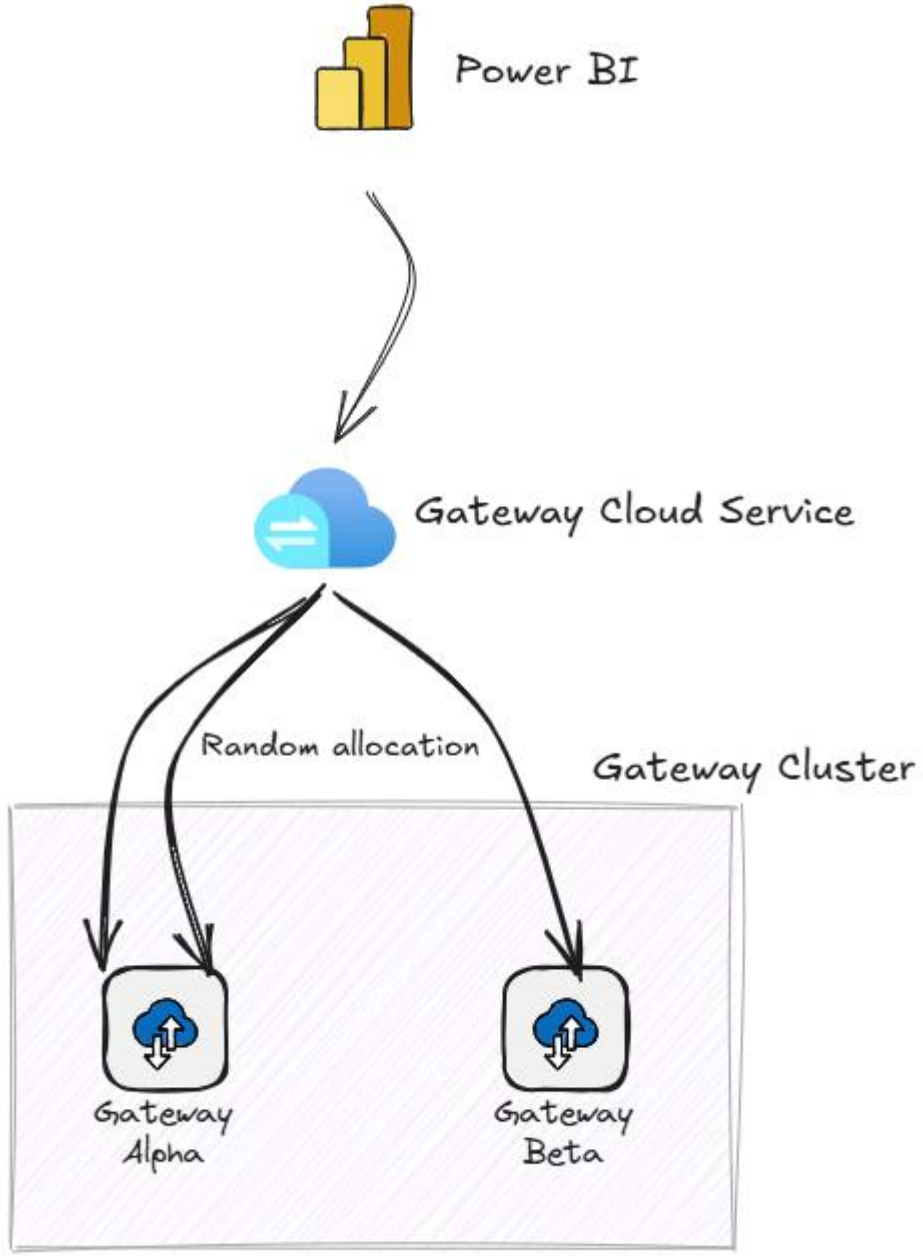
4. A cluster member is selected

Power BI Service sends a query to the Gateway Cloud Service

Based on the matching connection, cluster is identified

Depending on config, 3 options exist:

- Send to Alpha, unless unavailable
- Select a node at random (random load balancing)



Process Flow (Cluster)

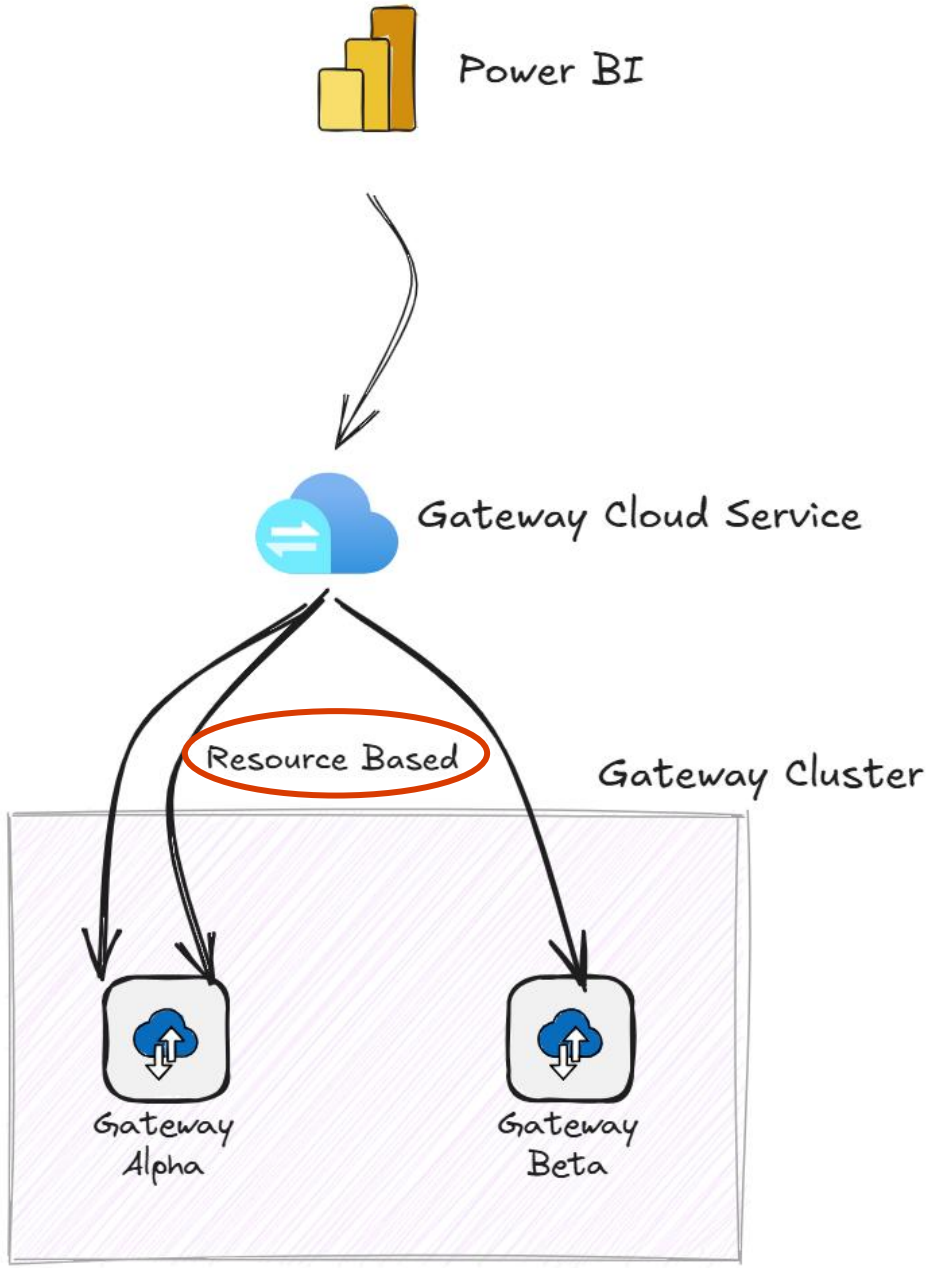
4. A cluster member is selected

Power BI Service sends a query to the Gateway Cloud Service

Based on the matching connection, cluster is identified

Depending on config, 3 options exist:

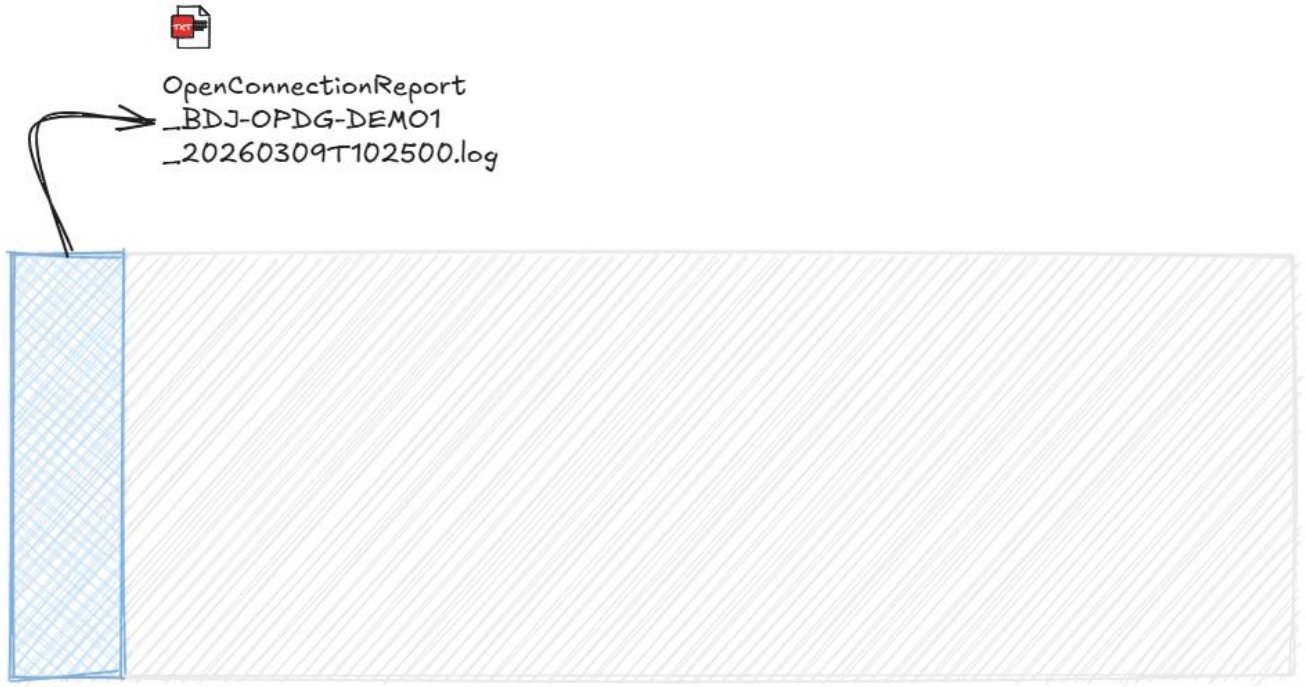
- Send to Alpha, unless unavailable
- Select a node at random (random load balancing)
- Select based on usage threshold



Process Flow (Cluster)

5. Create Connection

- Gateway decrypts credentials using a local key
- Creates a connection to data source, one per query



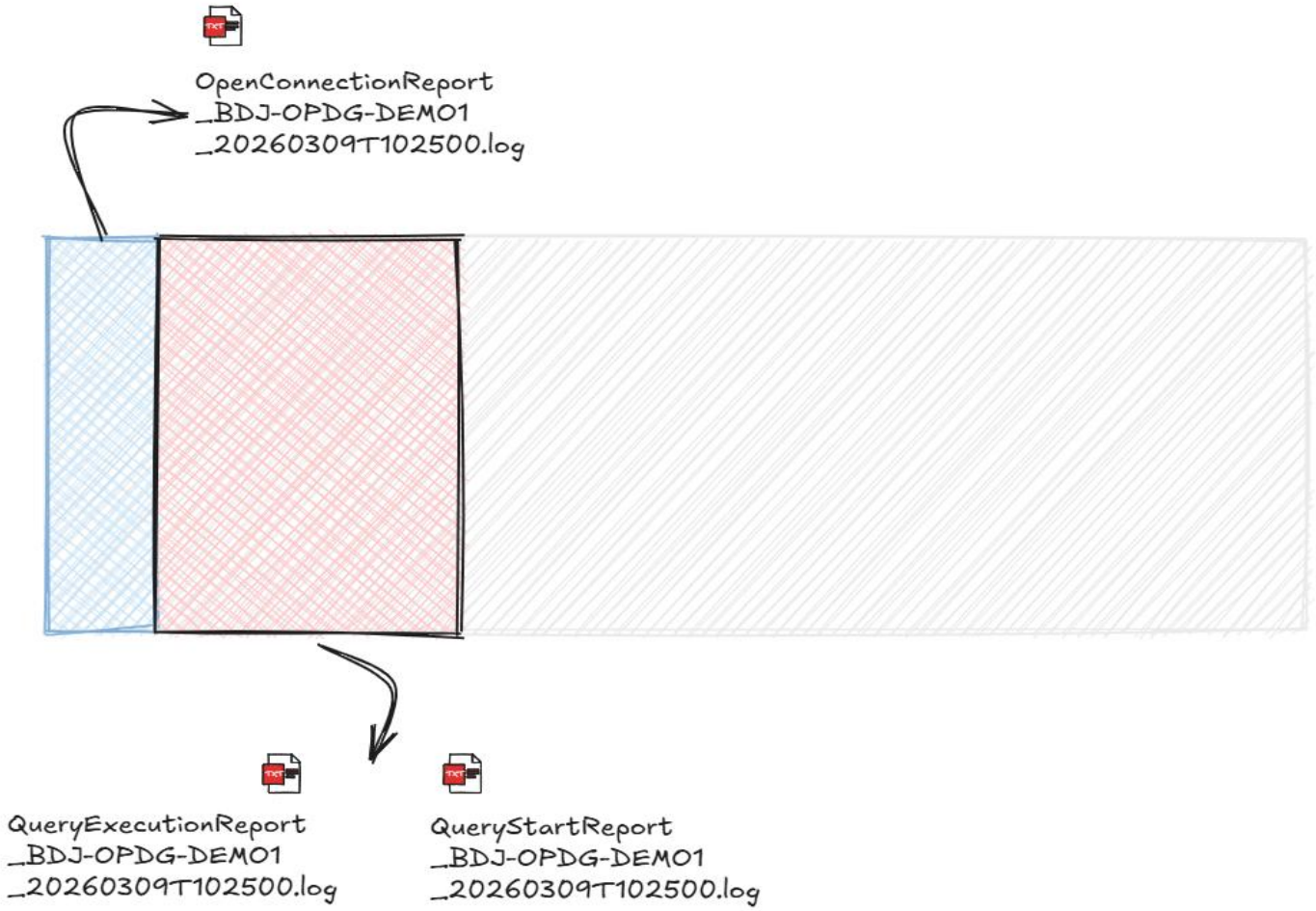
Process Flow (Cluster)

5. Create Connection

- Gateway decrypts credentials using a local key
- Creates a connection to data source, one per query

6. Execute query

- M Query is executed inside a Mashup container
- Query folding can happen at this stage



Process Flow (Cluster)

