DataOps as a Service

"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

ANTONI IVANOV

Software Engineer / Versatile Data Kit

aivanov@vmware.com
linkedin.com/in/antoni-ivanov

Agenda

Data Applications

API for Data

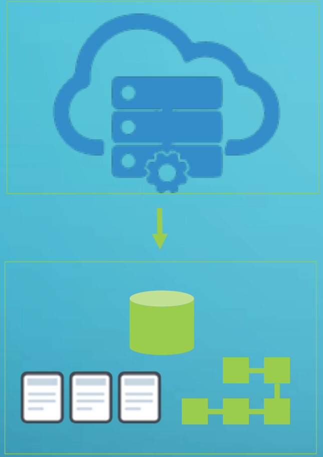
SLO and SLAs for Data

DevOps Cycle for Data

Versatile Data Kit

Applications

Examples



Application

- E-Commerce application
- Mobile app
- Customer relationship management
- Recommendation system

Data

- Databases
- Log files
- Click streams
- Metrics

Data Applications

Examples

Application

- E-Commerce application
- Mobile app
- Customer relationship management
- Recommendation engine

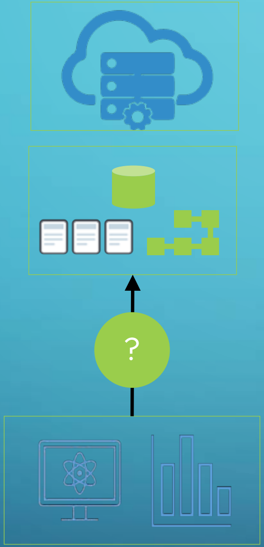
Data

- Databases
- Log files
- Click streams
- Metrics

Data Applications

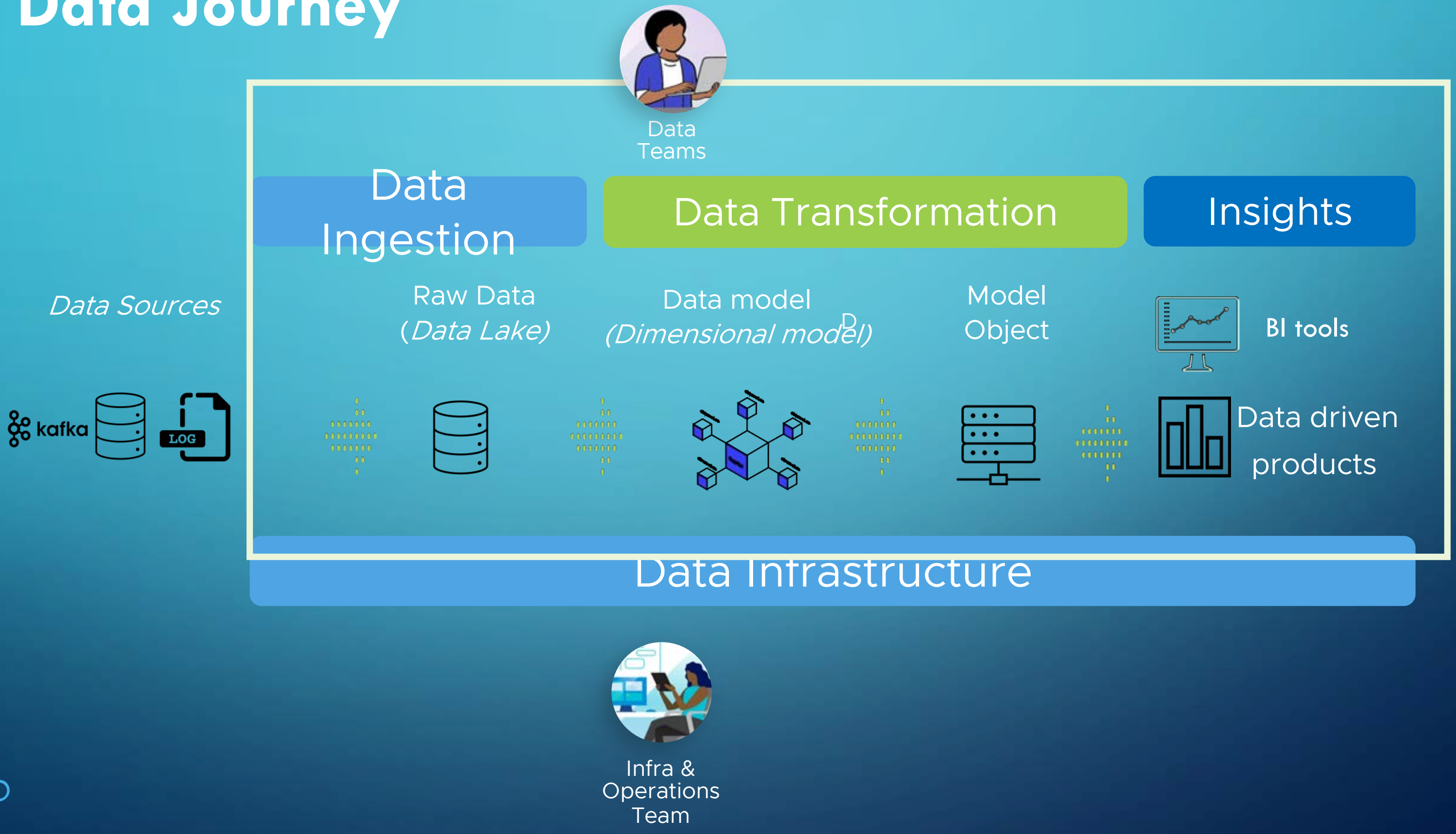
- Usage reporting
- Business intelligence
- Recommendation engine
- Forecasting model

Data Journey



Ecommerce app

The Data Journey



Data Teams

Data Ingestion

Data Transformation

Insights

Data Sources

Raw Data
(*Data Lake)*

Data model
(Dimensional model)
D

Model
Object

BI tools

Data driven
products

Data Infrastructure

Infra &
Operations
Team

How do multiple applications and data applications communicate between each other?

