apoIO

Architecture . Cloud . Big Data

DataOps Toolchain for Continuous Control Monitoring

🌐 www.apoio.fr

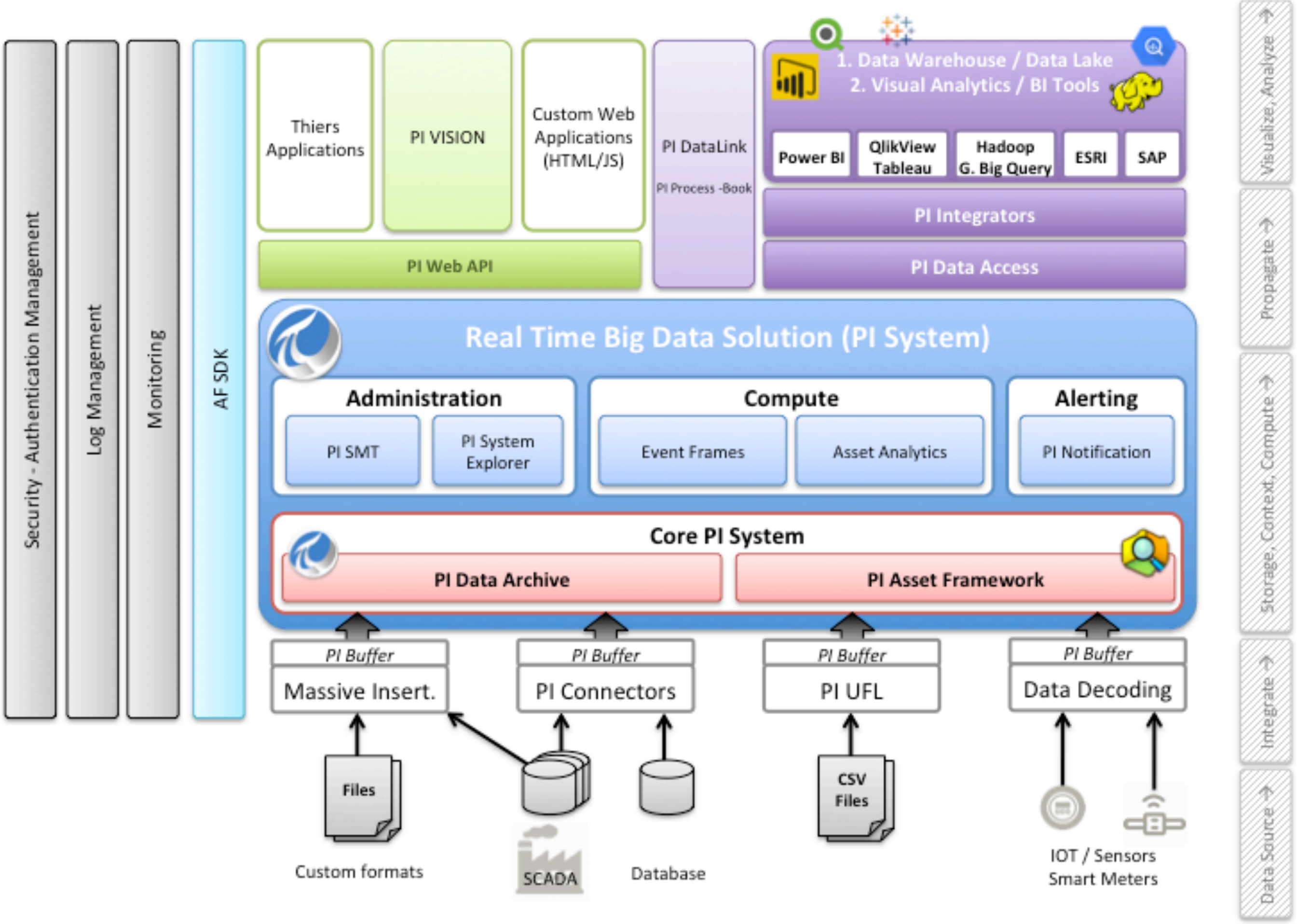
✉ info@apoio.fr

Introduction

All firms aim to transform their data in PI System to actionable information.
But many use cases take time to be implemented. Functional teams complain that their solutions can take months or even years to deploy or, worse, are never adopted.
Sometimes the reason is the complexity of the process or the algorithm but very often, the reason is a lack of methodology and technical capabilities to challenge the development of new features, test non regressions of existing ones, manage evolutions, fix bugs, monitor the data.
It is time to move the management of our Data within PI System From Artisanal Craft To Industrial Production.
From the delivery process perspective, the major issue is that the functional team is not fully qualified to test the data delivery by its own. They need help to challenge the data quality. They need some Continuous Control Monitoring.
Inspired from Agile, Lean, and DevOps, the DataOps methodology and its toolchains will enable to eliminate daily barrier.

Architecture Description

The architecture of all the MVP solutions is compliant
to this representation. It illustrates the Data Journey from Integration to Analysis stages and then its propagation to different reporting and dashboarding tools.



A Typical Client and related Use-Cases

In all of our experiences around PI System, the project is initialized and managed by a Business Project Manager.
Thus, our typical Client is Laura is a Project Manager, a Business Analyst.

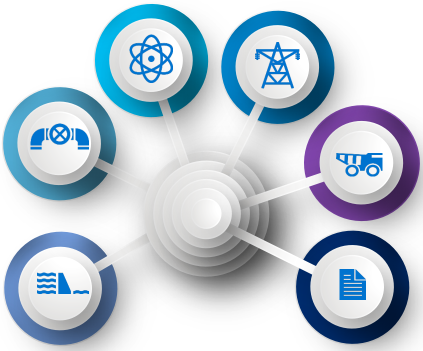
Laura has a great knowledge of her Business and has a clear idea of her goal on the project.

The objectives are Supervision, Dashboards, Reportings and are composed of several stages of Analysis.

For instance, the objectives can be Water Leak Detection, Energy consumption optimization, etc.

The objectives can be divided into several steps:

- Detect
- Learn
- Predict



Energy consumption optimization

Electrical disturbances

Water leaks

Mechanical breakdowns of devices

Water network issues

Behaviour of the consumers, etc.

Business Challenges on Data Quality

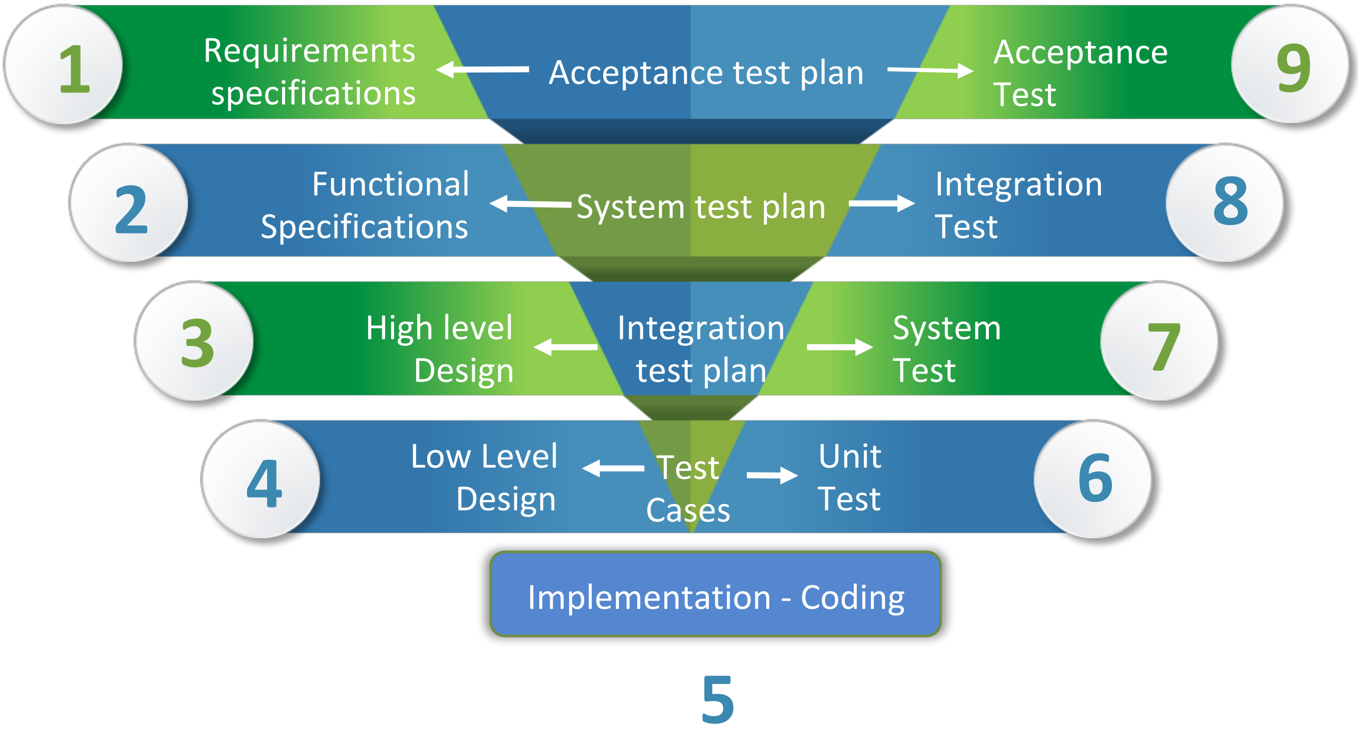
Sensors from the Water Network send by GSM values in a non chronological way : Invalidates all rules based on events

Quantity optimisation of substances used on a supply chain :

Many timeranges contain **exceptions** because of

Data Quality

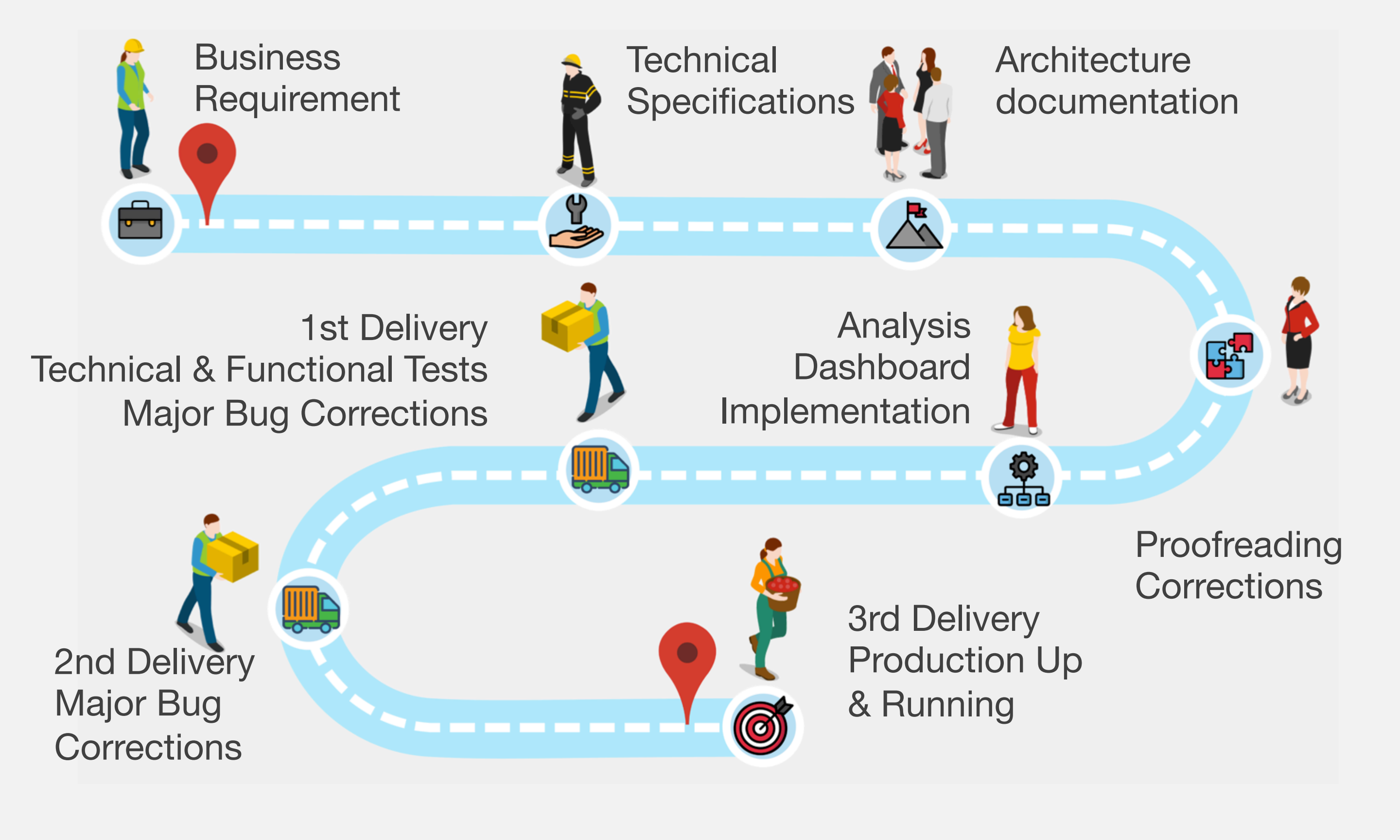
Classical V-Model Organization



The V-Cycle or V-Model Organisation begins with steps in

- Conception , Design, and writing documentations first: #1 to #4
- Implementation (#5) of the content which in PI System consists on configuring Asset Analytics, create Dashboards, etc.
- Test and validate every layer of specifications (#6 to #9)

V-Model from an Ideal Journey

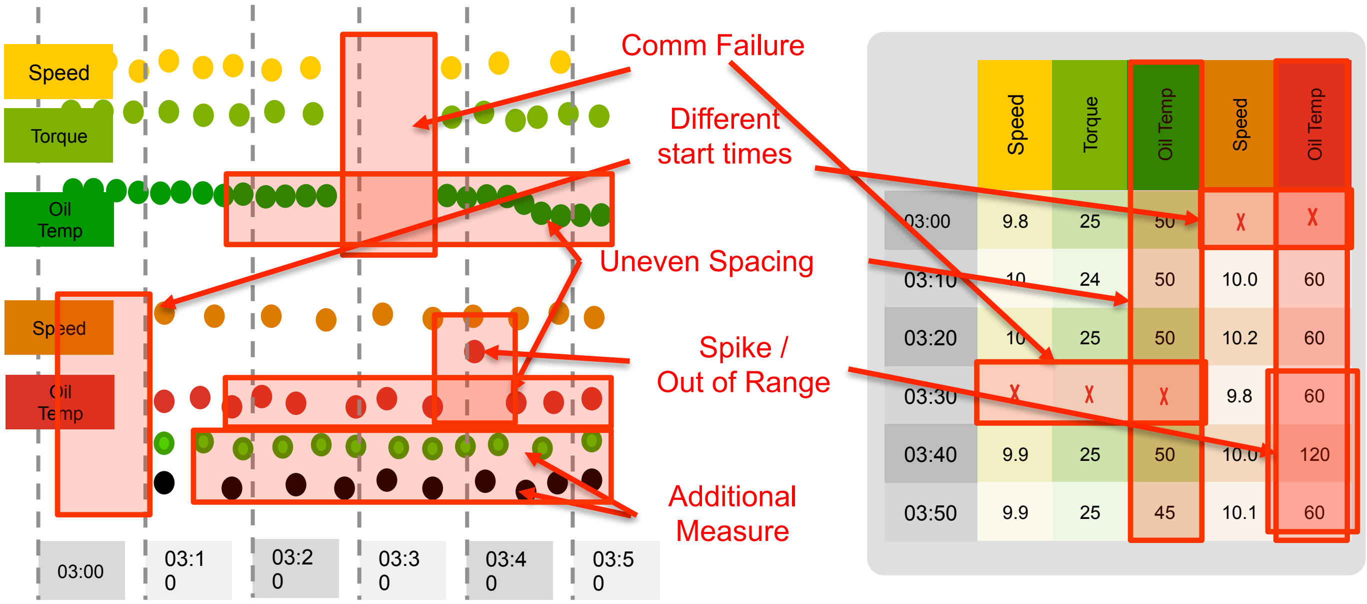


The V-Cycle or V-Model can be illustrated as a Journey. Here is an illustration of an ideal Journey when everyhting works as expected!

The V-Model as a Nightmare

The Data Quality invalidates the Business rules during their implementation and after production.

