SQL
KONFERENZ
2025

Integrating API Sources into your Data Lake

ADF, Synapse and Microsoft Fabric Pipelines

John Martin

Technology Partner and Alliances Manager

Redgate software

Nearly twenty years of experience working with data platform and cloud technologies. Working with SQL Server, Azure, AWS, Snowflake and Databricks to deliver OLTP and Analytics solutions.

in /in/johnqmartin

/johnmart82

https://blog.jqmartin.co.uk

john@jqmartin.co.uk

SQL KONFERENZ 2025

Why APIs are important to the data engineer

- Data Sources
- Sinks to store data
- Management APIs for services

API Data Sources

- SaaS platforms are prominent in the way that busineses buy line of business applications today.
- Different API options available.
  - REST & GraphQL typically
- Inconsistency between applications for how to interact with their API.
  - Authentication, Pagination, Rate Limits, etc.

SQL KONFERENZ 2025

Data Source APIs - Salesforce

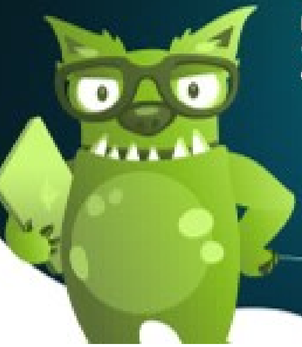
- 47 Different APIs that can be used to interact with the platform.
  - REST, SOAP, GraphQL, SOQL/SOSL, Bulk API 2.0, Metadata, Streaming
- Bulk API
  - Designed for large volumes ~ 2,000 records or more.
  - Asynchronous processing of SOQL queries.
- Explore using a combination of APIs to achieve an outcome.
  - Use the Metadata API to get column list for SOQL queries to avoid having to manage column lists as metadata.
- Leverage built-in connectors where possible to simplify the data extraction.

