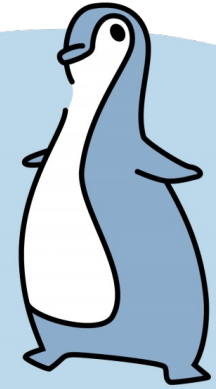


**SCALE**

**19x**



# DevOps for Data as a Service

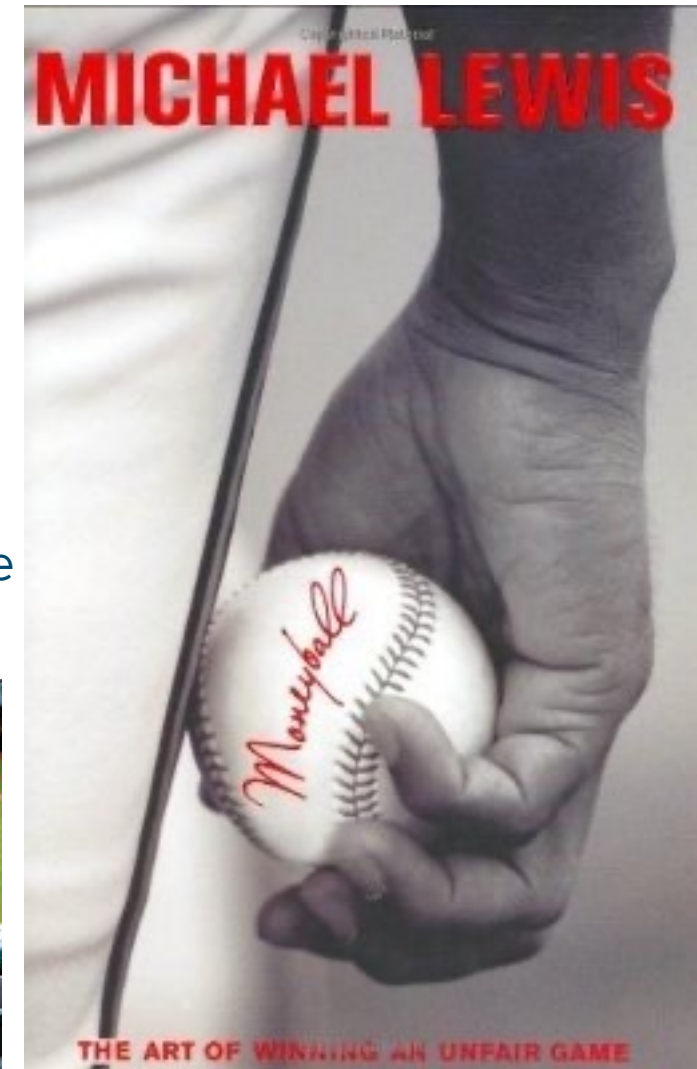
Antoni Ivanov

A lead maintainer of Versatile Data Kit 

30.07.2022

*“People in both fields operate with beliefs and biases.  
To the extent you can eliminate both and replace  
them with **data**, you gain a clear advantage.”*

Michael Lewis, Moneyball: The Art of Winning an Unfair Game



# Sneak Peek

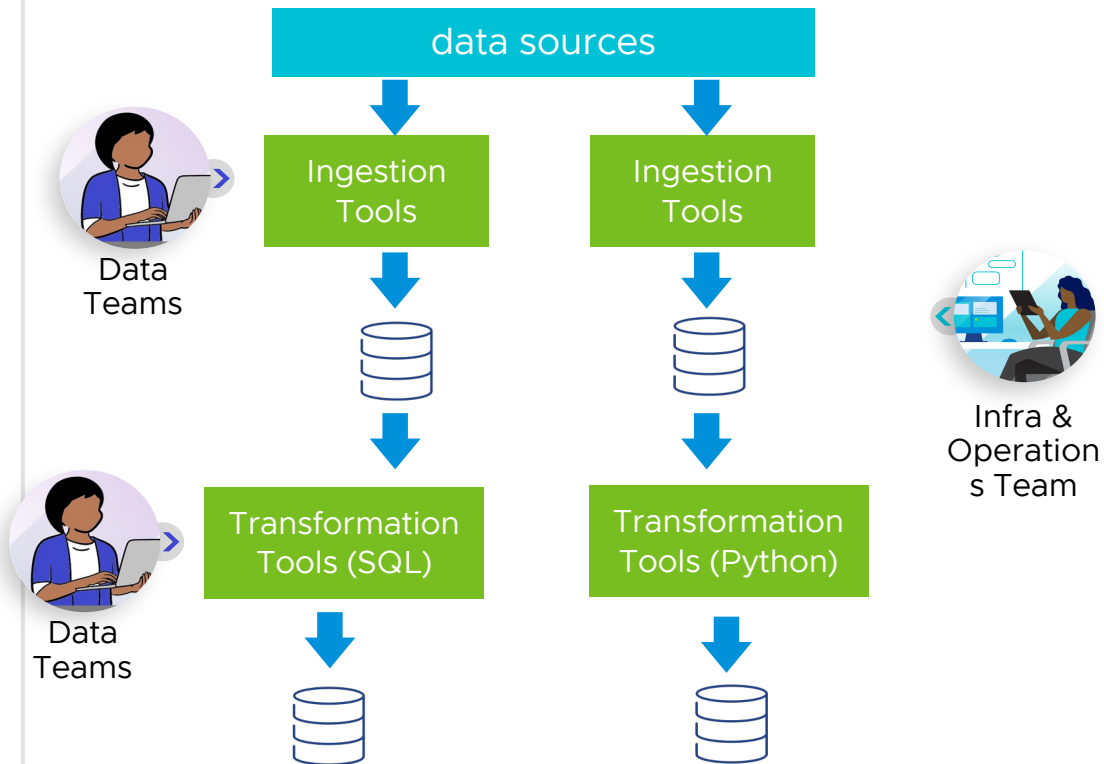
Enable everyone to focus on work that require their core skills