Data has imperfections : Welcome to Reality!



Comm Failure

Different start times

Uneven Spacing

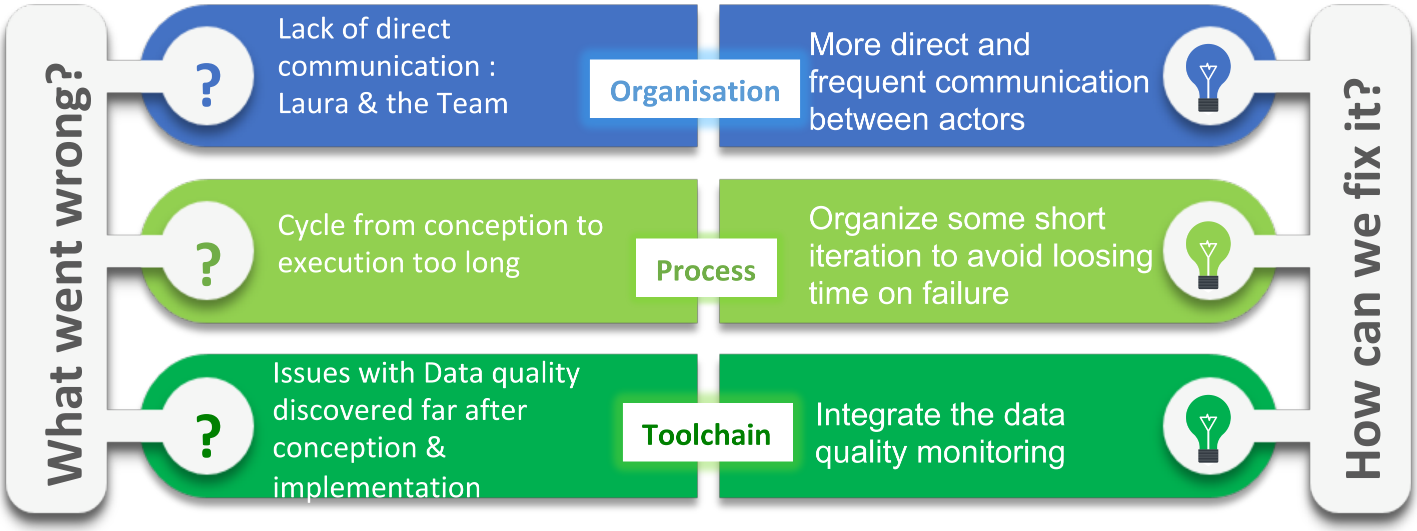
Spike / Out of Range

Additional Measure

	Speed	Torque	Oil Temp	Speed	Oil Temp
03:00	9.8	25	50	X	X
03:10	10	24	50	10.0	60
03:20	10	25	50	10.2	60
03:30	X	X	X	9.8	60
03:40	9.9	25	50	10.0	120
03:50	9.9	25	45	10.1	60

Data has imperfections.
How can we design rules on those?
Should we just pray that any of them will impact our rules?
Or should we define some parameters and quality rules to monitor?

Analysis of our project



What went wrong?

Lack of direct communication : Laura & the Team

Organisation

More direct and frequent communication between actors

Cycle from conception to execution too long

Process

Organize some short iteration to avoid loosing time on failure

Issues with Data quality discovered far after conception & implementation

Toolchain

Integrate the data quality monitoring

How can we fix it?

There are three axis of emprovements:

1. Organisation : the communication between actors were mostly based and organised around documentations. It has been proven that the direct and frequent communication between actors is the only way to understand the main issues. It does not replace the documentation but the volume of documentation does not replace the communication either.
2. Process : the duration of all tasks from conception, design to execution and testing took too much time. So any major failure has a great impact. We should organize some shorter iterations and deliver little by little and test the releases continuously to optimize the impact.
3. Toolchain : The Data Quality issues were not part of the project and were discovered far too late within the process, the toolchain should integrate a continuous way of testing non regressions, data quality, data exploration in general.

Data Quality Definitions

« Data Quality refers to the condition of a set of values of qualitative or quantitative variables. »

« Data Quality control is the process of controlling the usage of data for an application or a process. » https://en.wikipedia.org/wiki/Data_quality

Incompleteness
Precision
Severity of inconsistency
Accuracy
Missing / Unknown

Data Exploration

Approach similar to initial data analysis, whereby a data analyst uses visual exploration to understand what is in a dataset and the characteristics of the data

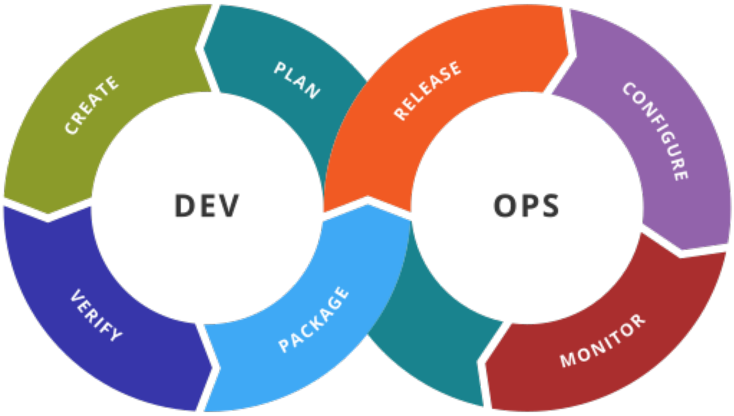
https://en.wikipedia.org/wiki/Data_exploration

Data Discovery

Process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

https://en.wikipedia.org/wiki/Data_mining

DevOps Approach : Continuous Improvement



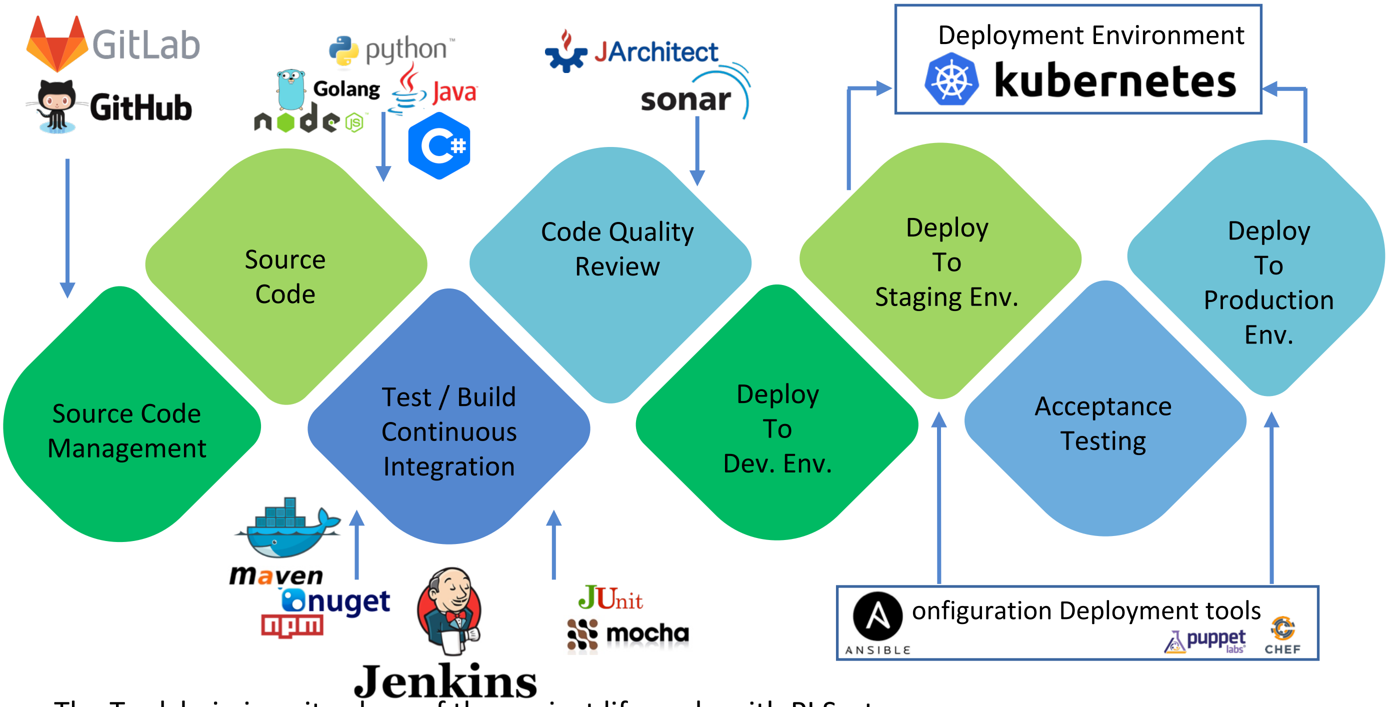
The Team contains

- Product Owner
- Developers
- Operationals

They work together in the same team, they share the same vision for a continuous improvement.