Data Source APIs - Xero

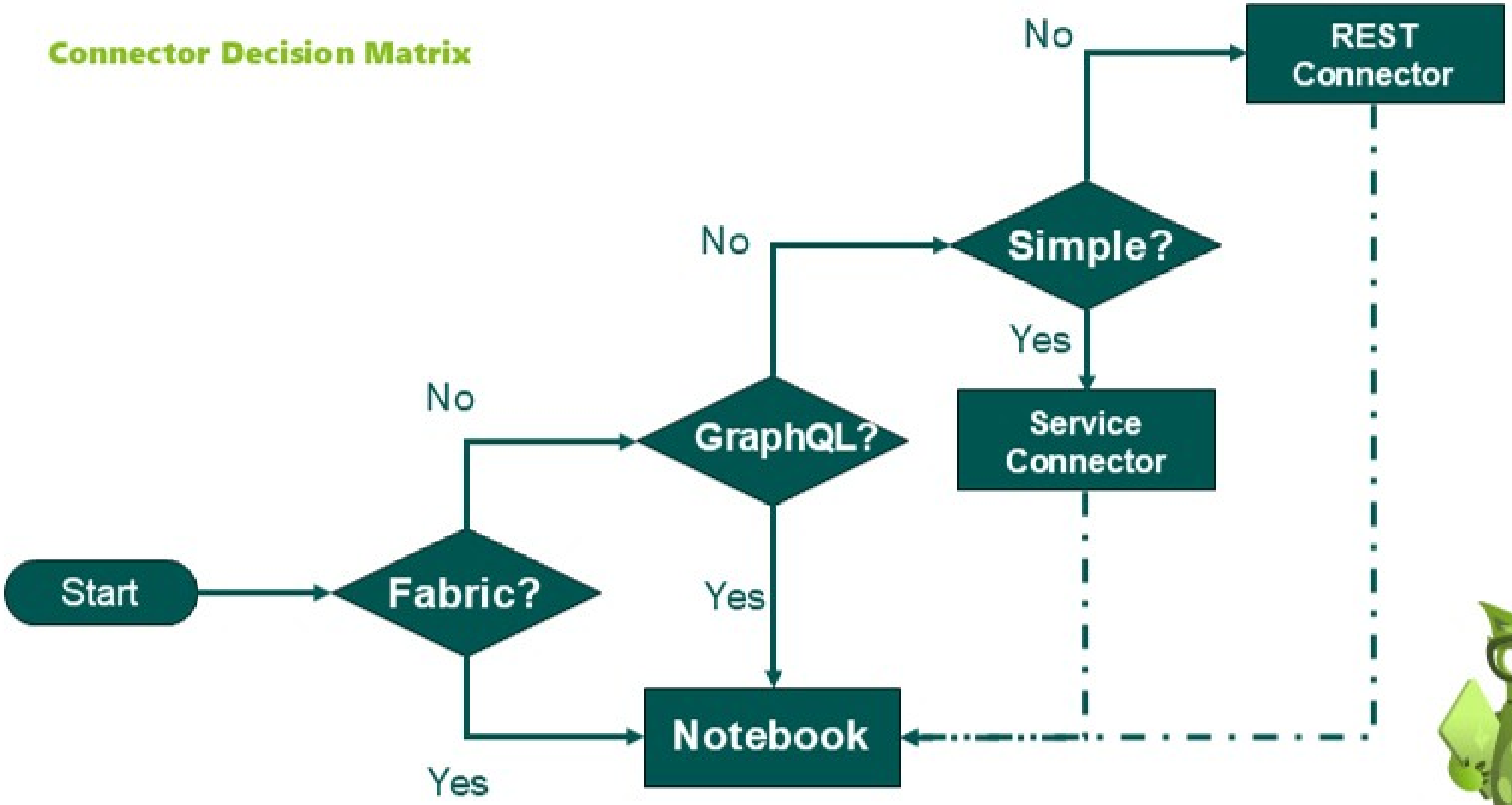
- Group of REST APIs focused on the parts of the product that organisations use.
    - Mainly geared towards application integrations not data engineering
- API Limts are quite low.
    - Up to 5 concurrent calls, 50 calls per-minute, maximum of 5,000 calls a day, 10 mb per-response
    - Limits are sent back in headers for all API responses
- Additional Costs for setting up app integrations.
- Token lifecycle management is difficult.
    - Access token lifespan is 30 minutes, refreshtoken is 60 days

Data Source APIs - GitHub

- REST and GraphQL APIs available for use.
- API Limits vary depending on factors such as authenticated Vs. not-authenticated.
  - 100 Concurrent requests, 1,000 requests per-hour/repo for standard users, 15,000 per-hour/repo for Enterprise Cloud users
  - GitHub OAuth apps have higher rate limits Vs. tokens
- REST API can be quite verbose and lacks filtering in areas, GraphQL can be more efficient but is more complex.



Connector Decision Matrix



Start

Fabric?

Yes

No

GraphQL?

Yes

No

Simple?

