Testing in DataOps

How to maximize your impact

Vincent Goris
Info Support Nederland

Vincent Goris

Analyst, chapter lead Data/AI for healthcare unit

vincent.goris@infosupport.com
www.linkedin.com/in/vincent-goris



We



- Data
- Sharing Knowledge
- Building Data savy organization
- Brick by brick



Agenda

- DevOps, Database DevOps and DataOps
- DataOps Testing
- Roadmap to DataOps
- Wrap up & Questions

DevOps vs Database DevOps vs DataOps

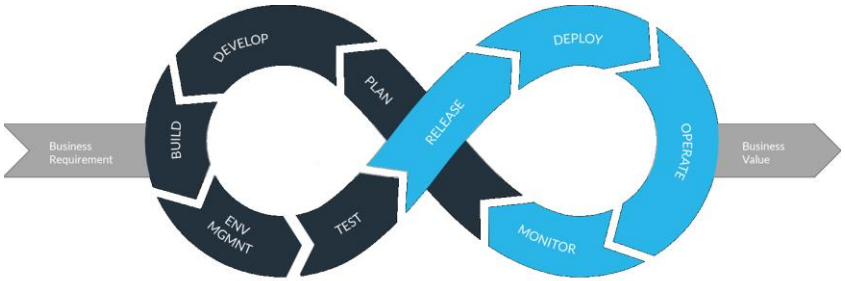




“DevOps is the ‘father’ of DataOps”

DevOps

- Shared Responsibility
- Automating, Monitoring and Testing
- Shortened release cycle
- Feedback cycle
- More value for customer

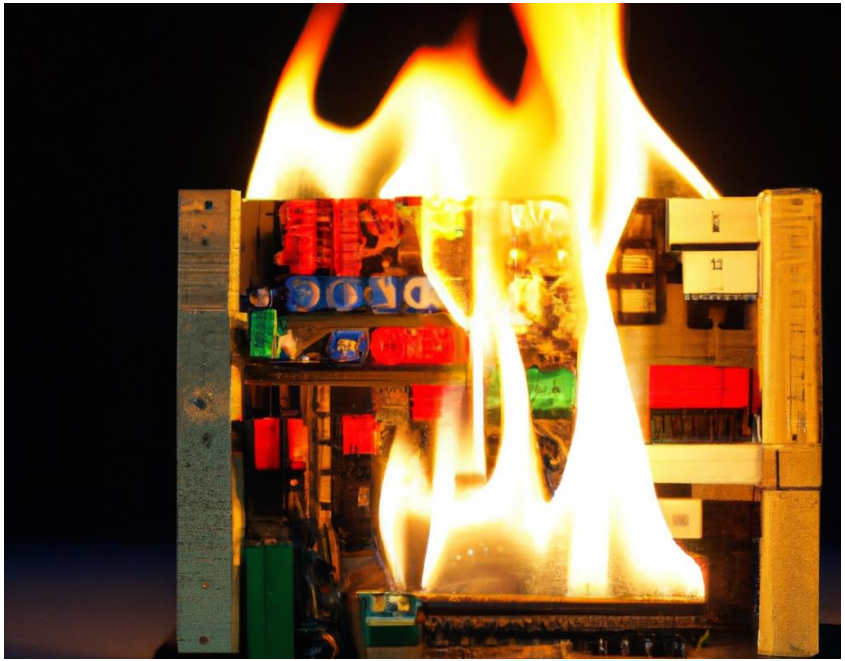


DevOps in practice

- Everything automated
- Git
- CI/CD
  - Multiple times a day
- DTAP
- Unittests

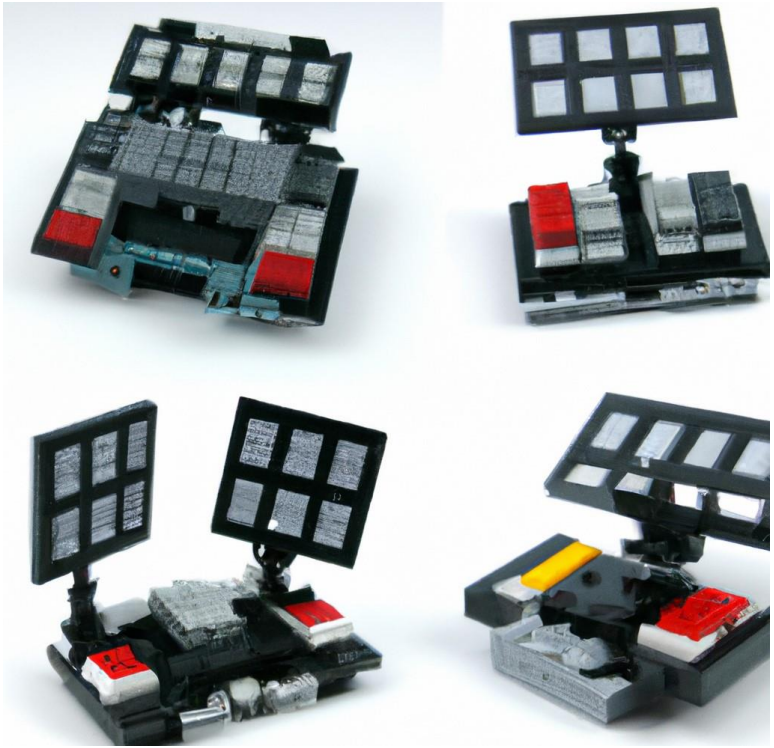
Database DevOps

- Changes in Production
- One time use solutions



Database DevOps

- Database1
- Database2
- Database2.final
- Database2.4
- Database2.reallyfinal
- Database2.final.final

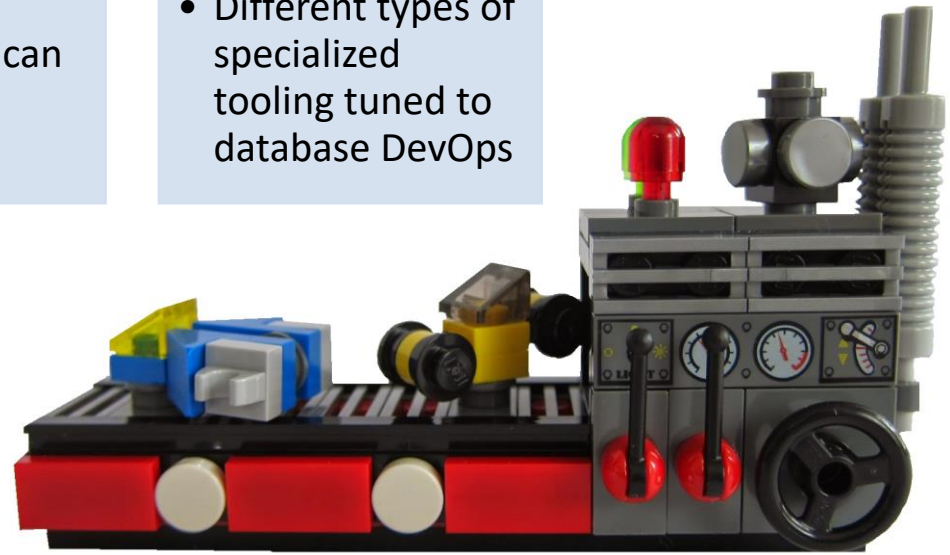


Database DevOps

- (Azure) DevOps
- Development of database through GIT
  - Version control
- Automated deployment
- Testing deployment
  - Make sure nothing unexpected breaks
  - Unit tests