Github repo

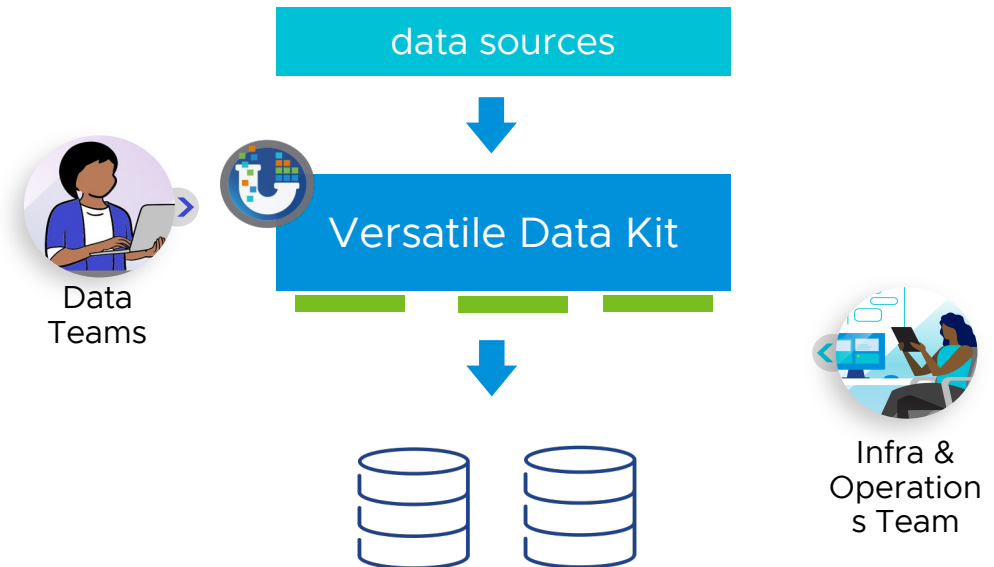
## Without Versatile Data Kit

- Fragmented Infrastructure
- Organization Silos
- Infra/Ops & Data Team tension



## With Versatile Data Kit

- Easier platform maintenance
- Self-service, fully automated data teams
- Improved collaboration



DevOps for Data: Why ?

DevOps for Data as a Service

Deliver analytics platform for your business quickly (demo)

Improve data infrastructure security (demo)

Improve data infrastructure stability (demo)

A high-angle photograph of two people sitting at a long, light-colored wooden table in a modern office or co-working space. The person on the left is wearing a plaid shirt and is seen from the back, looking at a laptop. The person on the right is wearing a dark jacket and is looking at a tablet. On the table are several items: a laptop, a tablet, a pair of headphones, a green cup, and a can of soda. The background is a plain white wall.

Agenda



# DevOps for Data: Why ?

DevOps for Data as a Service

Deliver analytics platform for your business quickly (demo)

Improve data infrastructure security (demo)

Improve data infrastructure stability (demo)



# Why do we care about DevOps for Data ?

Infra/Ops Team and Data Teams tension and conflict

*Inefficient Operations    Stalled development.*

Domain knowledge

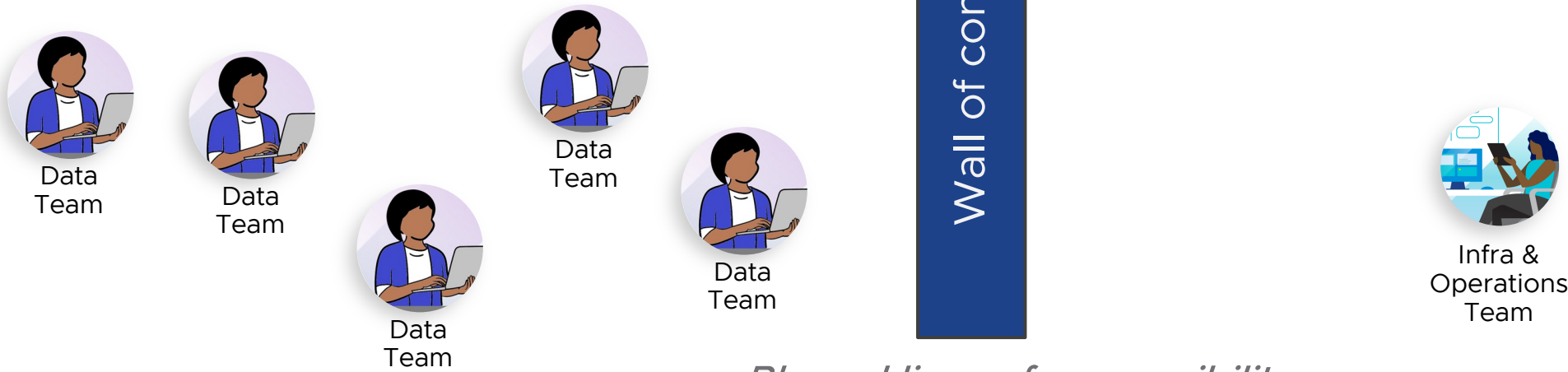
Implement business logic

Optimizes for agility and speed

DevOps & Infrastructure knowledge

Maintain infrastructure

Optimizes reliability, availability and security



*Blurred lines of responsibility*

DevOps for Data: Why ?