Yes

No

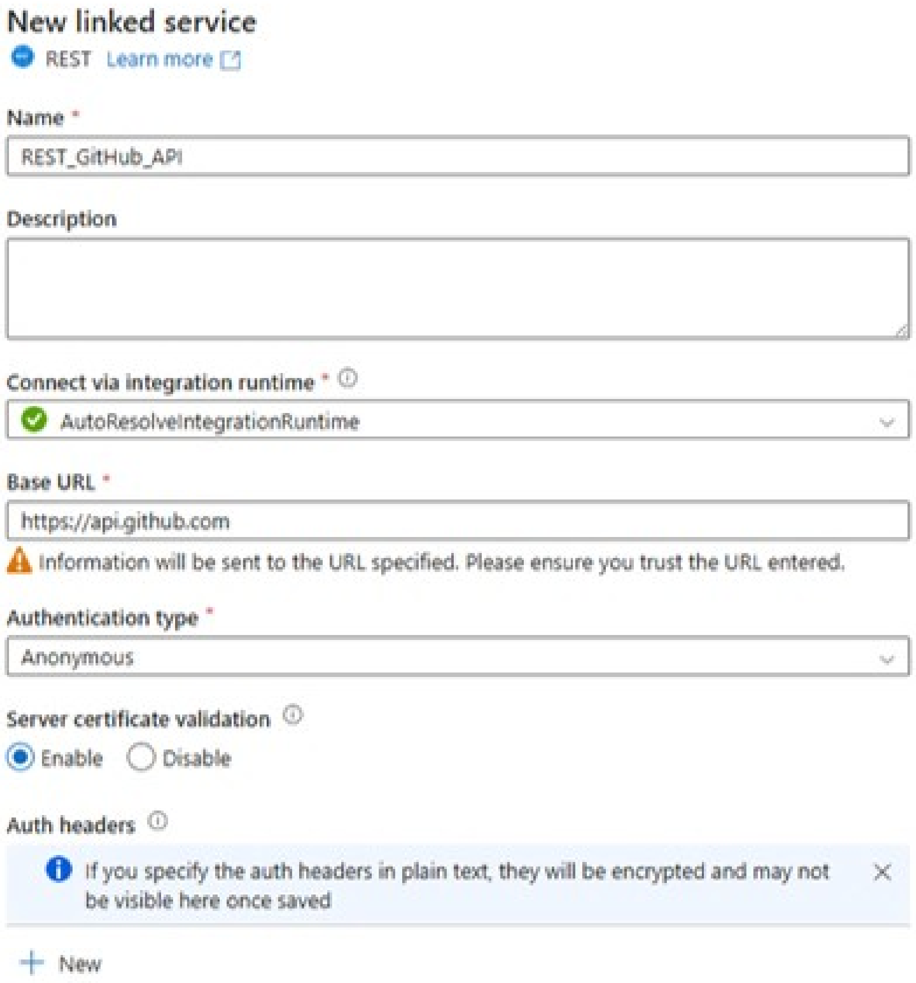
Notebook

Service Connector

REST Connector

API Data Source Principles

- Keep as generic as possible and parameterise, look at setting values at runtime.



New linked service

REST Learn more

Name *

REST_GitHub_API

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Base URL *

https://api.github.com

⚠ Information will be sent to the URL specified. Please ensure you trust the URL entered.