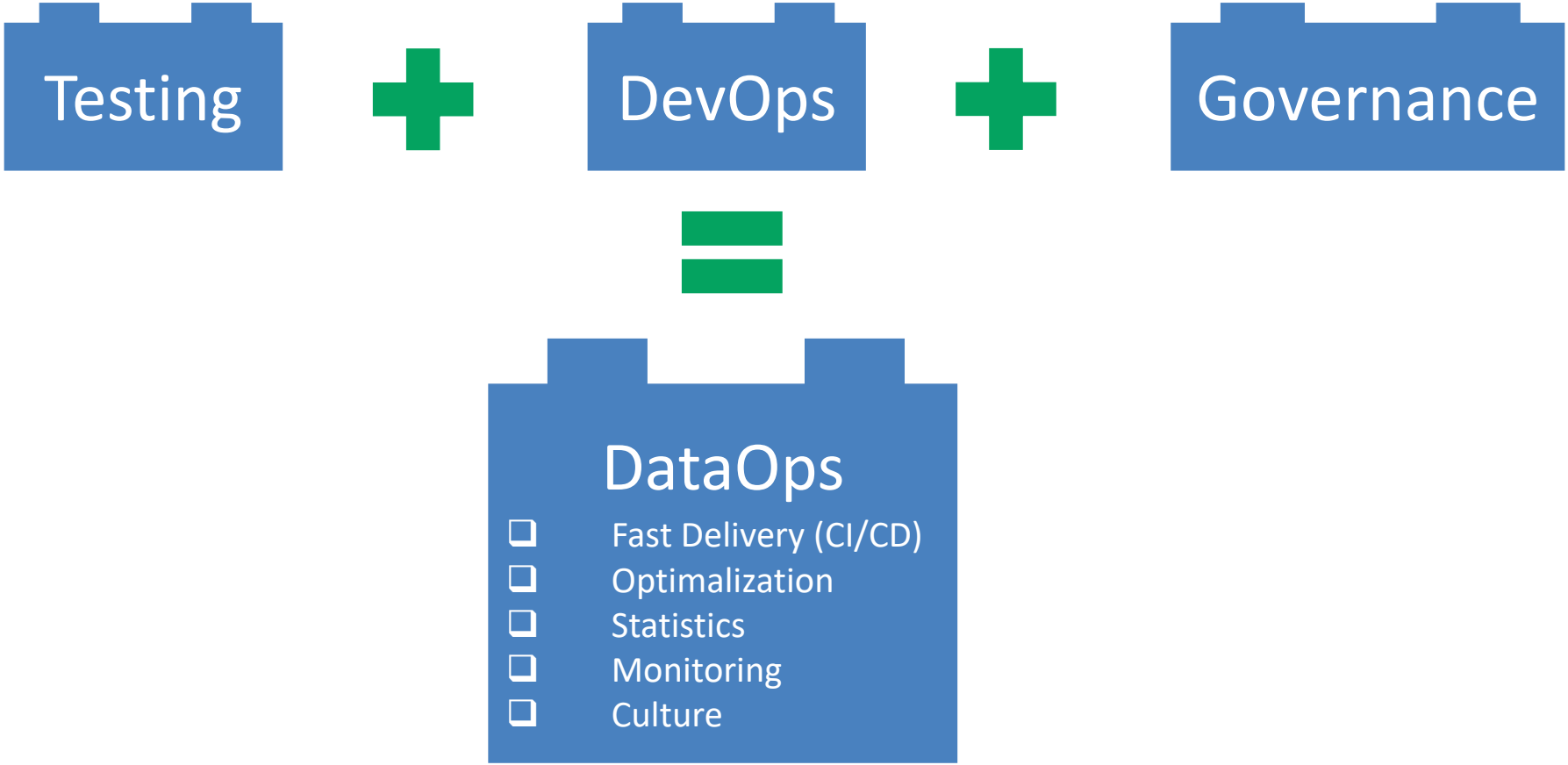
CI/CD

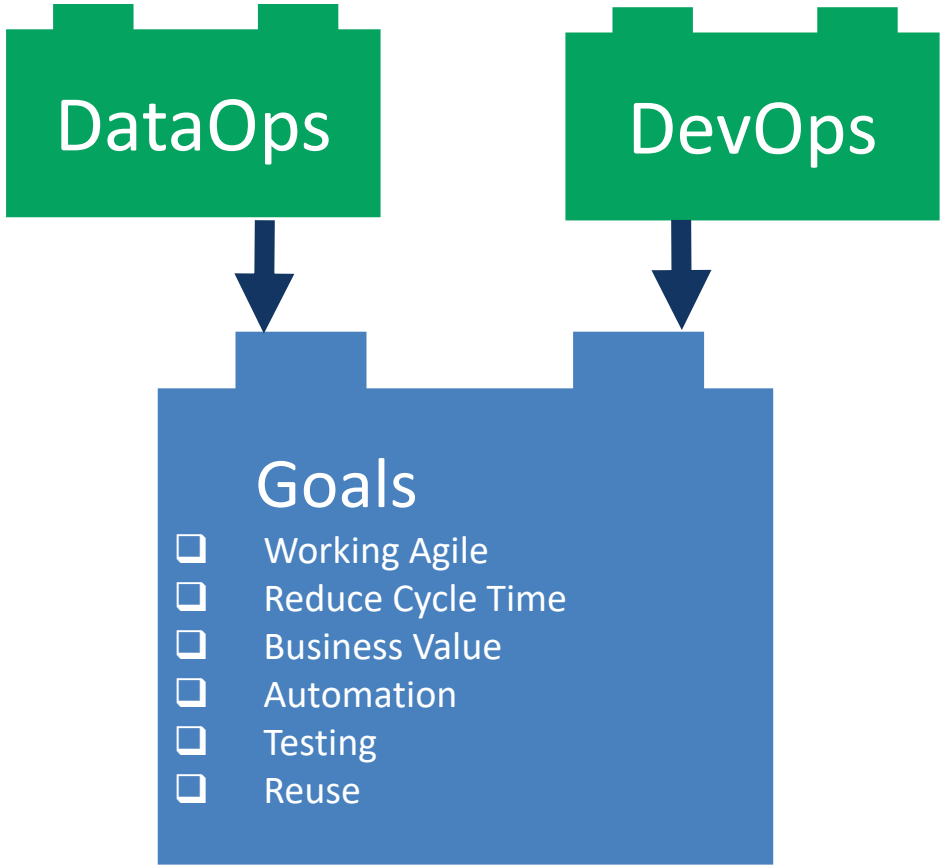
Versioning	Re-use	Tooling
• Version your data and solutions	• Make sure components can be re-used	• Different types of specialized tooling tuned to database DevOps