5. Create Connection

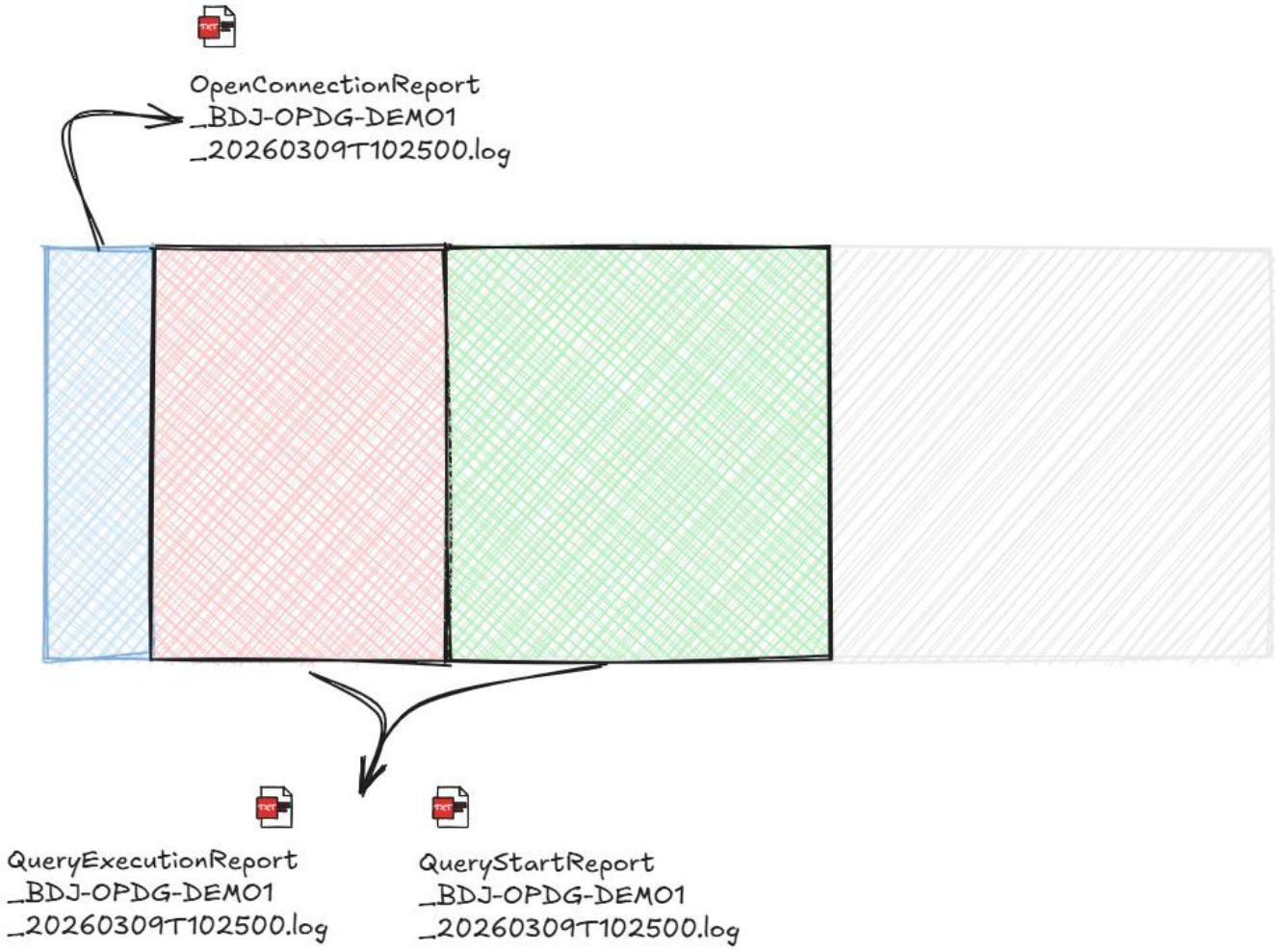
- Gateway decrypts credentials using a local key
- Creates a connection to data source, one per query

6. Execute query

- M Query is executed inside a Mashup container
- Query folding can happen at this stage

7. Process the extracted data

- All extracted data is saved (spooled) to disk



Process Flow (Cluster)

5. Create Connection

- Gateway decrypts credentials using a local key
- Creates a connection to data source, one per query

6. Execute query

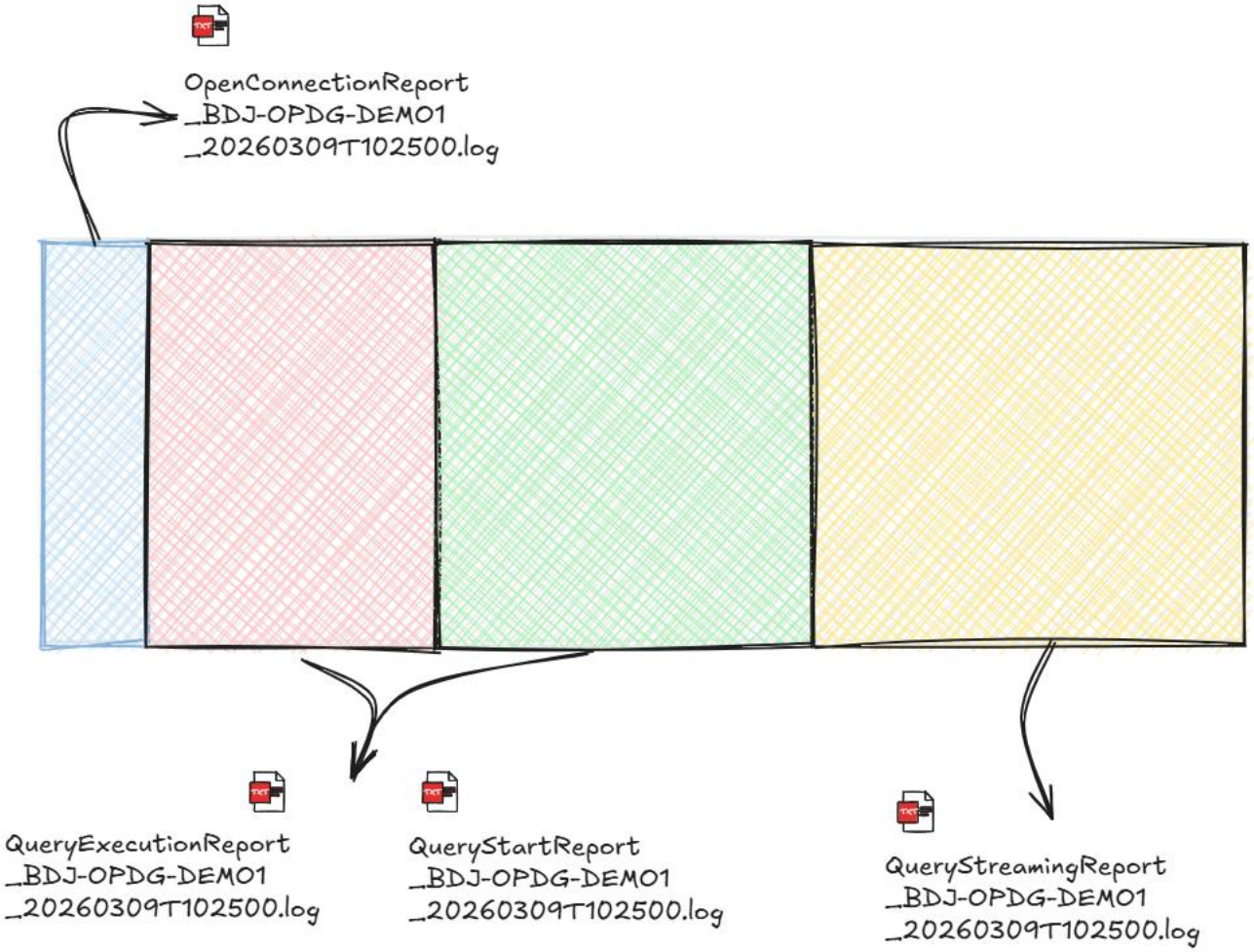
- M Query is executed inside a Mashup container
- Query folding can happen at this stage

