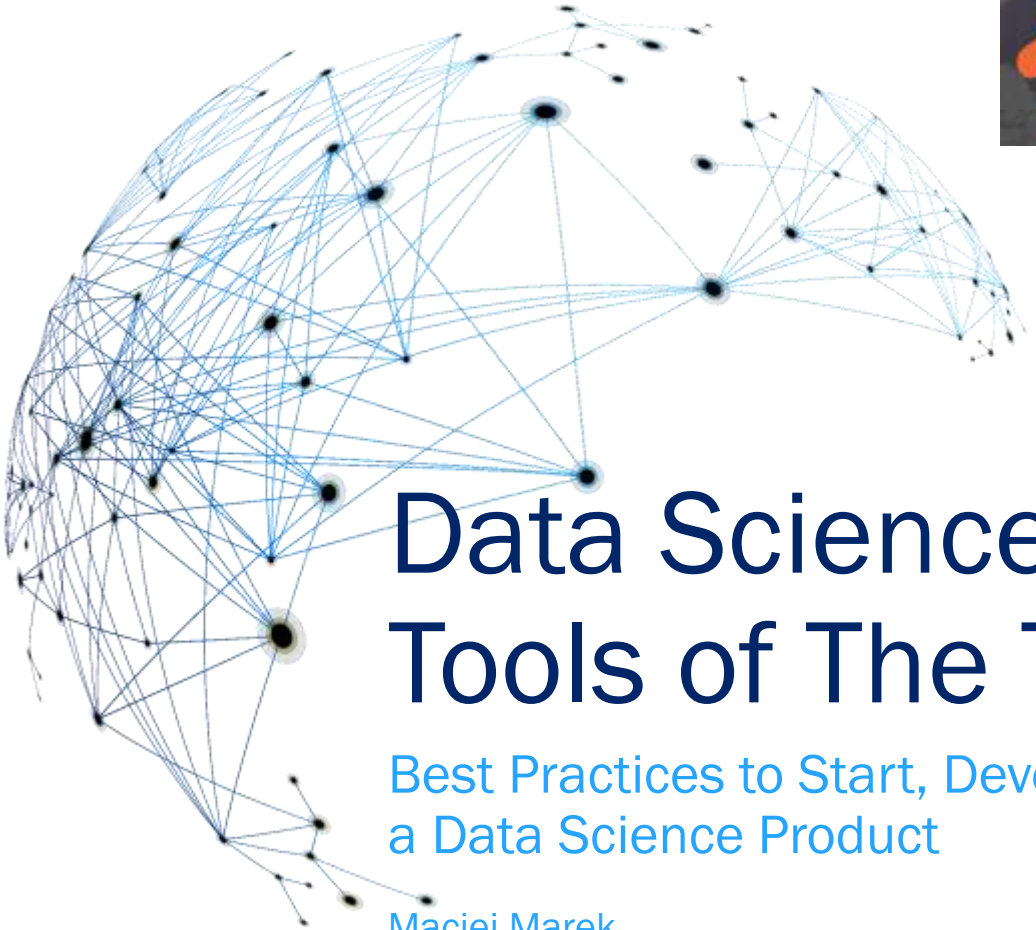


AI UKRAINE 19
INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND DATA SCIENCE APPLICATIONS

EAD | DIGITAL

Data Science @ PMI
Tools of The Trade

Best Practices to Start, Develop and Ship a Data Science Product

Maciej Marek

AI Ukraine, 21st September 2019 Kyiv

PHILIP MORRIS INTERNATIONAL

About me

- *Now*: Enterprise Data Scientist at PMI in Krakow
- *Education*: Computer Science
- *Every day*: Data Science best practices ambassador
- *Likes*: Motorbike trips, aviation

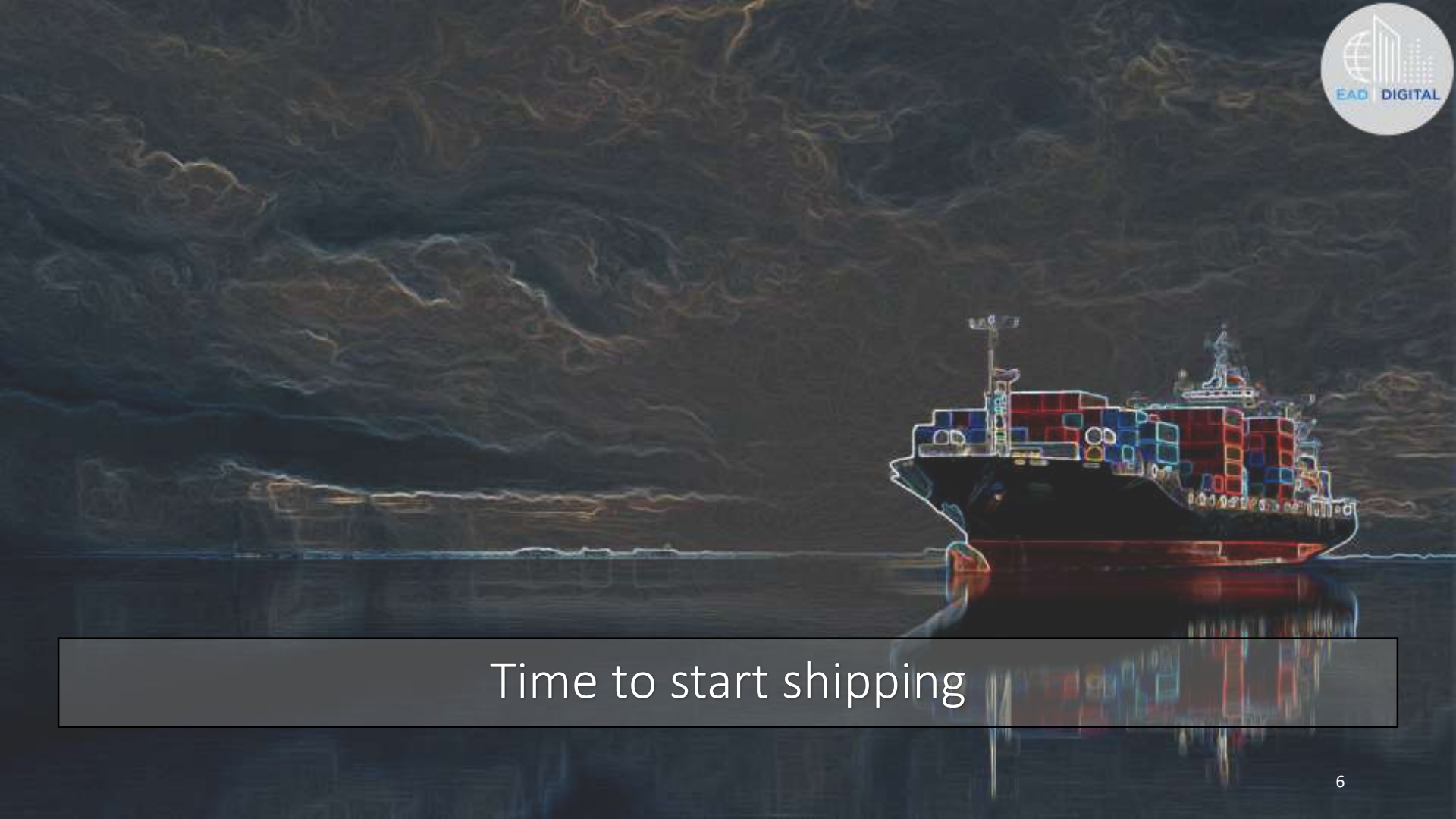
PMI & Digital

Machine Learning & AI @Enterprise level

Challenge: fast delivery of reproducible models to productions!

