Versatile Data Kit

VERSATILE
DATA KIT

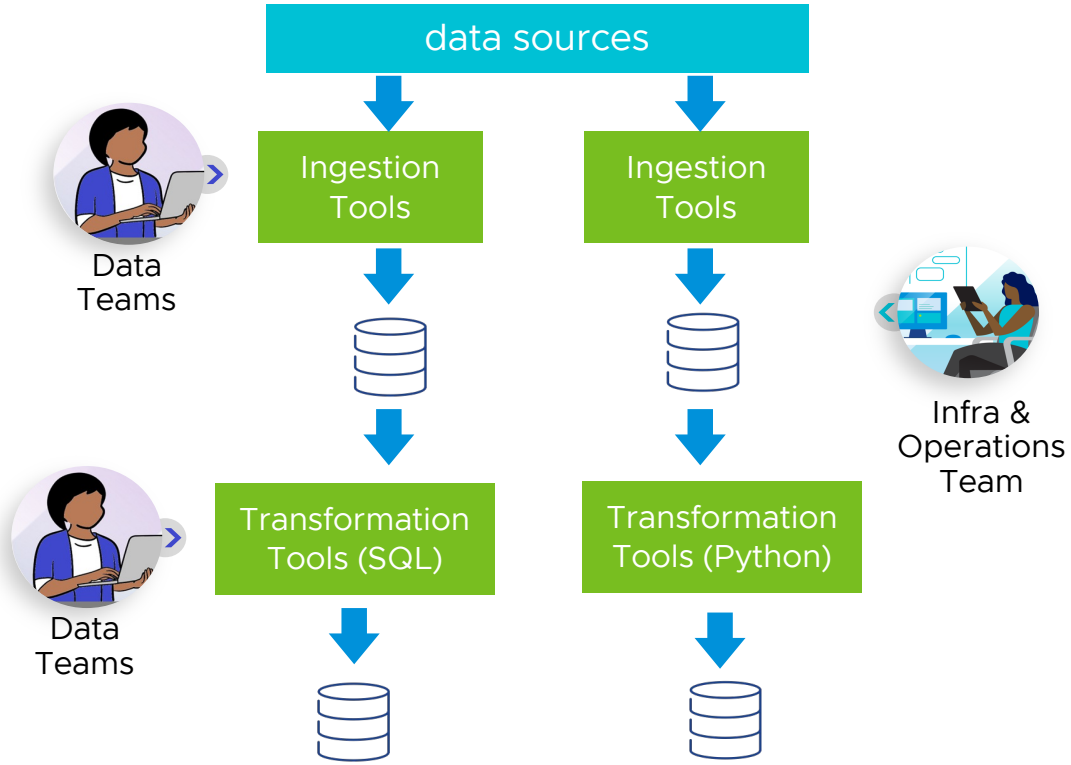
Preview

Enable everyone to focus on work that require their core skills

Github repo

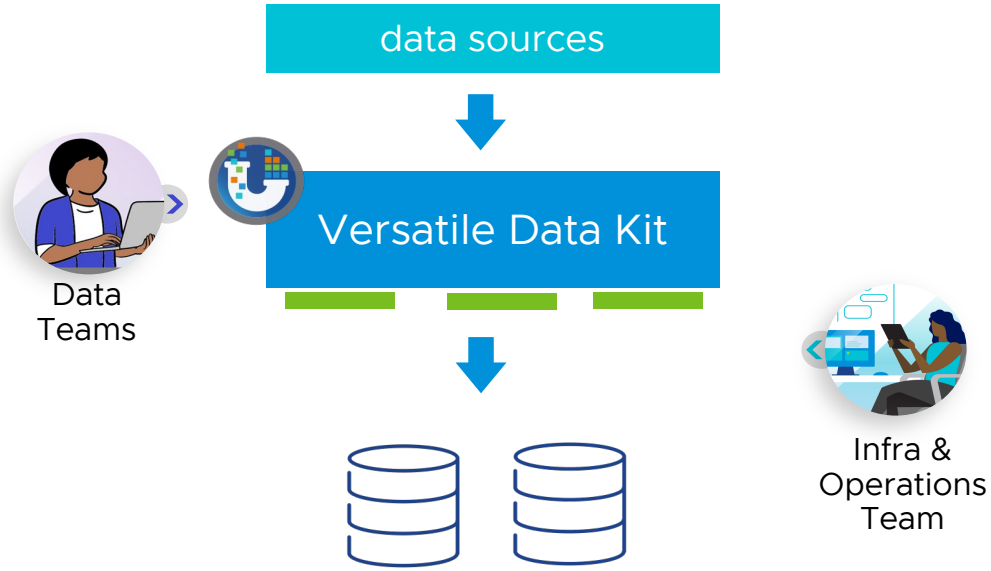
Without Versatile Data Kit

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



With Versatile Data Kit

- Easier maintenance using unified managed interface
- Self-service, fully automated data teams
- Improved collaboration



What are the DataOps challenges?