API

API is a set of rules, protocols, and tools that allow different software applications to communicate with each other

API Components

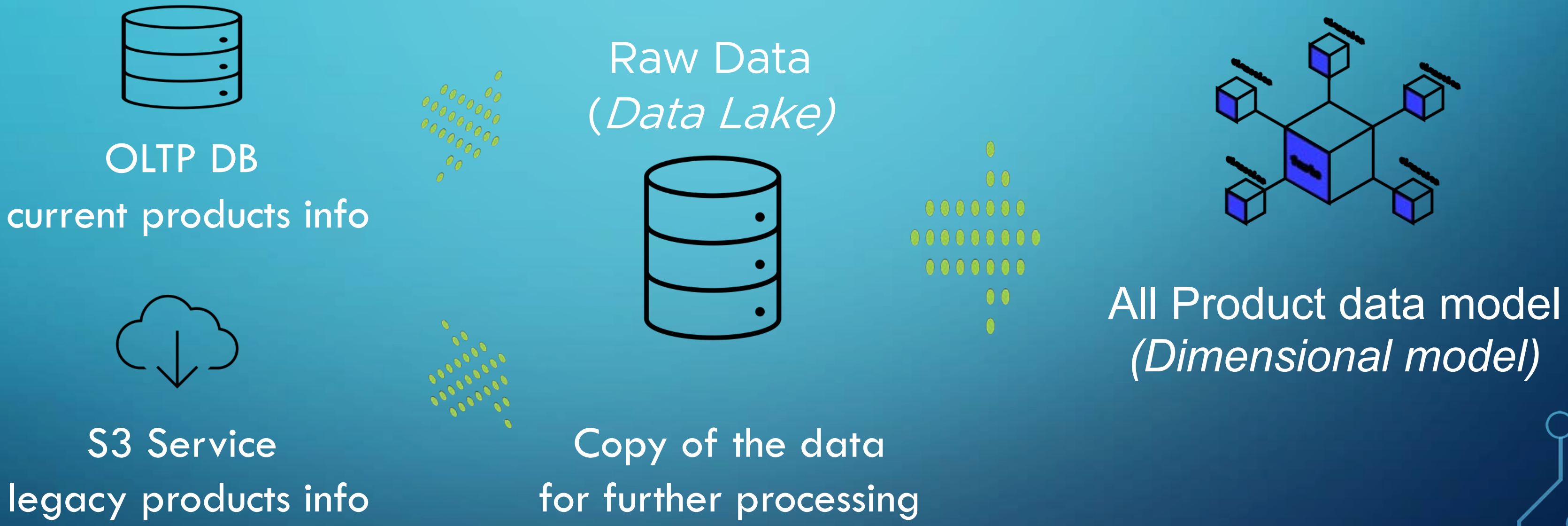
Interface and Contract

Security and Access Control

Usability and Documentation

Monitoring and Operations

API for Data Example



OLTP DB
current products info

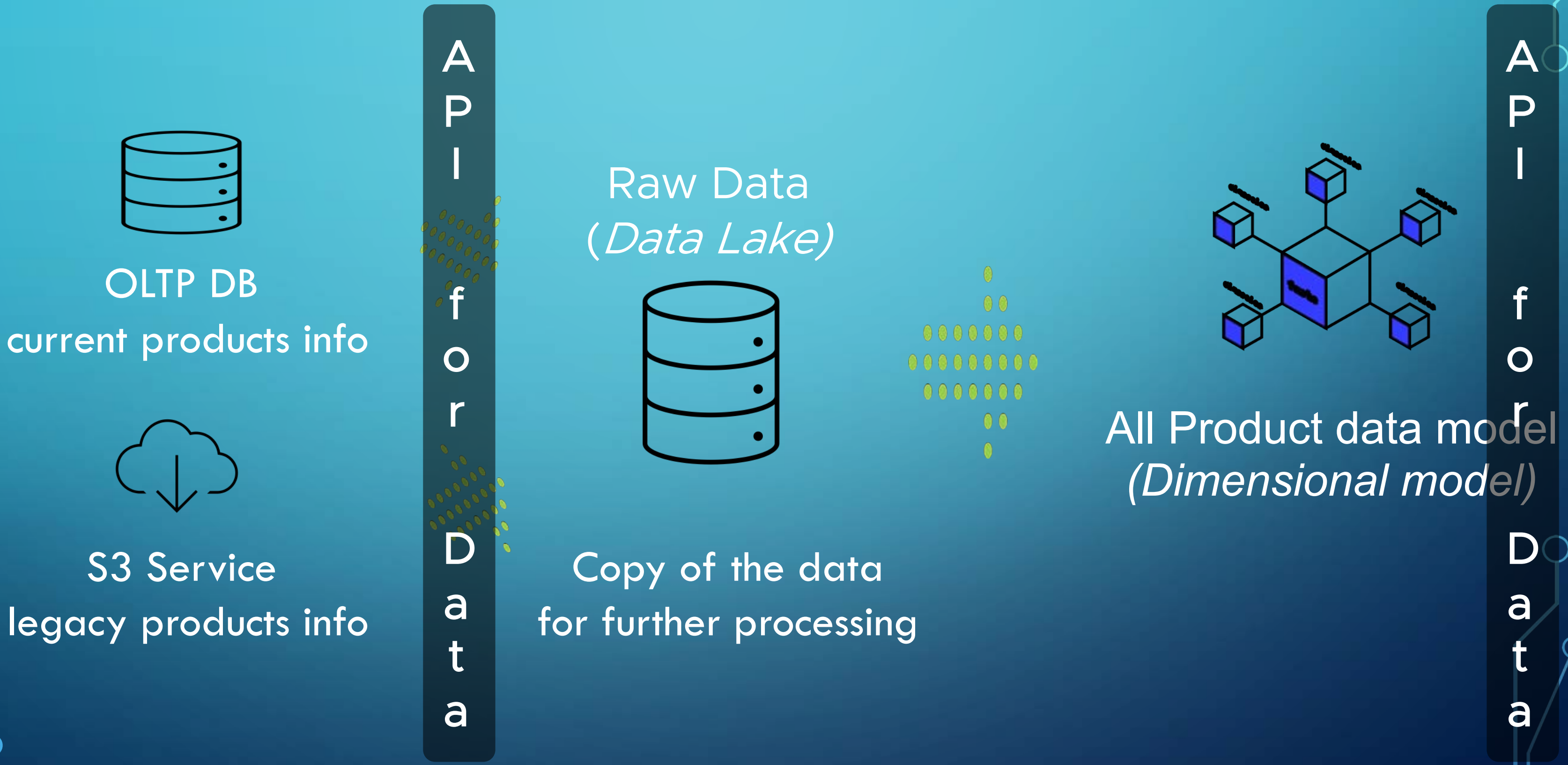
S3 Service
legacy products info

Raw Data
(Data Lake)