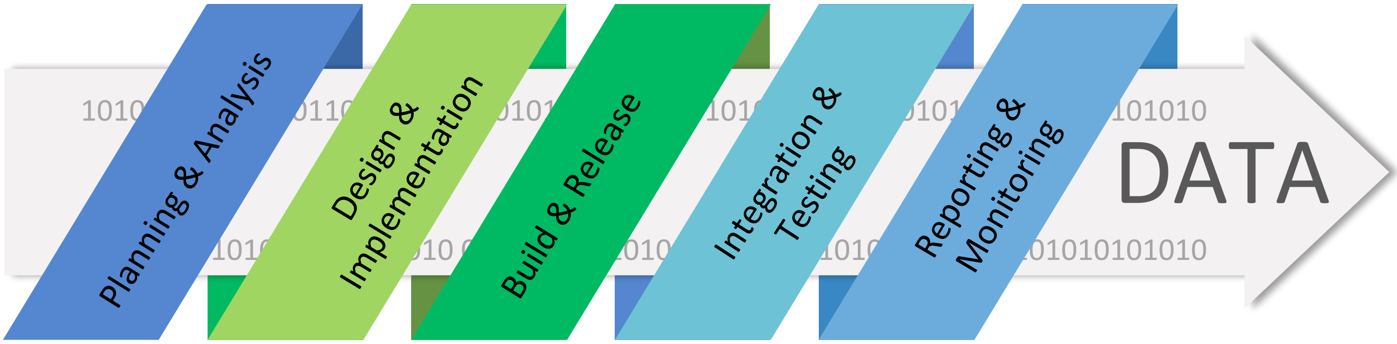
© https://en.wikipedia.org/wiki/DevOps_toolchain

Process : DevOps toolchain



The Toolchain is quite close of the project life-cycle with PI System.
The workflow starts with the Source Code, the the program is tested (unit test, non regression tests, performance) then it is built. The code qualiy can be challenged and if the whole workflow and assertions are validated, it can be deployed into differents environments.
The whole toolchain is automated.

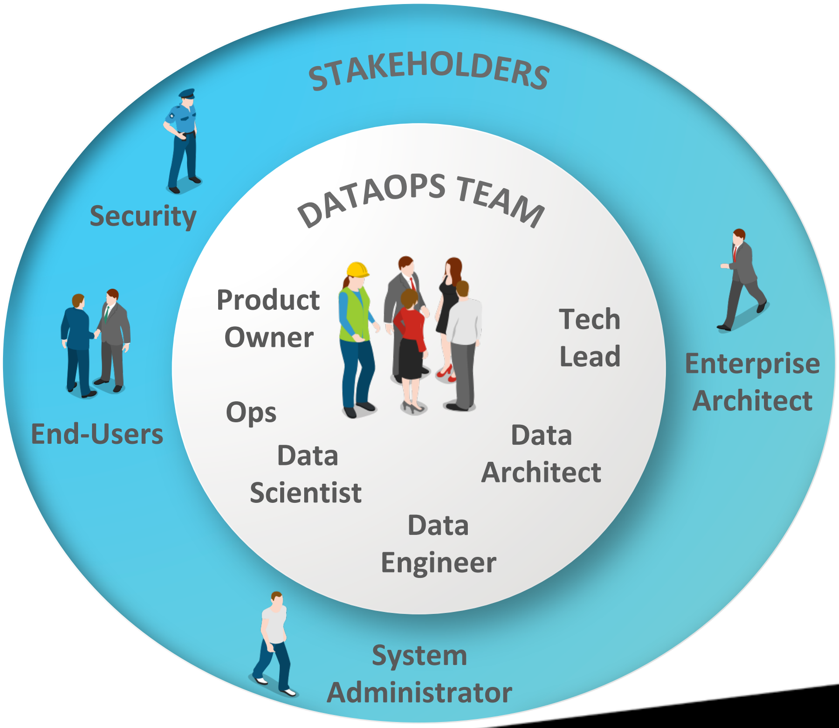
PI System Project life-cycle:
Data Driven Processes



But there is a huge difference between activities around PI System and the DEVOPS process. The PI System process contains many similar steps but the Data drives the whole PI System Project life-cycle.

DataOps & orgranisation chart

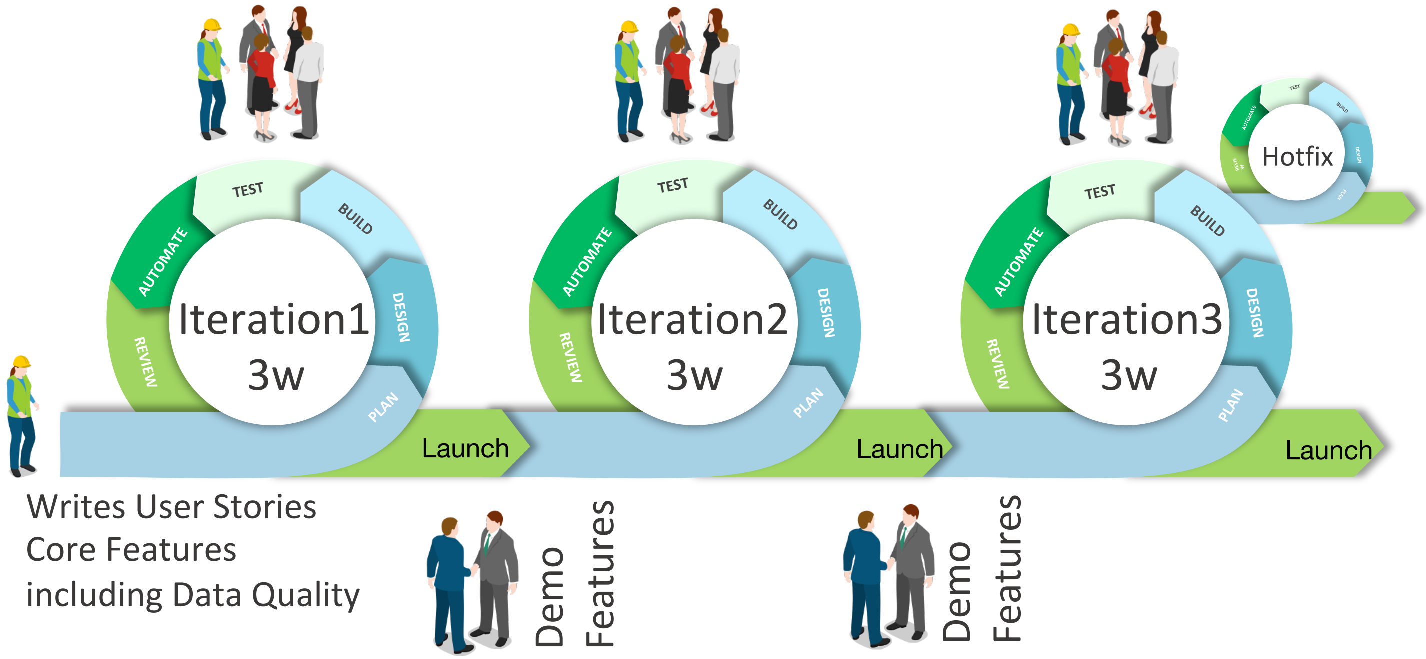
Laura has a new role:
Product Owner



"**DataOps** combines
- Agile development
- DevOps
- Statistical process controls