Authentication type *

Anonymous

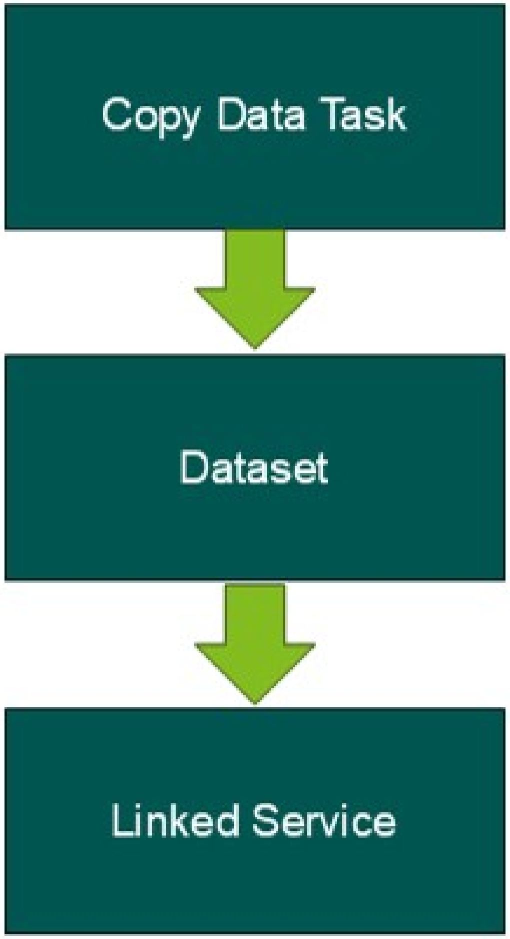
Server certificate validation

● Enable ○ Disable

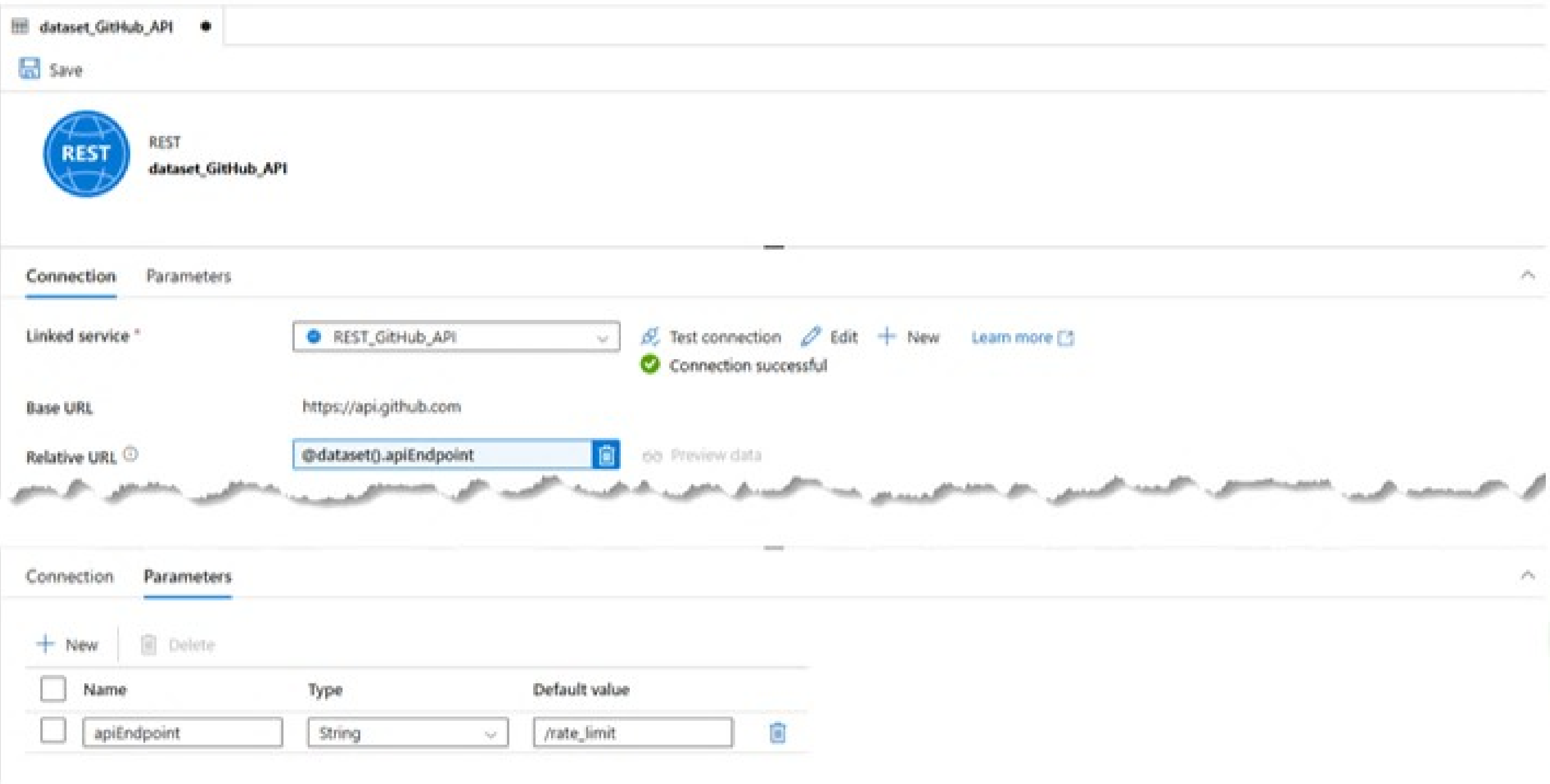
Auth headers

ⓘ If you specify the auth headers in plain text, they will be encrypted and may not be visible here once saved