Infra/Ops Team and Data Teams tension and conflict

Ineffcient Operations *Stalled development.*

Domain knowledge

Implement business logic

Optimizes for agility and speed


Data Team


Data Team


Data Team


Data Team


Data Team

Wall of conflict

DevOps & Infrastructure knowledge

Maintain infrastructure

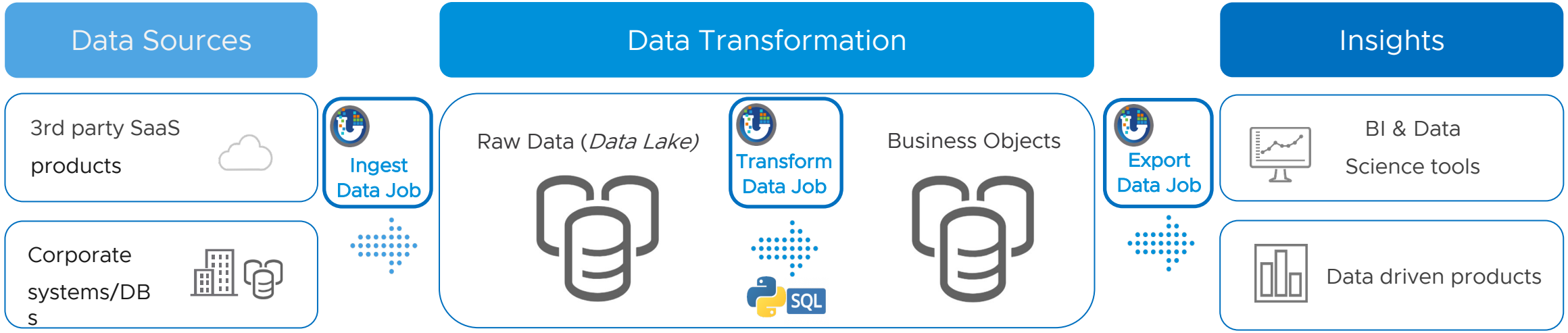
Optimizes reliability, availability and security


Infra & Operations Team

Blurred lines of responsibility

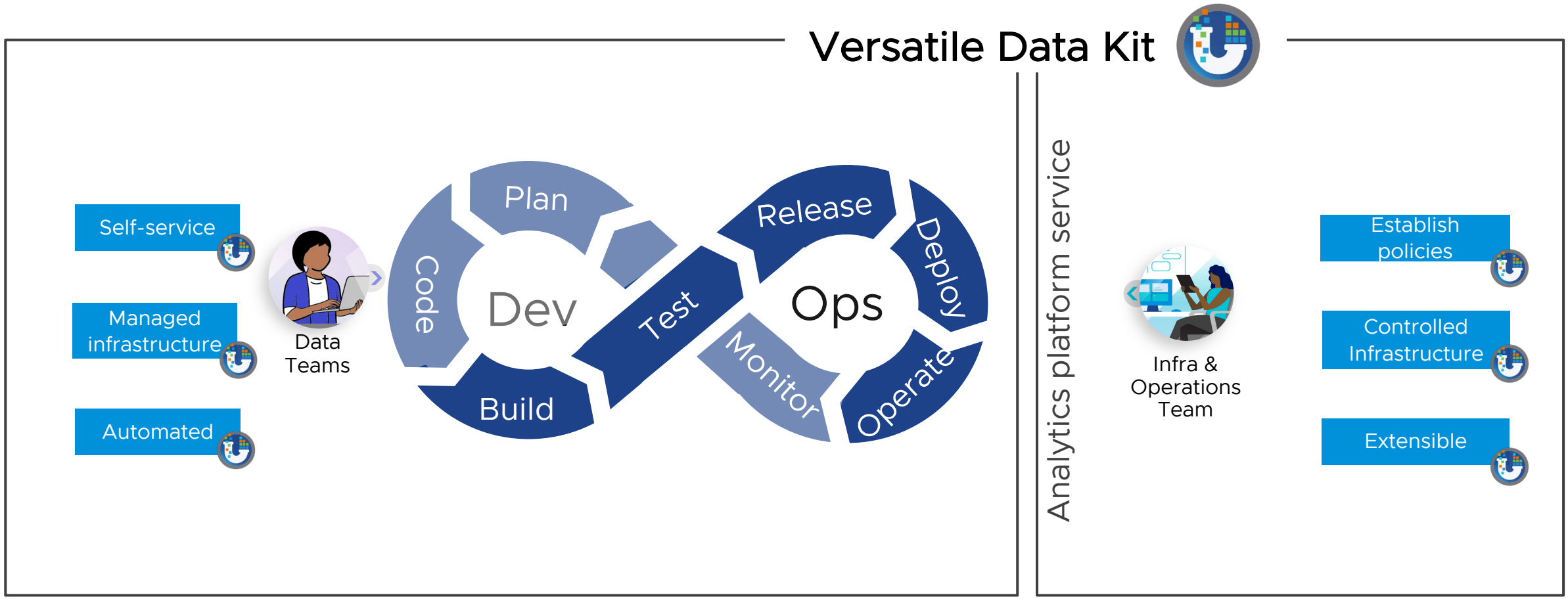
Versatile Data Kit

Data lifecycle (BI Journey) and where VDK fits in



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?