What is a Data Product?

- A software system with a ML/AI component,
  part of a large business system
- Formally defined, it is a system that:
  - takes raw data as input,
  - applies a machine-learned model to it,
  - produces data as output
- Additionally, a data product must be
  - dynamic and maintainable, allowing periodic updates
  - Responsive, performant and scalable



Time to start shipping

Culture Conflict

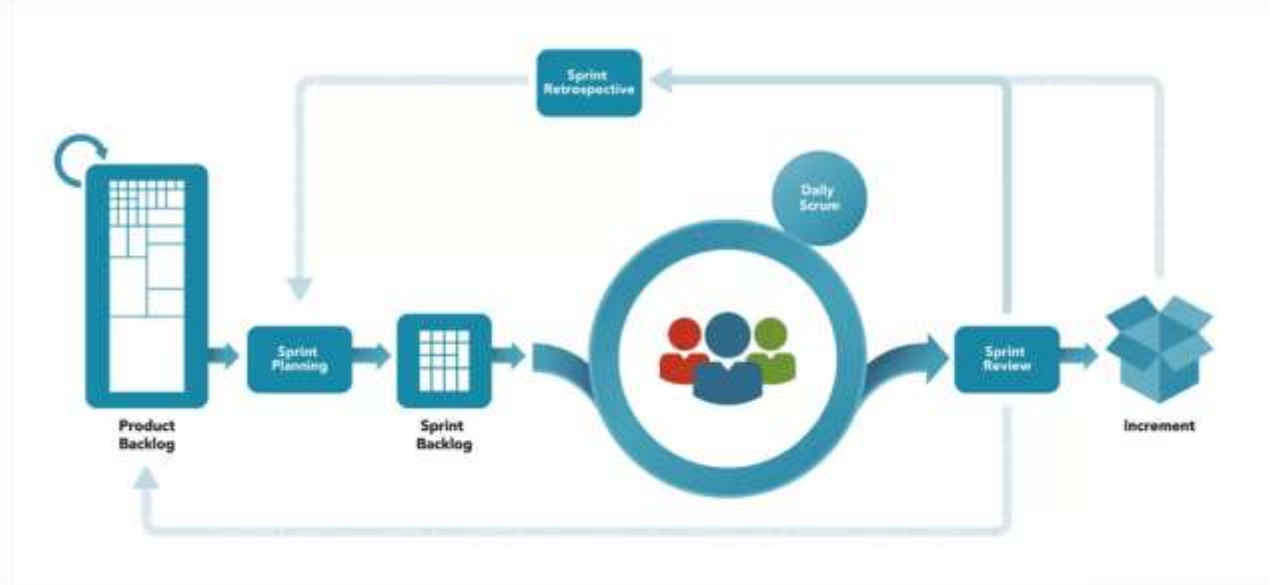
© Josh Wills
http://www.slideshare.net/g3illaIII/dataeng-conf-sf16-bridging-the-gap-between-data-science-and-data-engineering

- Companies structure DS divisions into
  Data Scientists – *build models*,
  Data Engineers – *put models into productions*,
  DevOps - *maintain the platform*

- Tendency to go into silos and do their owe thing.

- You hear things like '*But it work on my machine*'
  '*Here's my Jupyter Notebook, please industrialize it*'

- Everyone get caught in a vicious cycle of frustration



SCRUM certified

Data Science @ PMI

- We are part of PMI's **Enterprise Analytics and Data** (EAD) group
- **40+** Data Scientists across **4** hubs
- Offices in **Amsterdam** (NL), **Kraków** (PL), **Lausanne** (CH) and **Tokyo** (JP)

- **Profiles**
  - Education: **30%** PhD, **70%** MSc/BSc
  - Data Science Experience: **7.4 yrs** on average
  - Experience in PMI: **88% under 2yrs**

- Expertise in **Machine Learning**, **Big Data Engineering**, **Insights Communication**
- **SCRUM** certified

OCT

2012

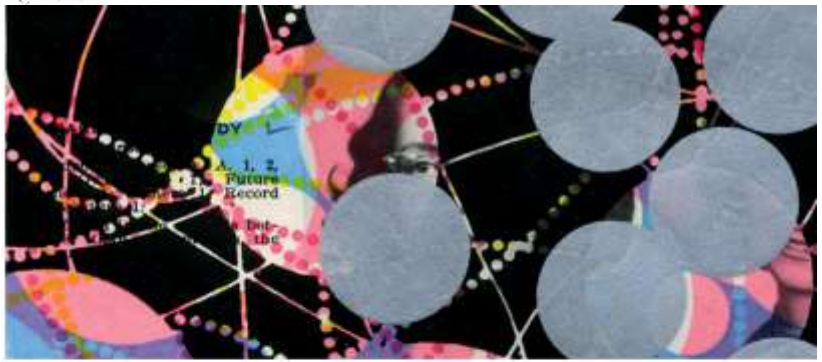
EAD | DIGITAL

Thomas H. Davenport and Dhanurjay Patil

OCT

2012

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT BUY COPIES

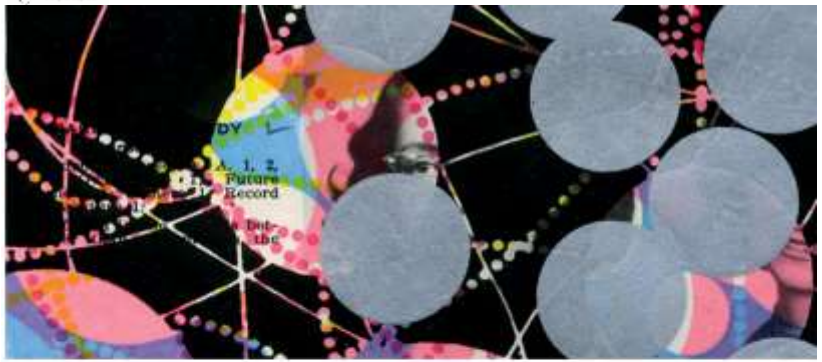
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

OCT

2012

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT TEXT SIZE PRINT BUY COPIES

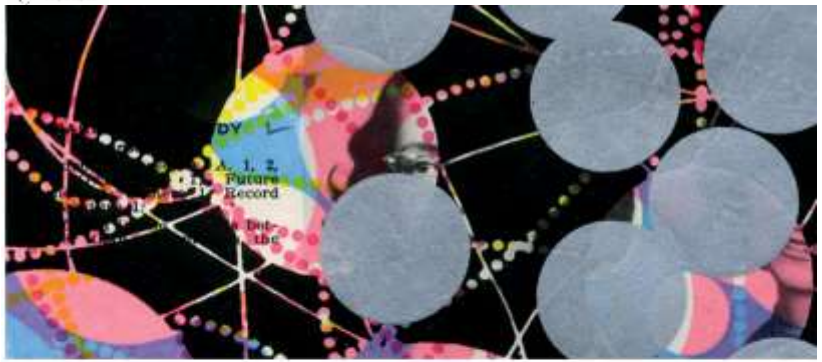
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