Copy of the data
for further processing

All Product data model
(Dimensional model)

API for Data Example



OLTP DB

current products info

S3 Service

legacy products info

API for Data

Raw Data

(Data Lake)

Copy of the data

for further processing

All Product data model

(Dimensional model)

API for Data

API for Data Example

Table/Entity Name: All Products

Most recent information about each product

Data Access

Tables in Database / SQL

Pandas DataFrames

Parquet/Arrow Data Format

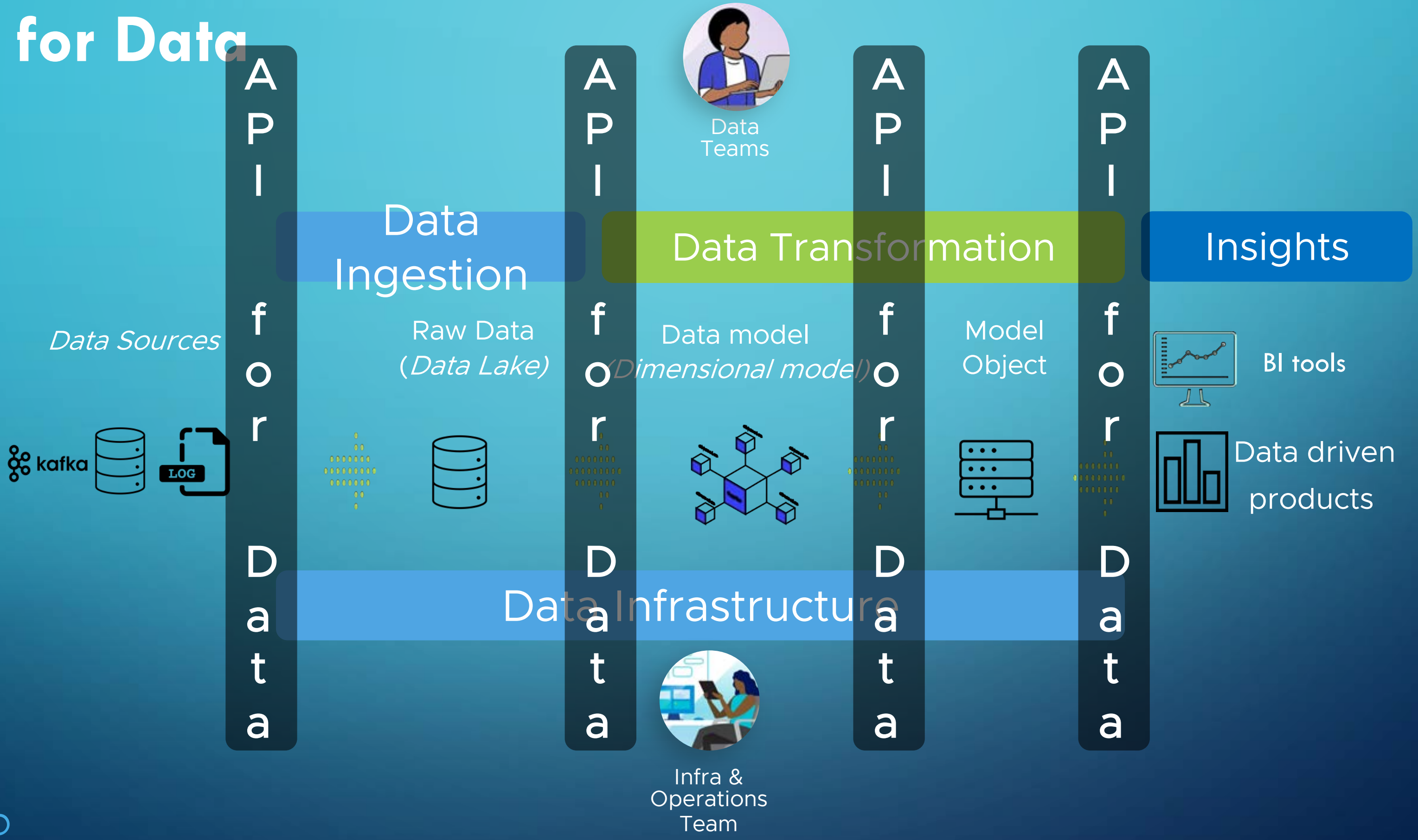
Data Schema

product_id: UUID
name: String
category: String
price: Decimal

Data Semantics

product_id: UUID and unique across all records.
name: non-empty string representing latest official name
category: Must belong to a predefined list of categories.
price: Must be a positive decimal number in currency XXX

API for Data


API for Data
API for Data
API for Data
API for Data
Data Teams
Data Ingestion
Data Transformation
Insights
Data Sources
Raw Data (Data Lake)
Data model (Dimensional model)
Model Object
BI tools
Data driven products
kafka
LOG
Data Infrastructure
Infra & Operations Team

SLOs and SLAs for Data

Data Semantics

Most recent information about each product

product_id: UUID and unique across all records.
name: non-empty string representing user facing name
category: Must belong to a predefined list of categories.
price: Must be a positive decimal number in currency XXX

SLOs and SLAs for Data

Data Accuracy SLOs

product_id: UUID and unique across all records.
name: non-empty string representing user facing name
category: Must belong to a predefined list of categories.
price: Must be a positive decimal number in currency XXX

Data Availability SLOs

The "products" table should be queryable 99.9% of the time

Data Freshness SLOs

Any changes in the inventory system should be updated in the "products" table within 1 hour