+ New



SQL KONFERENZ 2025



dataset_GitHub_API ●

Save

REST

dataset_GitHub_API

Linked service *: REST_GitHub_API

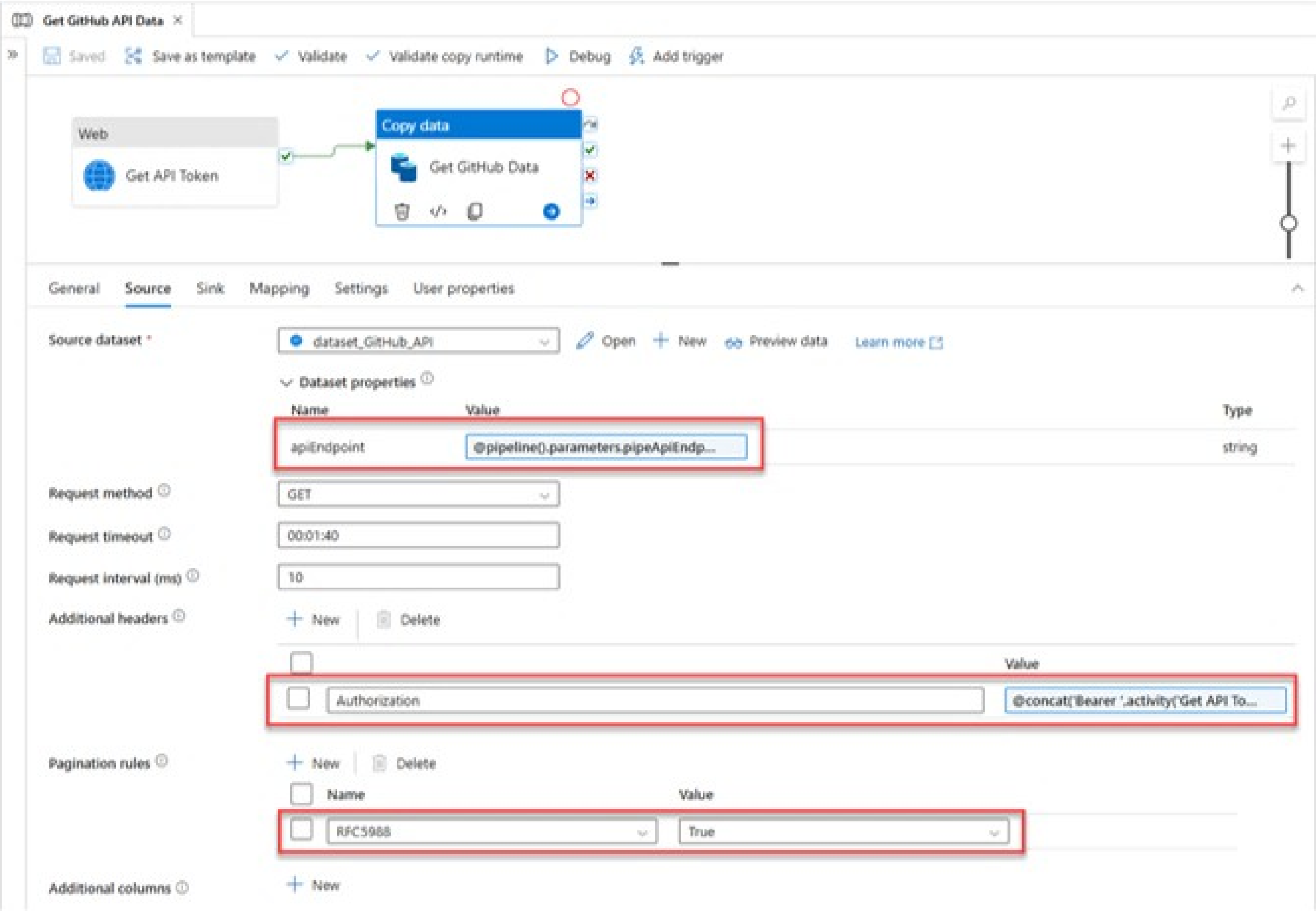
Connection successful

Base URL: https://api.github.com

Relative URL: @dataset().apiEndpoint

□	Name	Type	Default value
□	apiEndpoint	String	/rate_limit

SQL KONFERENZ 2025



Source dataset * dataset_GitHub_API

Dataset properties

Name	Value	Type
apiEndpoint	@pipeline().parameters.pipeApiEndp...	string

Request method: GET

Request timeout: 00:01:40

Request interval (ms): 10

Additional headers

□		Value
□	Authorization	@concat('Bearer ',activity('Get API To...