Automate and abstract the Data Journey

Versatile Data Kit Data SDK

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

python™

Data Teams

Automate and abstract the DevOps Cycle

Versatile Data Kit Control Service

REST API

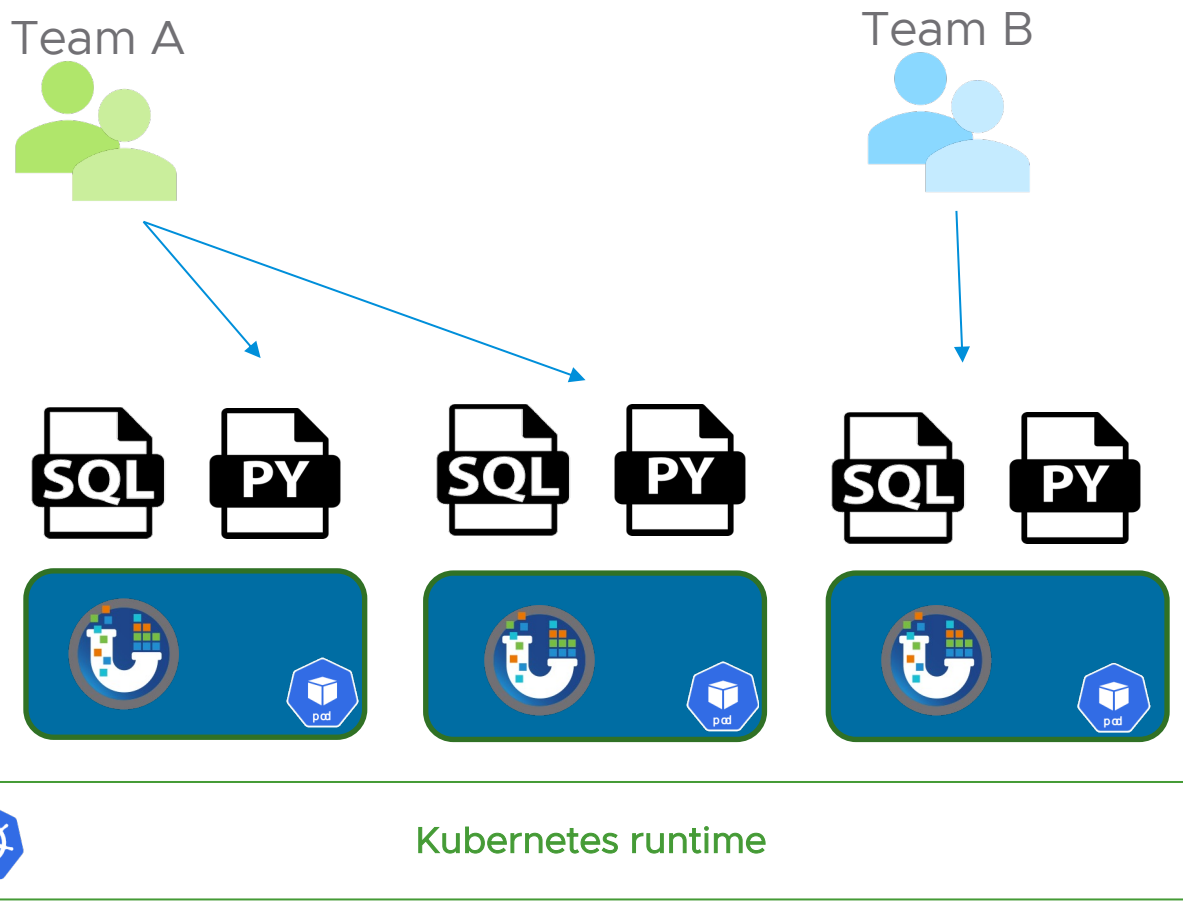
- Runtime for data jobs on schedule or manual
- Automatic Versioning & Deployment
- Monitoring & Alerting
- Extensibility and Governance

kubernetes

Infra & Operations Team

VDK SDK and VDK Runtime (Control Service)