Data Contracts

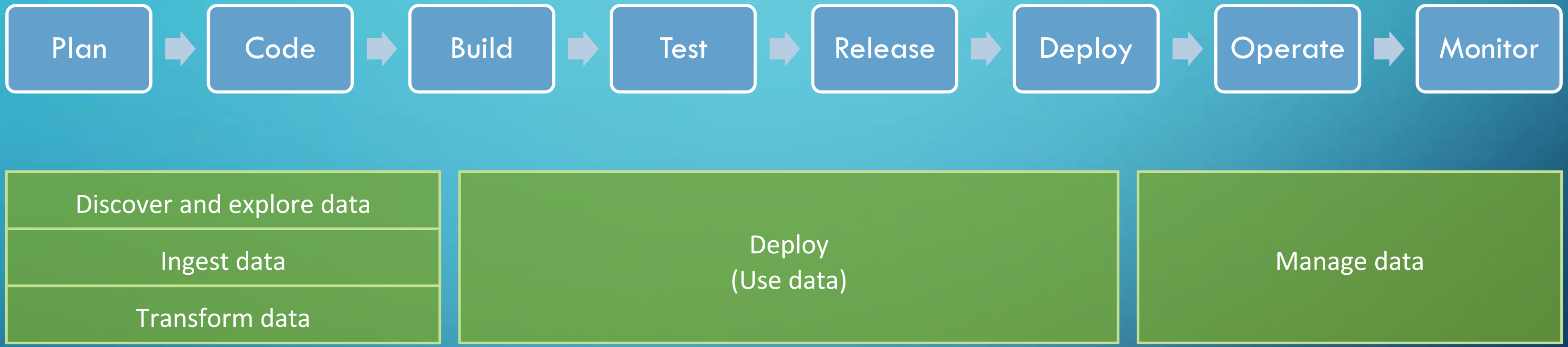
API for Data and SLO/SLAs for data

https://bit.ly/data-contract

Blog posts by Chad Sanderson

DevOps Cycle for Data

Data Teams



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

Discover and explore data

Ingest data

Transform data

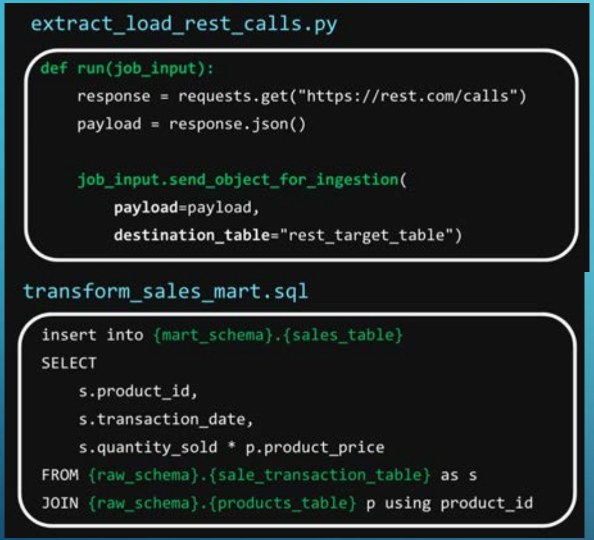
Deploy
(Use data)

Manage data

Versatile Data Kit

Develop Data Jobs

VDK SDK



extract_load_rest_calls.py

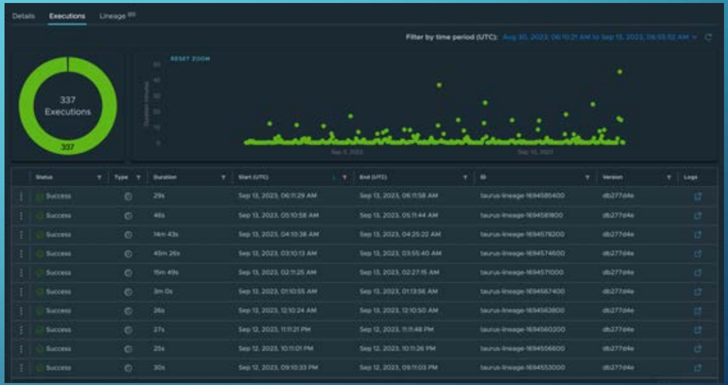
```python
def run(job_input):
    response = requests.get("https://rest.com/calls")
    payload = response.json()

    job_input.send_object_for_ingestion(
        payload=payload,
        destination_table="rest_target_table")
```

transform_sales_mart.sql

```sql
insert into {mart_schema}.{sales_table}
SELECT
    s.product_id,
    s.transaction_date,
    s.quantity_sold * p.product_price
FROM {raw_schema}.{sale_transaction_table} as s
JOIN {raw_schema}.{products_table} p using product_id
```