OCT

2012

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

OCT

2012

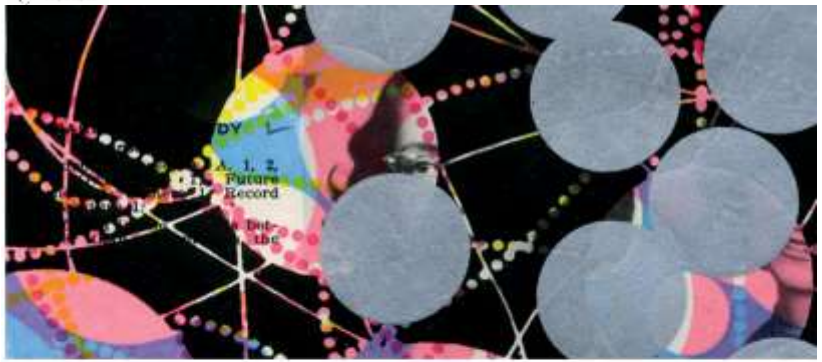
EAD | DIGITAL

THE TRUTH

Harvard Business Review

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

OCT

2012

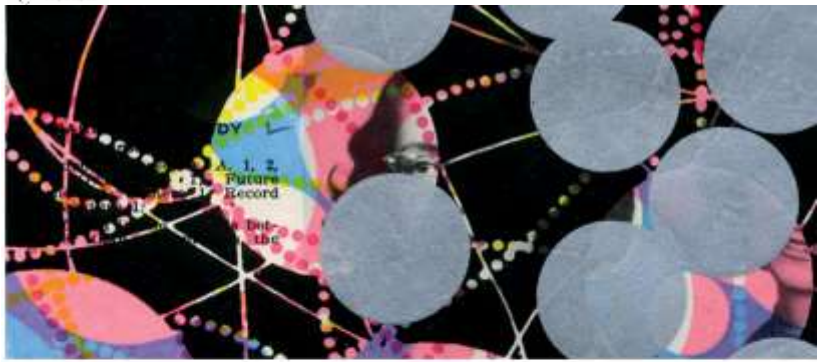
EAD | DIGITAL

THE TRUTH?

Harvard Business Review

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

OCT

2012

EAD | DIGITAL

THE TRUTH?