## DevOps for Data as a Service

Deliver analytics platform for your business quickly (demo)

Improve data infrastructure security (demo)

Improve data infrastructure stability (demo)

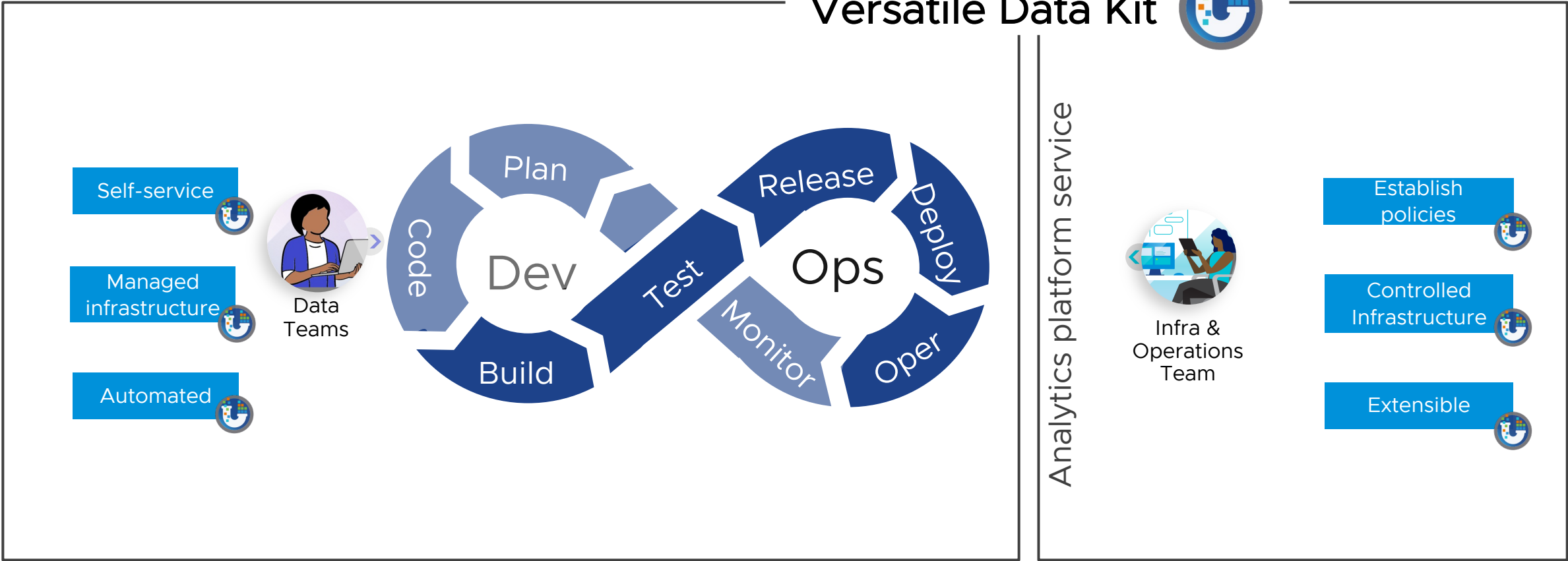




# DevOps for Data as a Service

Adopt and adapt DevOps to deliver value from data efficiently

## Versatile Data Kit



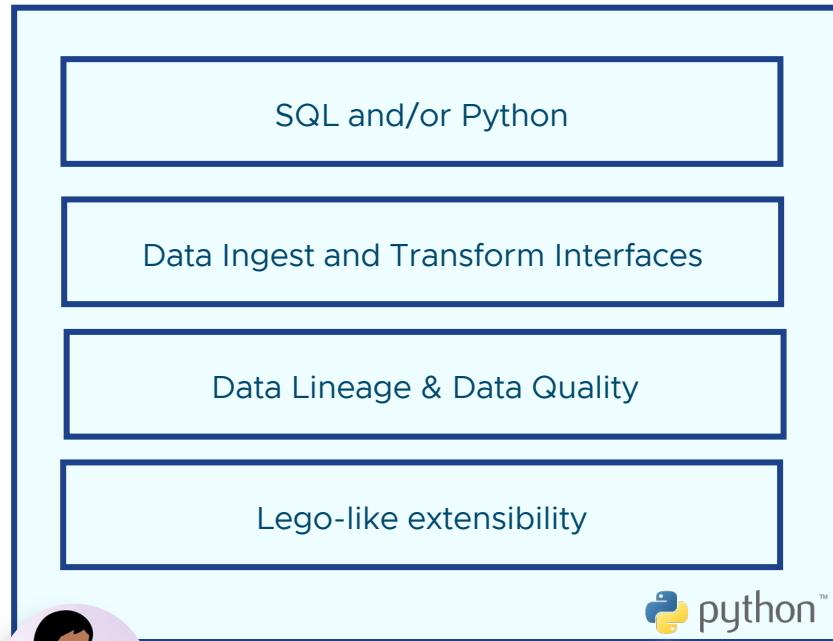




# Components

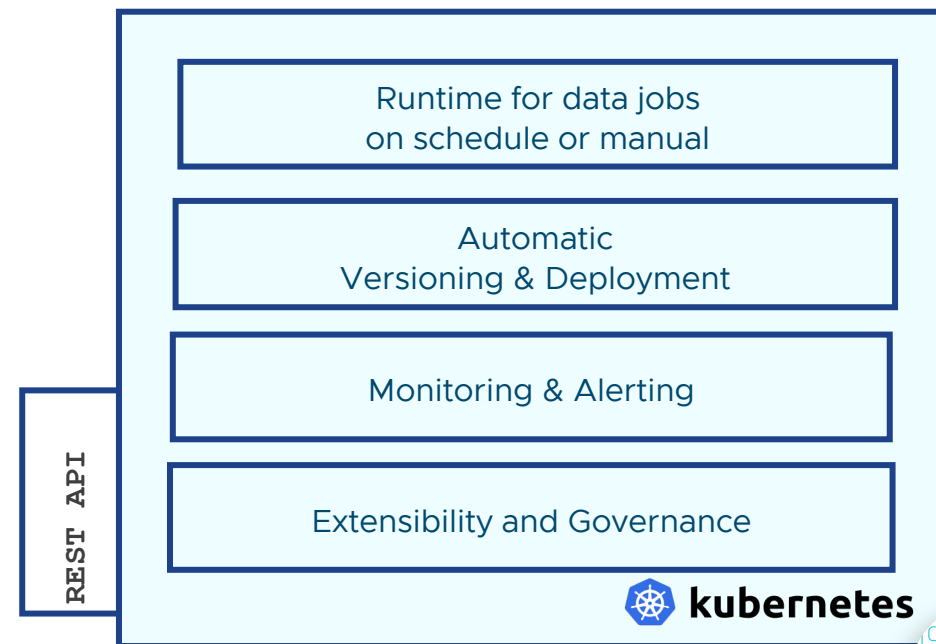
What does it take to run Versatile Data Kit and start deploying data jobs?