Deploy and Monitor

Control Plane and Operations UI

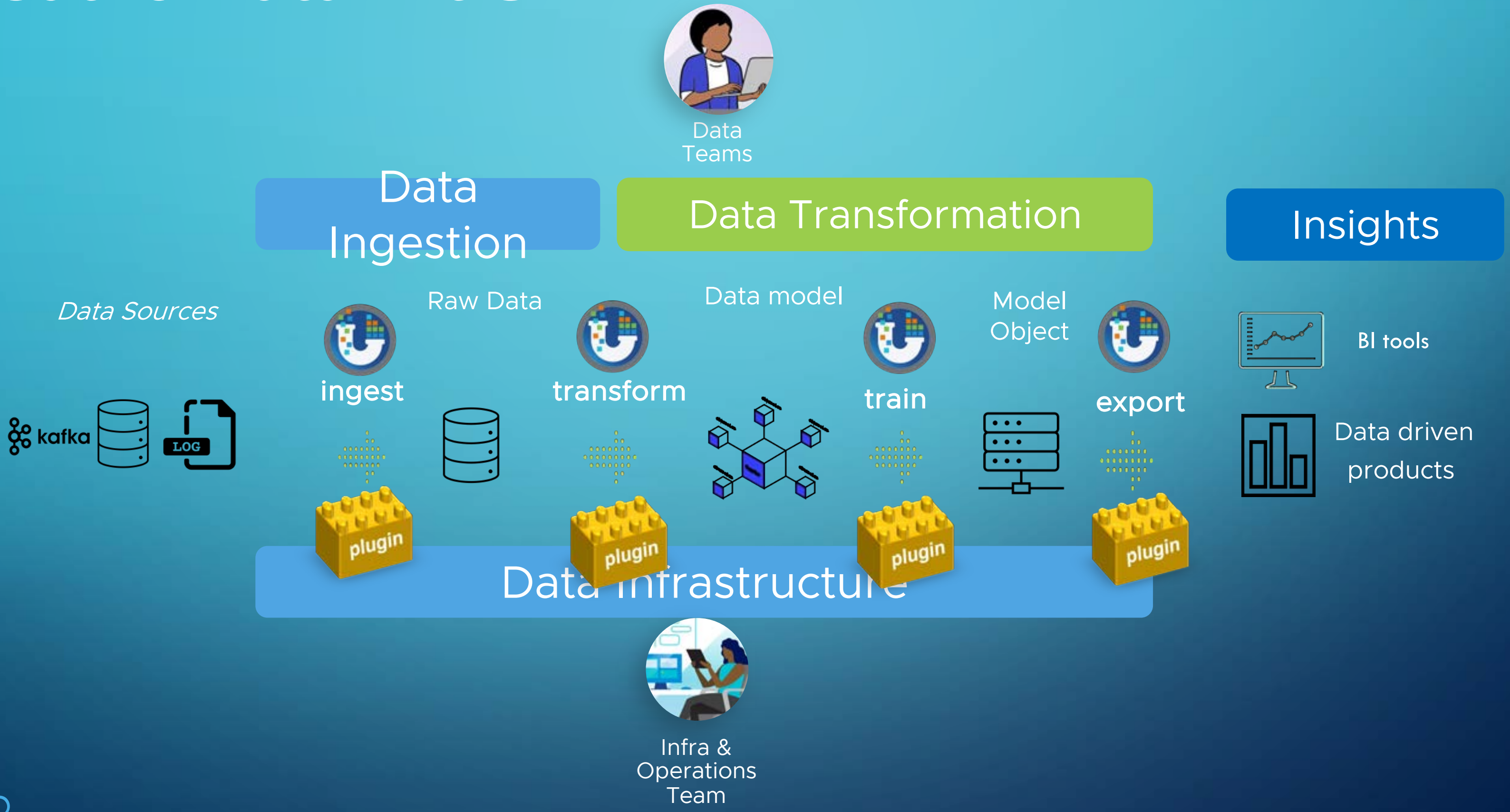


Filter by time period (UTC): Aug 30, 2023, 06:10:21 AM to Sep 13, 2023, 06:55:52 AM

Status	Type	Duration	Start (UTC)	End (UTC)	ID	Version	Logs
Success		29s	Sep 13, 2023, 06:11:29 AM	Sep 13, 2023, 06:11:58 AM	taurus-lineage-1694585400	db277d4e	
Success		46s	Sep 13, 2023, 05:10:58 AM	Sep 13, 2023, 05:11:44 AM	taurus-lineage-1694581800	db277d4e	
Success		14m 43s	Sep 13, 2023, 04:10:38 AM	Sep 13, 2023, 04:25:22 AM	taurus-lineage-1694578200	db277d4e	
Success		45m 26s	Sep 13, 2023, 03:10:13 AM	Sep 13, 2023, 03:55:40 AM	taurus-lineage-1694574600	db277d4e	
Success		15m 49s	Sep 13, 2023, 02:11:25 AM	Sep 13, 2023, 02:27:15 AM	taurus-lineage-1694571000	db277d4e	
Success		3m 0s	Sep 13, 2023, 01:10:55 AM	Sep 13, 2023, 01:13:56 AM	taurus-lineage-1694567400	db277d4e	
Success		26s	Sep 13, 2023, 12:10:24 AM	Sep 13, 2023, 12:10:50 AM	taurus-lineage-1694563800	db277d4e	
Success		27s	Sep 12, 2023, 11:11:21 PM	Sep 12, 2023, 11:11:48 PM	taurus-lineage-1694560200	db277d4e	
Success		25s	Sep 12, 2023, 10:11:01 PM	Sep 12, 2023, 10:11:26 PM	taurus-lineage-1694556600	db277d4e	
Success		30s	Sep 12, 2023, 09:10:33 PM	Sep 12, 2023, 09:11:03 PM	taurus-lineage-1694553000	db277d4e	