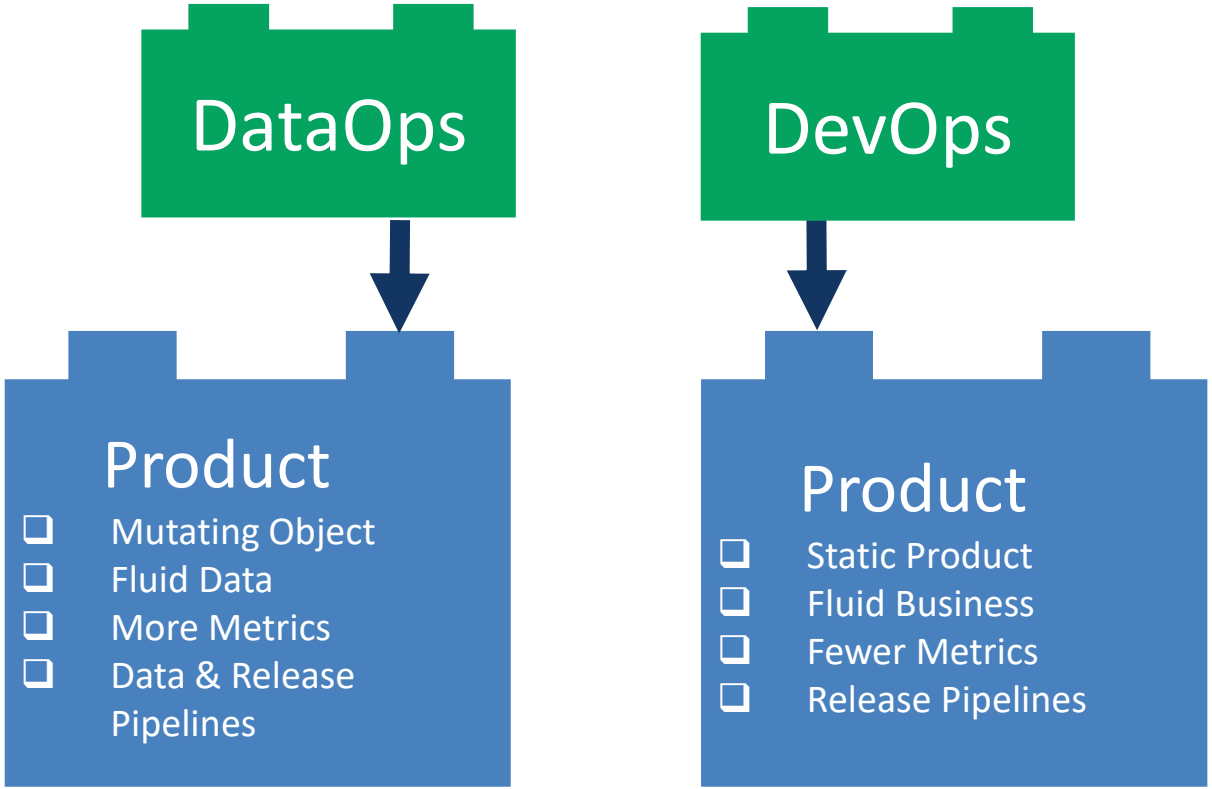


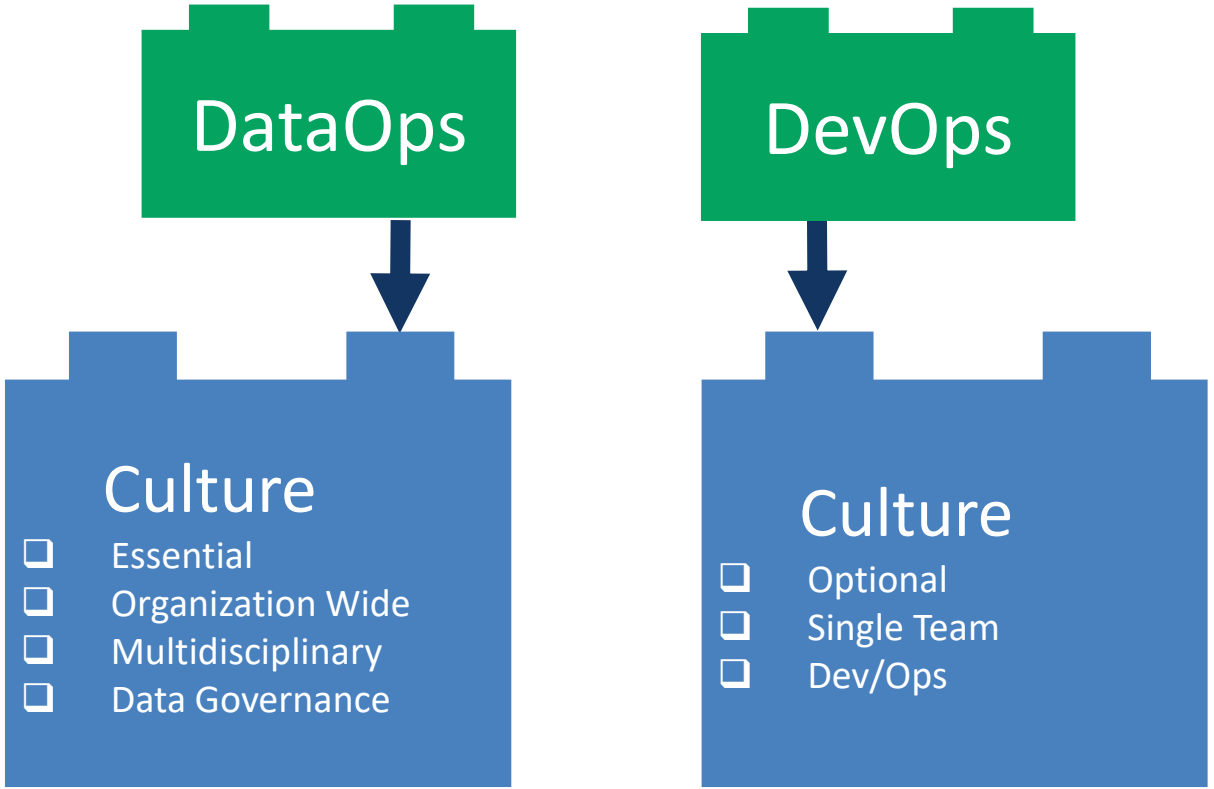
What is DataOps?