7. Process the extracted data

- All extracted data is saved (spooled) to disk

8. Stream data back to Power BI Service

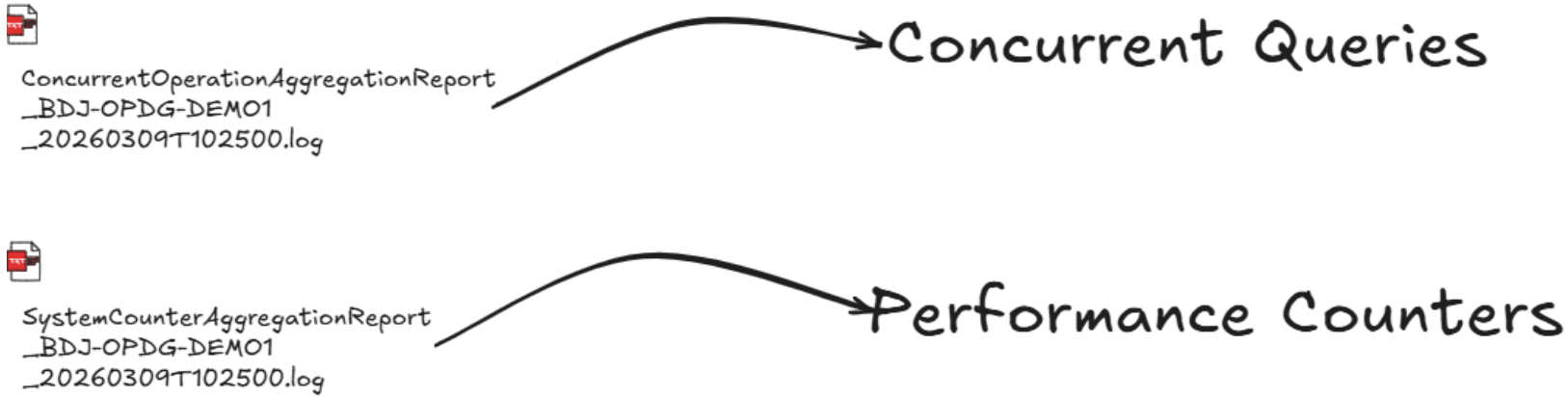
- Transferred, compressed data is streamed back
- Depending on settings, streaming might start early



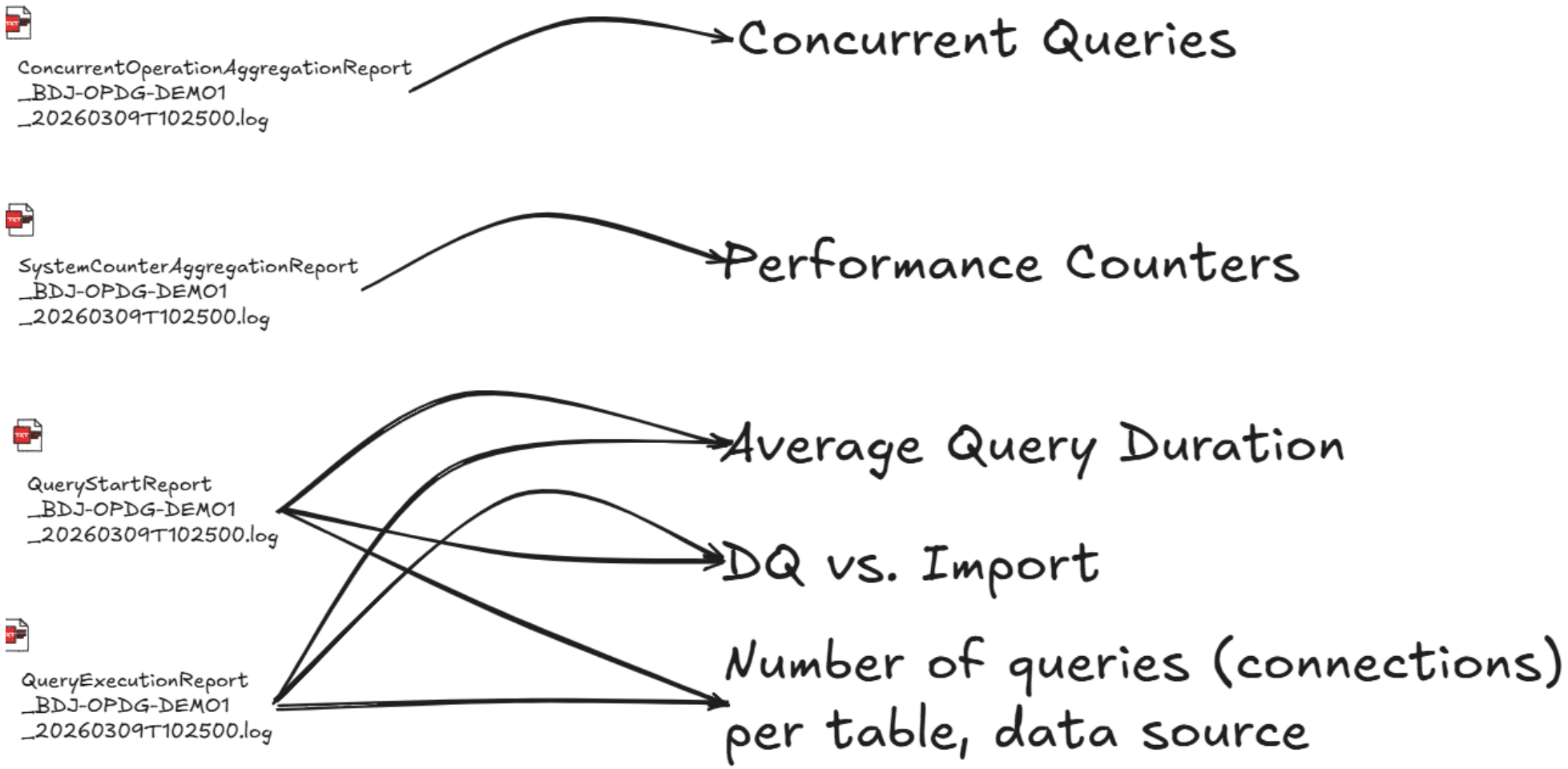
Is this thing on?
(Which logs do I interpret, and how?)