## Versatile Data Kit Data SDK



Data Teams

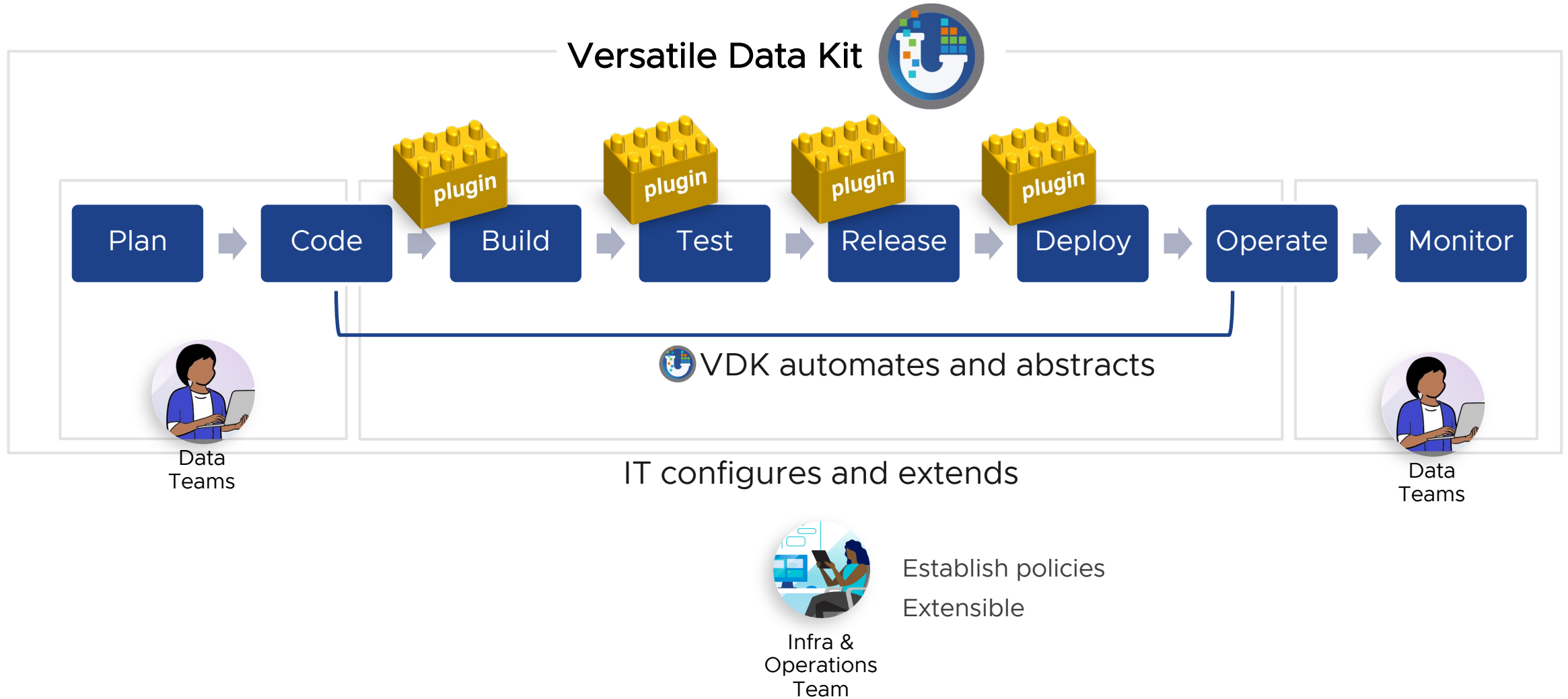
## Versatile Data Kit Control Service



Infra & Operations Team

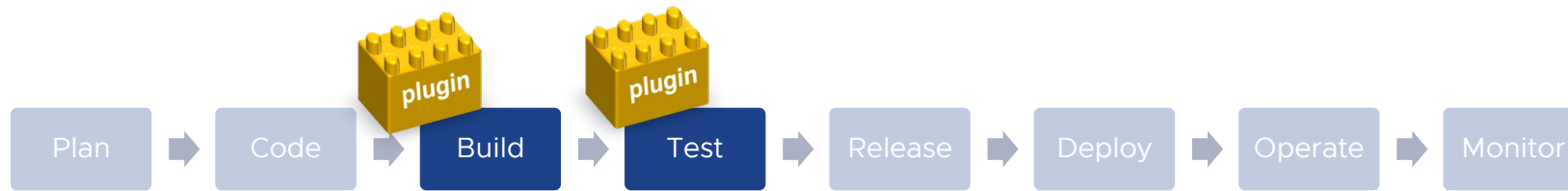
# Automate and Abstract the Development Process

Give power to Operators to establish best dev practices; Ease data job development