Pagination rules

□	Name	Value
□	RFC5988	True

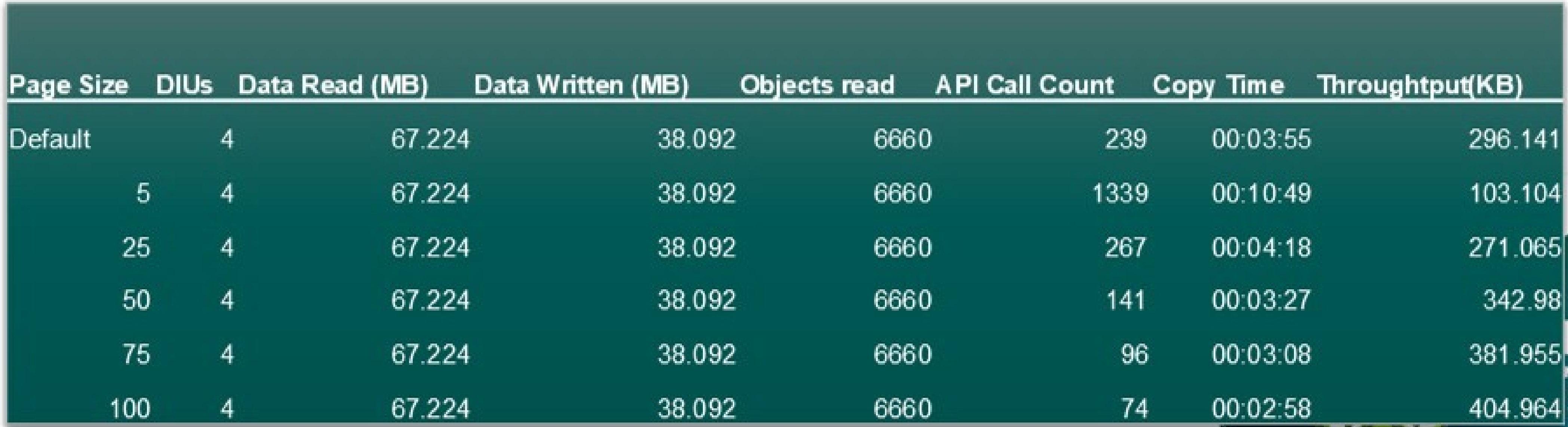
Additional columns

SQL KONFERENZ 2025

API Data Source Principles

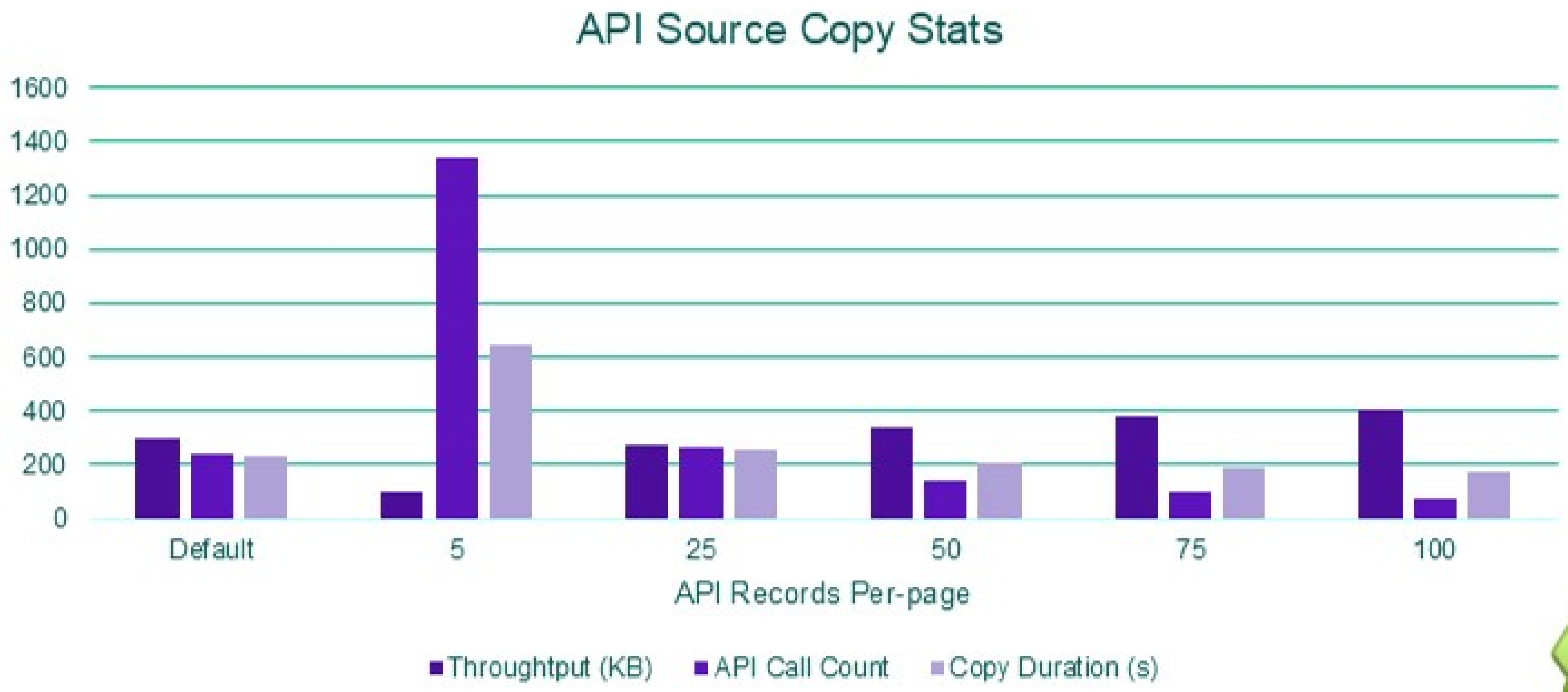
- Keep as generic as possible and parameterise, look at setting values at runtime.
- Think about iterator placement.
  - Minimise the number of activities which are executed with each iteration
- API Configuration Options.
  - Page Size can impact API call count and data size

Impact of API Page Size



Page Size	DIUs	Data Read (MB)	Data Written (MB)	Objects read	API Call Count	Copy Time	Throughtput(KB)
Default	4	67.224	38.092	6660	239	00:03:55	296.141
5	4	67.224	38.092	6660	1339	00:10:49	103.104
25	4	67.224	38.092	6660	267	00:04:18	271.065
50	4	67.224	38.092	6660	141	00:03:27	342.98
75	4	67.224	38.092	6660	96	00:03:08	381.955
100	4	67.224	38.092	6660	74	00:02:58	404.964

Impact of API Page Size



REST Data Sinks

- Data is sent to the API endpoint in a JSON Array, not an object.
- Similar to the REST source.
  - Reduced method set; PUT, PATCH, POST
  - Paramterise and make generic where possible
- Write Batch Size will determine number of calls to the API when sending data.
- There are several service specific linked servcies/connections which can be used if needed.

SQL
KONFERENZ
2025

Management APIs

- Platform Management is largely API driven now.
- Determine demarkation point between wher the platform ends and the data engineering begins.
  - Use internal Source Control features or manage externally
  - RAW API Calls or Terraform

SQL
KONFERENZ
2025

Management APIs

- Azure Data Factory
  - Manage Azure resources with IaC, objects internally under Git
- Azure Syanpse Analytics
  - Manage Azure resources with IaC, objects internally under Git
- Microsoft fabric
  - Still figuring this out...

John Q. Martin

Thanks for your time and feedback

SQL
KONFERENZ
2025