- Beyond Database DevOps
- Statistical process control
- Testing, testing and more testing
- Data Quality
- Governance

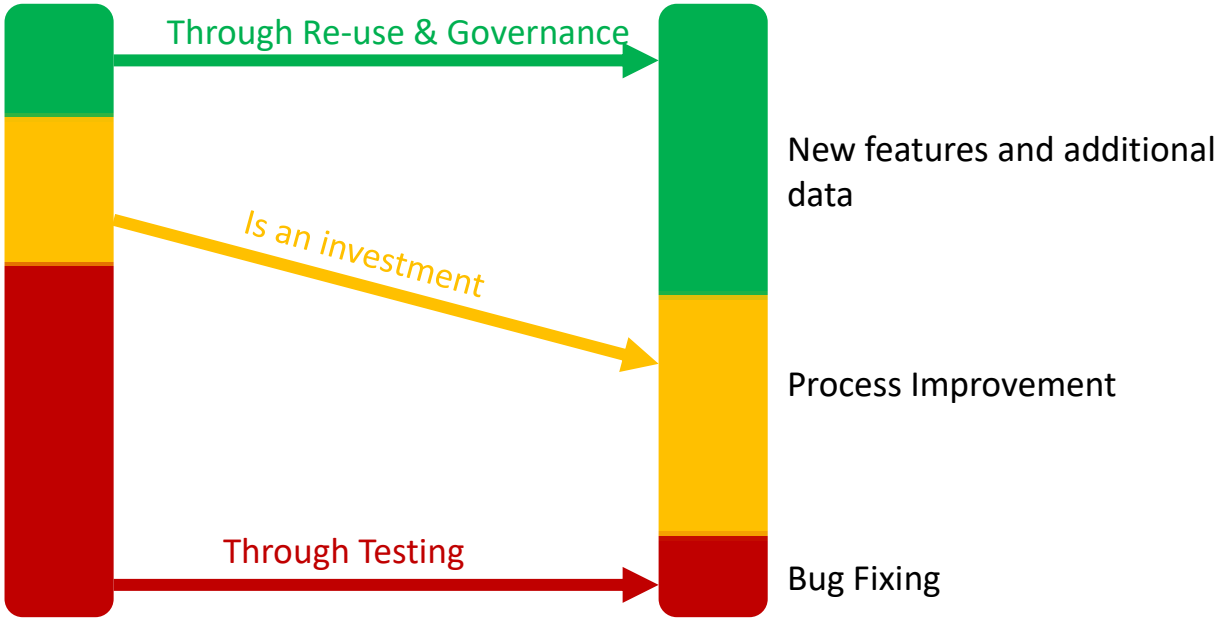

Testing
DevOps
Governance
DataOps
Fast Delivery (CI/CD)
Optimalization
Statistics
Monitoring
Culture


DataOps
DevOps
Goals
Working Agile
Reduce Cycle Time
Business Value
Automation
Testing
Reuse

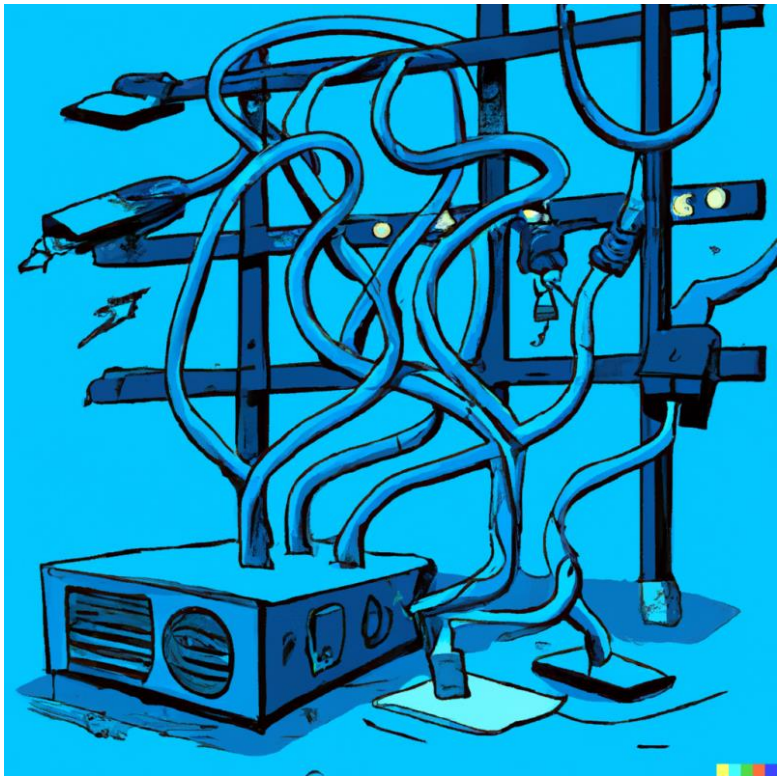

DataOps
DevOps
Product
Mutating Object
Fluid Data
More Metrics
Data & Release Pipelines
Product
Static Product
Fluid Business
Fewer Metrics
Release Pipelines


DataOps
Culture
Essential
Organization Wide
Multidisciplinary
Data Governance
DevOps
Culture
Optional
Single Team
Dev/Ops

Impact!


Through Re-use & Governance
Is an investment
Through Testing
New features and additional data
Process Improvement
Bug Fixing

Testing in DataOps



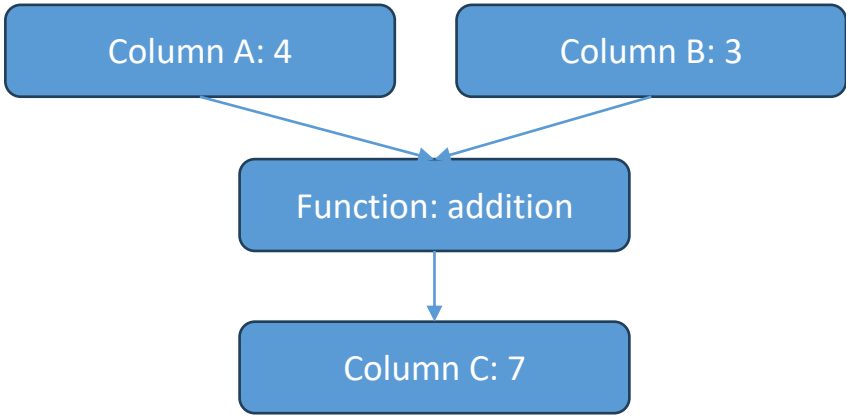
Trust is nice, Testing is better

Testing Basics

- Unittest
- Reducing dependencies
- Docker
- AAA
- Naming Conventions

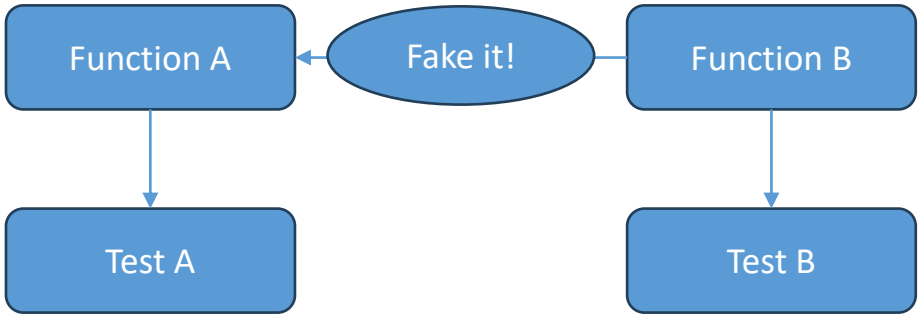
Unittests

- Test pieces of your code
- Independent
- Maintainable
- Coverage
- Fully automated
- Predictable