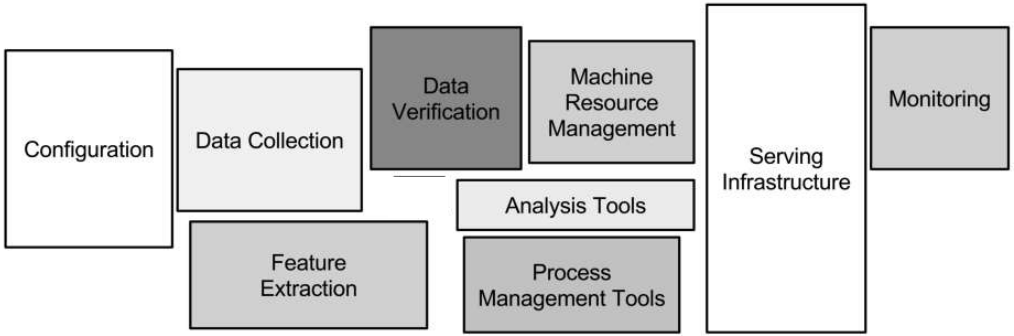
Harvard Business Review

http://reportajede.news/wp-content/uploads/2016/09/Graduados.jpg

http://reportajede.news/wp-content/uploads/2016/09/Graduados.jpg

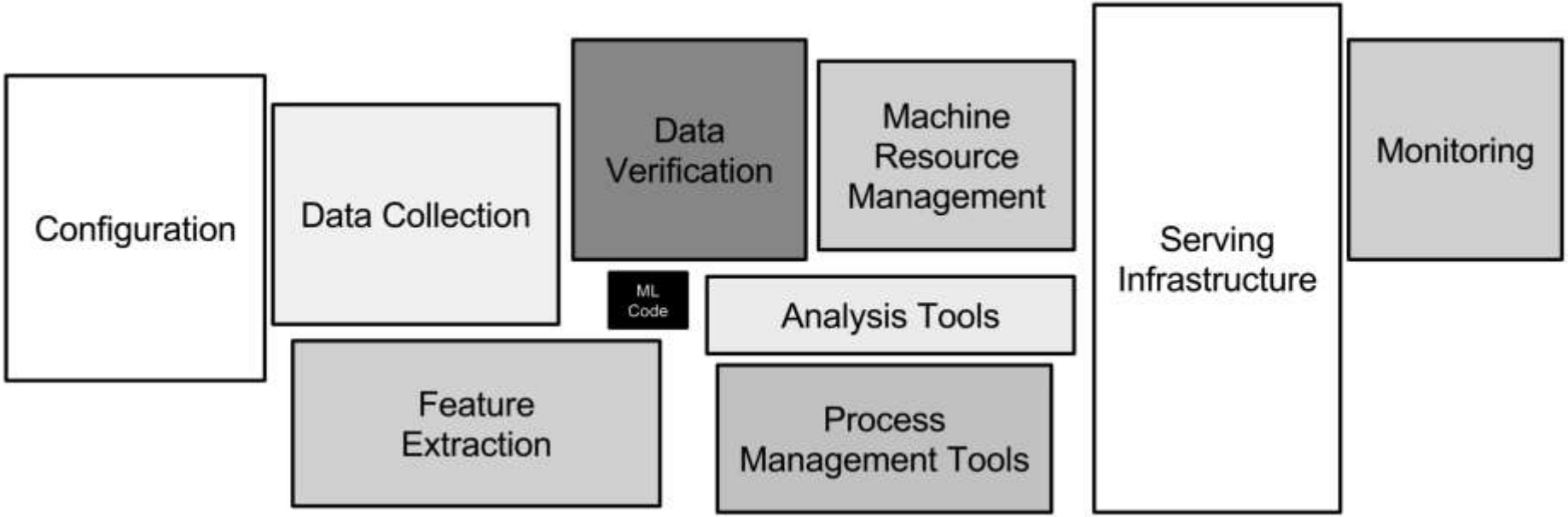
Perception

ML
Code



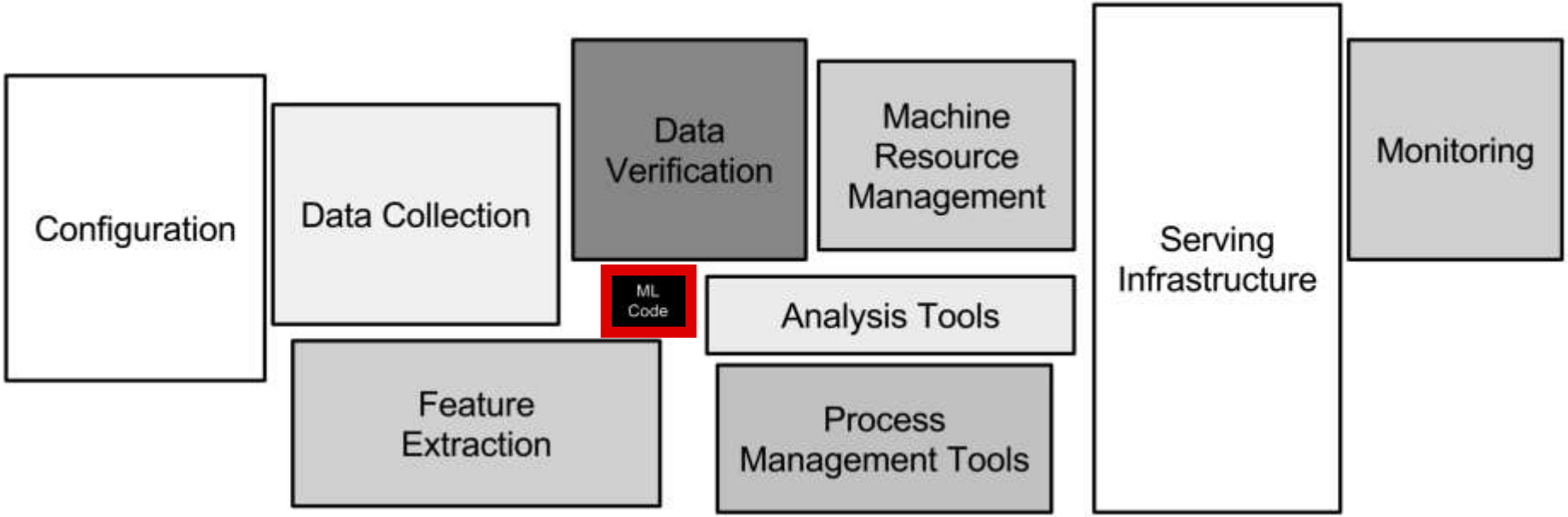
https://developers.google.com/machine-learning/crash-course/production-ml-systems

Reality



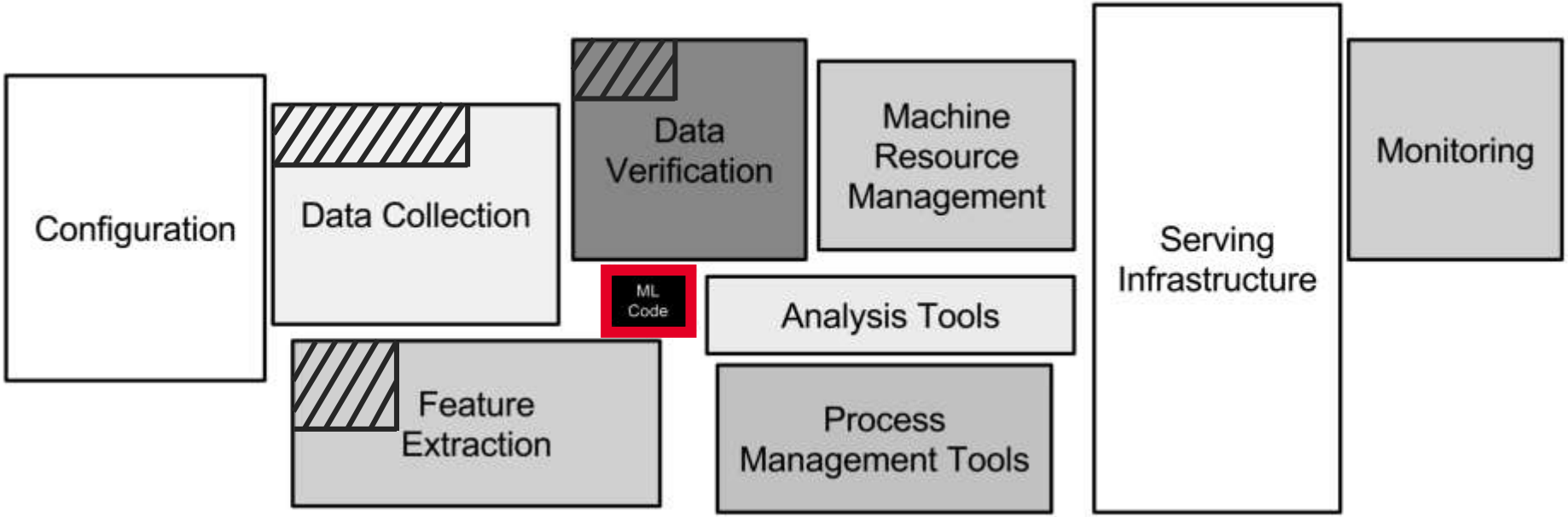
https://developers.google.com/machine-learning/crash-course/production-ml-systems

Reality



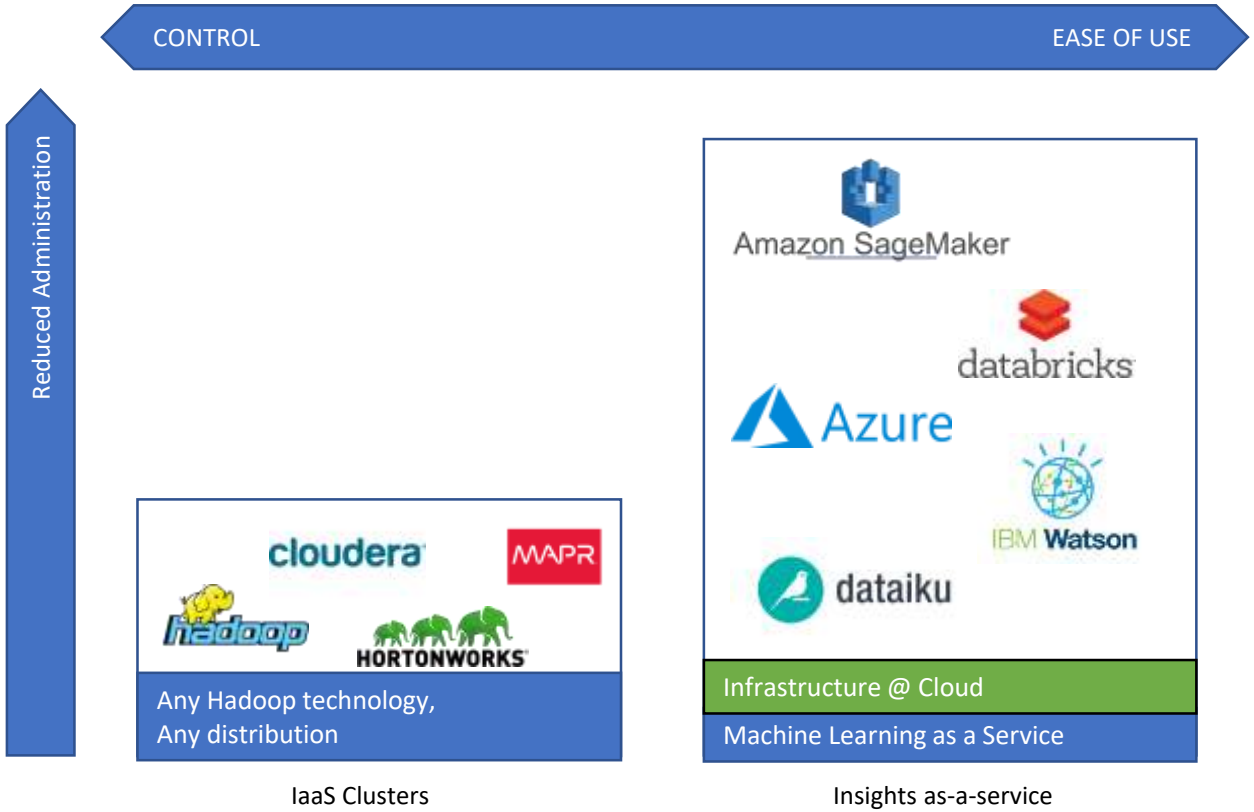
https://developers.google.com/machine-learning/crash-course/production-ml-systems

Reality



https://developers.google.com/machine-learning/crash-course/production-ml-systems

Machine Learning as a Service



IaaS Clusters

Insights as-a-service

IaaS Clusters

PMI Data Ocean

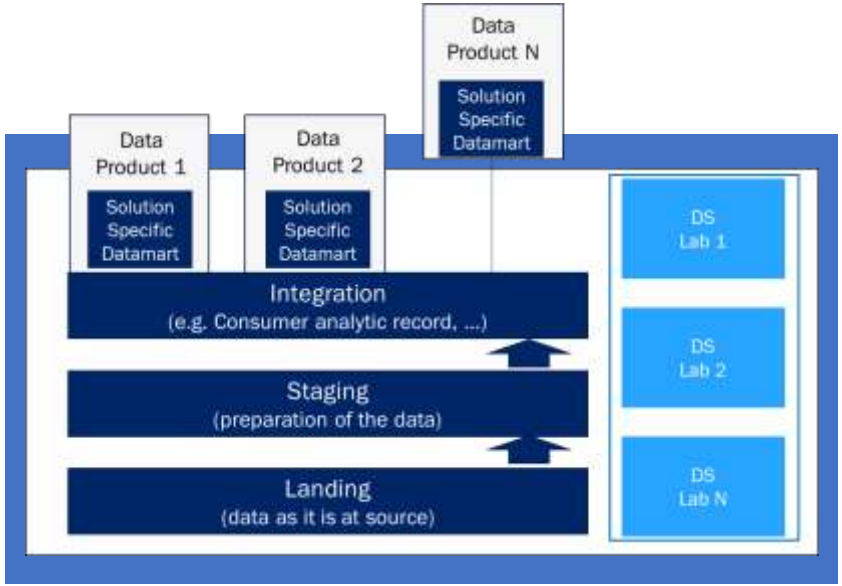
Insights as-a-service

IoT/streaming data

Cloud storage

Hadoop storage

Data warehouses



PMI DATA OCEAN

Data Products

BI tools

Data exports

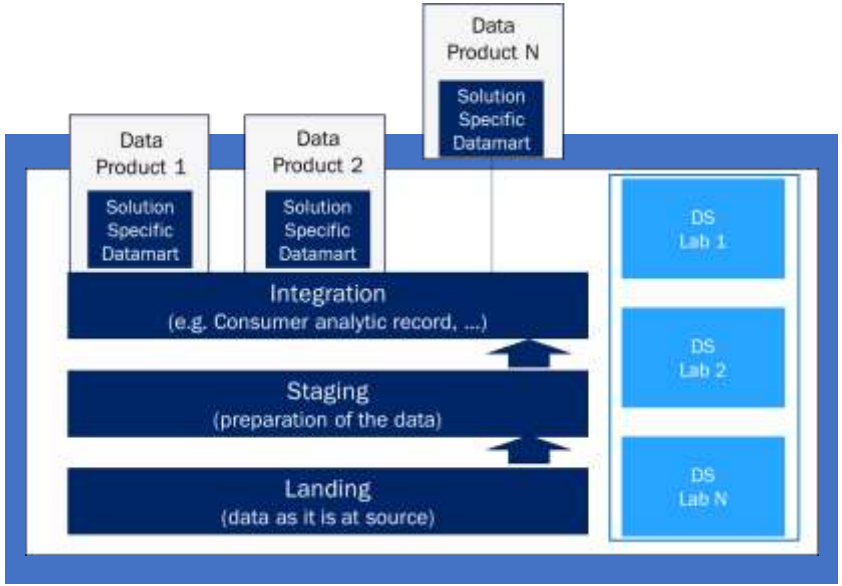
Data warehouses

IoT/streaming data

Cloud storage

Hadoop storage

Data warehouses



PMI DATA OCEAN

Data Products

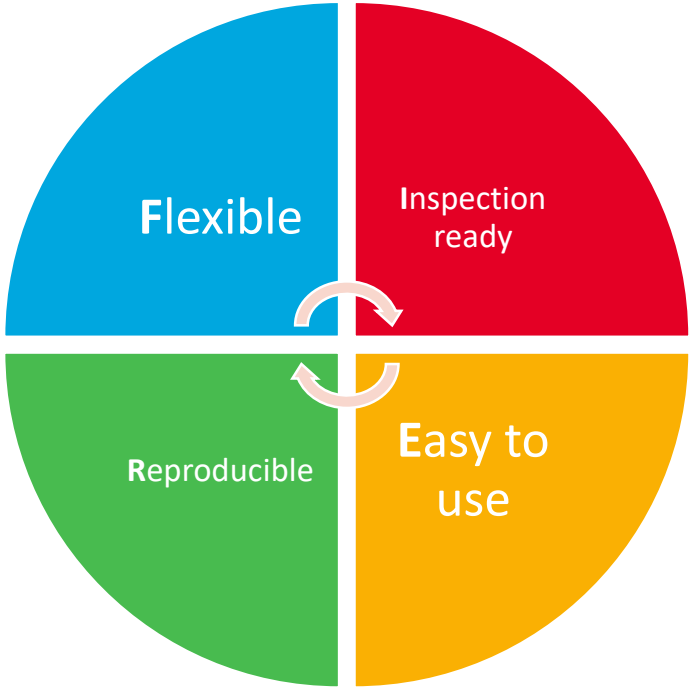
BI tools

Data exports

Data warehouses

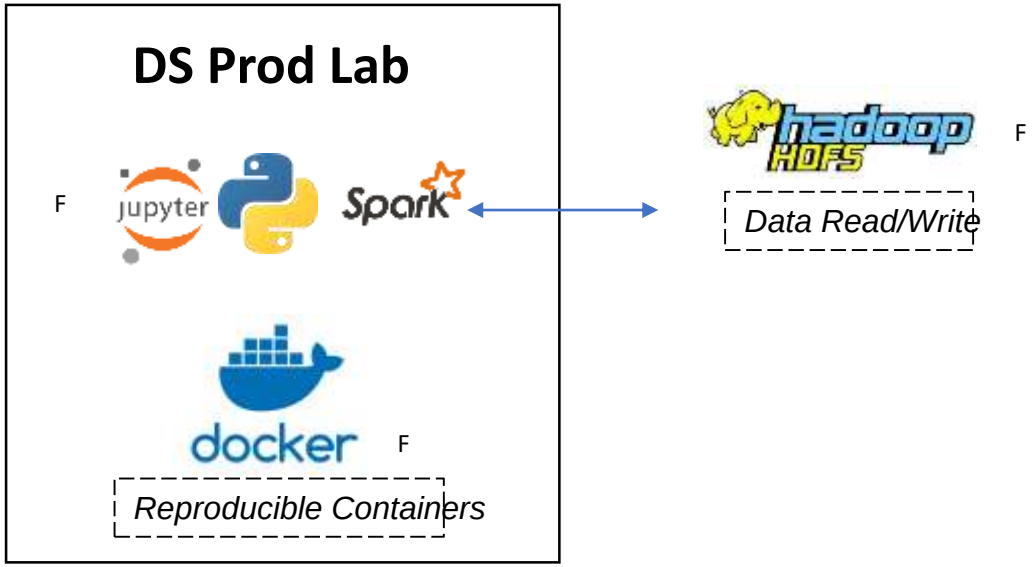
Our Vision

To create a workflow that is …



Data Product Architecture @PMI

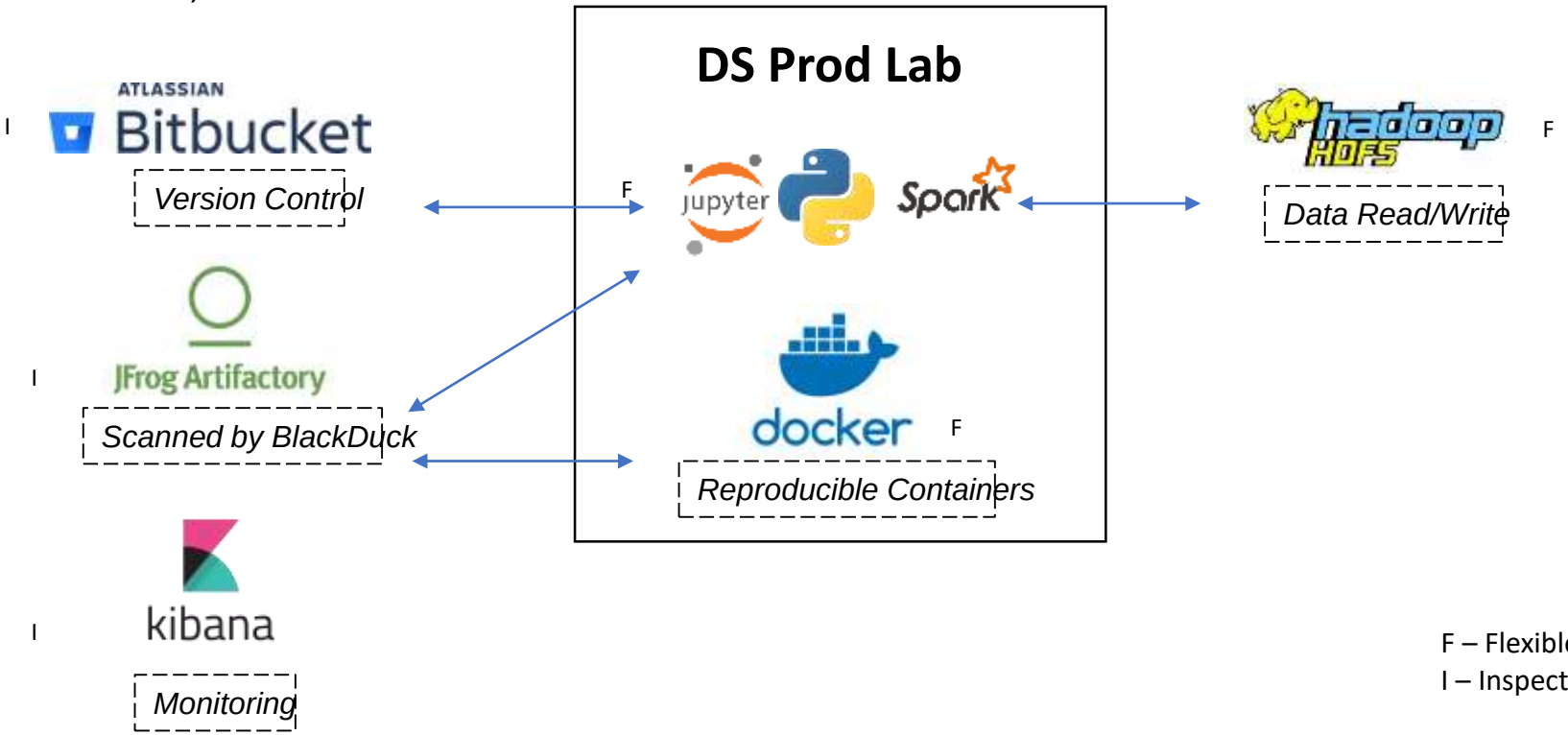
- The dots, connected.



F – Flexible

Data Product Architecture @PMI

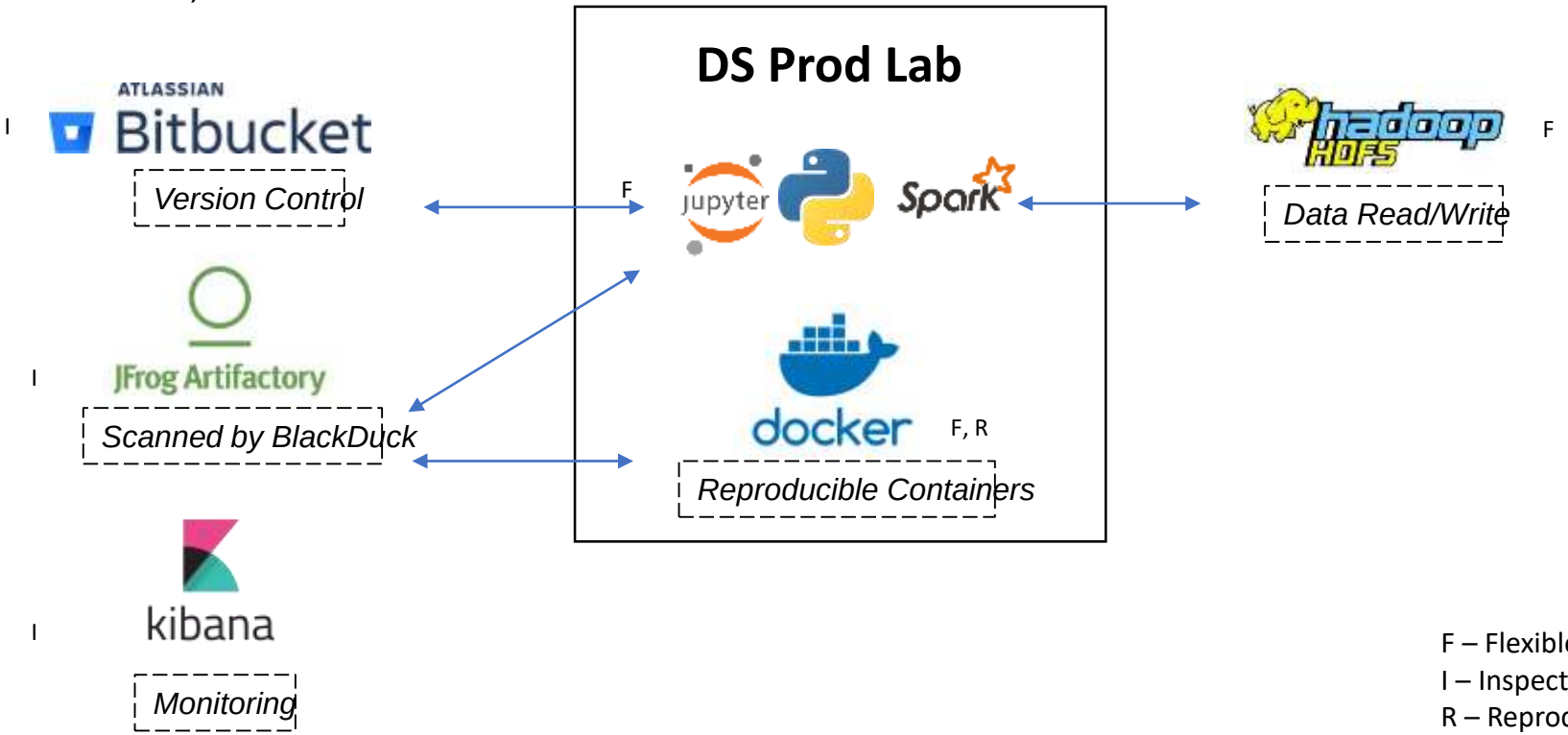
- The dots, connected.



F – Flexible
I – Inspection ready

Data Product Architecture @PMI

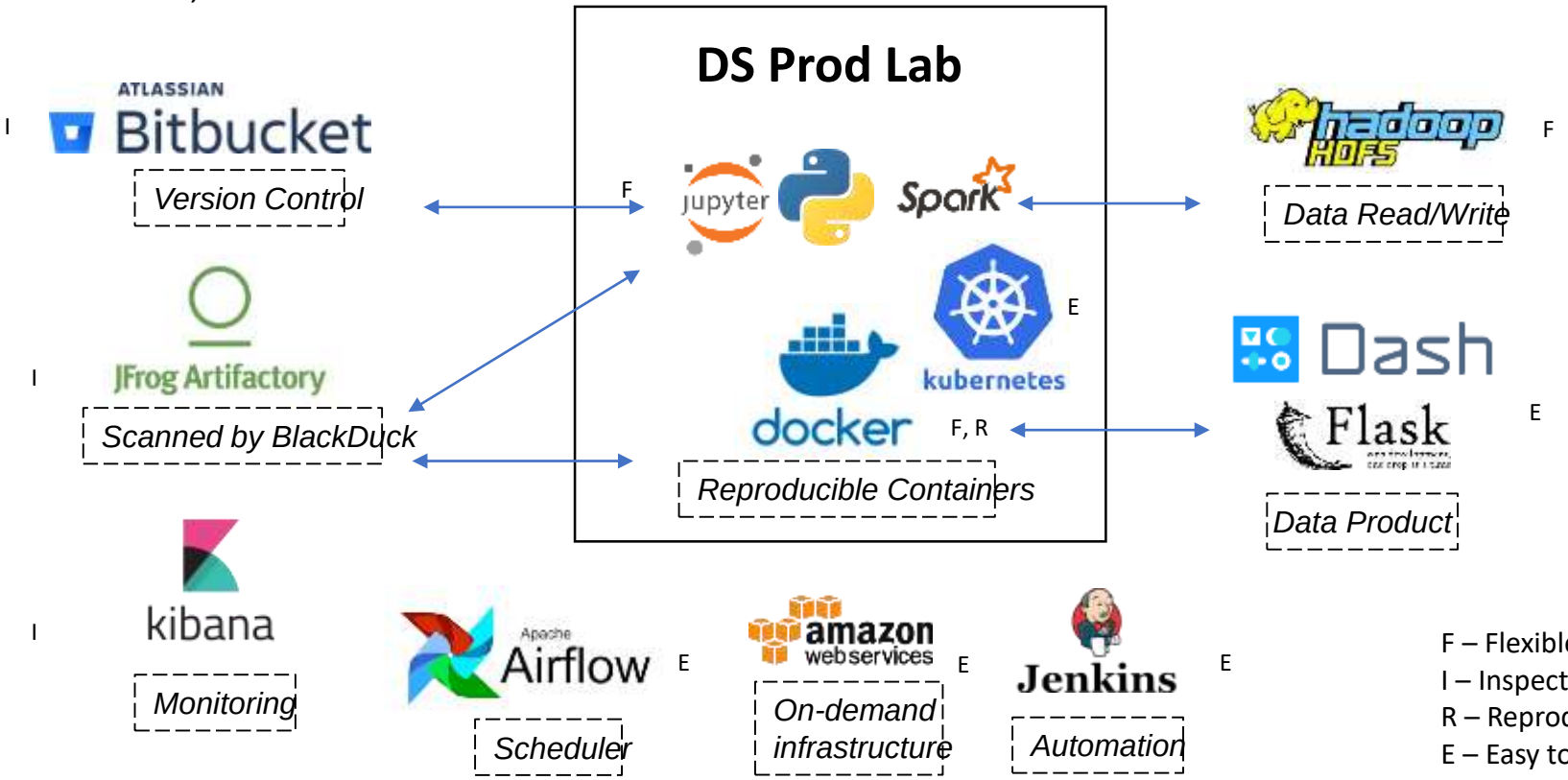
- The dots, connected.



F – Flexible
I – Inspection ready
R – Reproducible

Data Product Architecture @PMI

- The dots, connected.



F – Flexible
I – Inspection ready
R – Reproducible
E – Easy to use

Data Science Best Practices @ PMI

EAD | DIGITAL

- Docker
- CICD
- Code Reviews
- Project Templates
- Python Style Guides
- Testing
- Version Control
- Notebooks to Modules

Data Science Best Practices @ PMI

EAD | DIGITAL

- Docker
- CICD
- Code Reviews
- Project Templates

- Python Style Guides
- Testing
- Version Control
- Notebooks to Modules

For Reproducibility

Docker Containers

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Ease of installation

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Ease of installation

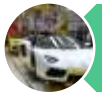
Reproducibility

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Ease of installation



Reproducibility

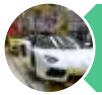
Isolation

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Ease of installation



Reproducibility

Isolation

Speed

Docker for Containerized Data Science

All your dependencies in one place.
Code guaranteed to run anywhere.

Freedom

Ease of installation

Reproducibility

Isolation

Speed

>100GB

APACHE Spark™

docker

For organization and predictability

Project Templates

CookieCutter

Everything has a place and a purpose

https://wallpapershome.com/images/wallpapers/holiday-cookies-3840x2160-baking-food-holiday-cookie-shape-409.jpg

CookieCutter

Everything has a place and a purpose

```
ds-cookiecutter
  data
    interim
    processed
    raw
  figures
  models
  notebooks
      01-Obtain.ipynb
      02-Scrub.ipynb
      03-Explore.ipynb
      04-Model.ipynb
      05-iNdustrialize.ipynb
  references
  reports
  src
    explore
    industrialize
    model
    obtain
    scrub
      __init__.py
    README.md
    DOCKERFILE
    Jenkinsfile
    environment.yml
```

https://wallpapershome.com/images/wallpapers/holiday-cookies-3840x2160-baking-food-holiday-cookie-shape-409.jpg

For integration and deployment

CICD for Data Science

“It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change.”

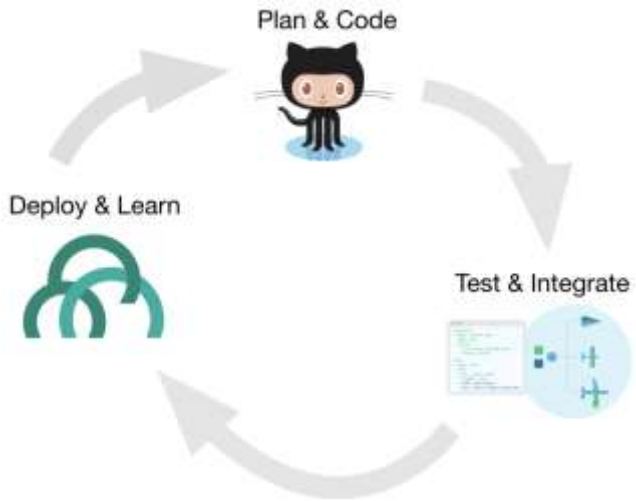
“It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change.”