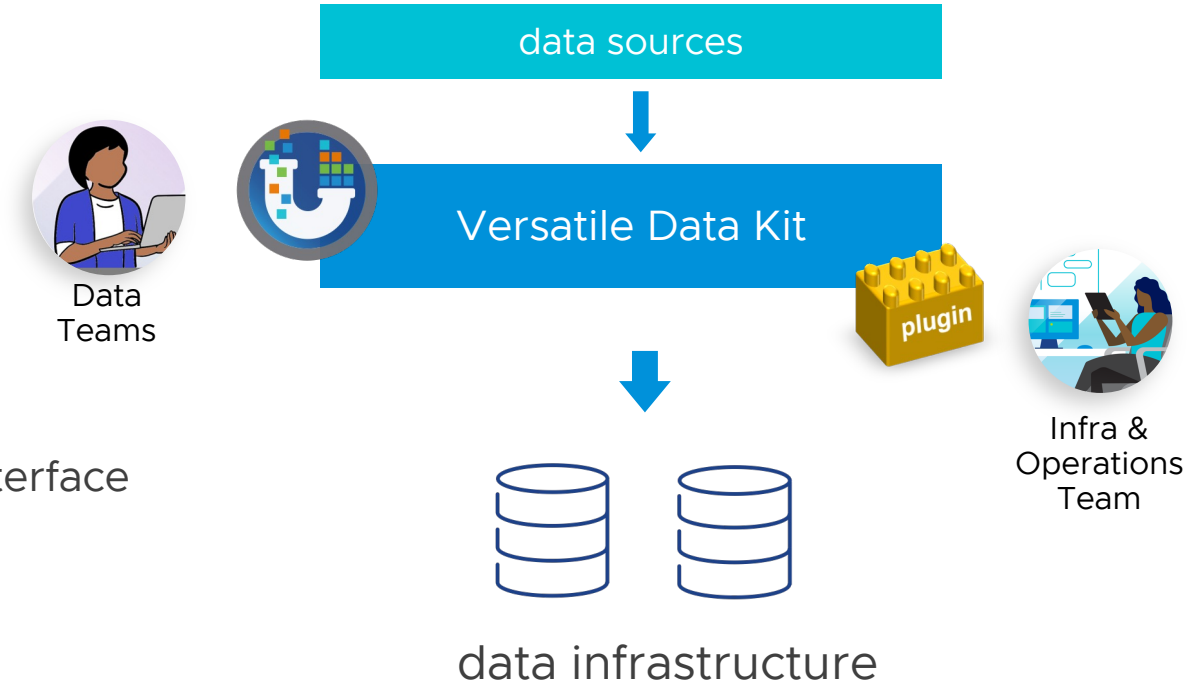
Data library and Job runner



- Single **SDK** version used across all job.
- Consistent system configuration
- Automated instantaneous upgrade of the sdk.
- Job versioning, rollback.

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

Control Infrastructure

Extensible

Quick example: vdk-impala, vdk-trino

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

https://github.com/vmware/versatile-data-kit/tree/main/projects/vdk-plugins/vdk-impala

Data Teams

execute query

data sources

Versatile Data Kit

automatic recovery

data infrastructure

Infra & Operations Team

plugin

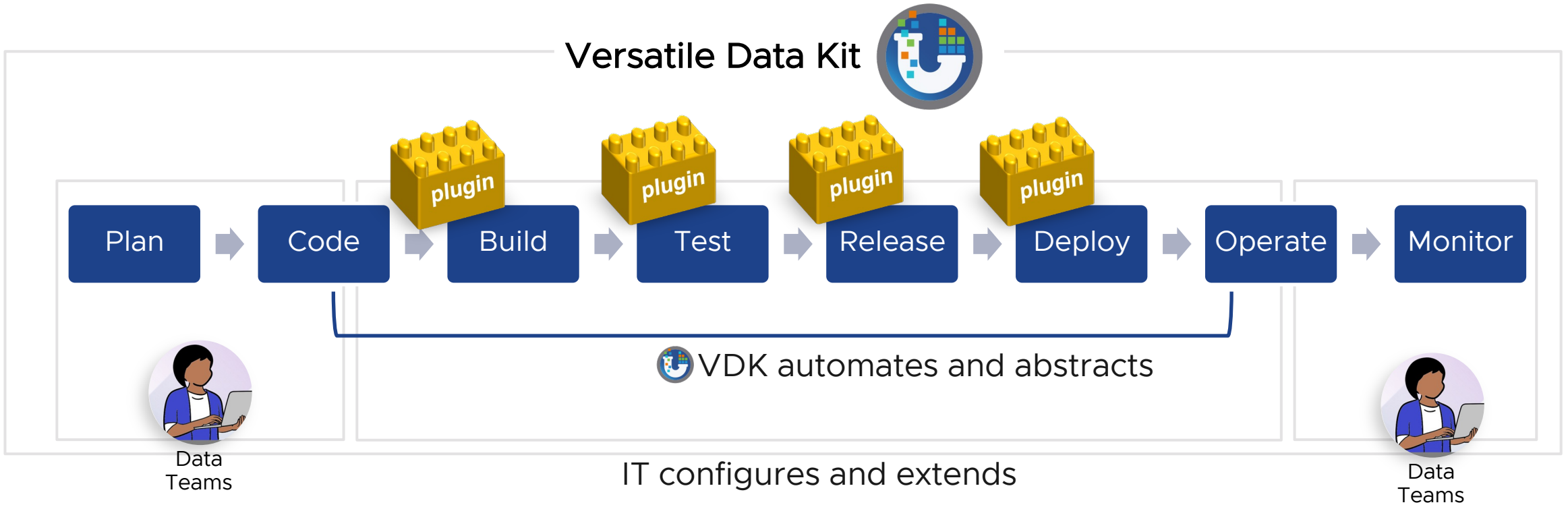
vdk-impala

```
@hookimpl
def db_connection_recover_operation(
                                    recovery_cursor):
    if impala_error_handler.handle(recovery_cursor):
        logging.getLogger(__name__).info(
            "Error handled successfully!
             Query execution has succeeded."
        )
    else:
        raise recovery_cursor.get_exception()
```