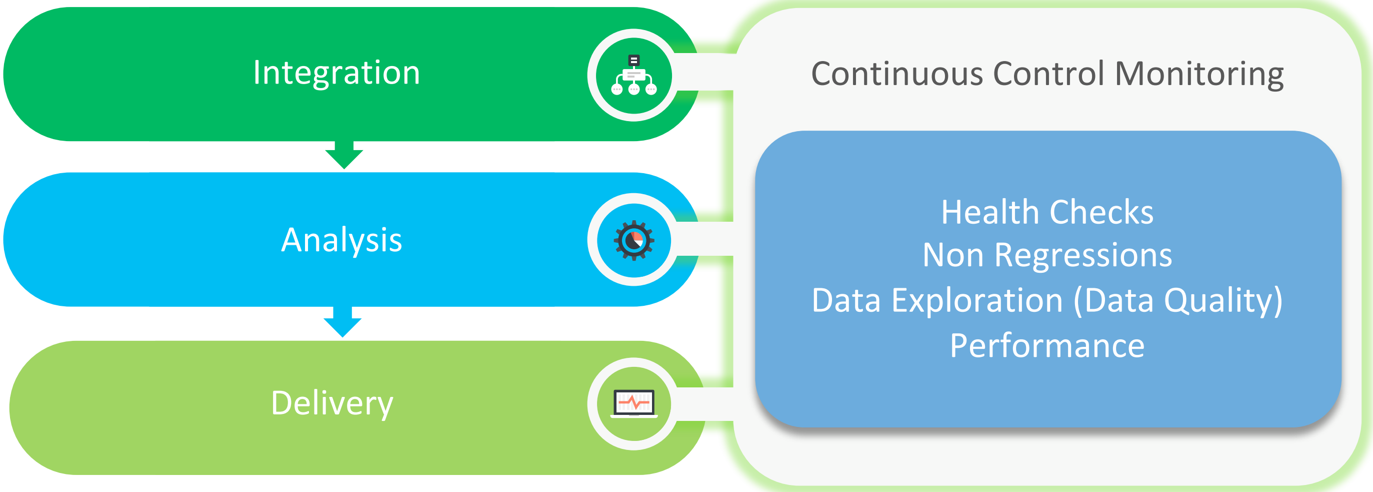
and
applies them to data analytics."
Wikipedia

Agile Management : Rethink the Journey



1. The Product Owner and the Tech Lead define core features the Business need that need to be done first. Those core features are converted in to separate testable units named User Stories.
2. The Data Quality is part of these core features.
3. The Testing process will be implemented in a continuous way.
4. The team begins an iteration by implementing the user stories. (3 weeks)
5. A Demo will prove or will invalidate every user story works.
6. The definition of DONE of a User Story will include the Continuous testing.
7. The Data Quality testing process will be part of the Continuous Control Monitoring (CCM).
8. The short iteration and its continuous testing should prevent as soon as possible if any Data Exploration or Data Quality invalidates a rule.
9. Once a bug is found in a rule. Regarding to its severity, we should be able to organize its correction with a Hotfix during an iteration or add the correction in a further iteration.
10. The End-Users can be invited into the periodic Demos, so that they know in advance the next evolution in production.

DataOps missions for PI System : Monitoring



Integration

Analysis

Delivery

Continuous Control Monitoring

Health Checks
Non Regressions
Data Exploration (Data Quality)
Performance

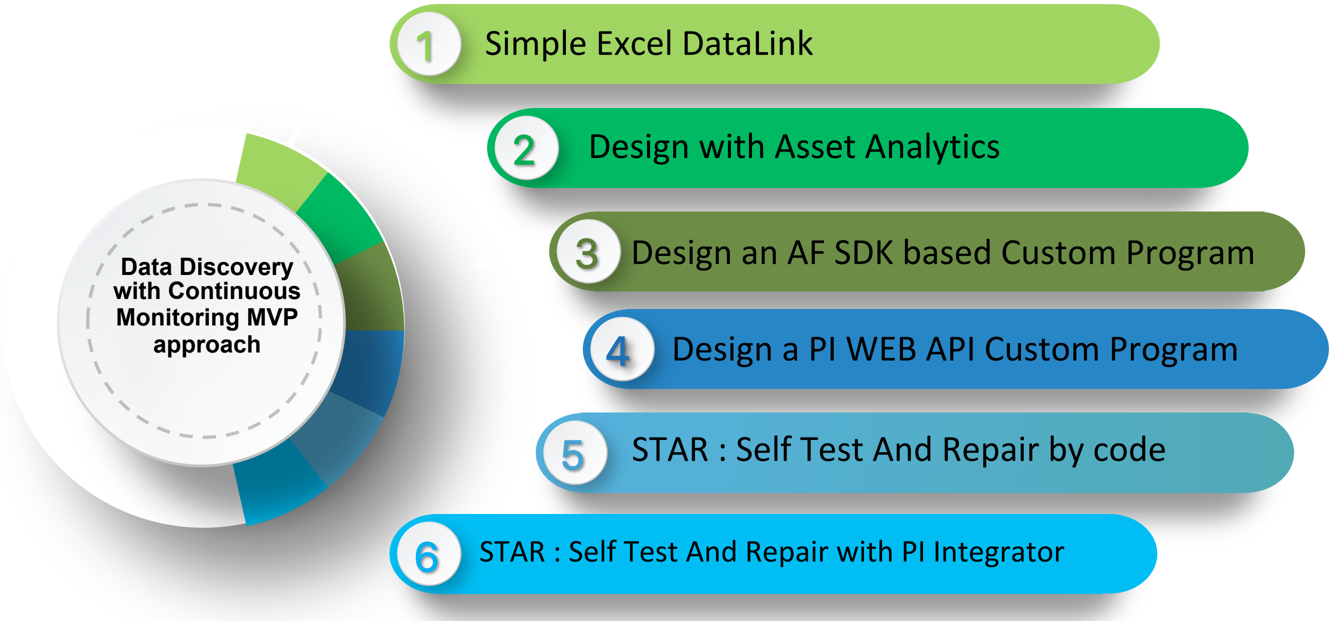
The Continuous Control Monitoring is one major mission of the DataOps Team.
It should be done during the whole stages of the Data transfer from Integration to Analysis to Delivery.
There are different type of control that this tool should do.:

- Health Checks : are the data sources and their propagation tools alive?
- Non Regressions : between different types of testing (unit testing, performance, etc.) Non Regressions are the most important. This should check if any evolution within the process (rules, data, devices, etc.) has caused a failure in a process that used to work properly.
- Data Exploration (Data Quality) : the Data Quality (missed data, precision, errors, etc) should be monitored. We should be able to have statistics and monitor them. The Data Exploration or Data Discovery contains the concept of Data Quality. We may need to explore the data to have some other statistics on our Data than the Quality.
- Performance: we should be able to check if our process takes more and more time for the same volume of data. If not, we will be able to know when this information has changed.

Agile Management & the MVP approach

The DataOps Team and its Product Owner defines core features including data quality. The DataOps Team thinks the incoming features as **Minimum Viable Product** (MVP).

Continuous Monitoring via MVP Toolchains



Data Discovery with Continuous Monitoring MVP approach

1 Simple Excel DataLink

2 Design with Asset Analytics

3 Design an AF SDK based Custom Program

4 Design a PI WEB API Custom Program

5 STAR : Self Test And Repair by code

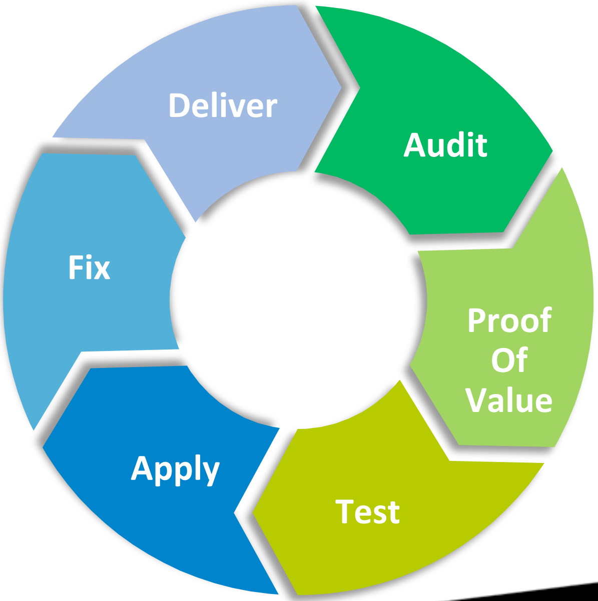
6 STAR : Self Test And Repair with PI Integrator

Regarding to few criterias that we will detail later, there are many different types of Continuous Control Motnitoring solutions for PI System projects.