Reducing dependencies

- Large solution with dependencies
- Dependency Injection
- Mocking Framework



Docker

- Temporary Databases for testing purposes



AAA

Arrange → Act → Assert





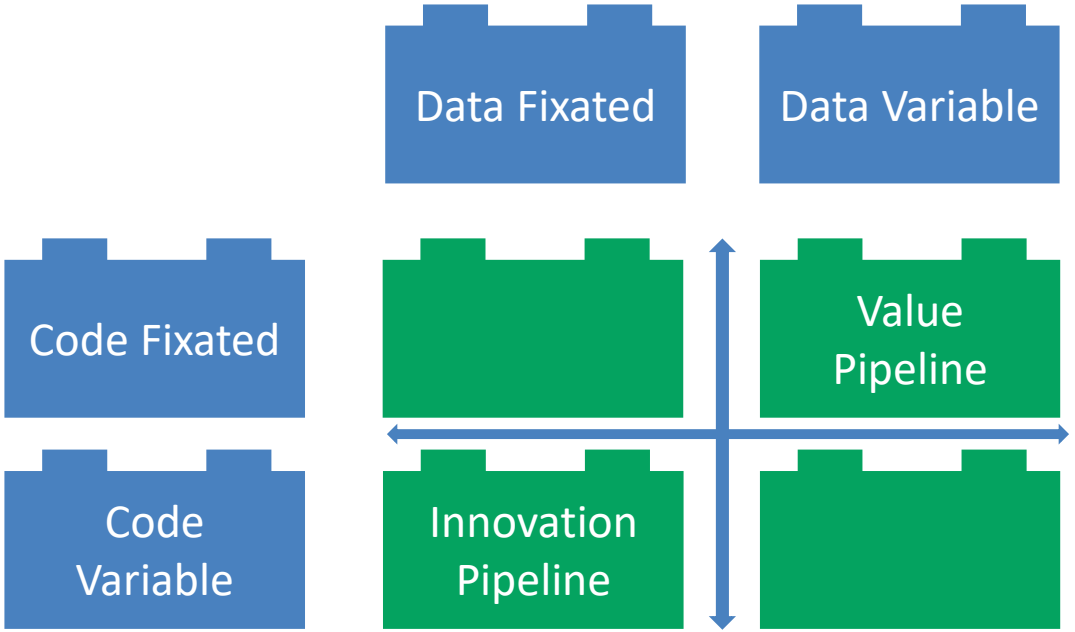


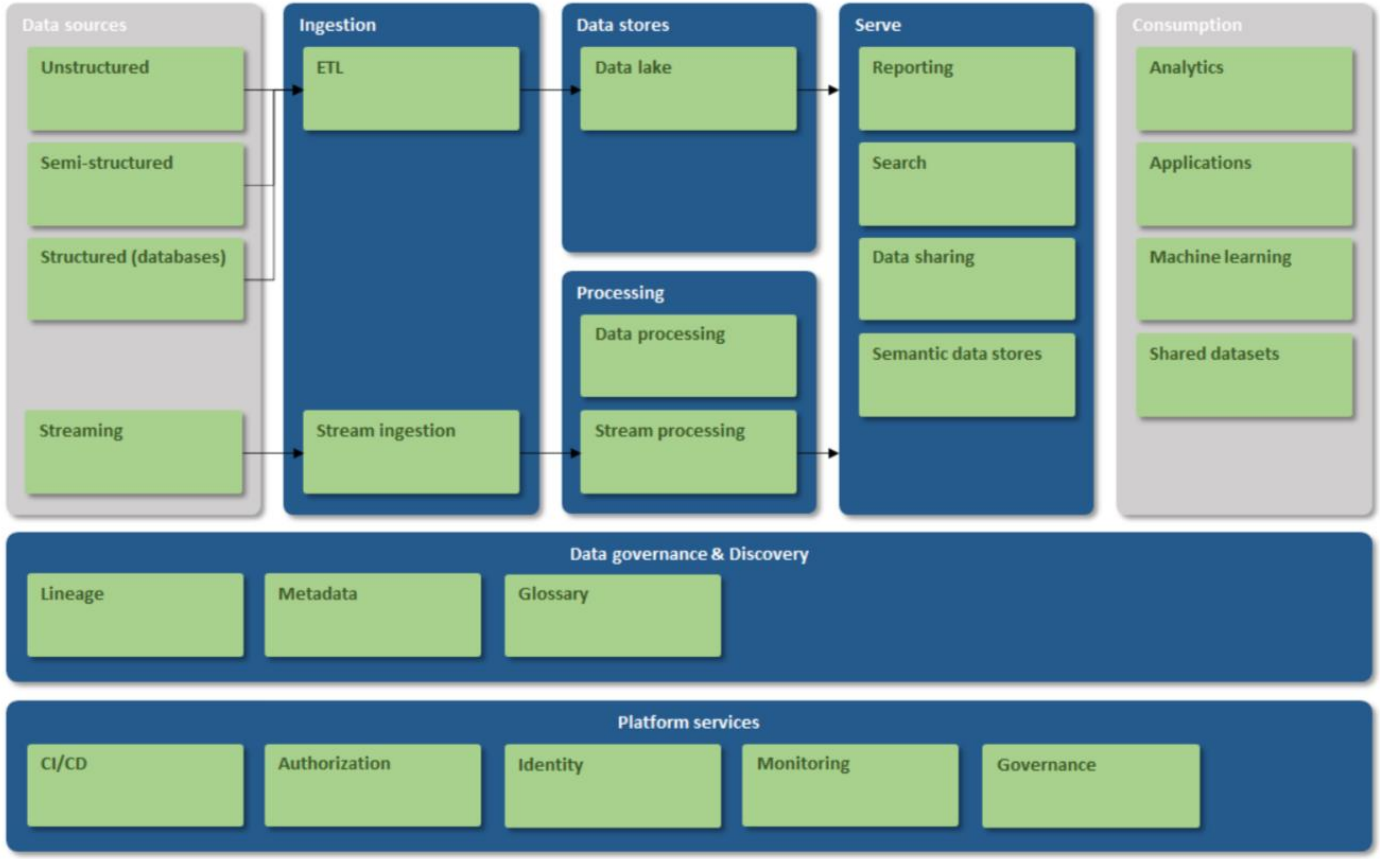

sqlpass.de
PASS

Perspective

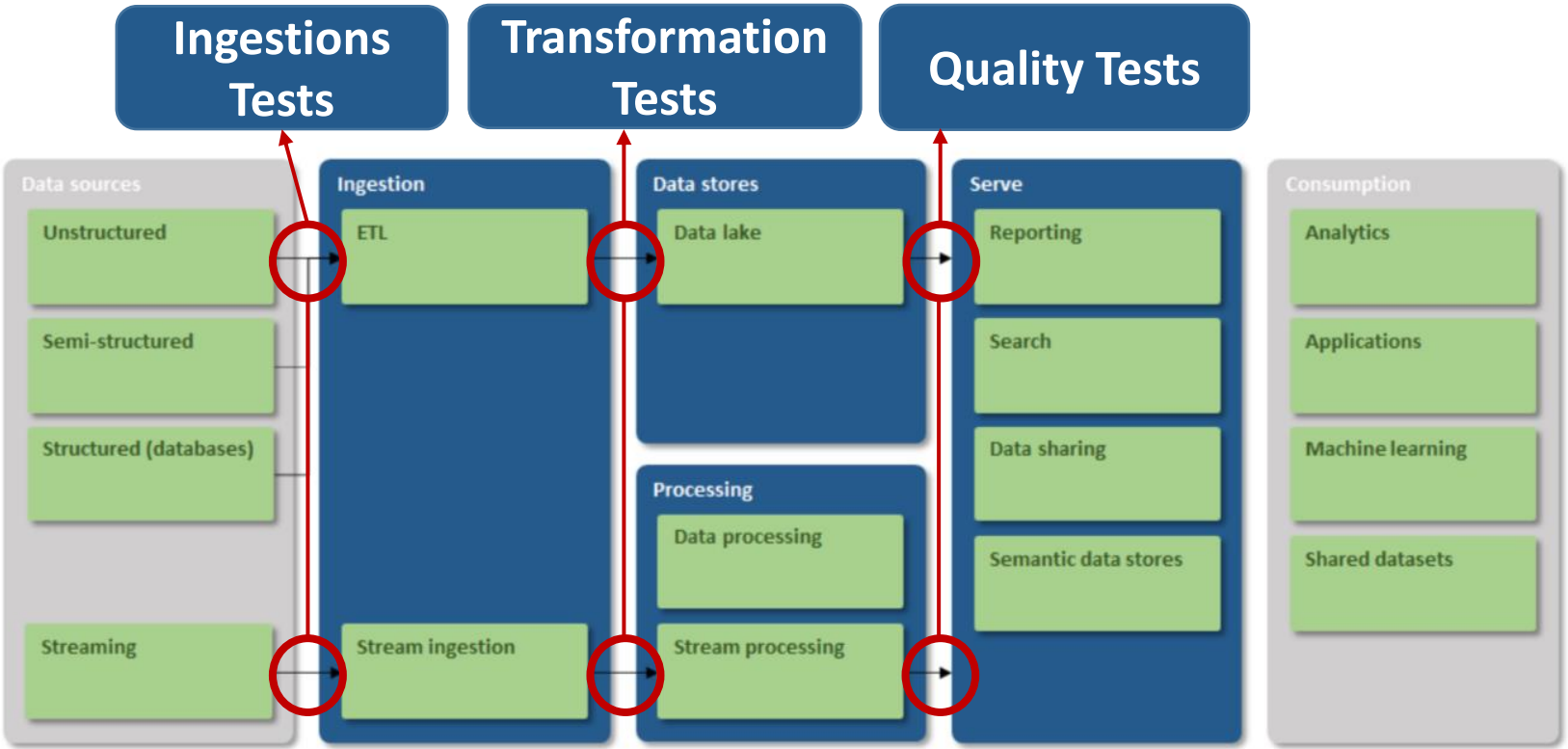
- Innovation Pipeline
  - Database DevOps
  - Does it work?
  - Changes in code don’t break your solution
- Value Pipeline
  - DataOps
  - Does it work with the data I receive?
  - The big bad world outside does not break your solution

Data Quality and Testing




Data sources
Unstructured
Semi-structured
Structured (databases)
Streaming
Ingestion
ETL
Stream ingestion
Data stores
Data lake
Processing
Data processing
Stream processing
Serve
Reporting
Search
Data sharing
Semantic data stores
Consumption
Analytics