What to care about

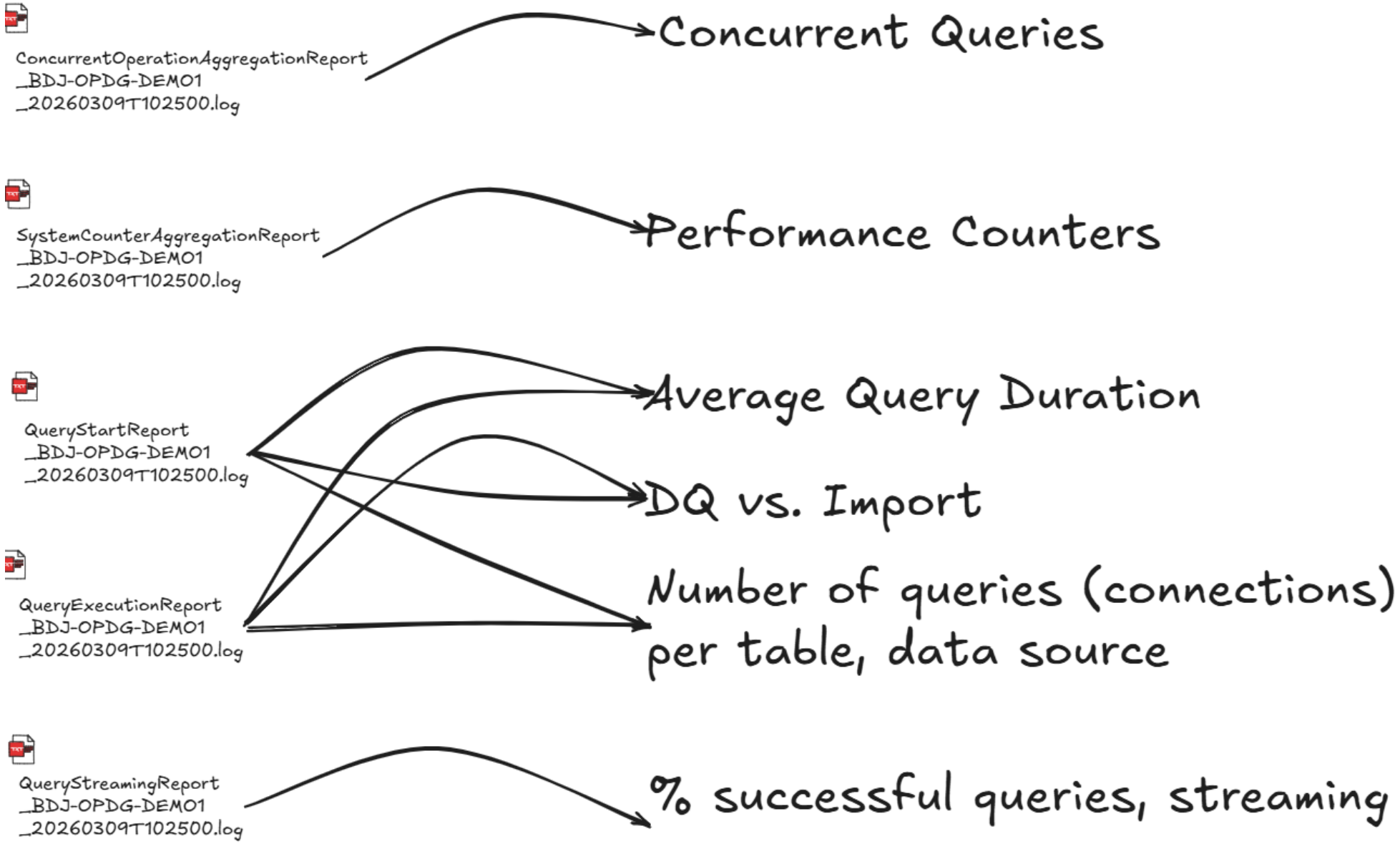
What to care about



What to care about



What to care about



Gateways are infrastructure too, yo.

Like all infrastructure, this needs to be monitored too

SystemCounterAggregationReport logs data for gateway processes

System Metrics, can be done through tool of choice (i.e. PerfMon)

- CPU, Memory, Network, Disk IO, Disk Size, ..
- Consider setting up alerts according to internal standards

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-log-files

Gateway Logs – GatewayInfo.log

[03/06/2026 10:15:42.123] [Info] [GatewayService] Service starting...
[03/06/2026 10:15:42.456] [Info] [Configuration] Loaded configuration from: C:\Program Files\On-premises data gateway\Microsoft.PowerBI.EnterpriseGateway.exe.config
[03/06/2026 10:15:42.789] [Info] [Network] Listening on port 443 for outbound connections.
[03/06/2026 10:15:43.012] [Info] [Auth] Successfully authenticated with Azure Service Bus.
[03/06/2026 10:15:43.345] [Info] [Cluster] Joined gateway cluster: "FinanceCluster01"
[03/06/2026 10:15:43.678] [Info] [DataSource] Registered data source: SQLServer01\PROD
[03/06/2026 10:15:44.001] [Info] [Heartbeat] Heartbeat sent to cloud service. Latency: 85ms
[03/06/2026 10:20:44.002] [Info] [Heartbeat] Heartbeat sent to cloud service. Latency: 82ms
[03/06/2026 10:25:44.003] [Info] [Query] Executed query for dataset: SalesAnalytics, Duration: 1.23s
[03/06/2026 10:30:44.004] [Info] [Heartbeat] Heartbeat sent to cloud service. Latency: 80ms

Gateway Logs – GatewayErrors.log

[03/06/2026 14:22:15.487] [Error] [ActivityID: 12345678-abcd-4ef0-9123-abcdef123456]
Source: MashupEngine
Message: Failed to connect to SQL Server database.
Details:
Server: mydbserver.database.windows.net
Database: SalesDB
ErrorCode: -2146232060
ExceptionType: System.Data.SqlClient.SqlException
ExceptionMessage: Login failed for user 'readonly_user'.
StackTrace:
at System.Data.SqlClient.SqlInternalConnection.OnError(SqlException exception, Boolean breakConnection)
at System.Data.SqlClient.TdsParser.ThrowExceptionAndWarning(TdsParserStateObject stateObj, Boolean callerHasConnectionLock, Boolean asyncClose)

Gateway Logs – GatewayNetwork.log

[2026-03-06 10:15:42.123] [INFO] Connection established to region: westeurope
[2026-03-06 10:15:42.456] [DEBUG] Sending heartbeat to service endpoint: https://service.powerbi.com
[2026-03-06 10:15:43.002] [INFO] Data channel opened (SessionId: 8f2a1c3b-4d5e-4a9f-9b2e-1c4f7a8d9e2f)
[2026-03-06 10:16:01.789] [WARN] Network latency detected: 452 ms
[2026-03-06 10:16:05.321] [ERROR] Failed to send packet (ErrorCode: 0x80072EE7 - Name resolution failure)
[2026-03-06 10:16:07.654] [INFO] Retrying connection to endpoint: https://service.powerbi.com
[2026-03-06 10:16:09.210] [INFO] Connection re-established successfully
[2026-03-06 10:16:15.998] [DEBUG] Received 2048 bytes from cloud service
[2026-03-06 10:16:20.432] [INFO] Session closed (SessionId: 8f2a1c3b-4d5e-4a9f-9b2e-1c4f7a8d9e2f)

Log Retention

Controlled by *ApplicationFileTraceListener* property

February 2023 introduced Age Based Retention

- **# Days Retention**, default 30 days
- **Maximum disk space** for combined log files, default 5 GB.
- **Maximum File Size**, default 100MB

Existing gateways leverage File Based Retention, unless changed

- **Number of files** to be retained per log type
- **Maximum disk space** to be consumed per log type

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-log-files

Config Options

2

An on-premises data gateway cluster
is a group of gateway instances
installed on separate machines to provide
high availability, fault tolerance, and load balancing
for secure data connectivity
between on-premises sources and cloud services.

Leveraging Gateway Clusters

Clusters can contain up to 10 nodes

Intended for High Availability, segregation of load

By default:

- Send all requests to Primary, until no longer available (active)
- Even when primary is overwhelmed, secondary remains on standby
- Retry failed request(s), send to secondary
- Rinse, repeat

UI based - *Distribute Requests across active gateways*

- *PowerShell – Set-DataGatewayCluster -LoadBalancingSelectorType*
- Determined at random, no decision whether nodes are overwhelmed or not

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-high-availability-clusters

Resource-based distribution

Requests are sent to nodes based on *round robin*

- No factual resource-based load balancing

Config based option for *resource-based distribution*

- **CPUUtilizationPercentageThreshold** (0..100, default 0 = disabled)
- **MemoryUtilizationPercentageThreshold** - (0..100, default 0 = disabled)
- **ResourceUtilizationAggregationTimeInMinutes** – (default 5, minutes)

Nodes will accept requests until a threshold is met

- Accepted queries can still go over configured threshold

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-high-availability-clusters

Spooling

Queried results are *materialized* to temporary file(s)

Default location catches many by surprise (C:\..)