## Quick example: DevOps Plugin

Establish standard system tests and security hardening



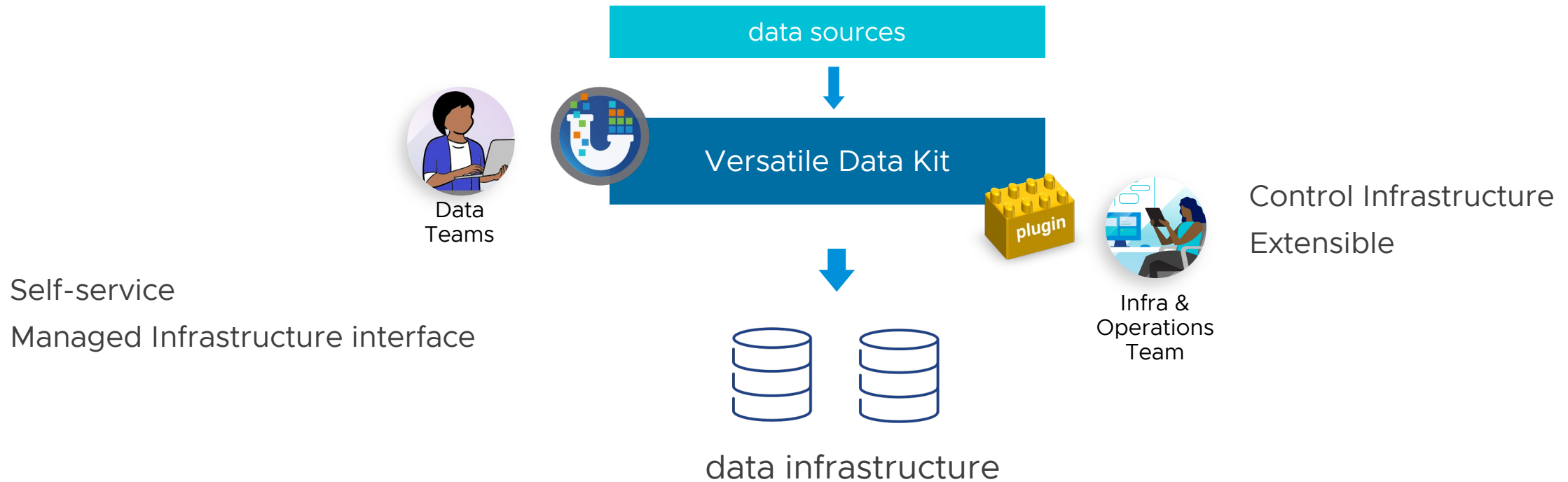
```
helm install --set job-builder=my-job-builder-image
```

```
2  >> FROM versatiledatakit/job-builder
3
4  # Run system test before accepting the new job code
5  RUN pytest system_test.py || die 'Failed system test'
6
7  # Remove execution privileges from files during container build
8  RUN chmod -R -x $job_name/
9
```