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

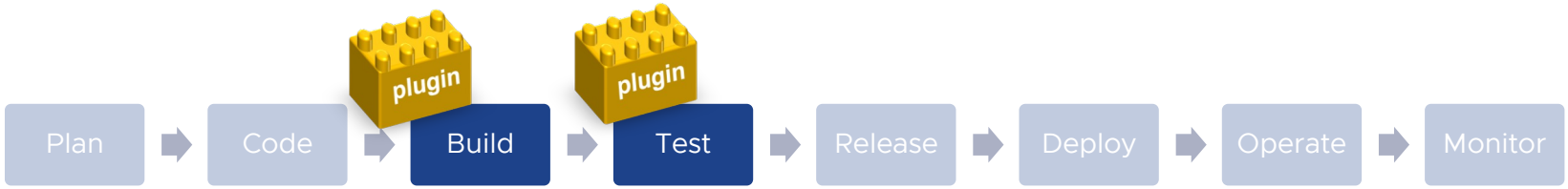
Automate and Abstract the Development Process

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                Xxx      -/tree/main
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
```

Data Journey : Ingestion

Data Teams

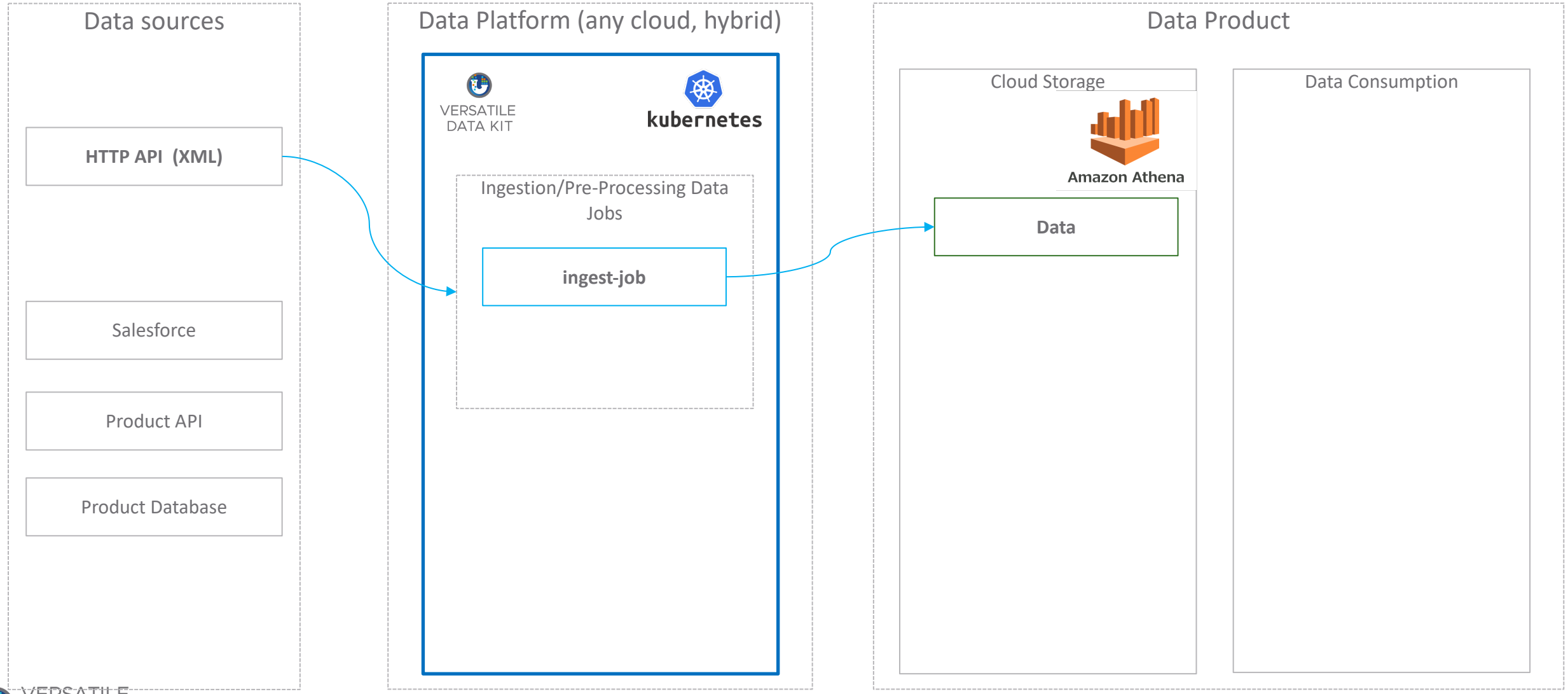
Data Sources

3rd party SaaS products

Corporate systems/DBs

Ingest Data Job

Ingestion


Data sources
HTTP API (XML)
Salesforce
Product API
Product Database
Data Platform (any cloud, hybrid)
VERSATILE DATA KIT
kubernetes
Ingestion/Pre-Processing Data Jobs
ingest-job
Data Product
Cloud Storage
Amazon Athena
Data
Data Consumption

Ingestion Job

vdk run ingest-job

response =

```
<Rates>
 <Rate>
  <No>C/NBP/2</No>
  <Date>2020-10-05</Date>
   <Bid>4.347</Bid>
   <Ask>4.431</Ask>
 </Rate>
 ….
</Rates>
```

User code

