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

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)
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

Wall of conflict

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
Components

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

Versatile Data Kit Data SDK
- SQL and/or Python
- Data Ingest and Transform Interfaces
- Data Lineage & Data Quality
- Lego-like extensibility

Versatile Data Kit Control Service
- Runtime for data jobs on schedule or manual
- Automatic Versioning & Deployment
- Monitoring & Alerting
- Extensibility and Governance

Data Lineage & Data Quality
SQL and/or Python

Data Ingest and Transform Interfaces

Data Lineage & Data Quality
Lego-like extensibility

REST API

Infra & Operations Team

Data Teams

Versatile Data Kit Data SDK

Versatile Data Kit Control Service
Automate and Abstract the Development Process

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

Versatile Data Kit

Plan → Code → Build → Test → Release → Deploy → Operate → Monitor

VDK automates and abstracts

IT configures and extends

Data Teams

Establish policies
Extensible

Infra & Operations Team
Quick example: DevOps Plugin
Establish standard system tests and security hardening

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

```
FROM versatiledatakit/job-builder

# Run system test before accepting the new job code
RUN pytest system_test.py || die 'Failed system test'

# Remove execution privileges from files during container build
RUN chmod -R -x $job_name/
```
Automate and Abstract the Data journey

Simplify and hide complexity of data infrastructure for Data teams
Give power to establish best Infrastructure practices

Self-service
Managed Infrastructure interface

data sources
Versatile Data Kit
Control Infrastructure
Extensible

Infra &
Operations
Team

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

Affirmative and abstract the Data Journey

**Versatile Data Kit Data SDK**

- SQL and Python
- Ingest and Transform Data
- Data Lineage & Data Quality
- Lego-like extensibility

Affirmative and abstract the DevOps Cycle

**Versatile Data Kit Control Service**

- Runtime for data jobs on schedule or manual
- Automatic Versioning & Deployment
- Monitoring & Alerting
- Extensibility and Governance

Infra & Operations Team
Data Teams
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:

Key takeaways
Day 1 Operations Simplified

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)

Key Takeaways
Day 2 Operations Simplified

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?

COPY Paste...

```sql
1 INSERT INTO tableName (sddc_sk, active_from, active_to, sddc_id, updated_by_user_id, state, is_nsxt, cl
2   500'), (sddc_sk, active_from, active_to, sddc_id, updated_by_user_id, state, is_nsxt, cl
3  5208
```

???
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)
```python
class QueryValidationPlugin:
    @hookimpl(tryfirst=True)
    def vdk_configure(self, config_builder: ConfigurationBuilder) -> None:
        # Declare needed configuration, it will be injected automatically from file, env variables, etc.
        config_builder.add(
            key="max_query_size",
            default_value=10000,
            description="The maximum query size in bytes allowed to be executed."
        )
```
class QueryValidationPlugin:

    @hookimpl(tryfirst=True)
    def vdk_configure(self, config_builder: ConfigurationBuilderInterface) -> None:
        # Declare needed configuration, it will be injected automatically from file, env variables, etc.
        config_builder.add(
            key="max_query_size",
            default_value=10000,
            description="The maximum query size in bytes allowed to be executed."
        )

    @hookimpl
    def initialize_job(self, context: JobContext) -> None:
        # Now let's get the correctly configured value
        self._max_query_size = context.core_context.configuration.get_value("max_query_size")

    @hookimpl(trylast=True)
    def db_connection_validate_operation(self, self, operation, parameters):
        parameters_length = 0
        if parameters:
            parameters_length = len("",".join(map(str, parameters)))
        if len(operation) + parameters_length > self._max_query_size:
            raise Exception(f"Database operation has exceeded the maximum limit of \{(self._max_query_size)\} characters,"
setup =
    name="vdk-my-validation",
    version="1.0",
    packages=setuptools.find_namespace_packages(),
    entry_points={
        "vdk.plugin.run": [
            "vdk-my-validation-plugin = vdk_my_validation",
        ]
    }
import setuptools

setuptools.setup(
    name="my-org-vdk",
    version="1.0",
    install_requires=[
        "vdk-core",
        "vdk-plugin-control-cli",
        "vdk-postgres",
        "vdk-snowflake",
        "vdk-ingest-http",
        "vdk-ingest-file",
        "vdk-my-validation"
    ]
)
(demo) {15:14} ~/data-engineer $ vdk run example
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

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

Infra & 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
  o Test Phase, Build Phase, Configure Phase, Run Phase
  o 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!

Search, contact us & follow Versatile Data Kit

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

https://medium.com/versatile-data-kit
Thank you

Feedback form:
