Welcome to the Online Architect Workshop

DataOps in your Data Management Discipline

Samuel Kuruvilla – TCS
Remy Van Der Kleij – Informatica
Siddharth Rajagopal – Informatica

DATAOPS in Data Management

Sidd Rajagopal
Solution Architect - Informatica

Agenda

Why?

Why do Organizations need to consider DataOps?

What

What are some of the key components in DataOps?

How?

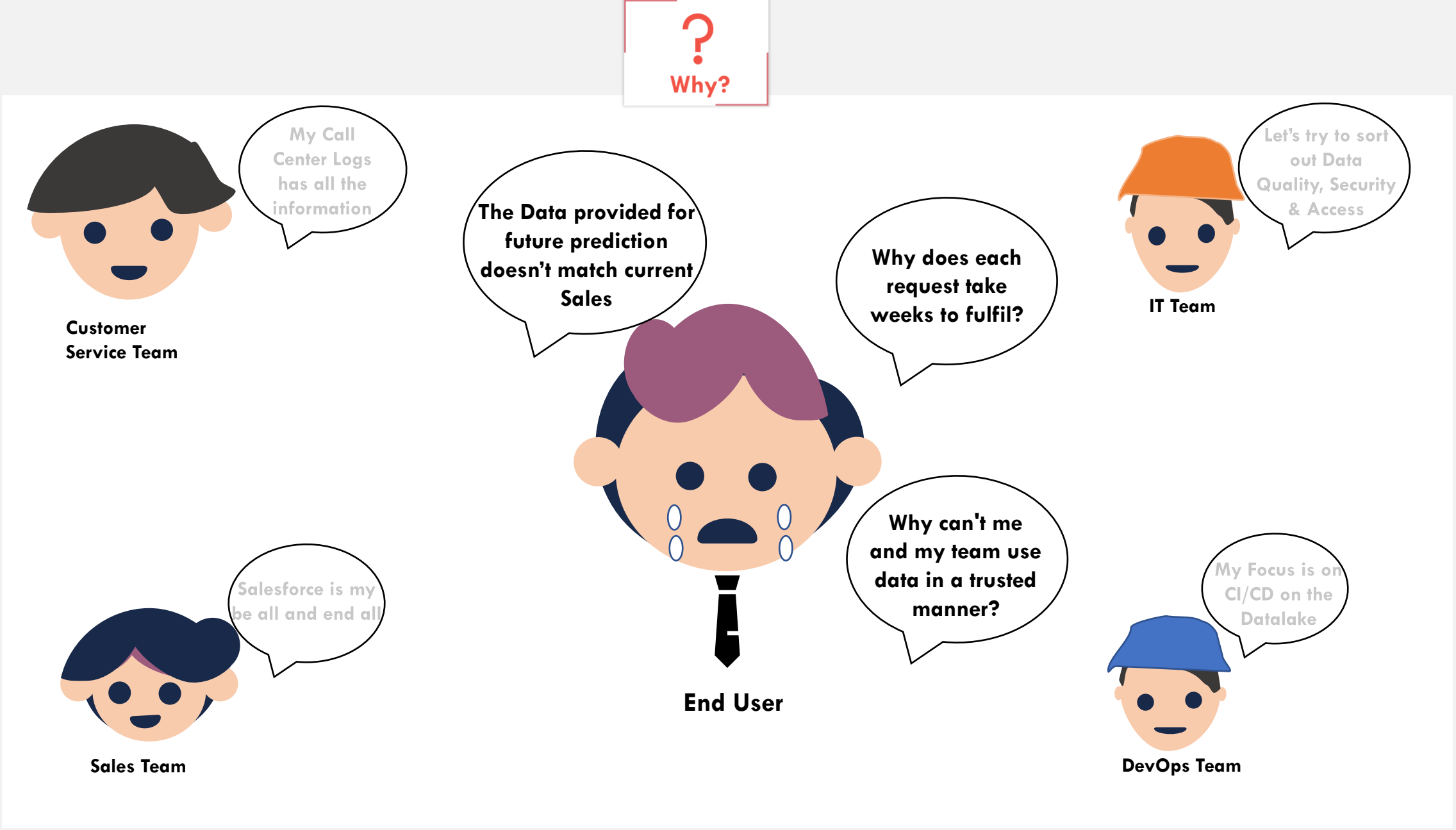
How do we go about designing & architecting?

Agenda

?

Why?

Why do Organizations need to consider DataOps?


Why?
My Call Center Logs has all the information
Customer Service Team
The Data provided for future prediction doesn't match current Sales
Why does each request take weeks to fulfil?
Let's try to sort out Data Quality, Security & Access
IT Team
End User
Why can't me and my team use data in a trusted manner?
Salesforce is my be all and end all
Sales Team
My Focus is on CI/CD on the Datalake
DevOps Team

Agenda

What

What are some of the key components in DataOps?

DataOps Myth

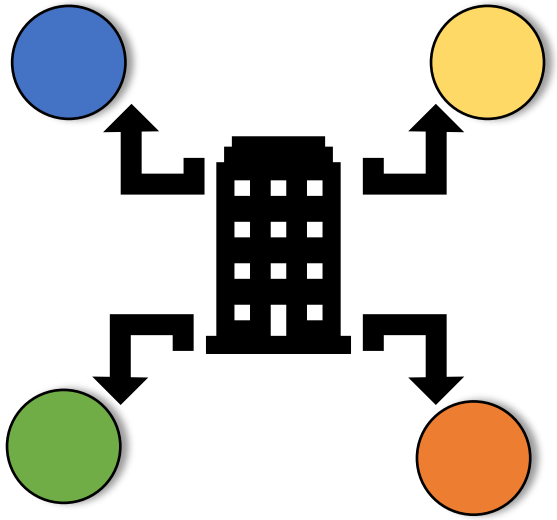
It’s not (only) a product

DataOps is a collaborative **data management** practice, really focused on improving **communication, integration and automation** of data flow between **managers and consumers of data** within an organization.

Photo Credit - https://www.dreamstime.com/

What

Key Components



Improving **communication**

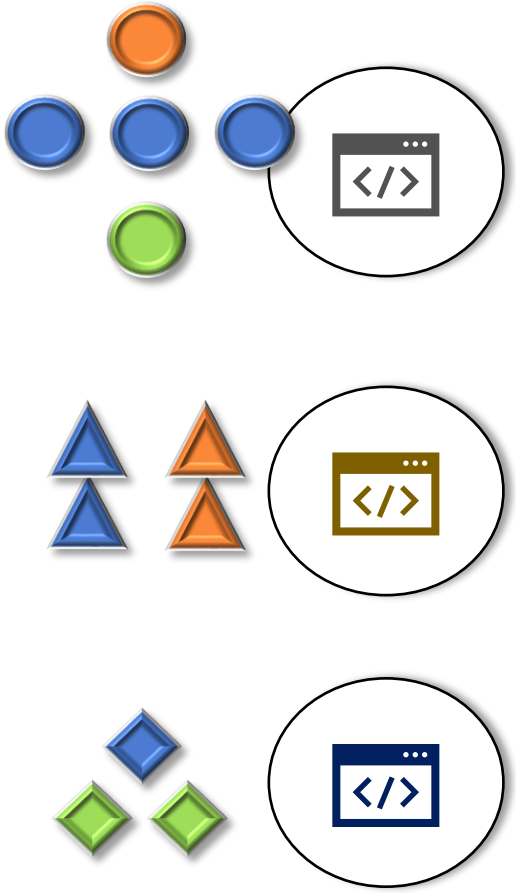
What

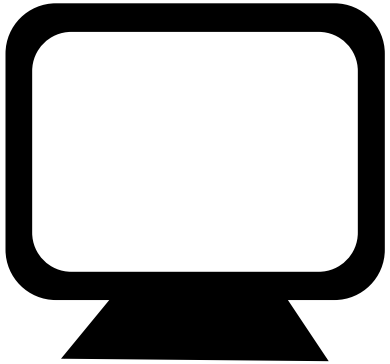
Agile Way of Working

Photo Credit - https://myva360.com/blog/how-to-become-an-agile-coach-in-2020

What

Metadata Driven Data Sharing

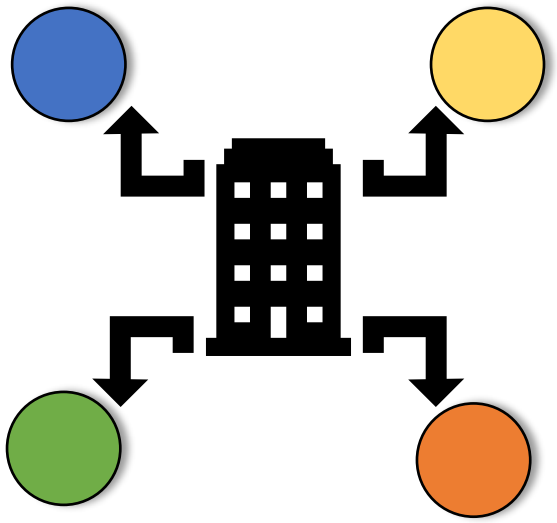




Metadata & Data Discovery

What

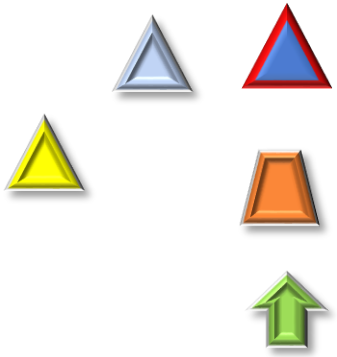
Key Components



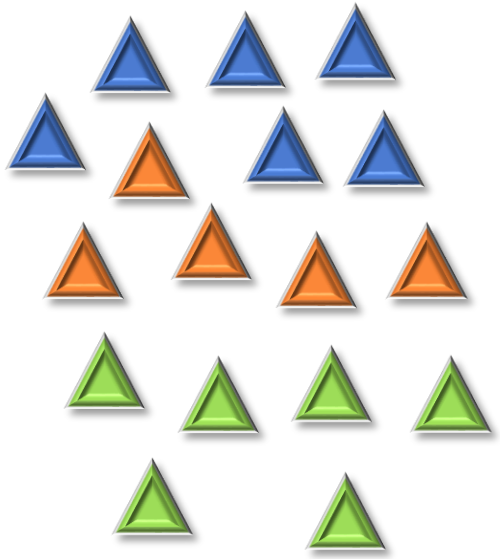
Improving **Integration**

What

Embedded Data Quality

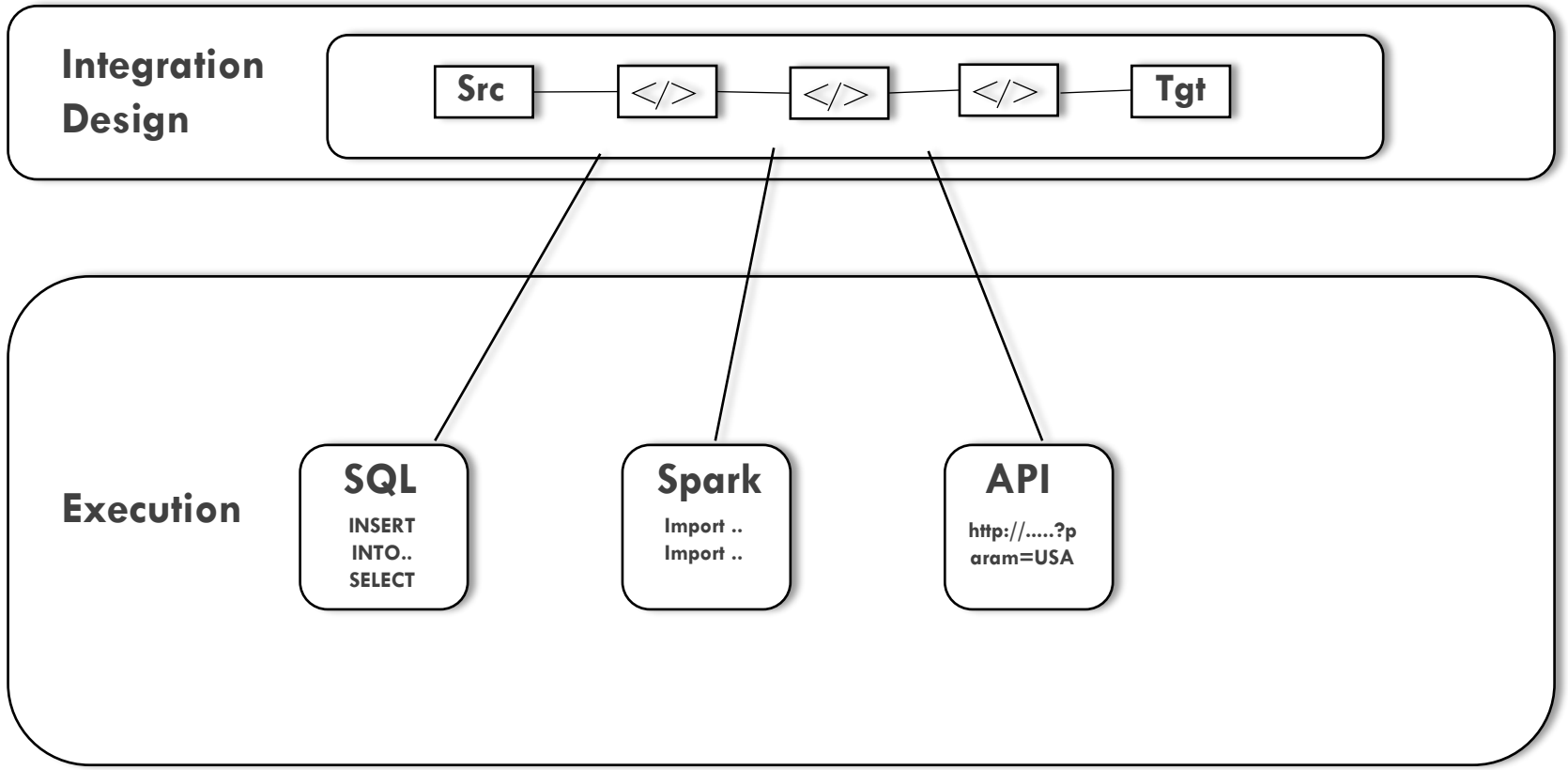


Data Quality as Design than a reaction



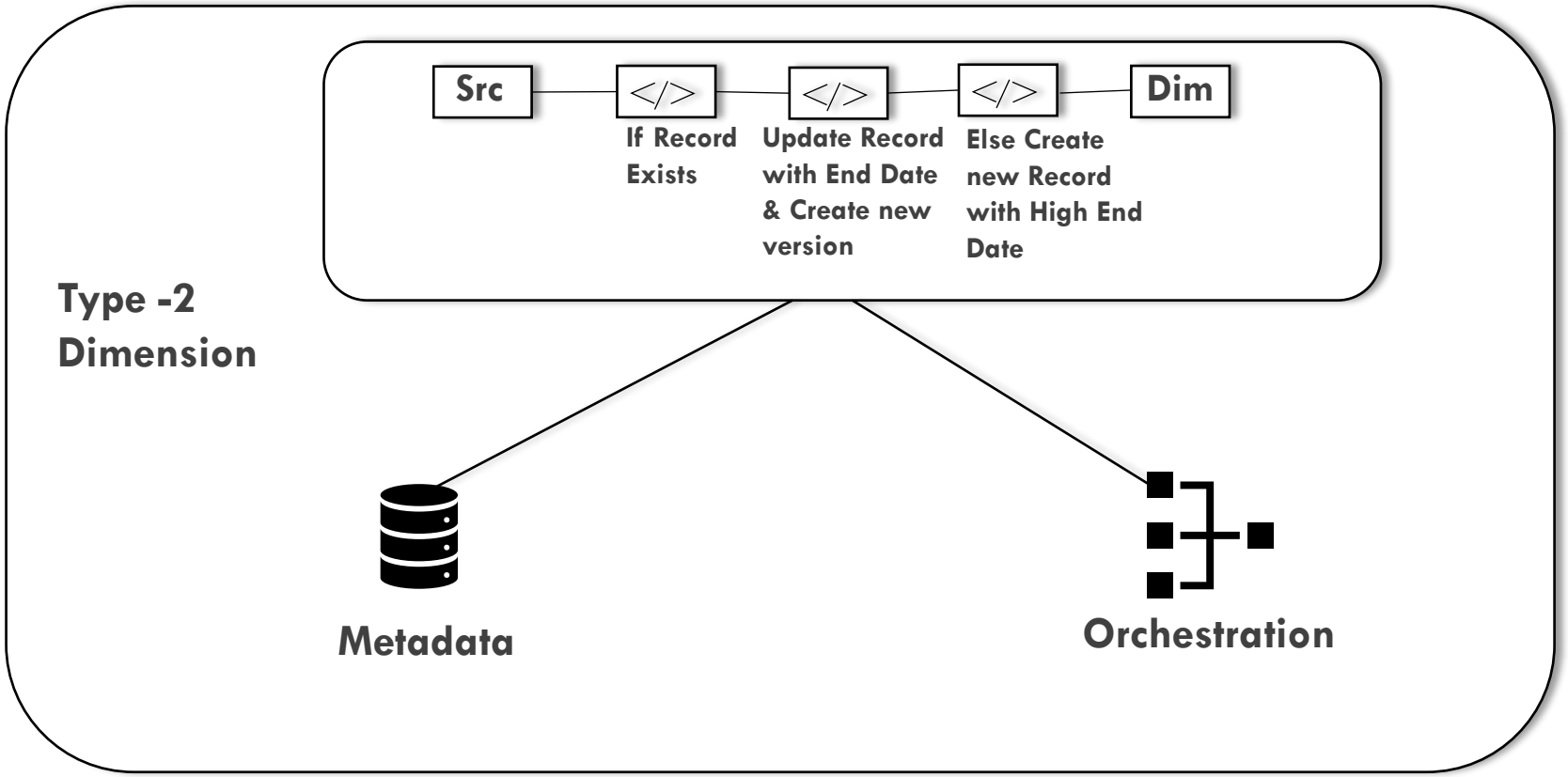
What

Design Driven Integration



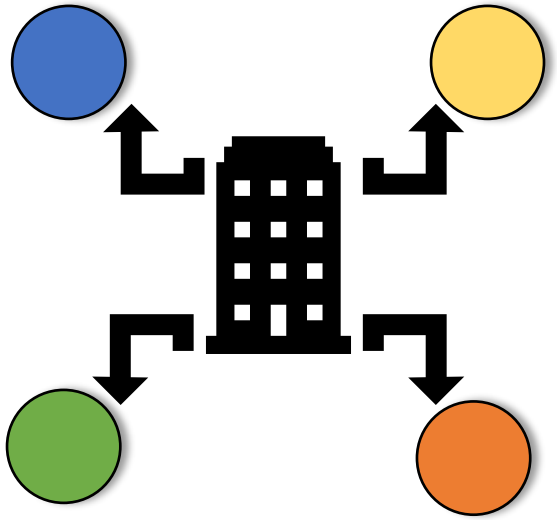
What

Templatize



What

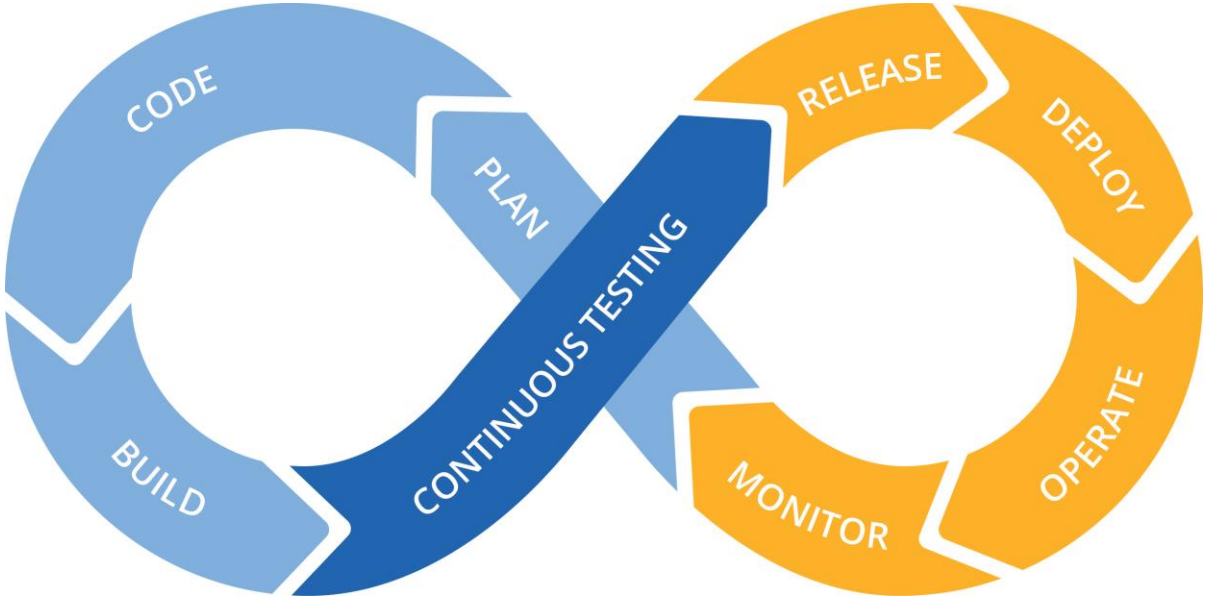
Key Components



Improving **Automation**

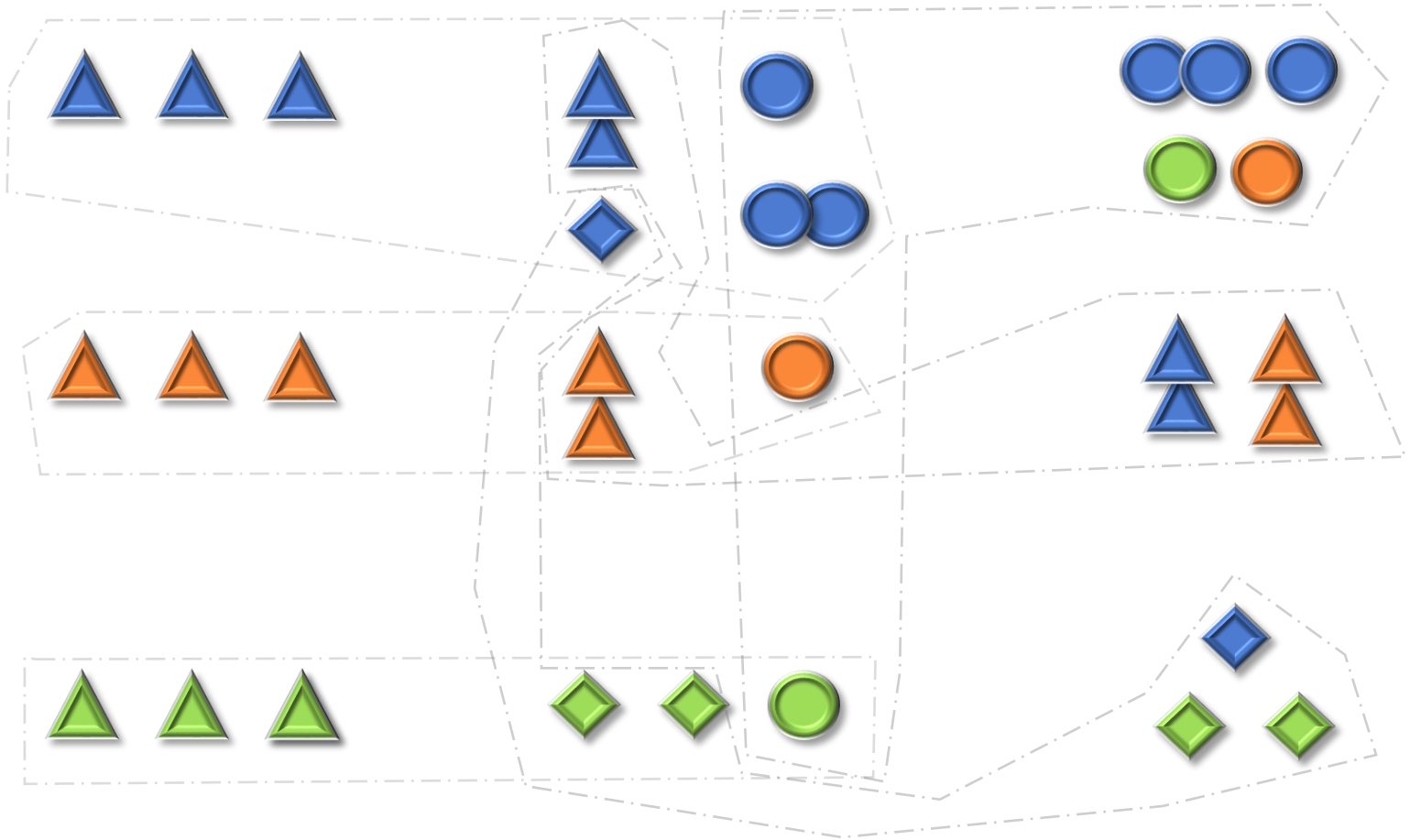
What

DevOps

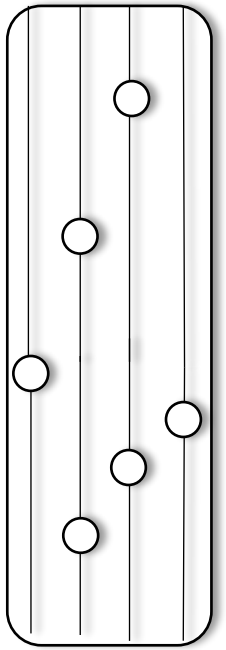

CODE
PLAN
BUILD
CONTINUOUS TESTING
RELEASE
DEPLOY
OPERATE
MONITOR

What

Infrastructure as a Platform



Infrastructure as a Platform



What

Augmented Data Management

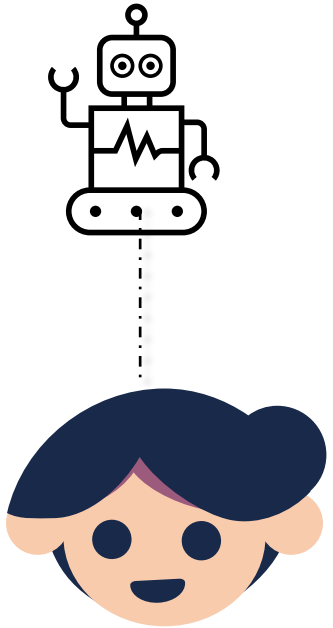
Inferences

Translations

Recommendations

Optimizations

Associations



Agenda

How?

How do we go about designing & architecting?

Agenda

How?

How do we go about designing & architecting?