# Automate and Abstract the Data journey

Simplify and hide complexity of data infrastructure for Data teams

Give power to establish best Infrastructure practices



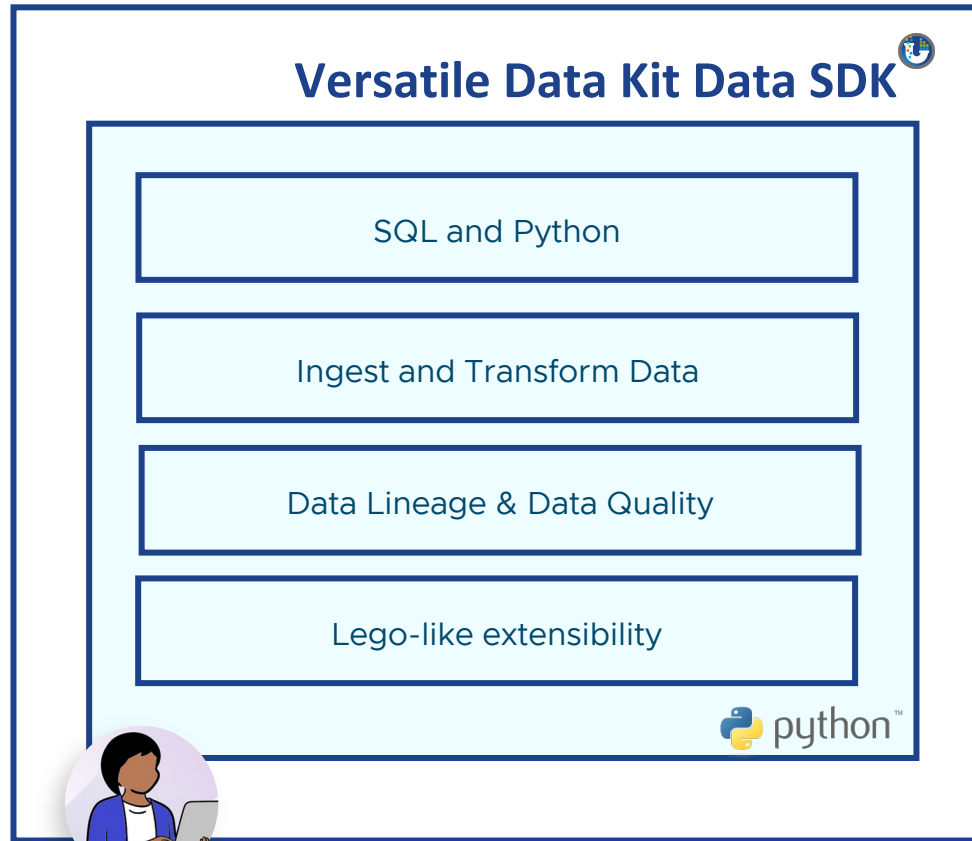




# Components

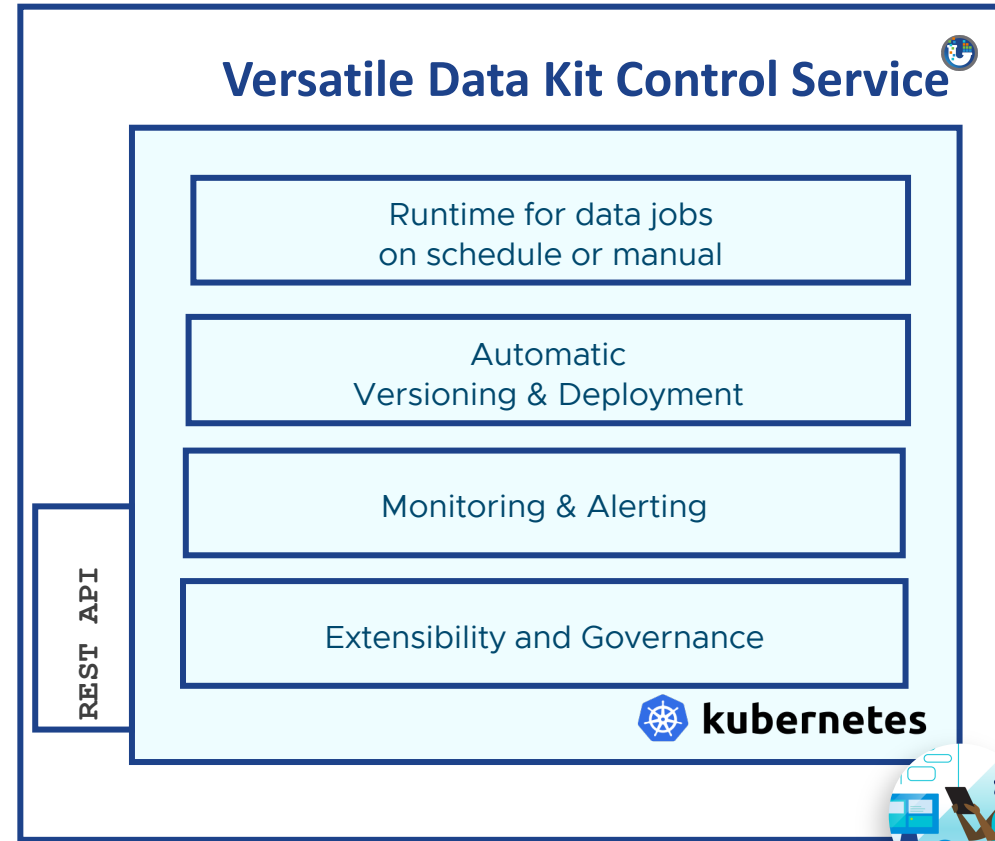
What does it take to run Versatile Data Kit and start deploying data jobs?

*Automate and abstract the Data Journey*



Data  
Teams

*Automate and abstract the DevOps Cycle*



Infra &  
Operations  
Team

DevOps for Data: Why ?

DevOps for Data as a Service

**Deliver analytics platform for your business quickly (demo)**

Improve data infrastructure security (demo)

Improve data infrastructure stability (demo)

Check out:



<https://github.com/vmware/versatile-data-kit/wiki/Install-VDK-Control-Service-with-custom-SDK>

# Key takeaways

## Day 1 Operations Simplified



Infra & Operations  
Team

Deliver data job as a service for users

Integrate with your existing infrastructure with configuration only

Cloud-native deployment on the Infrastructure of your choice



DevOps for Data: Why ?

DevOps for Data as a Service

Deliver analytics platform for your business quickly (demo)

**Improve data infrastructure security (demo)**

Improve data infrastructure stability (demo)

Check out:



<https://github.com/vmware/versatile-data-kit/tree/main/examples/ingest-and-anonymize-plugin>





# Key Takeaways

## Day 2 Operations Simplified



Infra & Operations  
Team

Enforce data security & governance policies across all jobs of all teams

Control over what happens throughout the whole data path

No changes required on the data engineering side

DevOps for Data: Why ?

DevOps for Data as a Service

Deliver analytics platform for your business quickly (demo)

Improve data infrastructure security (demo)

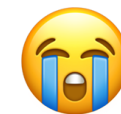
**Improve data infrastructure stability (demo)**



# What are we going to do?



```
1  INSERT INTO tableName (sddc_sk,active_from,active_to,sddc_id,updated_by_user_id,s
  •  '500'),(sddc_sk,active_from,active_to,sddc_id,updated_by_user_id,state,is_nsxt,cl
2  ....
3
  •  '2', 'RUNNING', 'TRUE', 'AWS', '497'),('sddc03-v01', '2.01.19', '3.01.19', '3',
5208 ,('sddc01-v01', '1.01.19', '2.01.19', '1', '9', 'STOPPED', 'FALSE', 'AWS', '500
  •  '2', 'RUNNING', 'TRUE', 'AWS', '497'),('sddc03-v01', '2.01.19', '3.01.19', '3',
5209 ,('sddc01-v01', '1.01.19', '2.01.19', '1', '9', 'STOPPED', 'FALSE', 'AWS', '500
  •  '2', 'RUNNING', 'TRUE', 'AWS', '497'),('sddc03-v01', '2.01.19', '3.01.19', '3',
```



????

# What is the problem

vdk run sql-job      cursor.execute(...)      vdk query -q "..."

```
select
    count(1) as uploads,
    trunc(arrival_ts, 'ww') week
from org
```

↓ intercepted



vdk-query-validation (plugin)





EXPLORER

EXAMPLE

- 10\_sql\_step.sql
- 20\_python\_step.py
- config.ini
- README.md
- requirements.txt