Applications
Machine learning
Shared datasets
Data governance & Discovery
Lineage
Metadata
Glossary
Platform services
CI/CD
Authorization
Identity
Monitoring
Governance


PASS

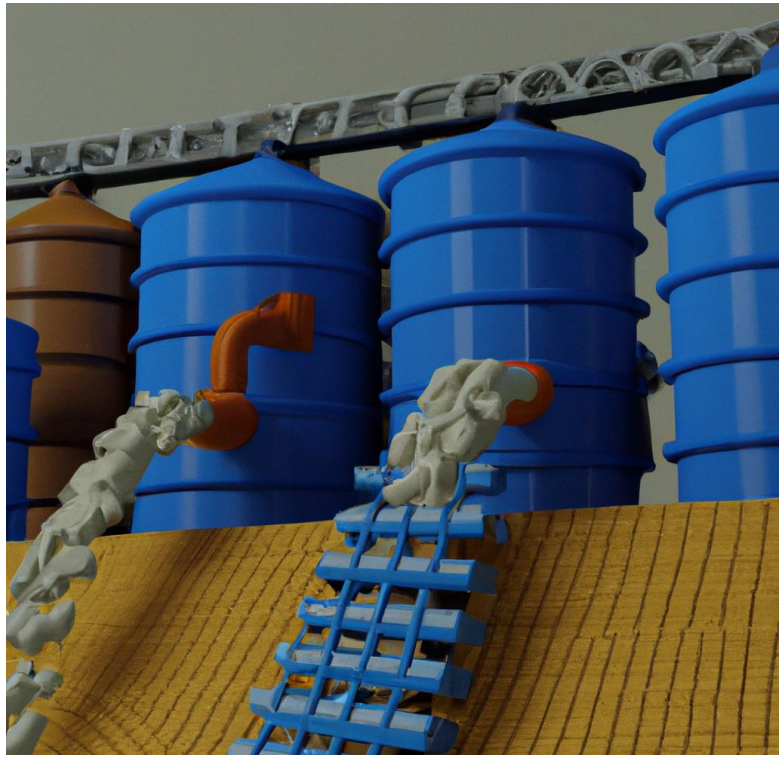

Ingestions Tests
Transformation Tests
Quality Tests
Data sources
Unstructured
Semi-structured
Structured (databases)
Streaming
Ingestion
ETL
Stream ingestion
Data stores
Data lake
Processing
Data processing
Stream processing
Serve
Reporting
Search
Data sharing
Semantic data stores
Consumption
Analytics
Applications
Machine learning
Shared datasets

Use Case


dit ijsje steunt u
roeiers van ORCA
Zopie is in 2005 opgericht door een
roeier met passie voor ijs en wordt volledig
door studenten van de roeivereniging Orca.
inkel is zo duurzaam mogelijk en als enige in Utrecht
open. Zo kunt u elke dag van biologisch ijs
koffie genieten.!!