```
job_input.send_object_for_ingestion(
        payload=xmltodict.parse(response),
        destination_table="rates"
)
```

Versatile Data Kit

No	Date	Bid	Ask
C/NBP/2	5/10/20	4.347	4.431
C/NBP/1	5/10/20	4.234	4.532

PostgreSQL presto snowflake GREENPLUM DATABASE trino Amazon Athena

Data Journey : Transformation

Data Teams

Data Transformation

Raw Data (*Data Lake)*

Transform Data Job

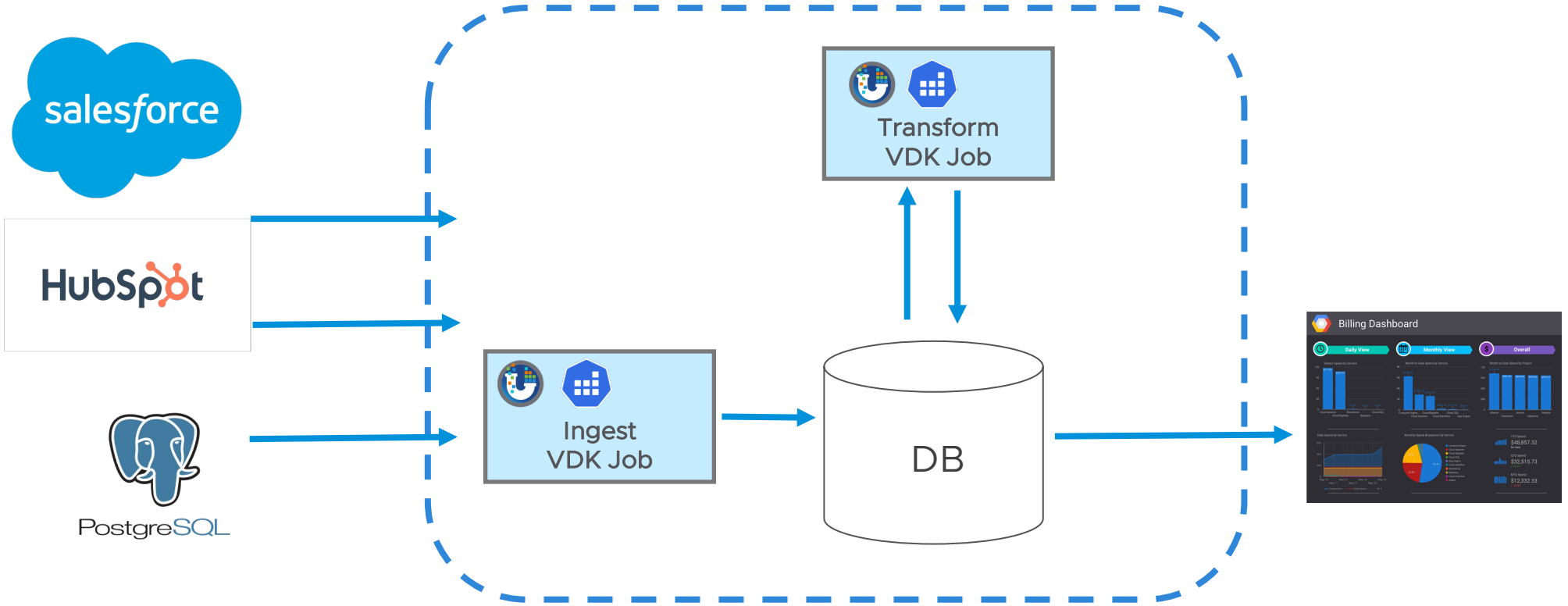
Business Objects

Data Transformation

Pre-process/ Clean big data quickly, create business KPIs

Out of the box functionality for pre-processing data in Star schema data model (fact/dimensions)

Data engineers focus on data transformations



Raw Data

Table:
vmc_console_org

id	sla	type
ab4234	core	partner
fh3hf3f	internal	internal
df3f422	core	customer

Table:
company_name

org	company	rn
ab4234	Amazon	1
1235ff	Amazon	2
df3f422	MIT	1

Table:
billing_start_date

org	date	rn
ab4234	11/10/20	1
df3f422	7/10/20	1
df3f422	5/10/20	1