> OUTLINE

> TIMELINE

> SONARLINT RULES

> SONARLINT ISSUE LOCATIONS

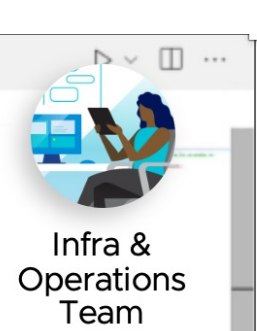
config.ini

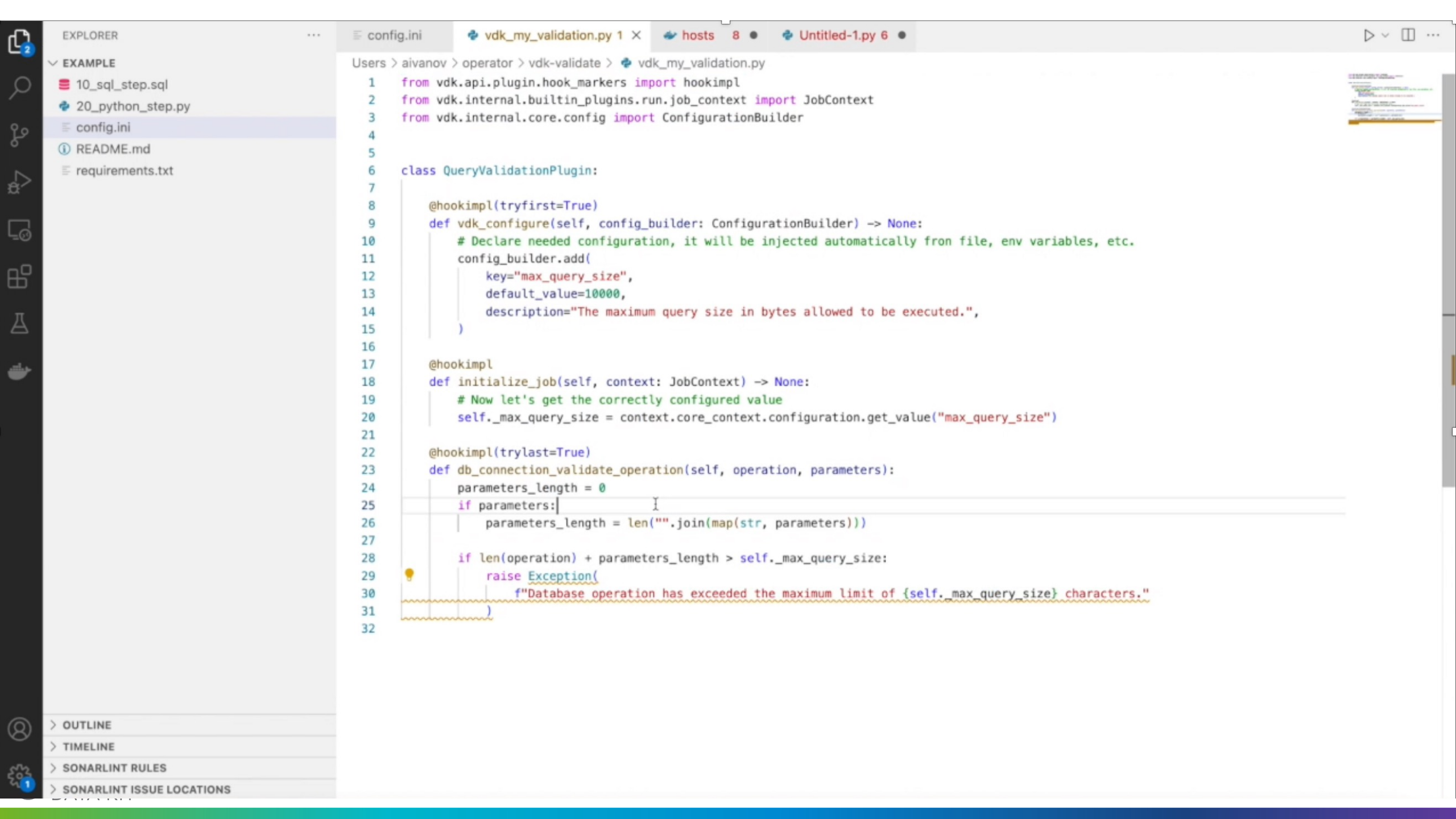
vdk\_my\_validation.py

hosts 8

Untitled-1.py 6

```
Users > aivanov > operator > vdk-validate > vdk_my_validation.py
1  from vdk.api.plugin.hook_markers import hookimpl
2  from vdk.internal.builtin_plugins.run.job_context import JobContext
3  from vdk.internal.core.config import ConfigurationBuilder
4
5
6  class QueryValidationPlugin:
7
8      @hookimpl(trystfirst=True)
9      def vdk_configure(self, config_builder: ConfigurationBuilder) -> None:
10         # Declare needed configuration, it will be injected automatically from file, env variables, etc.
11         config_builder.add(
12             key="max_query_size",
13             default_value=10000,
14             description="The maximum query size in bytes allowed to be executed.",
15         )
```





```
1  setuptools.setup(  
2      name="vdk-my-validation",  
3      version="1.0",  
4      packages=setuptools.find_namespace_packages(),  
5      entry_points={  
6          "vdk.plugin.run": [  
7              "vdk-my-validation-plugin = vdk_my_validation",  
8          ]  
9      }  
10 )  
11
```

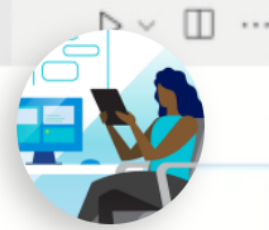
> Find

Aa ab \*

No results



Infra &  
Operations  
Team



Infra &  
Operations  
Team

Users > aivanov > operator > my-org-vdk > setup.py

```
1  import setuptools
2
3  setuptools.setup(
4      name="my-org-vdk",
5      version="1.0",
6      install_requires=[
7          "vdk-core",
8          "vdk-plugin-control-cli",
9          "vdk-postgres",
10         "vdk-snowflake",
11         "vdk-ingest-http",
12         "vdk-ingest-file",
13         "vdk-my-validation"
14     ]
15 )
```



```
(demo) {15:14}~/data-engineer ⇒ vdk run example
```