Ingestion Tests

- Frequency
- Data contract
- Hard checks
- Statistical distribution
- Privacy and personal data
- Error handling



Ingestion Tests

Icecream sales data

- Weekly
- Revenue, DateTime

Weather Data

- Weekly
- 49 Columns



Ingestion Tests

```python
import unittest
from datetime import datetime, timedelta

class TestDataIngestion(unittest.TestCase):
    def setUp(self):
        # Replace this with your data ingestion setup code if needed
        pass


    def test_weekly_data_ingestion(self):
        # Arrange: Get the latest data date from your data source
        latest_data_date = self.get_latest_data_date()

        # Act: Calculate the difference in days between the current date and the latest data date
        current_date = datetime.now().date()
        days_difference = (current_date - latest_data_date).days

        # Assert: Check if the data is ingested weekly (within 7 days of the current date)
        self.assertLessEqual(days_difference, 7, "Data is not ingested weekly.")

    # Replace this function with your actual data ingestion logic to get the latest data date
    def get_latest_data_date(self):
        # Replace this with your logic to fetch the latest data date from your data source
        # For demonstration purposes, we'll assume the latest data date is 7 days ago
        return datetime.now().date() - timedelta(days=7)

if __name__ == '__main__':
    unittest.main()
```

Transformation Tests

- Data Type
- Unit Test
- Transformation
- # Columns
- # Rows



Transformation Tests

- Ice cream sales

Date	Max temp celcius	Max temp Fahrenheit	Radiation (in J/cm2) /	Amount of rain (in 0.1 mm)
18-7-2018	23	73	1616	-1
19-7-2018	29	84	2719	0
20-7-2018	25	77	2332	0
21-7-2018	23	73	796	85
22-7-2018	27	80	2625	0
7-23-2018	27	80	2310	0
24-7-2018	29	84	2646	0
25-7-2018	32	89	2183	0
26-7-2018	30	86	1623	-1
27-7-2018	35	95	2497	-1
28-7-2018	35	95	2569	0

Transformation Tests

- Data Types
- Unit Tests

DateTime	Icecream	Coffee	Total
18-07-2018 12:48:45 CET	€ 5,50	€0	€ 5,50
18-07-2018 12:53:42 CET	€ 20,50	€ 4,50	€ 24,50
18-07-2018 12:55:45 CET	€ 15,90	€ 0	€ 15,90
18-07-2018 12:59:53 CET	€ 0	€ 2,00	€ 2,00
18-07-2018 13:10:10 CET	€ 15,90	€ 0	€ 15,90

Date	Day	Icecream	Coffee	Total
18-07-2018	Wednesday	€ 1.746,10	€ 47,90	€ 1.794,00
19-07-2018	Thursday	€ 1.105,40	€ 29,90	€ 1.135,30
20-07-2018	Friday	€ 256,60	€ 51,80	€ 308,40

```python
import unittest
import pandas as pd

class TestDataTransformation(unittest.TestCase):
    def setUp(self):
        # Create a sample DataFrame with transformed daily sales data
        self.sample_data = {
            'Date': ['2023-09-01', '2023-09-01', '2023-09-02', '2023-09-02'],
            'Revenue': [100, 50, 200, 150]
        }
        self.df = pd.DataFrame(self.sample_data)

    def tearDown(self):
        # Clean up any resources if needed
        pass

    def test_unique_dates(self):
        # Arrange: Convert the 'Date' column to datetime and get unique dates
        self.df['Date'] = pd.to_datetime(self.df['Date'])
        unique_dates = self.df['Date'].unique()

        # Act: Check if each date appears only once
        date_appearance_counts = self.df['Date'].value_counts()

        # Assert: Ensure that each date appears only once
        self.assertTrue(date_appearance_counts.max() == 1, "Dates do not appear only once.")

if __name__ == '__main__':
    unittest.main()
```

Quality Tests

- Business Rules
- Completeness
- Data Drift
- Metadata

Date	Day	Icecream	Coffee	Total	Employee	Max temp celcius	Max temp Fahrenheit	Radiation (in J/cm2) /	Amount of rain (in 0.1 mm)
17-7-2018	Tuesday	€ 1.168,20	-€ 169,20	€ 999,00			73	1616	-1
18-7-2018	Wednesday	€ 1.746,10	€ 47,90	€ 1.794,00			84	2719	0
19-7-2018	thursday	€ 1.105,40	€ 29,90	€ 1.135,30	2	25	77	23	
20-7-2018	Friday	€ 256,60	€ 51,80	€ 308,40	1	23	73	796	85
21-7-2018	Saturday	€ 1.109,75	€ 39,60	€ 1.149,35	1	27	80	2625	0
22-7-2018	Sunday	€ 1.030,25	€ 25,50	€ 1.055,75	1	27	80	2310	0
7-23-2018	Monday	€ 1.059,75				29	84	2646	0
24-7-2018	Tuesday	€ 1.128,95	€ 27,50	€ 1.156,45	1	32	89	2183	0
25-7-2018	Wednesday	€ 993,20	€ 36,50	€ 1.029,70	2	30	86	1623	-1
26-7-2018	Thursday	€ 783,55	€ 46,00	€ 829,55	2	35	95	2497	-1
27-7-2018	Friday	€ 1.050,15	€ 2.132,50	€ 1.082,65	1	35	95	2569	0
28-7-2018	Saturday	€ 812,75	€ 41,00	€ 853,75			75	1570	52

Data Drift

Business R

Completeness

Business Rules

```python
import unittest
import pandas as pd

class TestDataTransformation(unittest.TestCase):
    def setUp(self):
        # Create a sample DataFrame with transformed daily sales data
        self.sample_data = {
            'Date': ['2023-09-01', '2023-09-01', '2023-09-02', '2023-09-02'],
            'Revenue': [100, 50, 200, 150]
        }
        self.df = pd.DataFrame(self.sample_data)

    def tearDown(self):
        # Clean up any resources if needed
        pass

    def test_no_negative_revenue_after_transformation(self):
        # Arrange: Filter the DataFrame for daily sales and calculate daily revenue
        daily_sales = self.df.groupby('Date')['Revenue'].sum().reset_index()

        # Act: Check if any daily revenue is negative
        negative_revenue = daily_sales[daily_sales['Revenue'] < 0]

        # Assert: Ensure there are no negative daily revenues
        self.assertTrue(negative_revenue.empty, "Negative revenue values found after transformation.")

if __name__ == '__main__':
    unittest.main()
```



Let’s get started!

Roadmap to implement DataOps

1. Scan

- Data Driven Maturity
- DataOps Maturity

2. Educate

- People, Process & Technology
- Lean manufacturing

3. Pilot

- Business Use-cases
- CI/CD
- Showcase
- Data Quality

4. Productize

- Metadata, catalog and governance
- Automate & Smart
- Data Quality

5. Grow



infoSupport
Solid Innovator

Data-driven maturity model

Collect & Store

Focus on orchestrating efforts to **set up** analytics .

- ✓ Make data available for analyzing,
- ✓ Basic reporting.
- ✓ Analytics and a data-driven way of working must become official.

Explore

Analytics is getting more **formal** and **organized**.

- ✓ Insight shifts from the manager to the analytics department
- ✓ Many of the prerequisites have been taken care of the next stage.
- ✓ Management views predetermined reports and KPI 's for more informed decisions.

Stream & Operationalize

Focus on **developing** and **operationalizing** prescriptive analytics .

- ✓ Now what to do in the future to optimize our profitability
- ✓ Operationalizing insight without manual actions
- ✓ Recommendation services prescribe products to customers they will more likely respond to.

Predict

Focus on **empowering** everyone in the organization.

- ✓ Allowing rapid integration of analytical features into products or services.
- ✓ Everyone is warmed up for change during the previous stages.
- ✓ Continually showing the success of an analytical approach.