Samuel Kuruvilla
Director, Cloud and Data Services
TCS

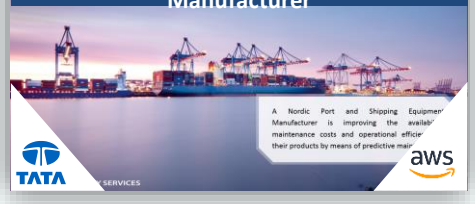
Data Ops Best Practices

A Practitioners Perspective

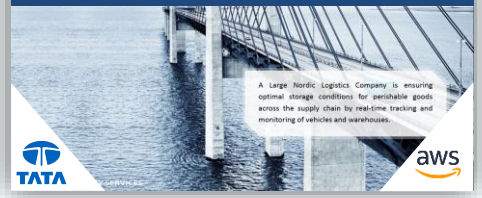
TATA CONSULTANCY SERVICES

We experience the benefits of DataOps through our experiences in multiple data led transformations of enterprises

Intelligent Logistics Data Platform for Nordic Cargo Handling Equipment Manufacturer



AI driven Operations Optimization for Nordic Logistics Company



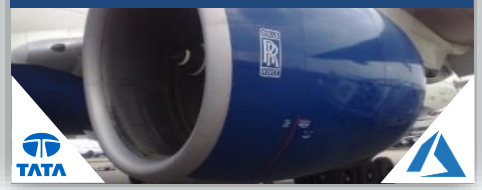
Remote Container Management for Nordic Logistics Company

Refinery 4.0 for Leading European Energy Company

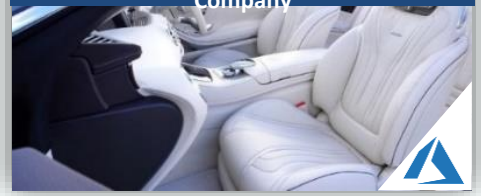
Predictive Maintenance Data Platform for Leading Indian Turbine Manufacturer

Fleet Management Data Platform for Leading US Engine Manufacturer

Power-by-the-Hour Platform for Leading British Power Systems Manufacturer



Connected Mobility DataPlatform for Leading European Automotive Parts Company



Digital Twins for Leading Global Oil & Gas Drilling Contractor

Device Management Platform for Leading Japanese Equipment Manufacturer

Pay-as-you-Go Telematics Solution for Leading Canadian Insurance Company

Data Platform for Nordic Engineering & Manufacturing Company

Connected Vessels Data Platform for Leading European Ship Building Company

AI based Combustion Tuning for Leading Japanese Power Plant Manufacturer

Connected Chillers Platform for Leading Asian HVAC Manufacturing Company

Connected Products & Solutions for Leading US Instrumentation Conglomerate

Connected Diagnostics for Leading Global Medical Equipment Manufacturing Company

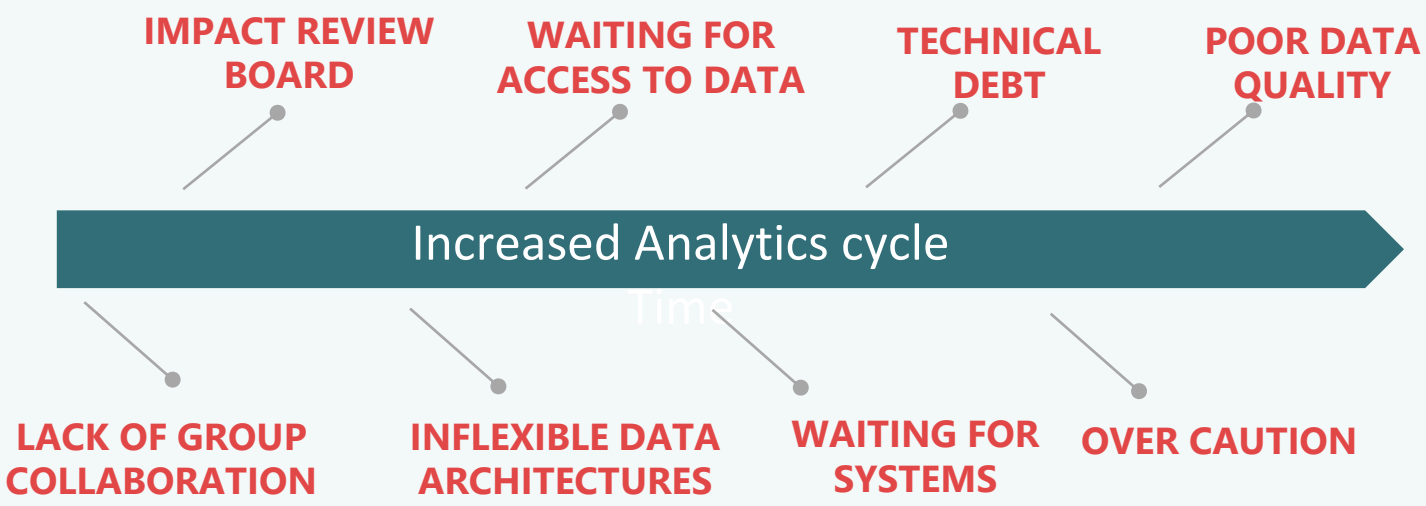
Intelligent Asset Management for a Leading Nordic Energy Company

Smart Factory Data Platformfor Large Asian Precision Engineering Company

Intelligent & Automated Operations Platform for Australian Mining Company

Challenges in Scaling AI Programs

In 2020, 80% of AI projects will remain alchemy, run by wizards whose talents will not scale in the organization ***Gartner***



SOURCE: https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

"We realized that the true problem, the true difficulty, and where the greatest potential is – is building the machine that makes the machine ***Elon Musk***

Focusing on the **How** is more important more than the **What** in data programs

What you do ?

- The Model
- Data Transformations
- Data Visualization
- Data Storage

How you do it ?

- Development
- Deployment
- Monitoring
- Iterating
- Collaborating
- Process Measurements

Change in **Mindset** and **Focus** is needed

Focus on People, Process and Operations

Decrease cycle time of change and continuous deployment

Lower error rates in production increasing customer data trust

Improved collaboration: Inter and Intra teams

Measure your progress and show productivity

AGILE

Speed

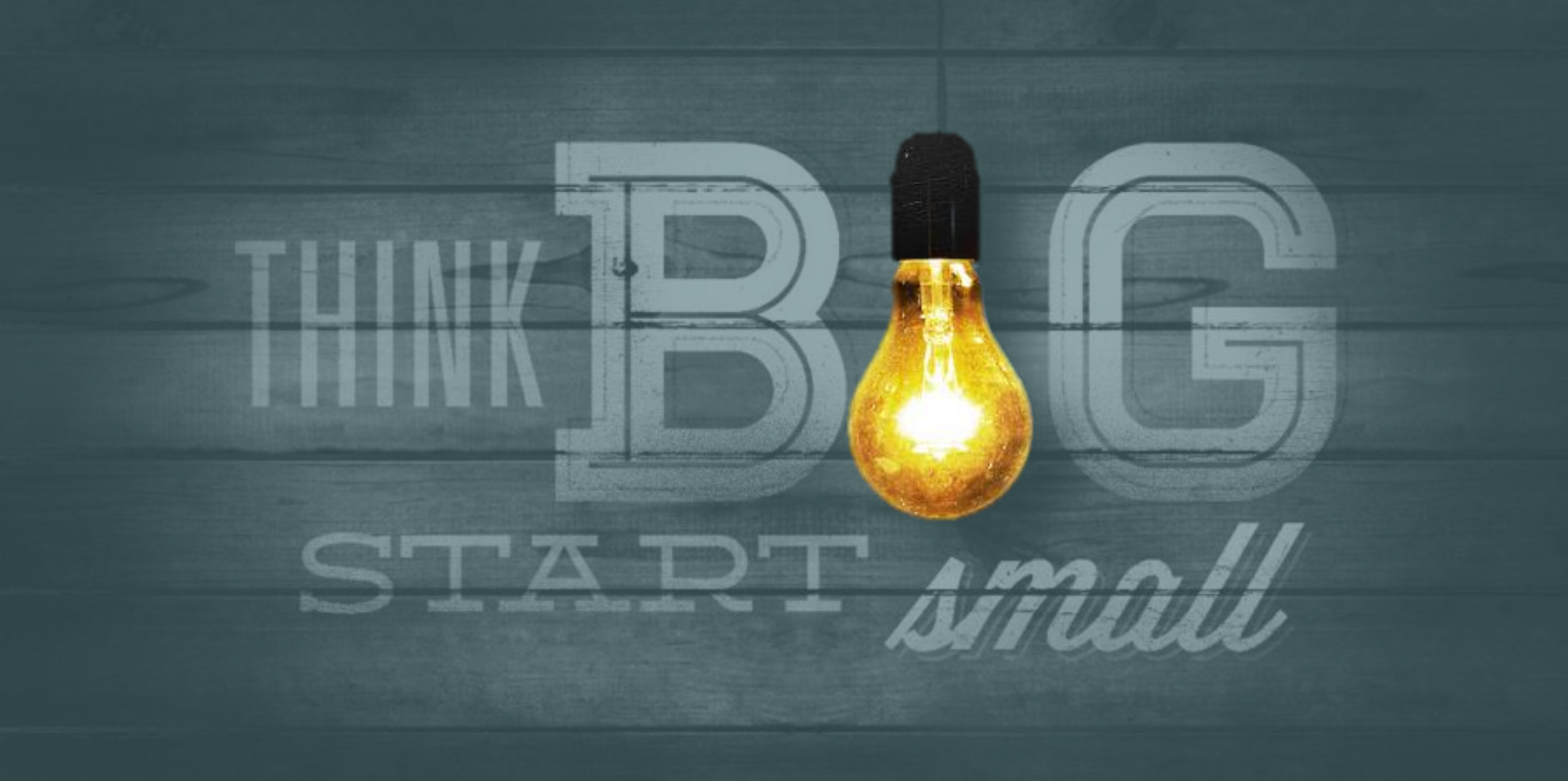
DEVOPS

Responsiveness

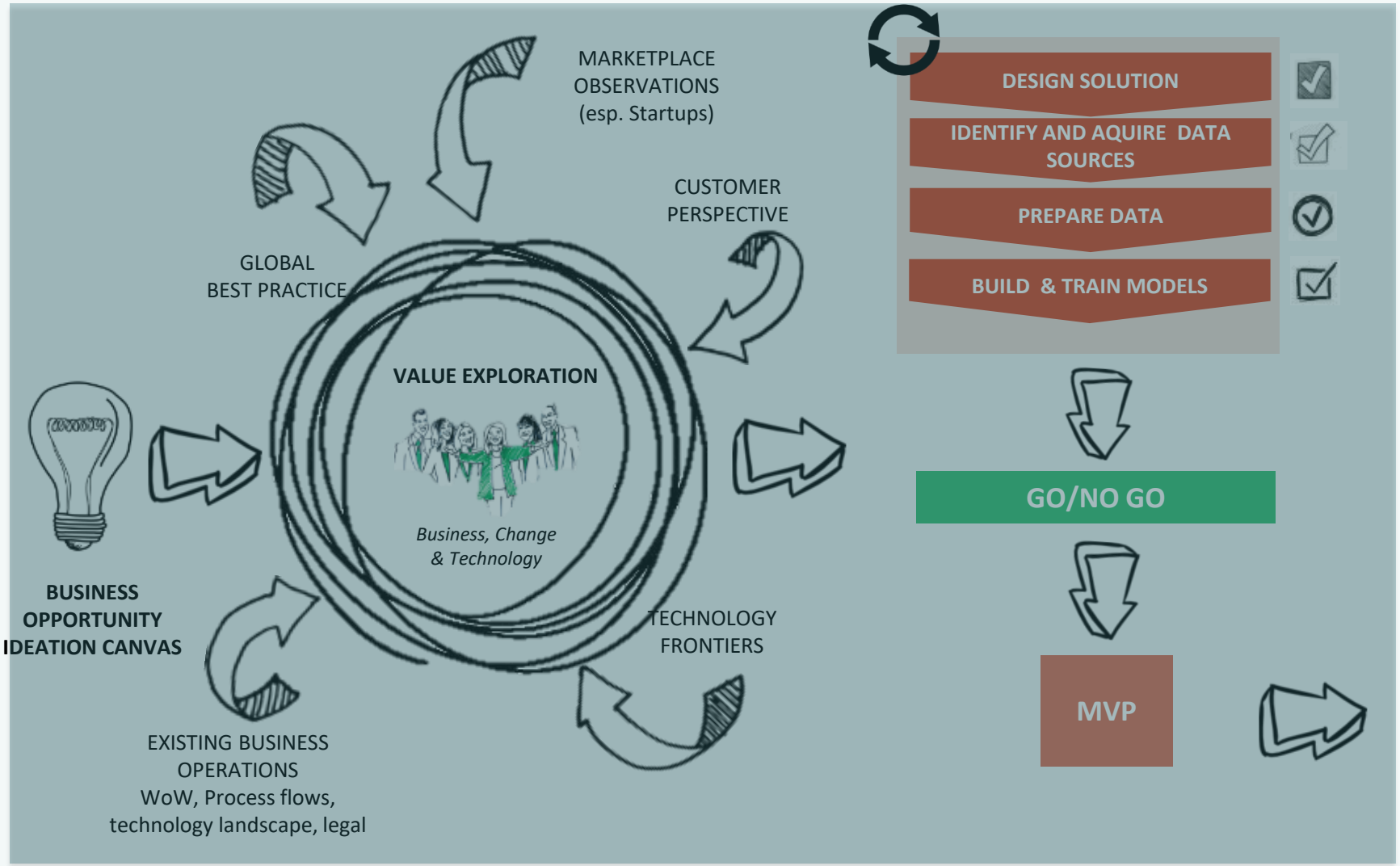
LEAN MANUFACTURING

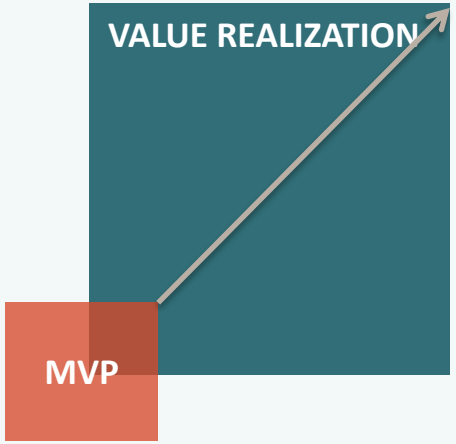
Statistical Process Control

DataOps


THINK BIG
START small

Value Realization through an **incremental Approach**





BUILD DATA PIPELINES → INTEGRATE END TO END WORKFLOWS → MONITORING & IMPROVMENETS

INNOVATE | INCUBATE & ITERATE | OPERATIONALIZE

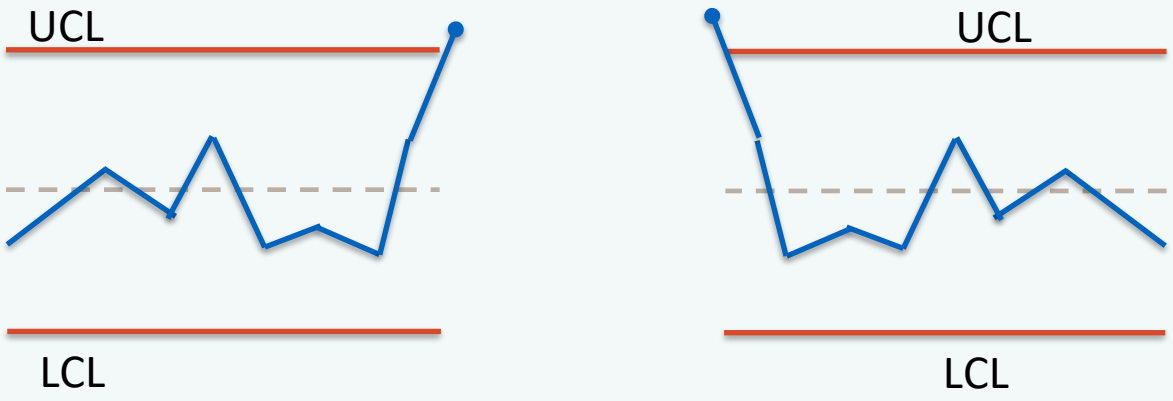
Best Practice 2 : Build Pipelines for Quality Data


Data is the Product in your Production Line
Data Consumers are your end Customers

Leverage Lean Manufacturing principles such as **Statistical Process Control** Measures to create data pipelines focused on Data Quality

ILLUSTRATIVE

Data engineering team listens and takes corrective action



Inputs

Verifying the inputs to an analytics processing stage

Count Verification - Check that row counts are in the right range

Conformity - Sweden Zip5 codes are five digits, US phone numbers are 9 digits

History - The number of prospects always increases,

Balance - Week over week, sales should not vary by more than 10%,

Temporal Consistency - Transaction dates are in the past, end dates are later than start dates

Application Consistency - Body temperature is within a range

Field Validation - All required fields are present, correctly entered,

Business Logic

Checking that the data matches business assumptions

Customer Validation - Each customer should exist in a dimension table

Data Validation - 90 percent of data should match entries in a dimension table

Output

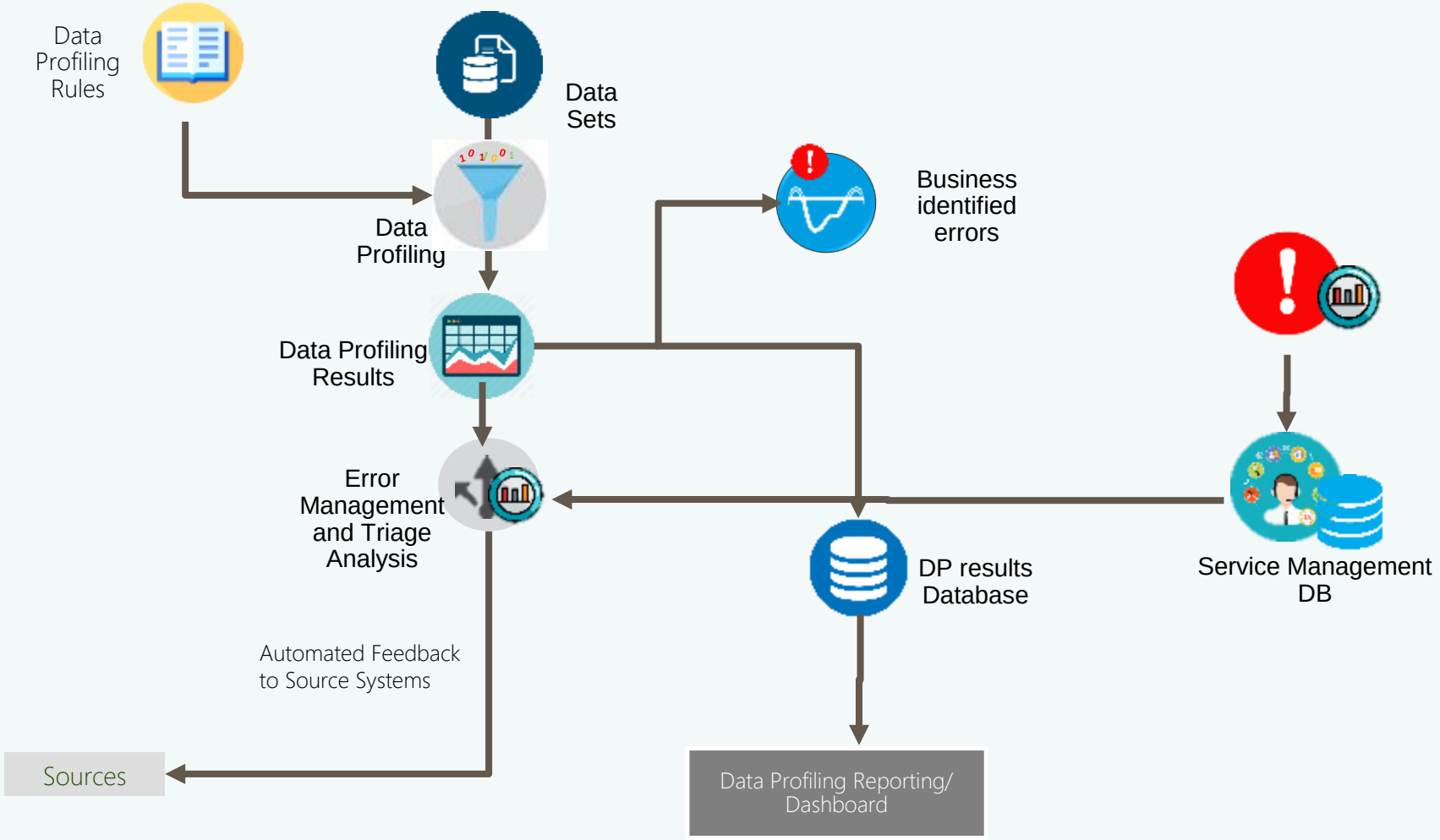
Checking the result of an operation, for example, a cross-product join

Completeness - Number of customer prospects should increase with time

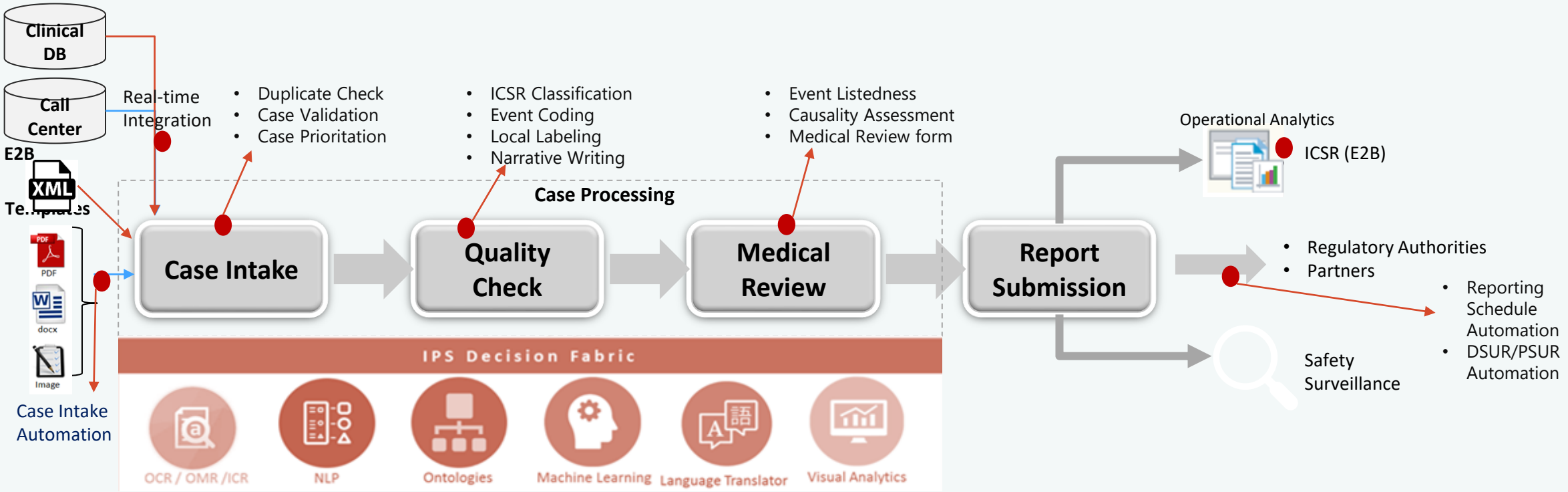
Range Verification - Number of physicians in the US is less than 1.5 millio

Data Pipelines for **Data Quality** management

Sequence of activities in Data Profiling



Data Pipelines built on **Business Data Rules** for Integrated Patient Safety Use Case