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

Versatile Data Kit SDK



Data Teams

Data Ingestion

Data Transformation

Insights

Data Sources

Raw Data

Data model

Model Object

ingest

transform

train

export

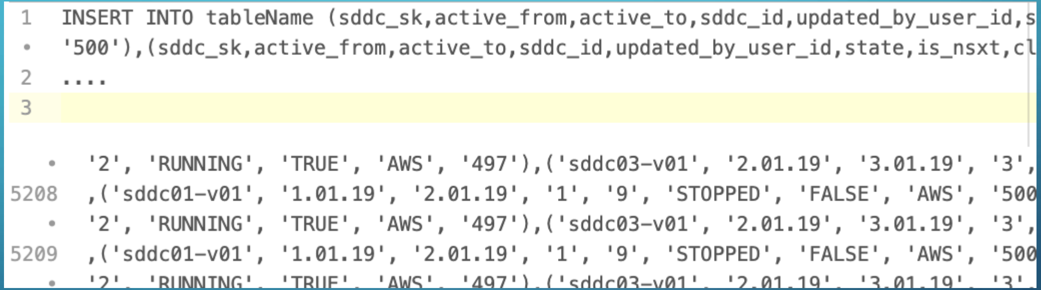
BI tools

Data driven products

Data Infrastructure

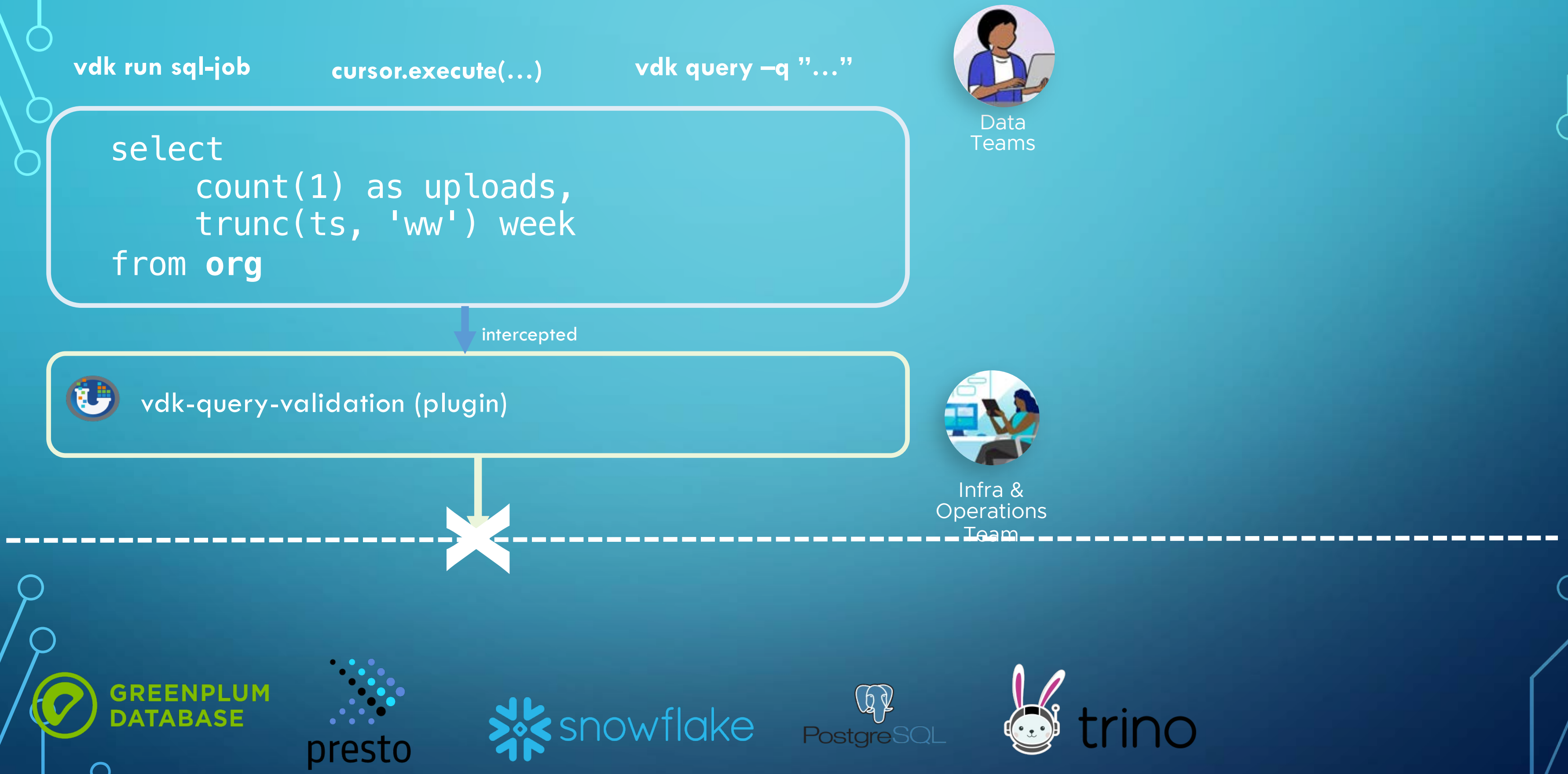
Infra & Operations Team

WHAT ARE WE GOING TO DO?



```
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', 'RUNNING', 'TRUE', 'AWS', '497'),('sddc03-v01', '2.01.19', '3.01.19', '3',
,('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',
,('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',
```

😭 ????



vdk run sql-job

cursor.execute(…)

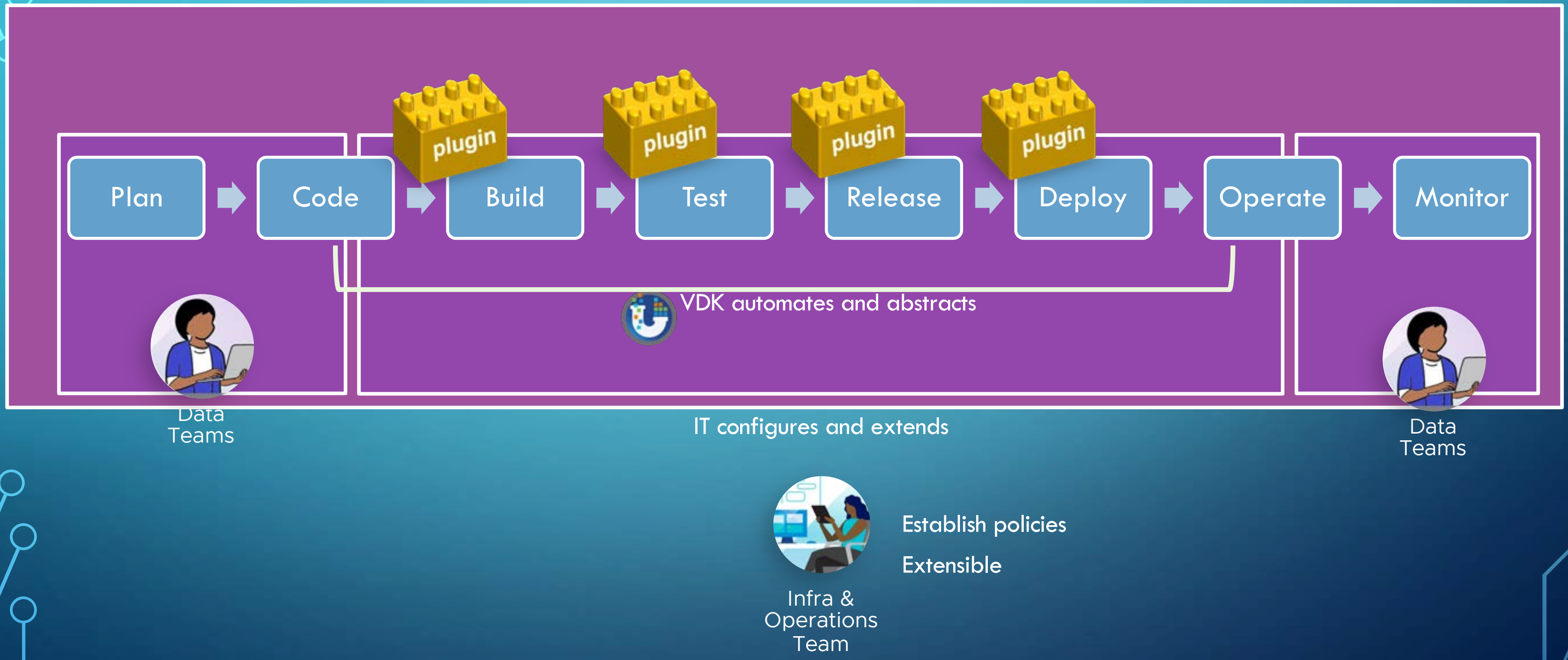
vdk query –q ”…”