C:\Users\PBIEgwService\AppData\Local\Microsoft\On-premises data gateway\Spooler

Disk IO (and virus scanners..) can be significant bottleneck

Reduce data that needs to be spooled by:

- Query Folding
- Data Reduction (Select what you need)

Refer to *QueryExecutionReport* for:

- **SpoolingDiskWritingDuration(ms)**: time to write all data to disk
- **SpoolingDiskReadingDuration(ms)**: time to read all data from disk
- **SpoolingTotalDataSize(byte)**: Compressed size of all data written/read

Persistent Cache still maps to %TMP location for a user

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-performance#understand-performance-logs
https://medium.com/@demerzel/power-bi-data-gateway-spooler-management-d8c32fad2d36

StreamRequestBeforeCompletes

Rather than spooling, results are streamed back to Power BI Service

In common cases, reducing the overall time taken

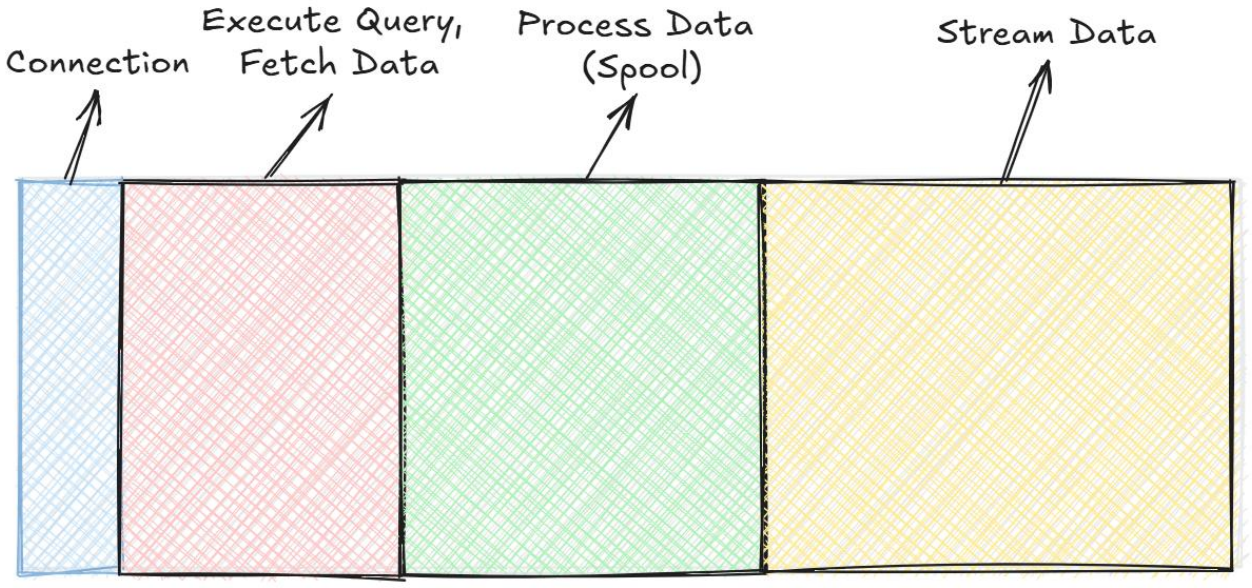
Queries will not be retried when:

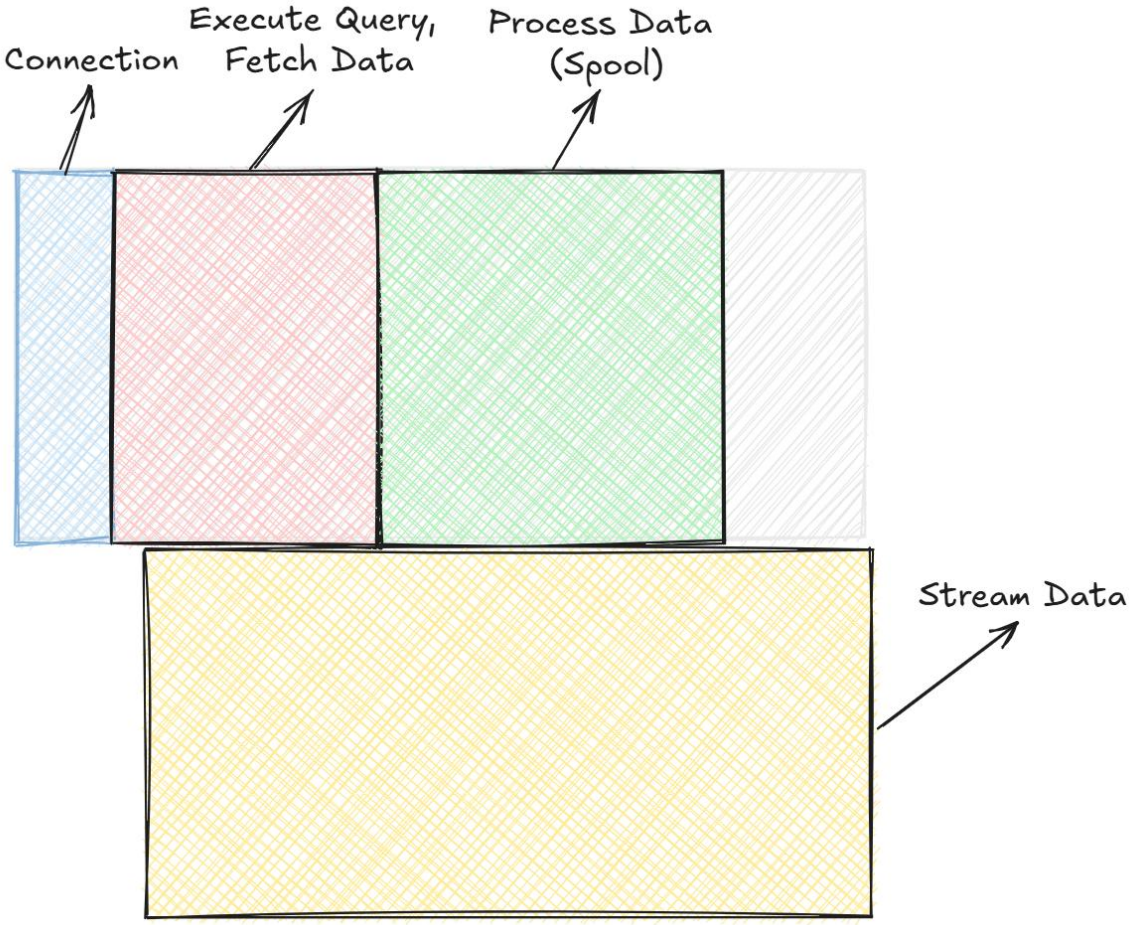
- Queries fail
- Queries are interrupted
- Queries time out

This is not a silver bullet solution

- Test, Benchmark, Validate to evaluate whether it makes sense for you

StreamRequestBeforeCompletes





Where to Install a gateway

Physical location of the gateway makes a difference

Proximity to data source(s) should be prioritized

- Gateway needs a *line of sight* to data sources

Balance with:

- Power BI Tenant Home Region
- Power BI Capacity region
- Azure Relay (maps to tenant home region)

When different region is needed

- Configure the Azure Relay (BYO Relay)

Admin Decisions

2

Who can install a gateway?