```sql
SELECT
    org.id AS org_id,
    org.org_sla,
    org.org_type,
    COALESCE(CASE
            WHEN ocn.comp_name_type = 'map_org_company_name'
                THEN ocn.company_name
            WHEN org.org_type IN ('INTERNAL_CORE', 'INTERNAL')
                THEN 'VMware'
            WHEN org.org_type = 'INTERNAL_AWS'
                THEN 'Amazon AWS'
            ELSE ocn.company_name
            END, org.org_display_name) AS company_name,
    bsd.billing_start_date
FROM vmc_console_org as org
LEFT JOIN company_name ocn ON ocn.dim_org_id = org.id
    AND ocn.rn = 1
LEFT JOIN billing_start_date bsd ON bsd.org_id = org.id
    AND bsd.rn = 1
```

transform and load into

Business objects

Table:
dimension vmc organization

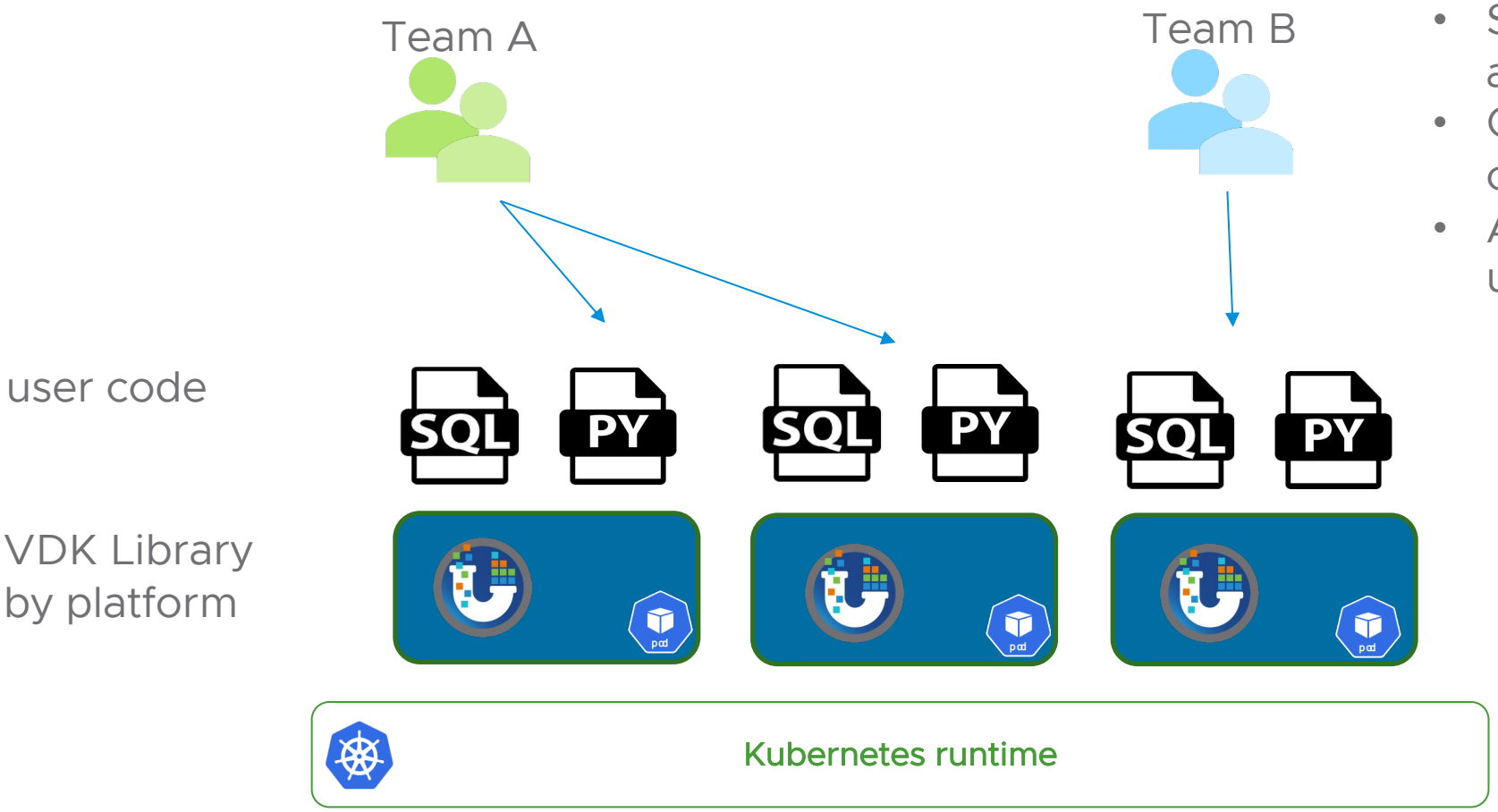
id	type	company	created
ab4234	partner	Amazon	11/10/20
df3f422	customer	MIT	5/10/20
fh3hf3f	internal	VMware	7/10/20

Deployment

Data Teams

VDK SDK and VDK Runtime (Control Service)

Data library and Job runner



- Single version used across all job.
- Consistent system configuration
- Automated instantaneous upgrade of the sdk.

Operations

Data Teams

Data Jobs statistics, status, notifcations, alerts

Data Jobs

Data Jobs help Data Engineers develop, deploy, run, and manage data processing workloads

Jobs with Failing executions - Last 24h

pmitev-test	48
e2e-cp-stg-test-job	8
e2e-cy-dp-failing	2
e2e-cp-stg-failing-job	2

Most recent failed executions - UTC time

pmitev-test-1662670800	Platform Error	Sep 8, 2022, 09:01:11 PM
pmitev-test-1662669000	Platform Error	Sep 8, 2022, 08:31:16 PM
pmitev-test-1662667200	Platform Error	Sep 8, 2022, 08:01:12 PM
pmitev-test-1662665400	Platform Error	Sep 8, 2022, 07:31:09 PM