1. One simple solution is to use Excel Spreadsheet to integrate and some data with DataLink and test it manually within Excel. It is a very good way of exploring the data but it remains a manual way.
2. It is also possible to configure PI Expressions or Event Frame Generations with Asset Analytics to check some Data Quality rules. It is very easy to implement but can be difficult because of the Data: for instance, if the Data arrives in a non chronological way, the processing won’t be effective.

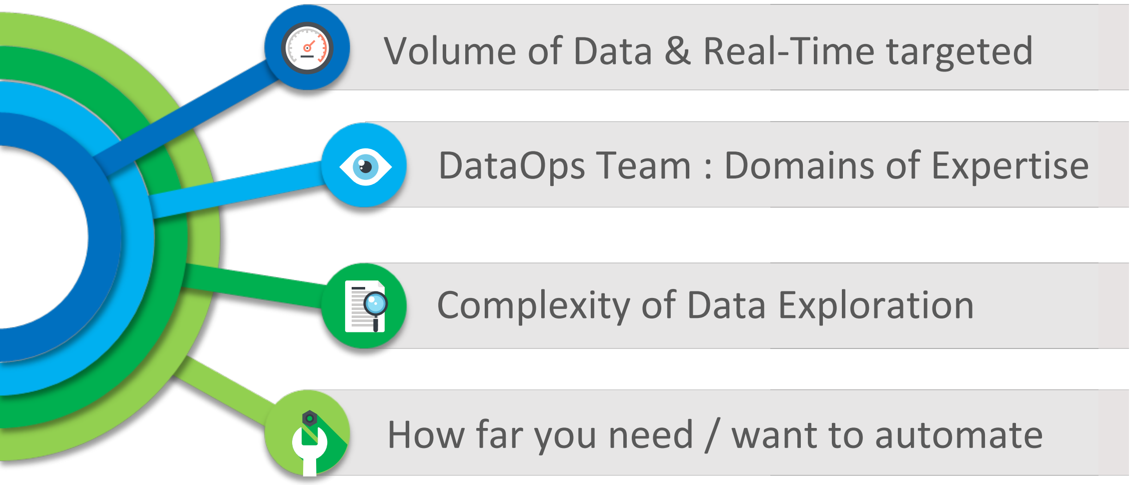
3. The Continuous Control Monitoring (CCM) Toolchain may be implemented in a Custom Program in C# for example, using AF SDK. It is possible to integrate with the Data by Event (PI Data Pipe) or in Batch Mode. It can be very performant and be used in Bulk mode. But the code will need to be maintained and follow a process (see DevOps Toolchain!) for delivery.
4. Another CCM solution for DataOps Toolchain can be based on PI WEB API. This is one efficient way to interact with PI System with any other language that is not .Net based. For instance, if you need to interact with some Python librairies for data structures, this solution can be the right candidate.
5. The next step for your CCM Toolchain is to be able to repair itself some input. This is what is called STAR : Self Test And Repair. We can use one program illustrated before and add some extra functionalities to create some related data. For instance, it can be usefull for some missed data, a precision issue, etc. It is also interesting to consider building statistics and pushing the result into PI System for some futur use.
6. A STAR Tool can be based on PI Integrator. If you need to interact with a Data Lake or a Business Intelligence tool. You can add your Data Exploration rules and then push the results to PI System for some futur use.

STAR : Self Test And Repair Toolchain



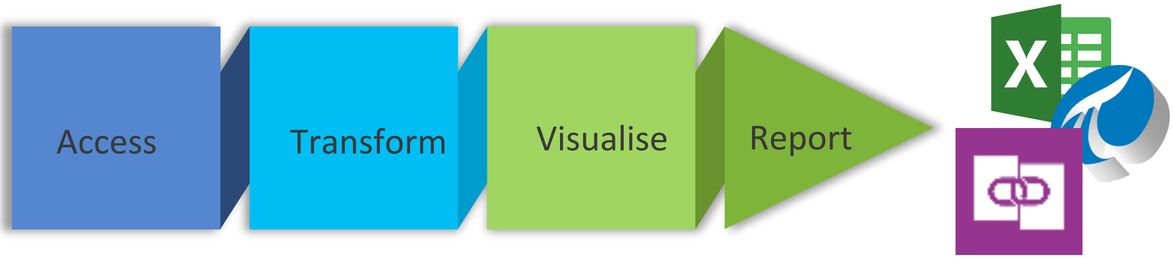
The Process is very close to the other Continuous Control Monitoring Toolchain but here the Step “FIX” has been added.

Rationales to select the right MVP Toolchain



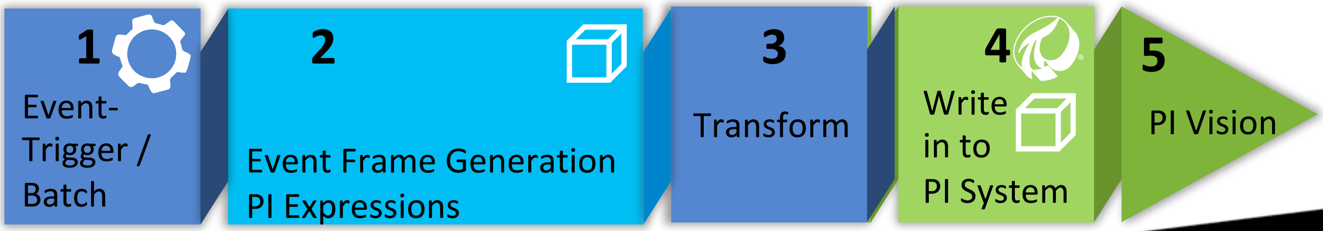
- Volume of Data & Real-Time targeted
- DataOps Team : Domains of Expertise
- Complexity of Data Exploration
- How far you need / want to automate

MVP : Simple Excel PI DataLink



Access → Transform → Visualise → Report

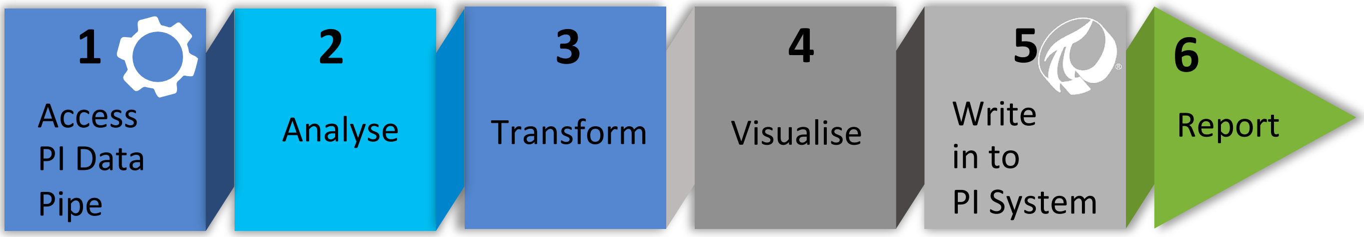
MVP : Asset Analytics



1 Event-Trigger / Batch → 2 Event Frame Generation PI Expressions → 3 Transform → 4 Write in to PI System → 5 PI Vision

Smart Meters in Utilities : AF SDK Custom Program

AF SDK



1 Access PI Data Pipe

2 Analyse

3 Transform

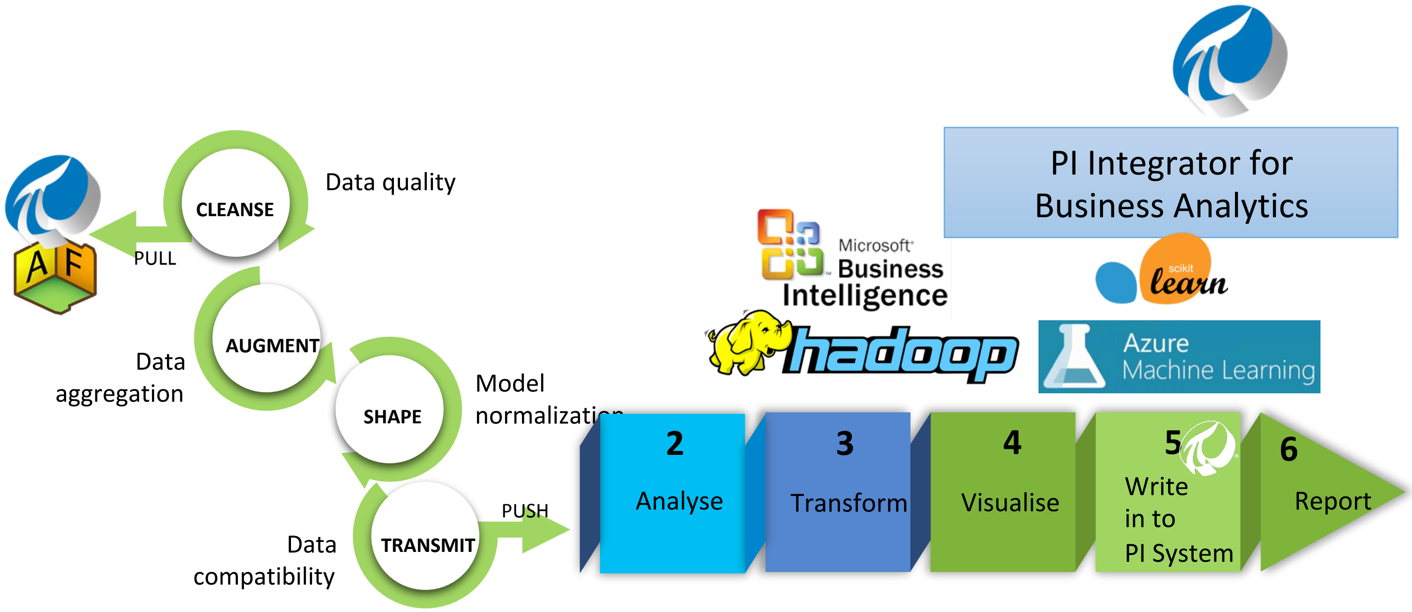
4 Visualise

5 Write in to PI System

6 Report

Google Big Query

MVP : STAR with Machine Learning



PI Integrator for Business Analytics

PULL

CLEANSE

Data quality

AUGMENT

Data aggregation

SHAPE

Model normalizati

TRANSMIT

Data compatibility

PUSH

Microsoft Business Intelligence

hadoop

scikit learn

Azure Machine Learning

2 Analyse

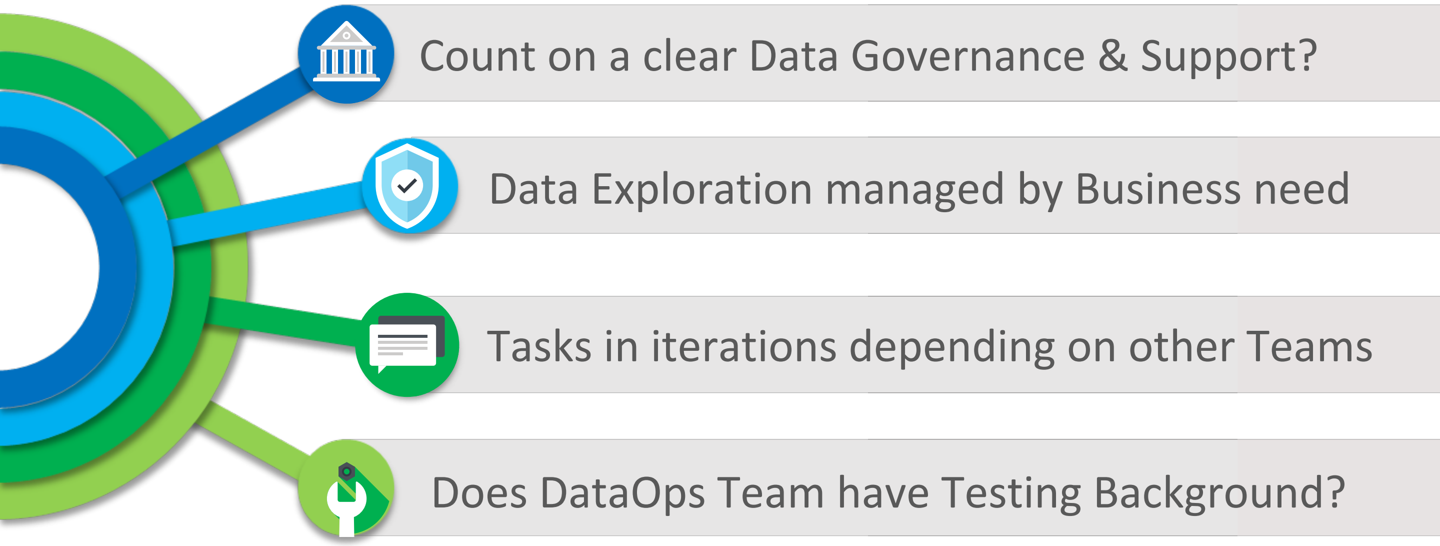
3 Transform

4 Visualise

5 Write in to PI System

6 Report

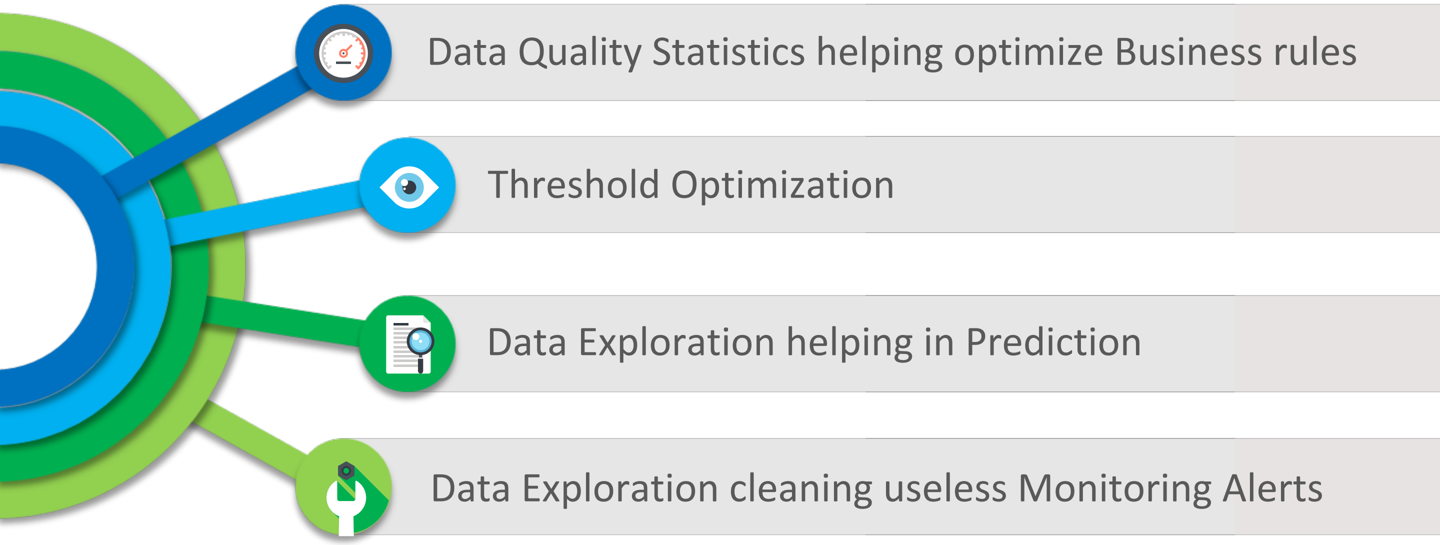
Risks of DataOps Approach



- Count on a clear Data Governance & Support?
- Data Exploration managed by Business need
- Tasks in iterations depending on other Teams
- Does DataOps Team have Testing Background?

CCM bringing Business values

From Data Quality to Data Exploration



- Data Quality Statistics helping optimize Business rules
- Threshold Optimization
- Data Exploration helping in Prediction
- Data Exploration cleaning useless Monitoring Alerts

Call for Participation!

As explained previously, apoIO company has experiences in many different domains of industry. We have already designed architecture including PI System in several scales.
After analyzing the Project Journey, we came to the conclusion that the common part is the data exploration. We also noticed that the Data Exploration rules including the Data Quality ones had many similarities.
This is why we are quite excited to announce we are starting an opensource project on GitHub named PI-Data-Exploration.
If interested on contributing to this project, please feel free to contact us!
The PI-Data-Exploration common Goal is:
Answer to some sharp questions about PI System Data Exploration.