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

intercepted

vdk-query-validation (plugin)

Versatile Data Kit Control Service



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

plugin

plugin

plugin

plugin

VDK automates and abstracts

Data Teams

IT configures and extends

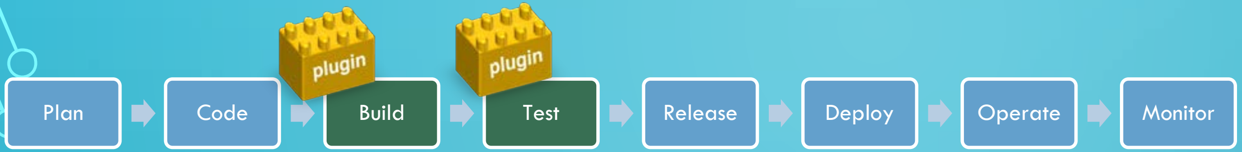
Data Teams

Infra & Operations Team

Establish policies

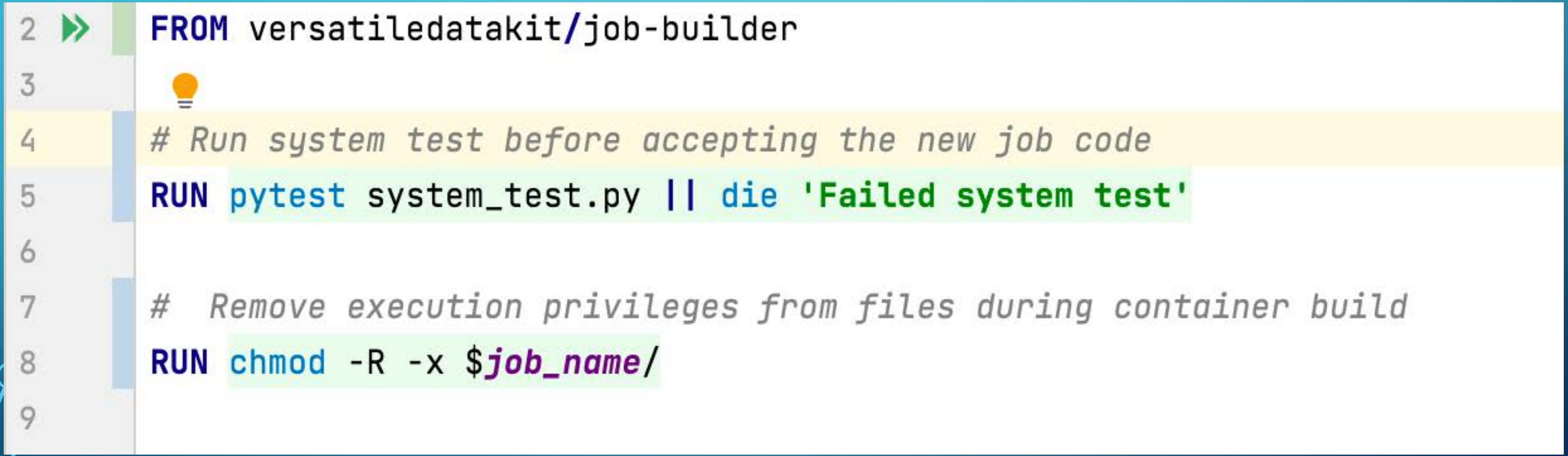
Extensible

Versatile Data Kit Control Service



plugin
plugin

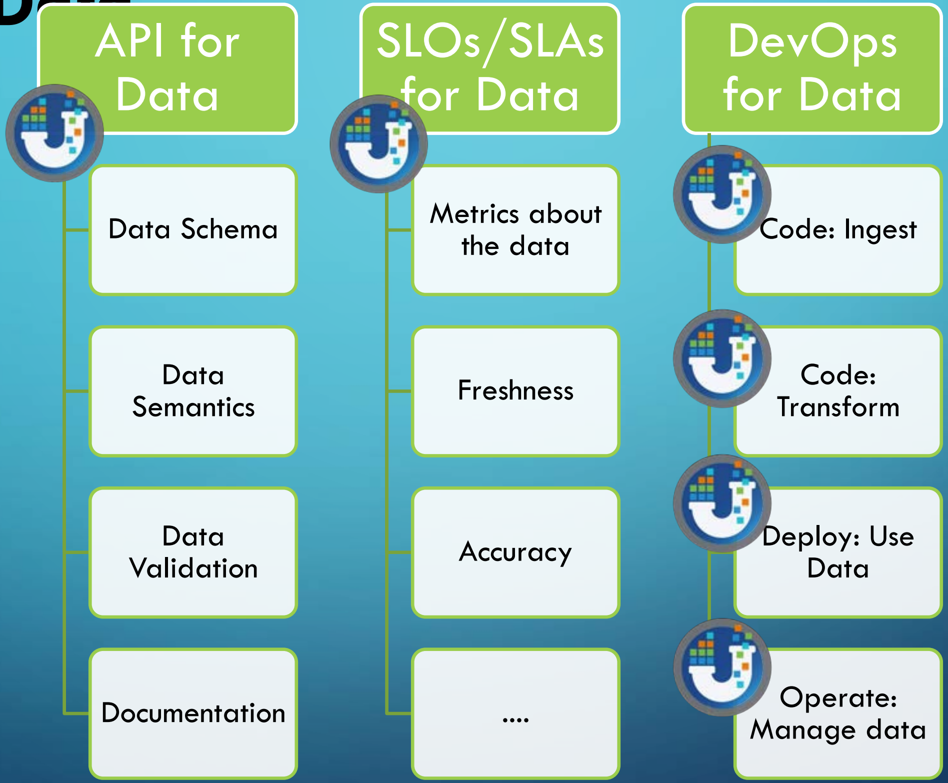
Plan → Code → Build → Test → Release → Deploy → Operate → Monitor



```
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/

```

Establish standard system tests and security hardening

DataOps for Data



API for Data
- Data Schema
- Data Semantics
- Data Validation
- Documentation

SLOs/SLAs for Data
- Metrics about the data
- Freshness
- Accuracy
- ....