Data  
Team

Data  
Team

```
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/builtin_plugins/run_job_input.py, line 56, in run_sql_step
    job_input.execute_query(sql)
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/builtin_plugins/run_job_input.py, line 115, in execute_query
    return connection.execute_query(query)
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/builtin_plugins/connection/managed_connection_base.py, line 116, in execute_query
    cur.execute(query)
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/builtin_plugins/connection/managed_cursor.py, line 90, in execute
    errors.log_and_rethrow(
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/core/errors.py, line 379, in log_and_rethrow
    raise exception
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/vdk/internal/builtin_plugins/connection/managed_cursor.py, line 86, in execute
    self.__connection_hook_spec.db_connection_validate_operation(
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/hooks.py, line 286, in __call__
    return self._hookexec(self, self.get_hookimpls(), kwargs)
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/manager.py, line 93, in _hookexec
    return self._inner_hookexec(hook, methods, kwargs)
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/manager.py, line 84, in <lambda>
    self._inner_hookexec = lambda hook, methods, kwargs: hook.multicall(
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/callers.py, line 208, in _multicall
    return outcome.get_result()
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/callers.py, line 80, in get_result
    raise ex[1].with_traceback(ex[2])
File /Users/aivanov/.pyenv/versions/3.9.1/envs/quickstart-vdk37/lib/python3.9/site-packages/pluggy/callers.py, line 187, in _multicall
    res = hook_impl.function(*args)
File /Users/aivanov/operator/vdk-validate/vdk_my_validation.py, line 37, in db_connection_validate_operation
    raise Exception(
Exception: Database operation has exceeded the maximum limit of 10 characters.
```

# What we did

vdk run sql-job    cursor.execute(...)    vdk query -q "..."

```
select
    count(1) as uploads,
    trunc(arrival_ts, 'ww') week
from org
```

↓ intercepted



vdk-query-validation (plugin)



# Key Takeaways

## Day 2 Operations Simplified



Infra & Operations  
Team

Ensure stability of the infrastructure and provide better service

Ensure data workflow stability and enforce best practices

Do not sacrifice user satisfaction

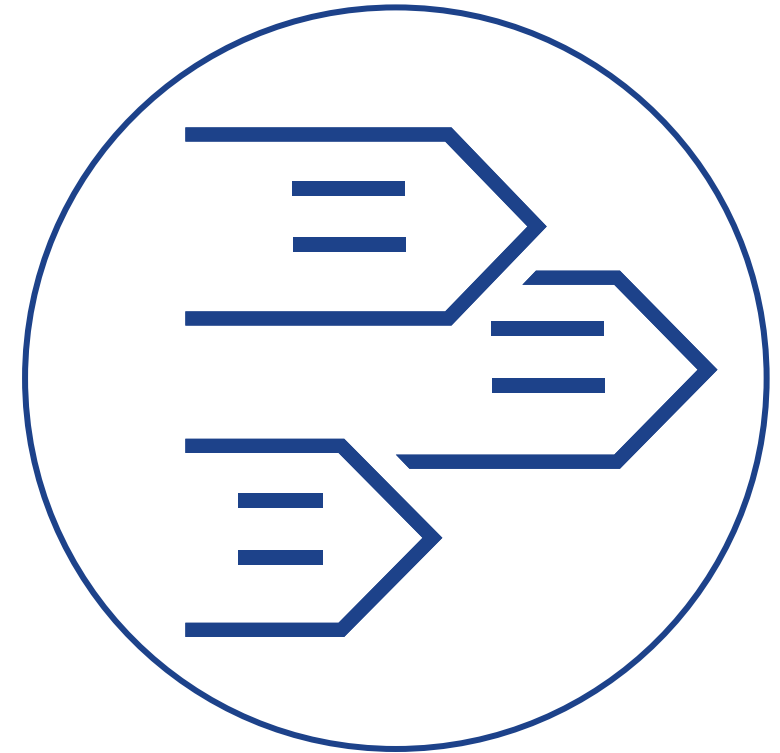
Reduce dependencies to Infrastructure & Operations Team

# Summary

Enable everyone to focus on work that require their core skills

## Make Data easy to consume and enable quick business value

- Control over what happens throughout the whole data path
- Ensure stability of the infrastructure without sacrificing user satisfaction
- Ability to interfere at each step of the DevOps lifecycle to ensure best practices
- Reduce dependencies between teams

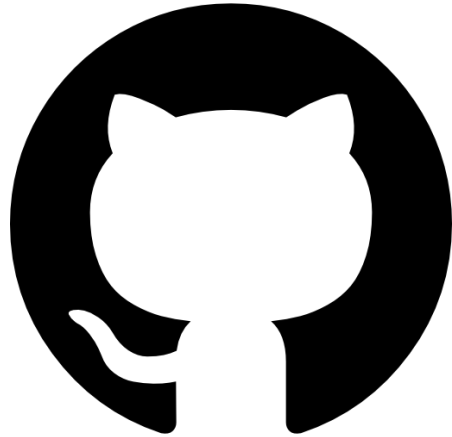




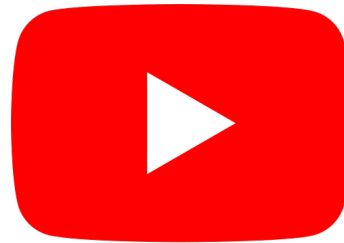
## Versatile Data Kit Offers More....

- DevOps Cycle Extensions
  - Test Phase, Build Phase, Configure Phase, Run Phase
  - Job Creation and Deletion
- Team ownership enabling collaborative multi-tenancy
- Automatic Job User provisioning (for Kerberos)
- Extensible Authorization: OAuth2 custom claims or WebHook
- Usage telemetry webhook for analytics purposes.
- Automatic and extensible error ownership categorization

# Search & Test Versatile Data Kit



Raise an issue in  
GitHub if you have any  
questions!



<https://medium.com/versatile-data-kit>

## Search, contact us & follow Versatile Data Kit

<https://github.com/vmware/versatile-data-kit/#contacts>



- <https://github.com/vmware/versatile-data-kit>
- <https://medium.com/versatile-data-kit>

# Thank you

Feedback form:



<https://bit.ly/vdk-scale>

## Q&A