Best Practice 3 : Build Operational Support through Automation

1

Strengthen Monitoring & Event Management

To eliminate application performance issues and downtime

2

Automate Maintenance & Health Checks

Autonomous Ready for Business health checks

3

Enable Self Help

For User Inquiry / FAQ through Chat bots & Self service portals

4

Implement Orchestrated Automation

For end to end automation of typical incidents / service requests

5

Self Heal

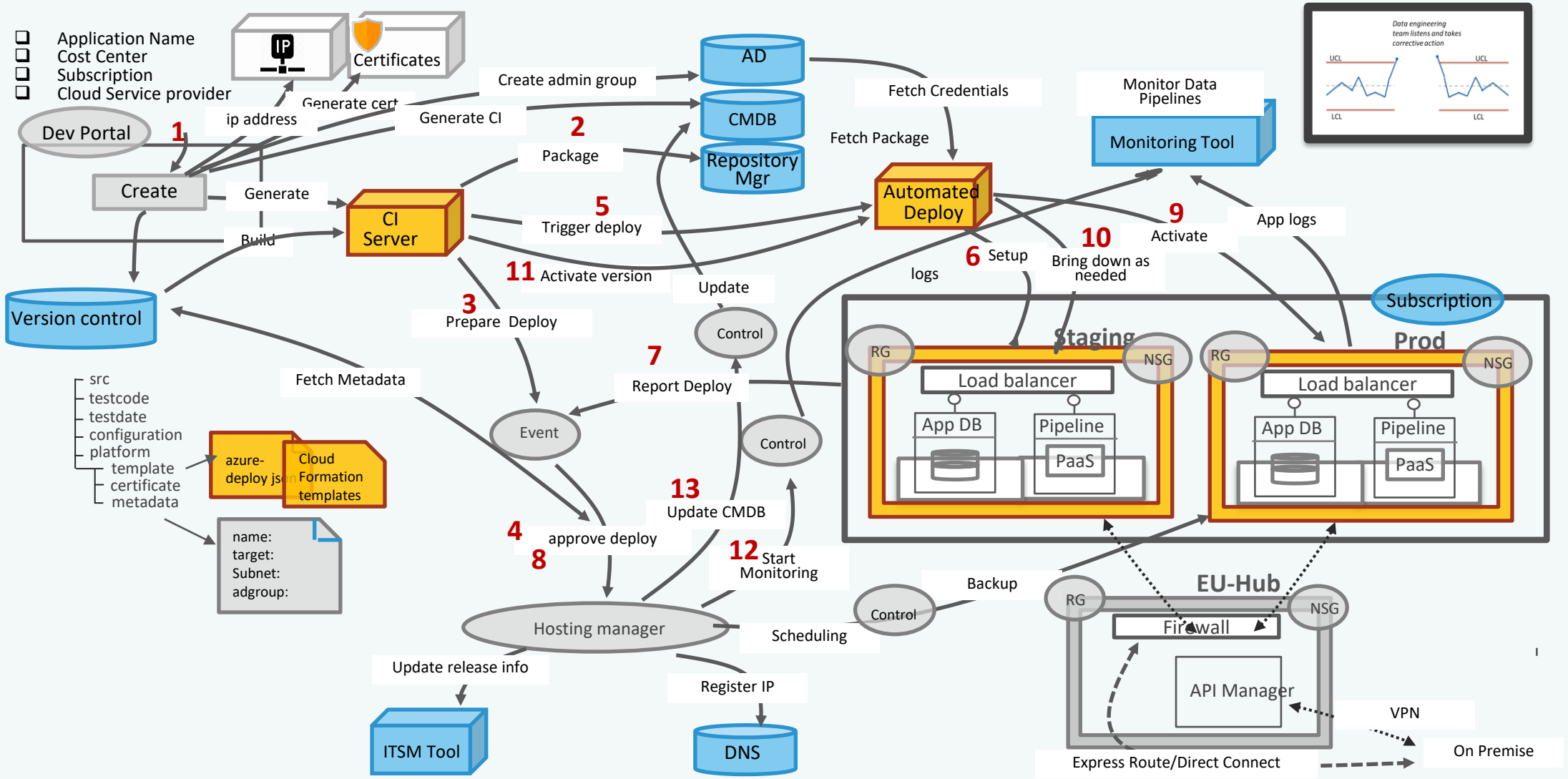
Automate and Triage incident resolution

Maturity

TATA CONSULTANCY SERVICES

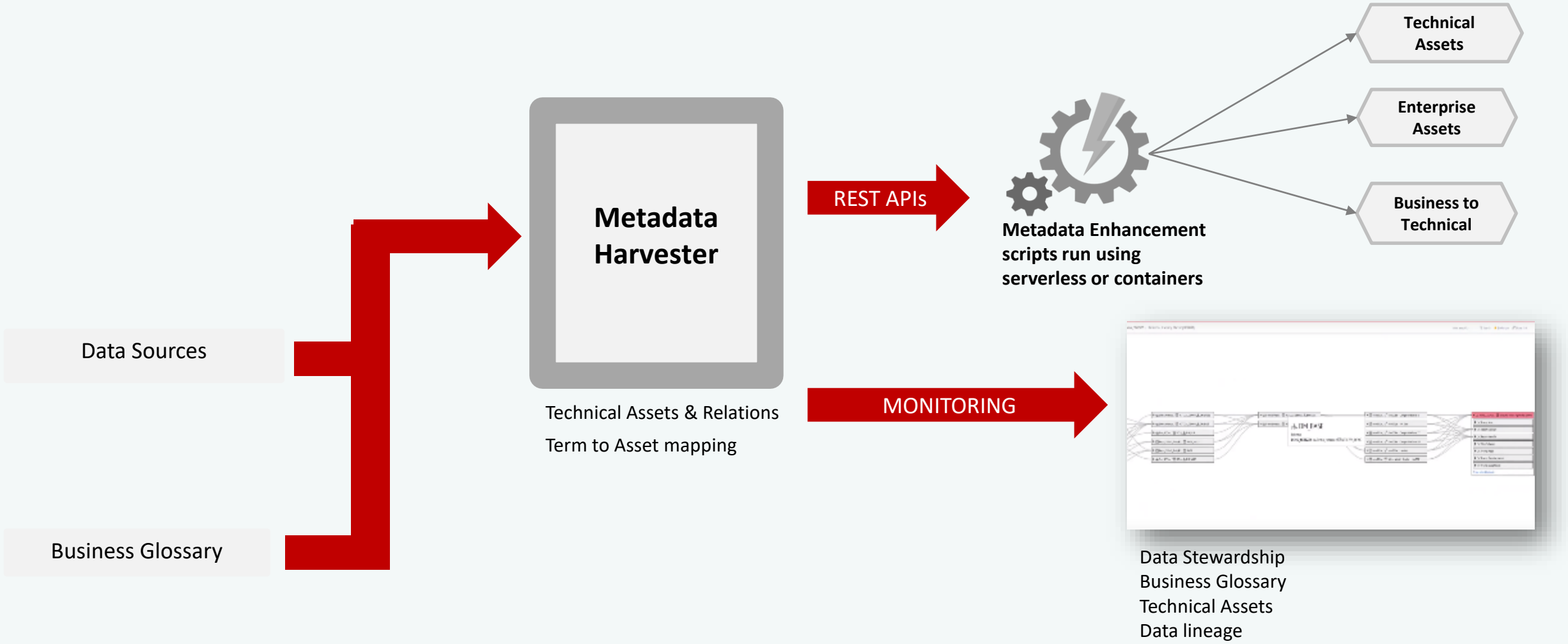
Automating Data Pipelines

ILLUSTRATIVE



Best Practice 4 : Create business data glossaries and catalogs for Self Service

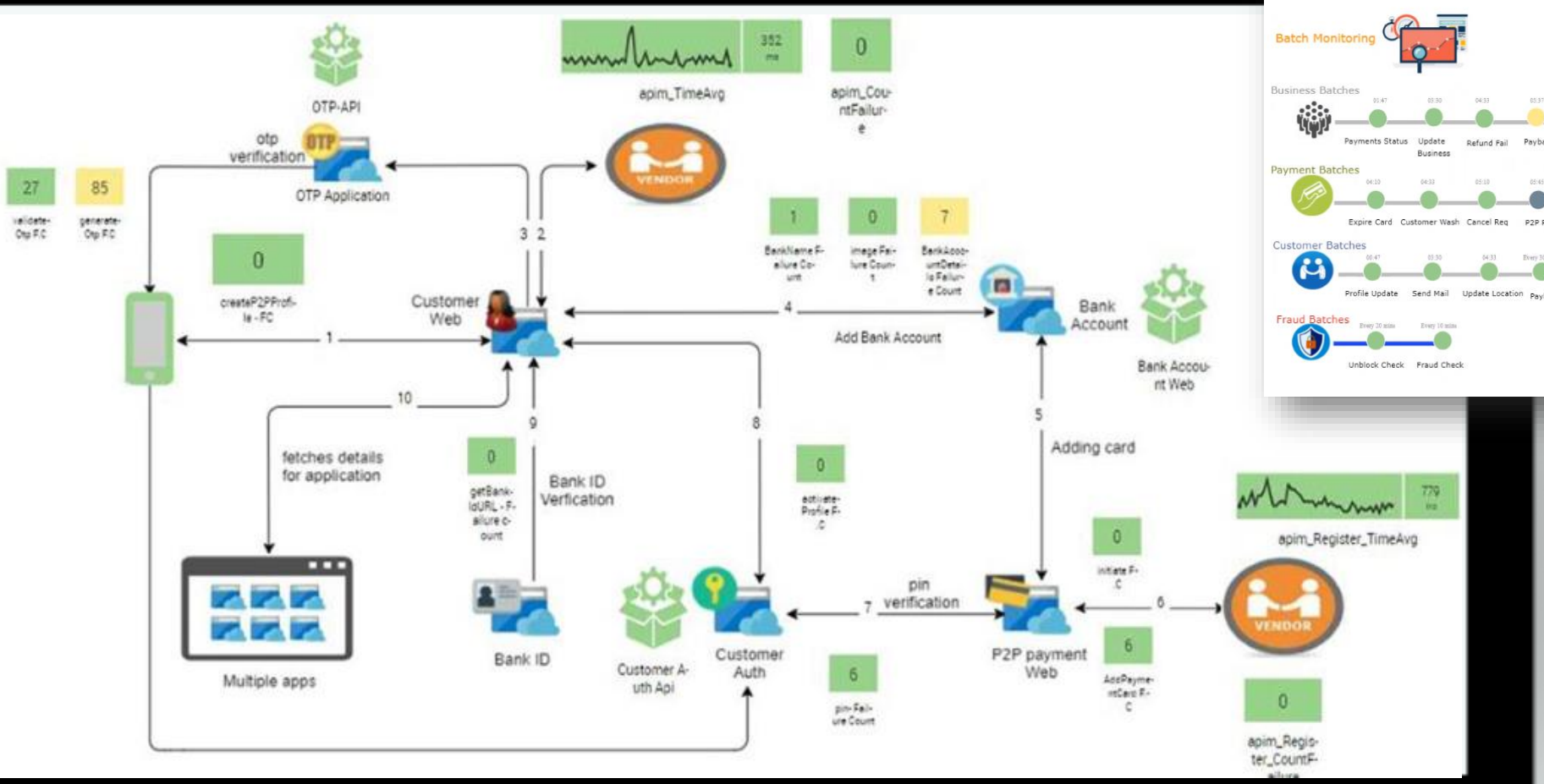
Build Integrations with your Business Glossary for Automated Updates

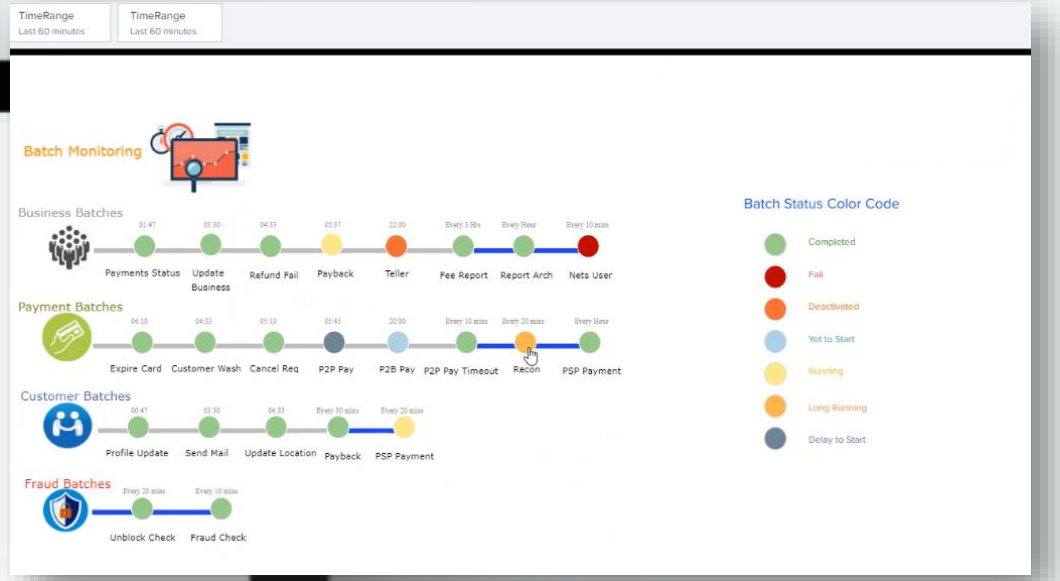


Ingest Metadata → Metadata Discovery → Metadata Management

Best Practice 5 : Monitor your data pipelines focused on Customer SLAs

Data Pipeline Monitoring needs to tie back to the **business value chain**





Agenda

How?