DevOps for Data
- Code: Ingest
- Code: Transform
- Deploy: Use Data
- Operate: Manage data

Thank you

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

DataOps: DevOps for Data

Ineffcient 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

Data Team

Data Team

Data Team

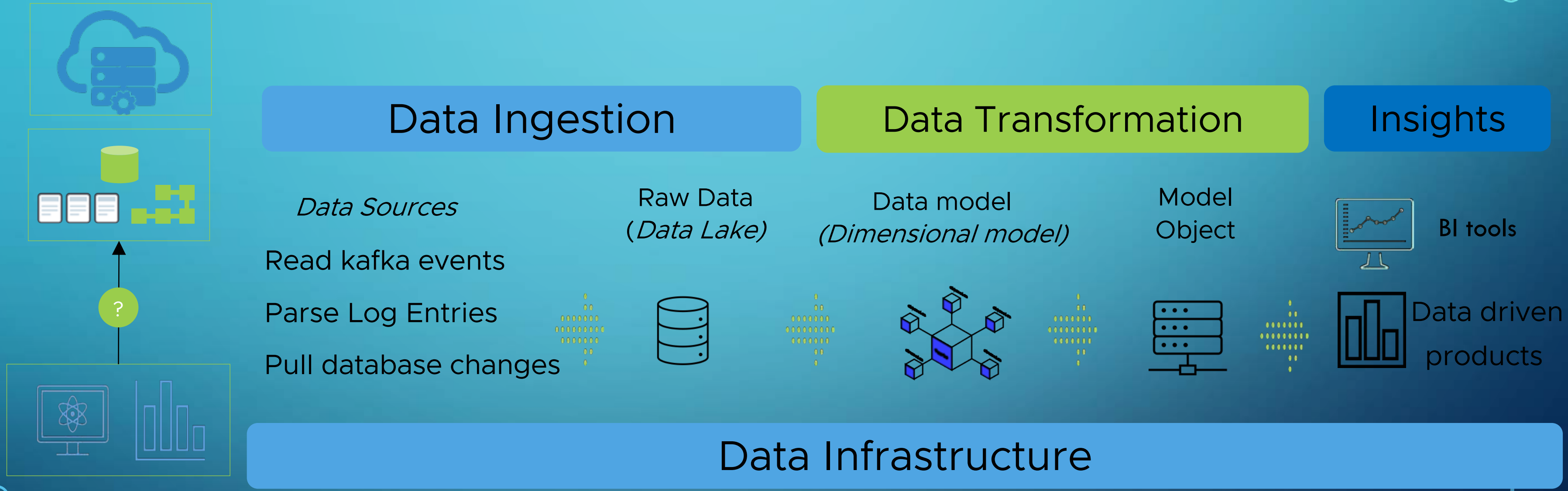
Data Team

Data Team

Infra & Operations Team

Blurred lines of responsibility

Building a data app is hard



Data Ingestion

Data Transformation

Insights

Data Sources

Read kafka events

Parse Log Entries

Pull database changes

Raw Data
(*Data Lake)*

Data model
(Dimensional model)

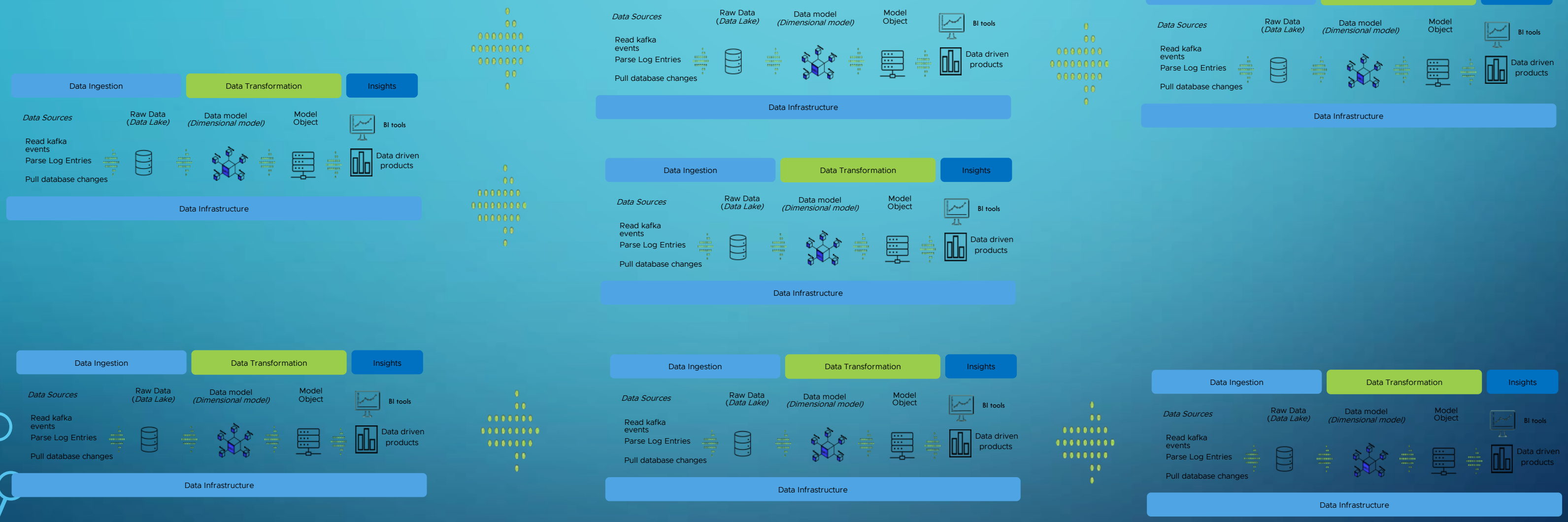
Model
Object

BI tools

Data driven
products

Data Infrastructure

Complexity increases



Data Ingestion | Data Transformation | Insights

Data Sources

Raw Data
(Data Lake)

Data model
(Dimensional model)

Model
Object

BI tools

Read kafka
events
Parse Log Entries
Pull database changes

Data driven
products

Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure

Data Ingestion | Data Transformation | Insights
Data Sources
Raw Data
(Data Lake)
Data model
(Dimensional model)
Model
Object
BI tools
Read kafka
events
Parse Log Entries
Pull database changes
Data driven
products
Data Infrastructure