By default, all users can install and configure a gateway

- Also bound by permissions on the machine

Restrict who can install gateways:

- Power Platform Admin Center (Manage Gateway Installers)
- Manage Connections and Gateways, Tenant Administration mode
- PowerShell cmdlets (Set-DataGatewayTenantPolicy, Set-DataGatewayInstaller)

Restrict which tenant(s) a gateway can connect to

- Registry Key **AllowedRegistrationTenants**

https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-tenant-registration
https://learn.microsoft.com/en-us/power-query/dataflows/data-exfiltration-best-practices
https://learn.microsoft.com/en-us/data-integration/gateway/data-connection-auditing

Private Connectivity

Architecture

Gateway Cloud Service

Azure Hosted Service that routes data requests between Power BI and gateway clusters

Azure Relay

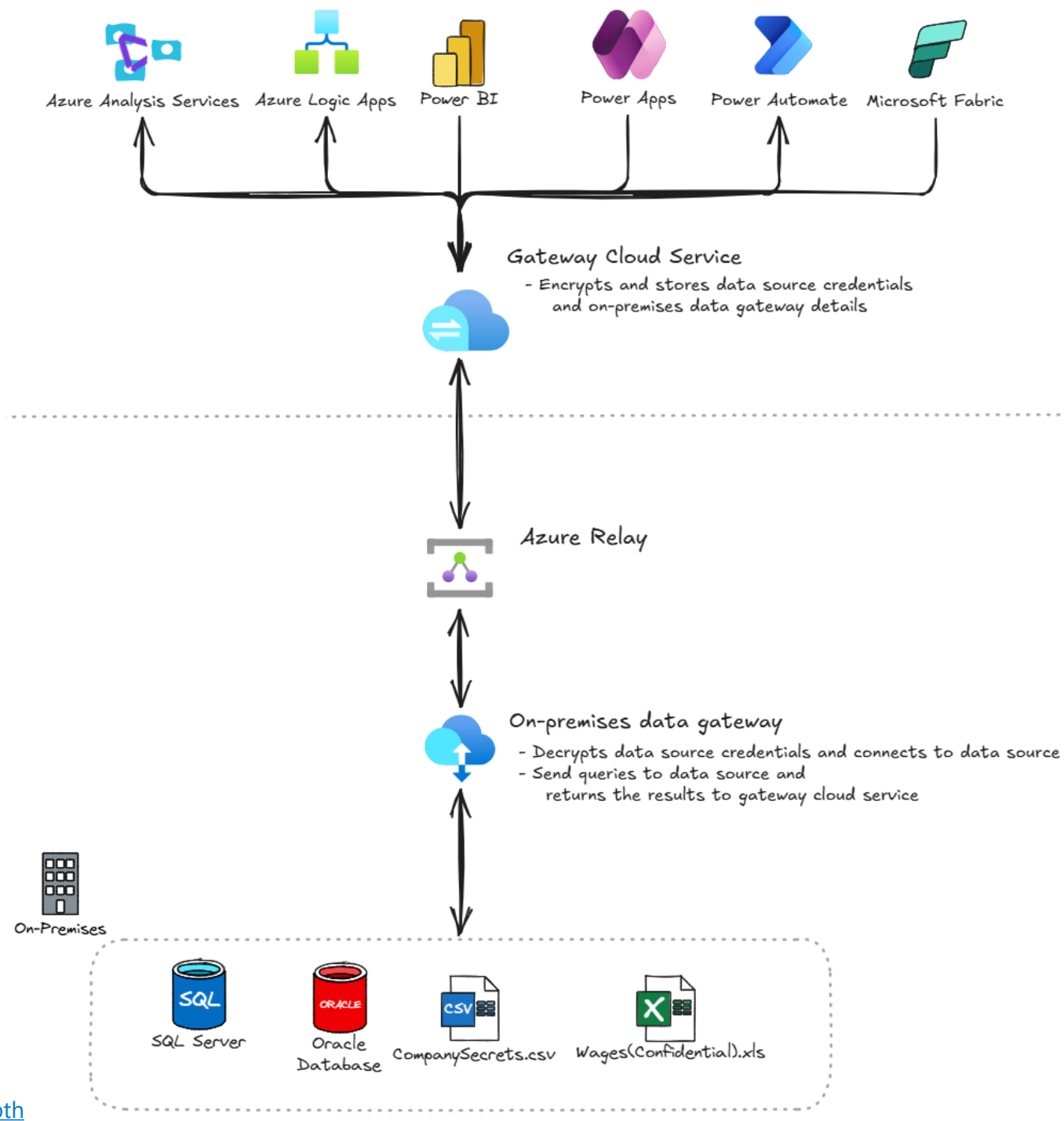
Azure Service that enables cloud services to exchange traffic with on-premises network, only through outbound firewall ports

Gateway Cluster

Group of gateways (nodes) working together to ensure high availability, and/or load balancing

On-premises data gateway

Software installed on corporate machines, enabling data transfer between isolated sources and Power BI Service



https://learn.microsoft.com/en-us/data-integration/gateway/service-gateway-onprem-indepth

Private Connectivity

Private Links are not supported for

- on-premises data gateway
- Azure Relay

Data from Azure Relay to Gateway Cloud Service uses public internet

- TLS 1.2 or above is used to encrypt data

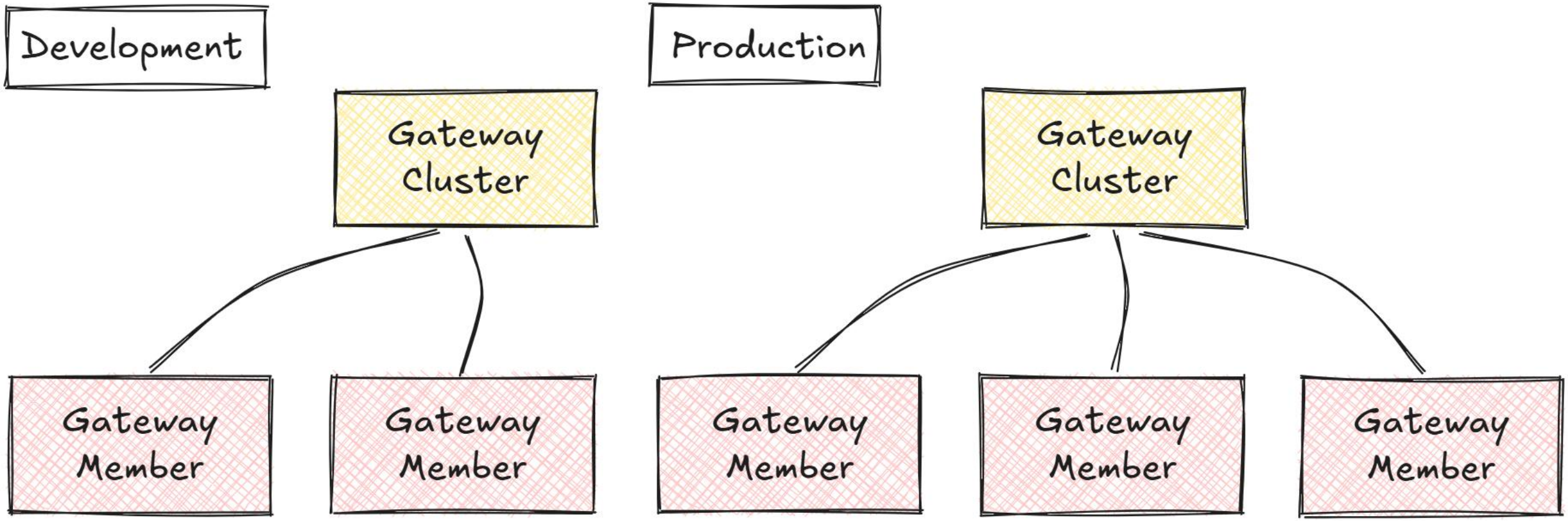
All Azure internal traffic goes over the Azure backbone (docs).