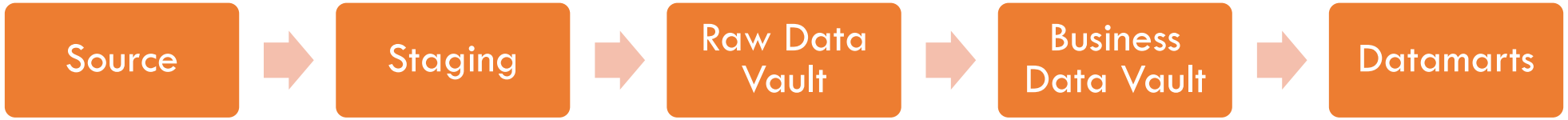
How do we go about designing & architecting?

Remy van der Kleij
Solution Architect, Informatica

Customer example

Financial institution - Enterprise Datawarehouse

Pipeline stages



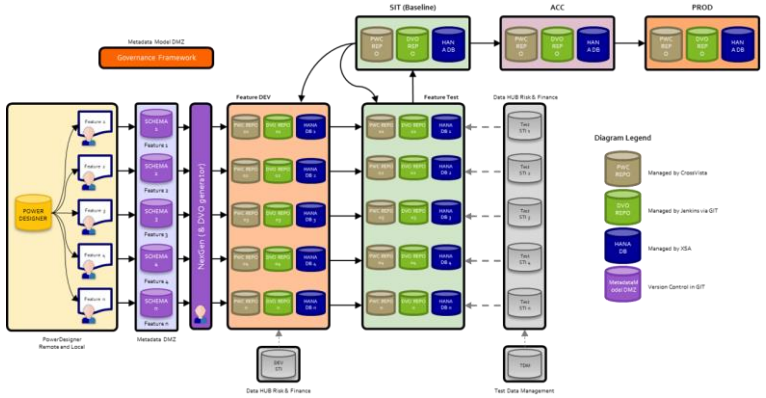
High-level summary of current assets/logic

- Mappings:
  - Staging: 350
  - Raw Data Vault: 2900
  - Business Data Vault: 250
  - Datamarts: 350
- Automatic data consistency validations across stages
- Job control through regular workflows

Customer example

Delivery model

- 25+ development teams working across 100+ private Dev/Test environments
- Generating DDL scripts and ETL assets from PowerDesigner models
- Challenge: not easy to branch & merge metadata-driven ETL components!



Approaches to address this

1. Minimize number of objects to deploy
2. Minimize size of objects to reduce risk of conflicts
3. Minimize number of teams working on same objects

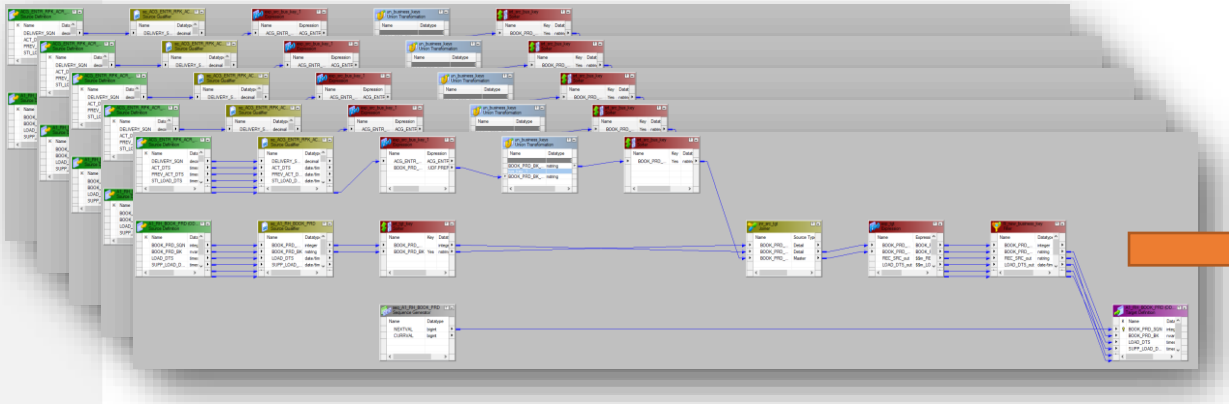
Customer example

Modernizing to template-based/metadata-driven integrations

- Primarily suits first pipeline stages - high similarity between processes
- Less suited for end of pipeline - custom business logic

Results in to-be situation

- Staging: 350 mappings → 1 template, 350 workflows → 1 template
- Raw Data Vault: 2900 mappings → <15 templates, 400 workflows → <10 templates



- Other layers: templatize common components only

Customer example

Technical benefits

- Minimal deployments of ETL objects: only parameter values/files
  - ➔ Much easier to branch & merge than entire ETL jobs
  - ➔ Many parameters automatically derived from data model at runtime
- Template mappings self-adjust to data model changes: only deploy DDL
- Job control through reusable template workflows
- Optimize scheduling based on data model dependencies

➔Total number of ETL-related objects to be managed reduced by **>75%**!

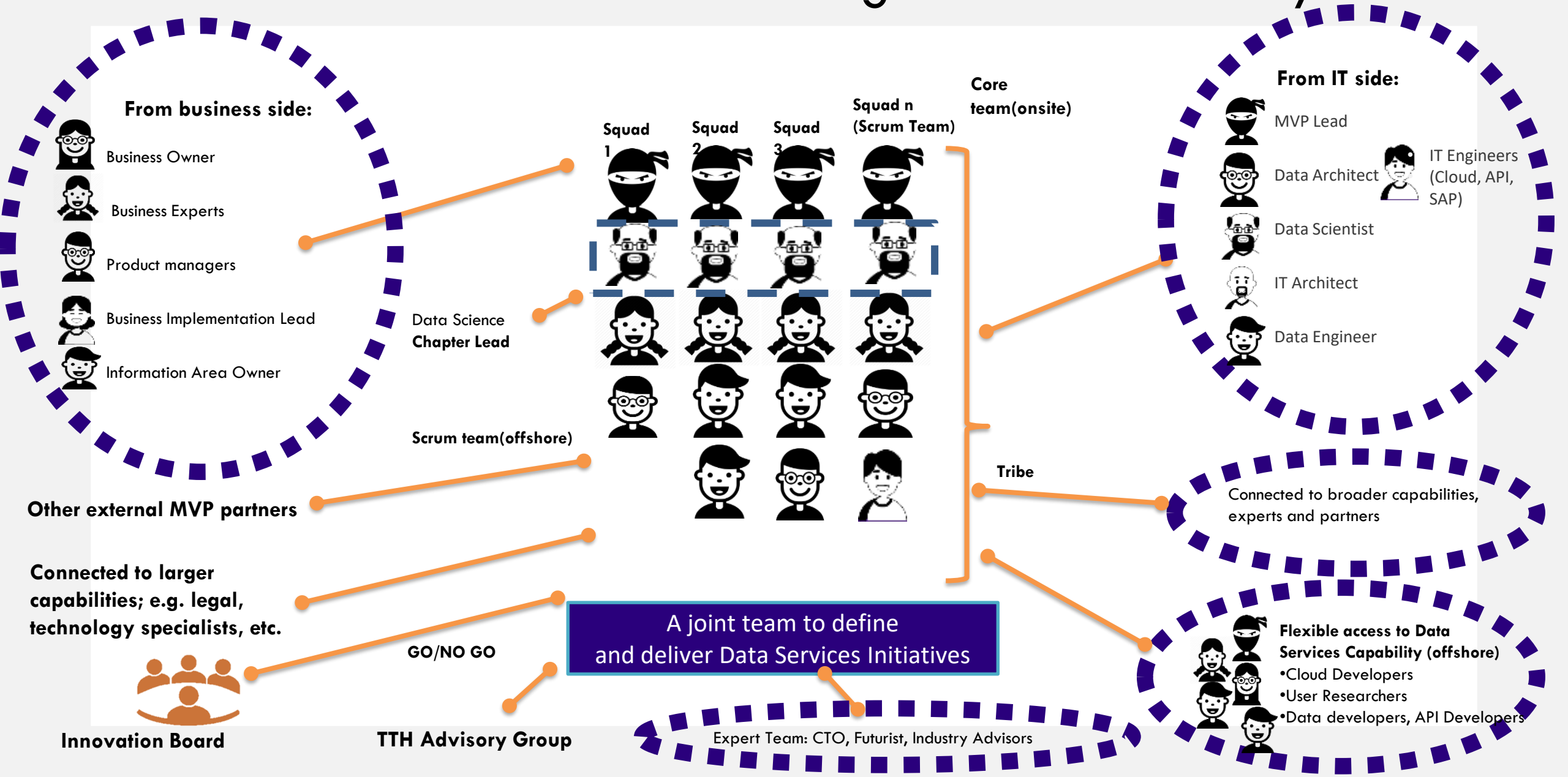
Business benefits

- Faster ingestion of sources into datawarehouse
- Increase delivery speed of backlog items by freeing up development resources

"I said I'd like to see the Burndown Chart, not burn down the chart!!!"

Minimize number of teams working on the same assets

Data Services – Structure : Leverage a Broader Ecosystem



Summary

Speed to Market

Time to respond to new Analytical Needs

1

Process

Lean, Agile, DevOps

3

Decentralized Teams

Enabling Different Teams to work on common assets

2

Metadata & Automation

Metadata Driven, Automation of Pipelines

4

DataOps is the foundation in terms of framework to enable the shiny Analytics

We have heavily invested in a Next Gen Analytical Platform

Thank You

tcs | TATA CONSULTANCY SERVICES

Informatica™