e2e-cp-stg-test-job-1662664523	User Error	Sep 8, 2022, 07:16:25 PM

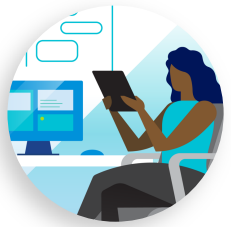
Manage Data Jobs

REFRESH ENABLE DISABLE EXECUTE NOW

All | Enabled | Disabled | Not Deployed | Search

	Job name	Deployment Status	Last Execution End (UTC)	Last Execution Duration	Last Execution Status	Success rate	Schedule (in UTC)	Logs	Details
○	tms-staging-vdk-1644321902	✓	Sep 8, 2022, 09:30 PM	12s	Success	100.00%	Every minute		
○	pmitev-test	✓	Sep 8, 2022, 09:01 PM	57s	Platform Error	0.00% (337 failed)	Every 30 minutes		
○	e2e-cp-stg-test-job	✓	Sep 8, 2022, 07:16 PM	1m 1s	User Error	0.00% (115 failed)	At 11:11 PM, on day 5 of the month, and on Monday, only in August		

Managed Infrastructure



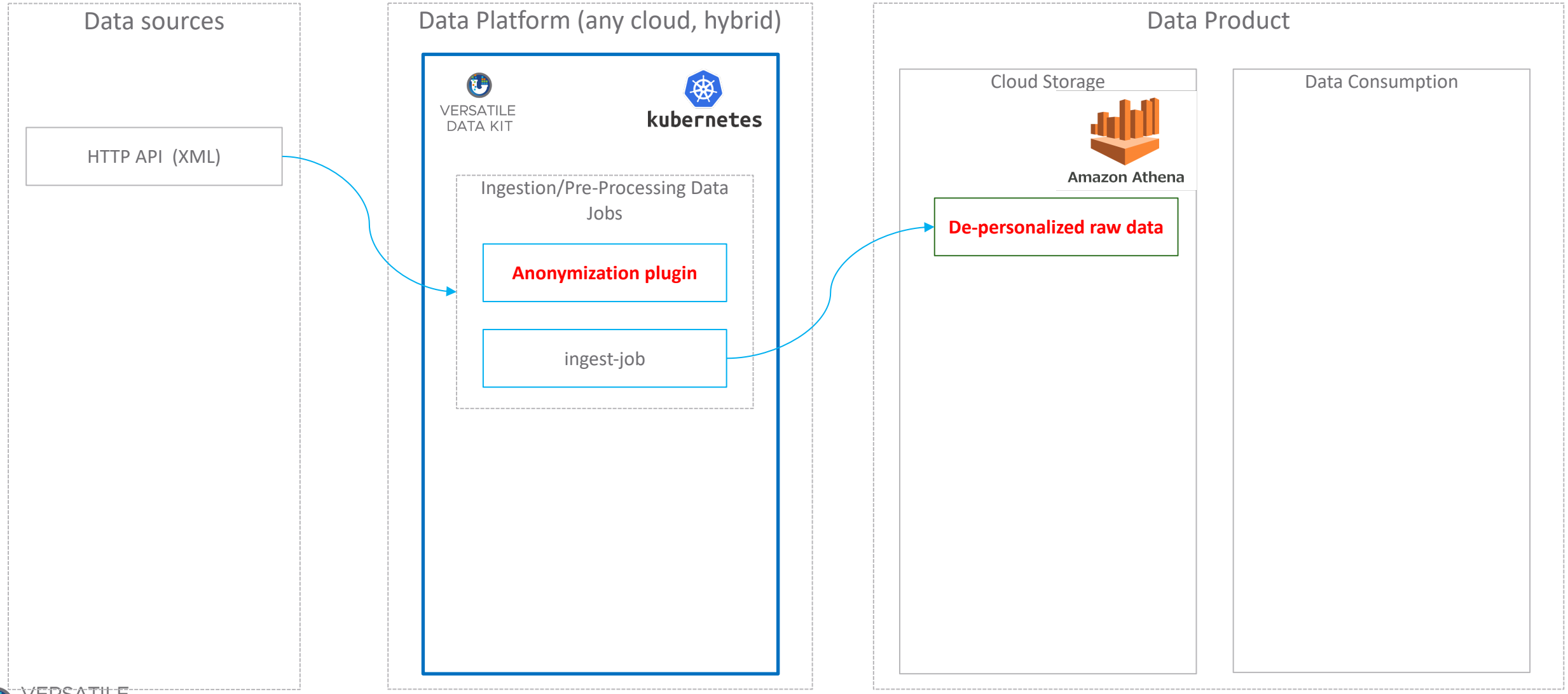
Infra &
Operations
Team

Improve data infrastructure security (demo)

Check out:

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

Data Platform – Ingest and Anonymization POC



What are we going to do?

vdk run ingest-job

```
job_input.send_object_for_ingestion(
        payload=response.json(),
        destination_table="rest_table"
)
```

data:

userId	id	title	completed
1	11	delectus	1
2	12	autem	0

intercepted

vdk-poc-anonymization (plugin)

userId	id	title	completed
1	11	325959af	1
2	12	f0a4cedd	0

PostgreSQL presto snowflake GREENPLUM DATABASE trino Amazon Athena

Improve data infrastructure stability (demo)

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)

PostgreSQL presto snowflake GREENPLUM DATABASE trino Amazon Athena

Deliver analytics platform for your business quickly

Infra & Operations Team

Check out:

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