Alternatives to leverage for Private Connectivity:

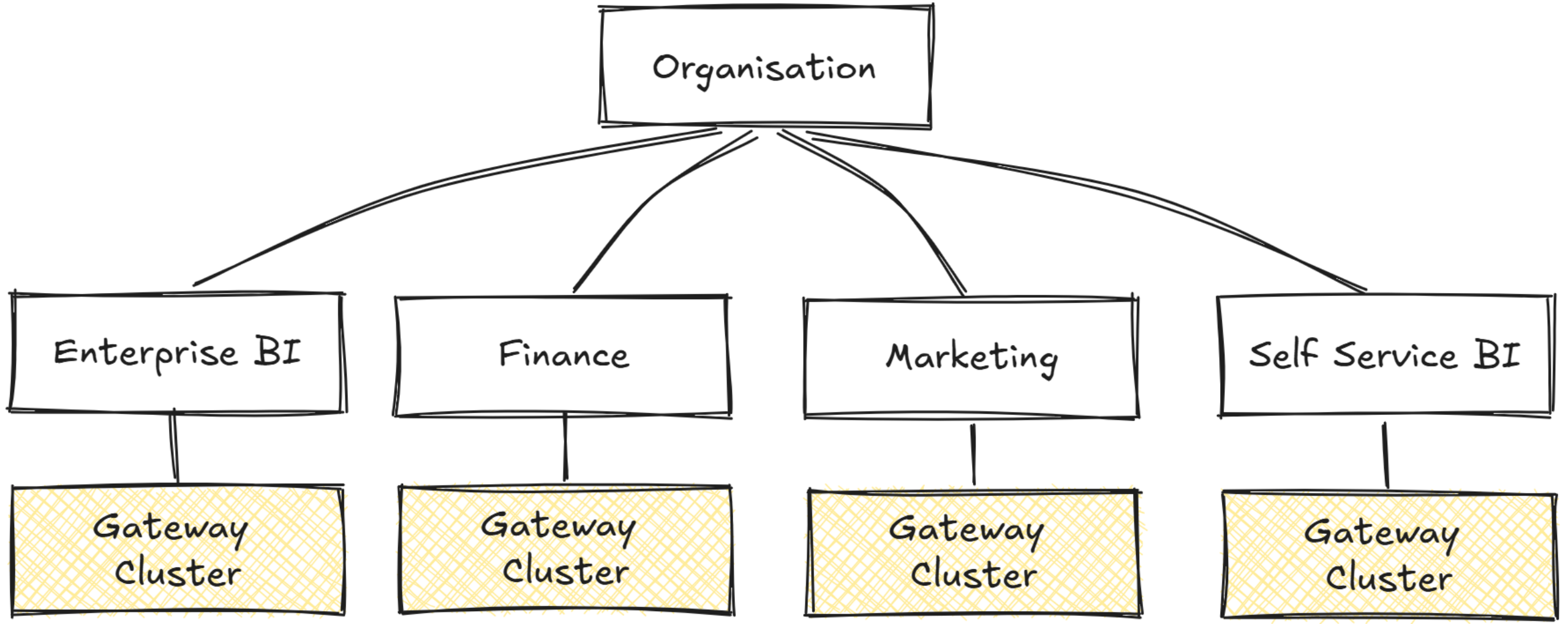
- VNET Data Gateway
- Azure ExpressRoute

Planning Decisions

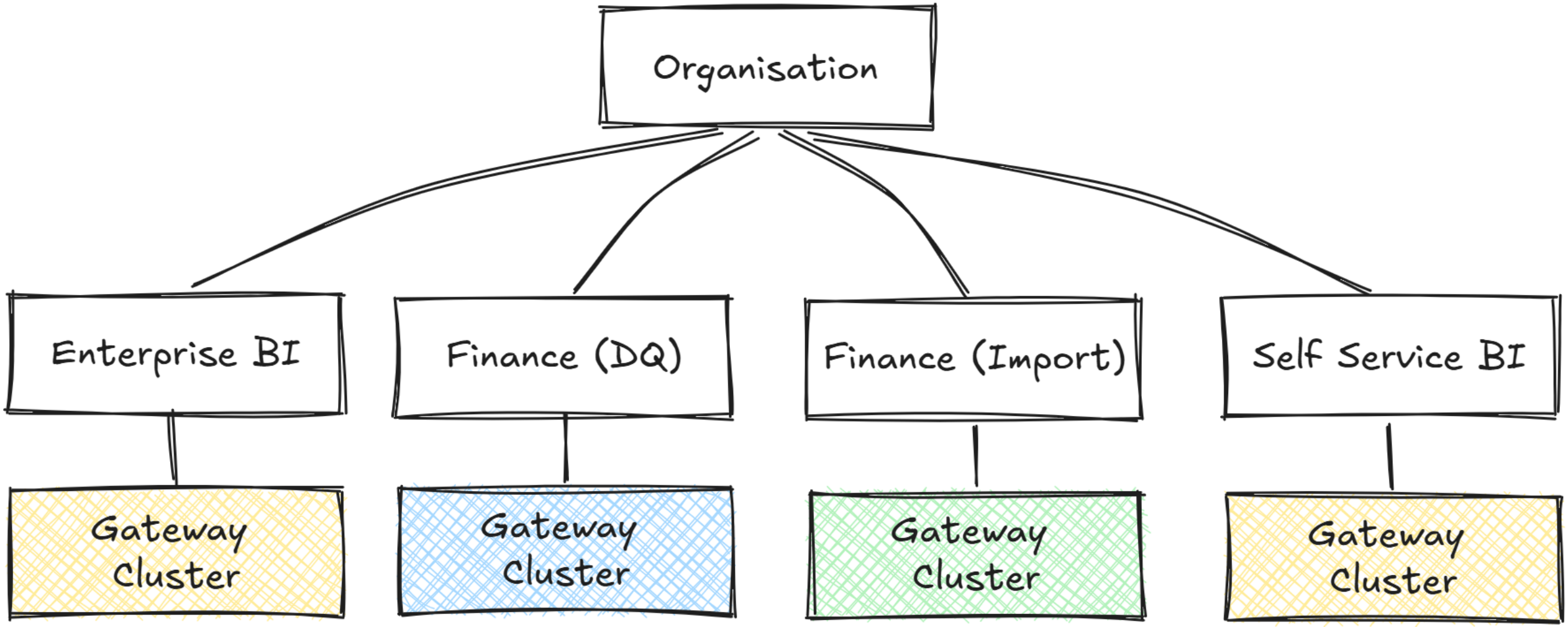
Patterns – By environment



Patterns – By department



Patterns – By environment, workload type



Recommendations for mission-critical workloads

Always make sure to have all recovery keys stored in a safe place

High availability and **query distribution** make a difference

Understand how workloads are influenced most by resource type

- Import: Memory, Disk, CPU
- DQ: CPU, Disk, Memory
- LiveConnection: CPU, Disk, Memory

Leverage resource monitoring, act on extended resource scarcity

- **CPU** ~80%
- **Memory** ~80%
- **Disk** ~5GB available
- **Concurrent queries** = 40

Understand when to scale up, out, or create new clusters

Will it fold? is the existential question that needs to be asked

https://learn.microsoft.com/en-us/data-integration/gateway/plan-scale-maintain

On-premises Data Gateways are not for Cloud sources, right?

Let’s show an example ☺

https://techcommunity.microsoft.com/blog/analyticsonazure/part-2-performance-configurations-for-connecting-pbi-to-a-private-link-adb-works/4396060


Fin.

Loved it?
Learned something?
Tell us!
Share your feedback
in just 1 minute

Demystifying the on-premises data gateway (for Power BI Semantic Models)

https://github.com/BenniDeJagere/Presentations/{Year}/{YYYYMMDD}_{Event}

PBIG
POWER BI GEBRUIKERSGROEP