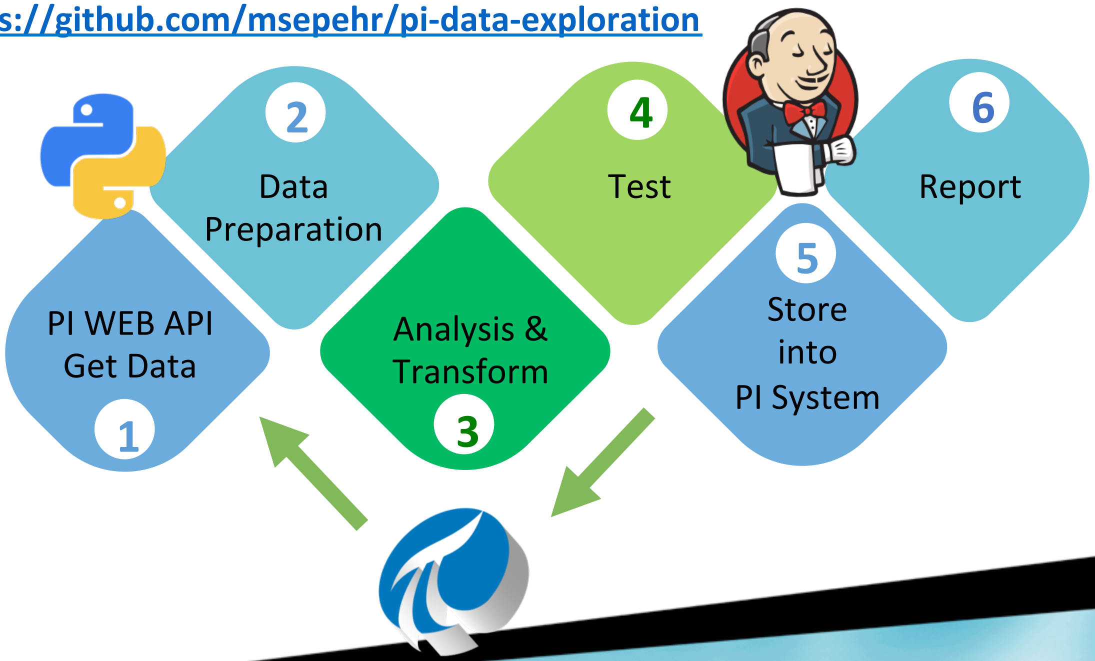
Sharp Questions:

- Classification Modeling as Leak Detection (no threshhold)
- Autofill on Missed Data
- Statistics over the data as Missed Data

Python Dependencies:

- PI-Web-API-Client-Python : PI Client for Python
- pandas : Data structure manipulation
- numpy : perform calculations over the data

https://github.com/msepehr/pi-data-exploration



Key Takeaways

Key Insights Lessons learned from our journey

01 Blueprint for success
Do not let your transformation initiative fail because of a lack of anticipation on the project due to the data

02 Data Exploration, Data Discovery
more & more vital in Businesses in any sector

03 Agility with a clear Roadmap
Iteration on core features is completed with a real vision in a long term

04 Empowering End-Users
Better and faster communication, interactions with DataOps Team

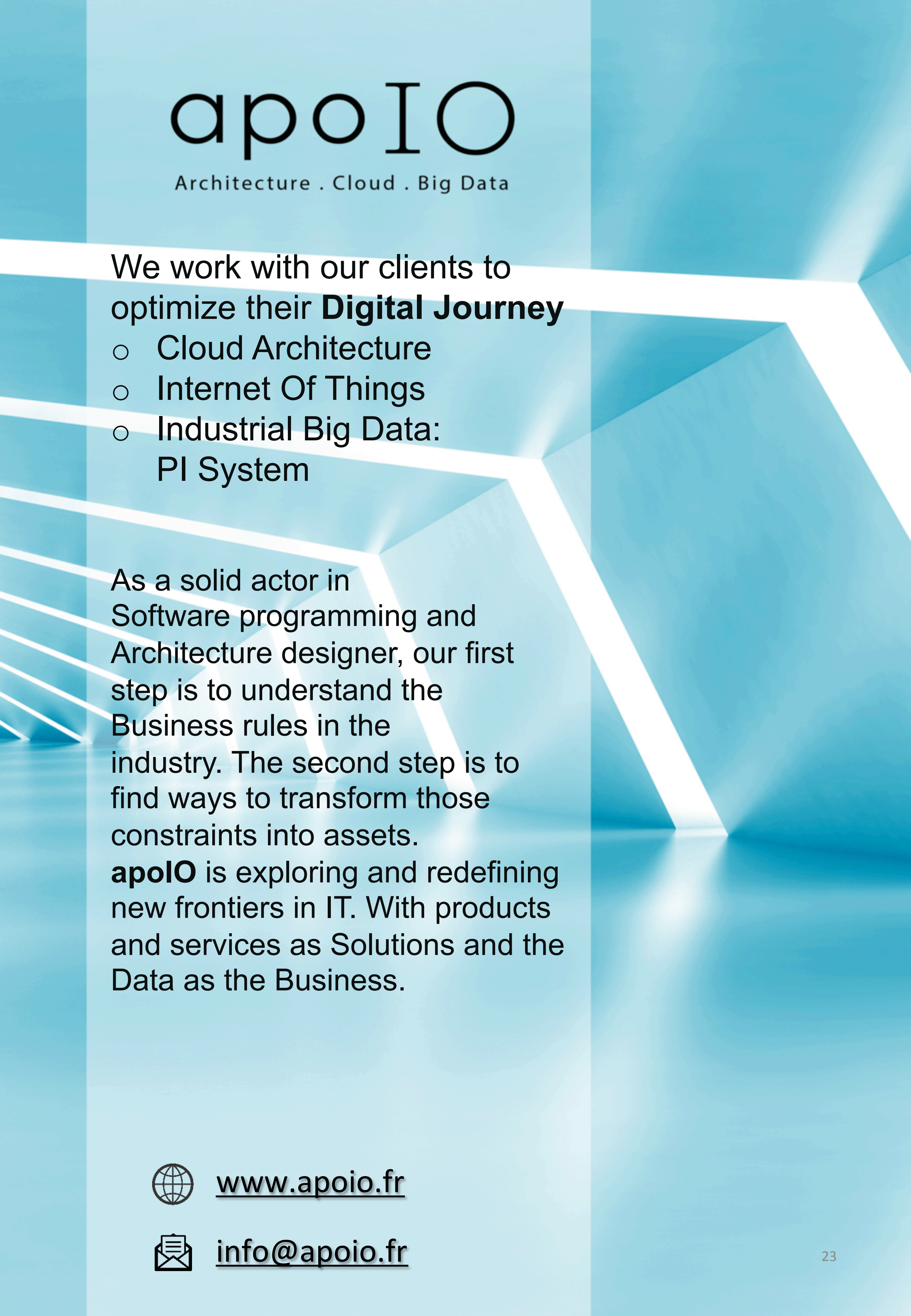
05 New/Classic Skills for DataOps
Data preparation, Machine Learning, Business Intelligence, Data Visualization

06 Automate, automate, automate!
From Data Exploration to Business Rules, Industrialize as much as you can

About the Author

Mahyar Sepehr is a DevOps with an IT architecture and software development background. He is the founder of **apoIO** company.

He is expert in Cloud architecture including Kubernetes and Docker (CaaS), AWS (IaaS, Functions), Openstack (IaaS). He has experiences on Scrum and agile methodologies as Scrum Master or Product Owner. He has worked in DevOps teams in Version Control-Based Deployment, Scale IT Automation and Deliver Value Continuously including Continuous Integration, Continuous Delivery. Mahyar Sepehr has several successful experiences in designing solutions with PI System in different industrial sectors from proof of concept to performance testing and production monitoring.



apoIO

Architecture . Cloud . Big Data

We work with our clients to optimize their **Digital Journey**

- Cloud Architecture
- Internet Of Things
- Industrial Big Data:
  PI System

As a solid actor in Software programming and Architecture designer, our first step is to understand the Business rules in the industry. The second step is to find ways to transform those constraints into assets.
apoIO is exploring and redefining new frontiers in IT. With products and services as Solutions and the Data as the Business.

🌐 www.apoio.fr

✉ info@apoio.fr