Prescribe

Main source of **competitive advantage**

- ✓ The way the business is done is defined
- ✓ Innovative products and services can redefine the business model
- ✓ Analytical transformation is a journey that never ends. The organization keeps progressing and keeps redefining itself in everything it does.
- ✓ Continuous improvement is so deeply ingrained in the organization, it becomes second nature.

Roadmap to implement DataOps

1. Scan

- Data Driven Maturity
- DataOps Maturity

2. Educate

- People, Process & Technology
- Lean manufacturing

3. Pilot

- Business Use-cases
- CI/CD
- Showcase
- Data Quality

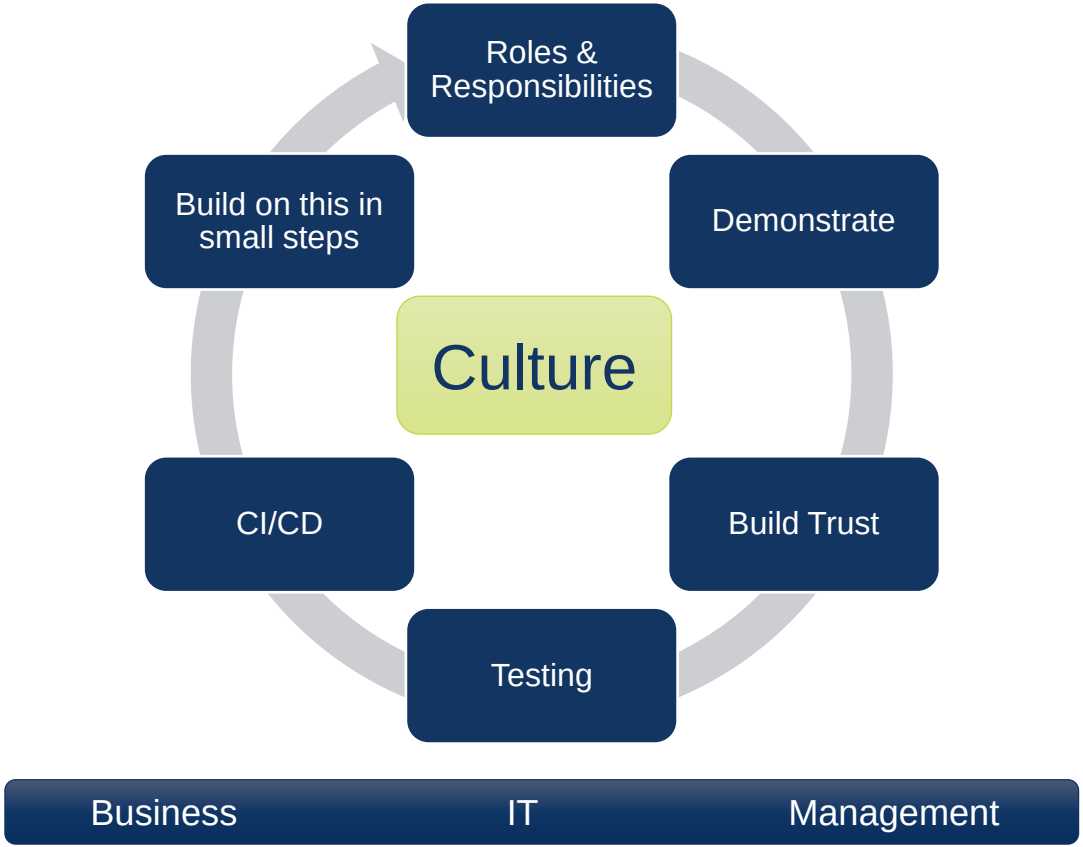
4. Productize

- Metadata, catalog and governance
- Automate & Smart
- Data Quality

5. Grow



Solid Innovator


Roles & Responsibilities
Demonstrate
Build Trust
Testing
CI/CD
Build on this in small steps
Culture
Business
IT
Management

Roadmap to implement DataOps

1. Scan

- Data Driven Maturity
- DataOps Maturity

2. Educate

- People, Process & Technology
- Lean manufacturing

3. Pilot

- Business Use-cases
- CI/CD
- Showcase
- Data Quality

4. Productize

- Metadata, catalog and governance
- Automate & Smart
- Data Quality

5. Grow



Solid Innovator

Pilot: start small with high impact

Data Quality

Start improving (Data) Quality

Showcase

Get the whole organization involved



CI/CD

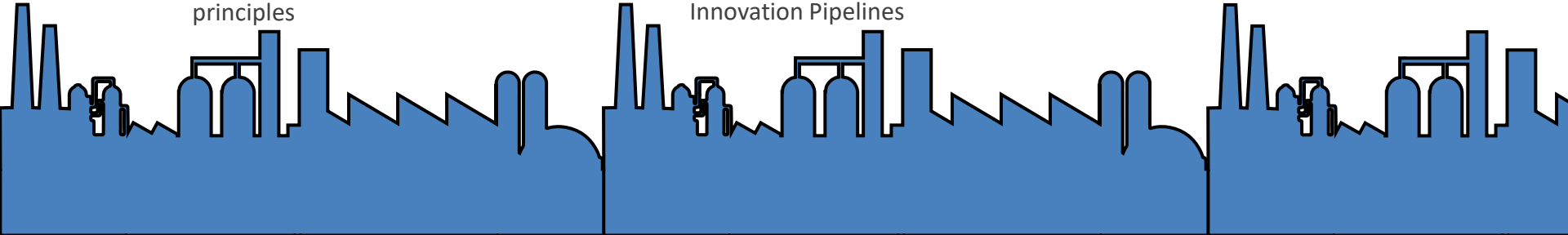
Start with basic CI/CD principles

Testing

First tests for Value and Innovation Pipelines

People

Culture is paramount



Roadmap to implement DataOps

1. Scan

- Data Driven Maturity
- DataOps Maturity

2. Educate

- People, Process & Technology
- Lean manufacturing

3. Pilot

- Business Use-cases
- CI/CD
- Showcase
- Data Quality

4. Productize

- Metadata, catalog and governance
- Automate & Smart
- Data Quality

5. Grow



Solid Innovator

Productize

- Policies & procedures
- Security & Privacy:
  - Infrastructure
  - Software
  - Data
- Monitoring and support



Roadmap to implement DataOps

1. Scan

- Data Driven Maturity
- DataOps Maturity

2. Educate

- People, Process & Technology
- Lean manufacturing

3. Pilot

- Business Use-cases
- CI/CD
- Showcase
- Data Quality

4. Productize

- Metadata, catalog and governance
- Automate & Smart
- Data Quality

5. Grow



Solid Innovator

It’s a marathon, not a sprint



THANK YOU TO OUR SPONSORS

PASS DEUTSCHLAND e.V.
Microsoft Data Platform Community

pmOne

SOLARWINDS

smart casual datadesign

Mainzer Datenfabrik

b.telligent
smart data. smart decisions.

HMS analytical software

ORAYLIS
turn your data into value

TIMEXTENDER

Microsoft

ITGAIN

SOLISYON
www.solisyon.de

GDS BUSINESS INTELLIGENCE GmbH

Thank you very much for your attention.

Vielen Dank für Eure Aufmerksamkeit.



sqlkonferenzeval.azurewebsites.net/8129818569561