Charles Darwin

Continuous Integration (CI), Delivery (CD) and Deployment

Development practices for overcoming integration challenges and moving faster to delivery

- **Continuous Integration** requires multiple developers to integrate code into a shared repository frequently. Requested merges are automatically tested and reviewed.
  - Enabled by git-flow, code standards and automated testing



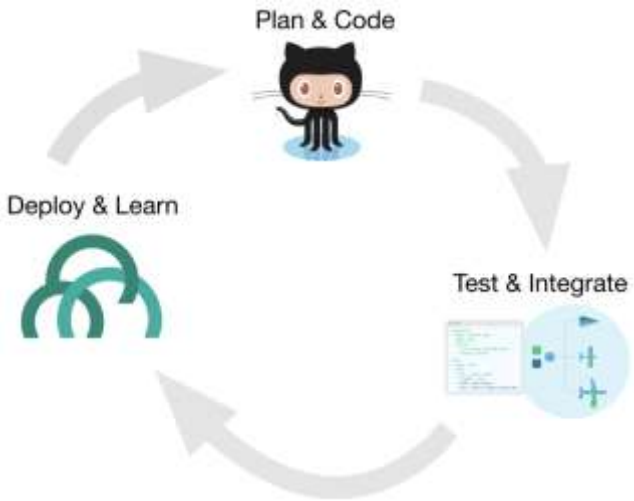
The CI/CD Cycle

Continuous Integration (CI), Delivery (CD) and Deployment

Development practices for overcoming integration challenges and moving faster to delivery

- **Continuous Integration** requires multiple developers to integrate code into a shared repository frequently. Requested merges are automatically tested and reviewed.
  - Enabled by git-flow, code standards and automated testing

- **Continuous Delivery** makes sure that the code that we integrate is always in a deploy-ready state.
  - Enabled by agile (iterative) methods, testing and build automation



The CI/CD Cycle

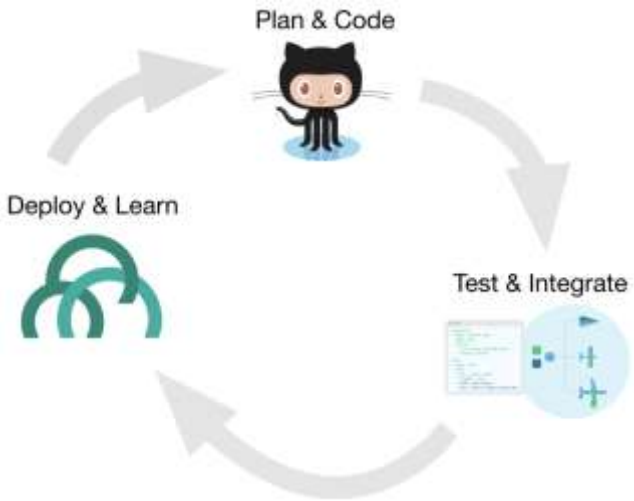
Continuous Integration (CI), Delivery (CD) and Deployment

Development practices for overcoming integration challenges and moving faster to delivery

- **Continuous Integration** requires multiple developers to integrate code into a shared repository frequently. Requested merges are automatically tested and reviewed.
  - Enabled by git-flow, code standards and automated testing

- **Continuous Delivery** makes sure that the code that we integrate is always in a deploy-ready state.
  - Enabled by agile (iterative) methods, testing and build automation

- **Continuous Deployment** is the actual act of pushing updates out to the user – think of your iPhone apps or Desktop browser that prompt for updates to be installed periodically.



The CI/CD Cycle

Continuous Integration (CI), Delivery (CD) and Deployment

From business needs to value

Exploration phase

Automation phase

Project Phase

Alpha

Beta

Industrializatio

Code Review

Jenkins

Understanding business needs

Web-based application

Generating value

For monitoring

Kibana and Airflow

kibana

- Discover
- Visualize
- Dashboard
- Timelion
- Dev Tools
- Management

application logins - metric

84

application logins

model runs - metric

671

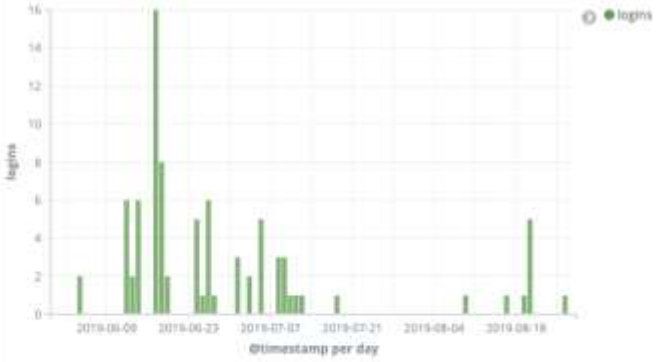
model runs

csv downloads - metric

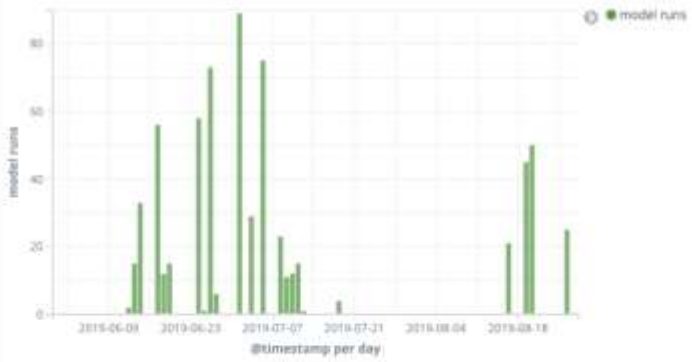
8

csv downloads

application logins - bar

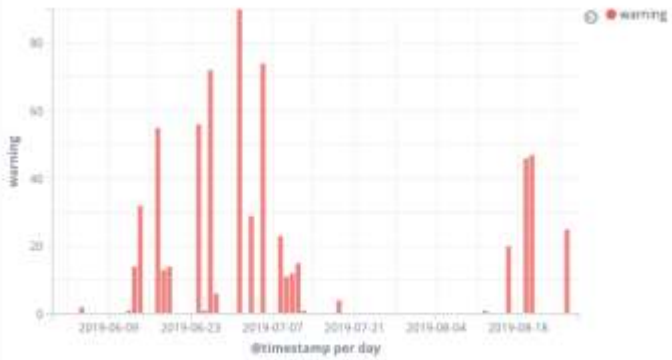


model runs - bar



model runs - cumulative

application warnings



EAD | DIGITAL

Airflow | DAGs | Data Profiling | Browse | Admin | Docs | 13:04 UTC

DAGs

Show entries

Search:

	DAG	Schedule	Owner	Recent Statuses	Links
On	example_bash_operator	0 0 * * *	airflow		
On	example_branch_operator	@daily	airflow	5	
On	example_http_operator	1 day, 0:00:00	airflow	1 1 4	
On	example_passing_params_via_test_command	*/1 * * * *	me	1 1	
On	example_python_operator	None	airflow		
On	example_short_circuit_operator	1 day, 0:00:00	airflow	4	
On	example_subdag_operator	@once	airflow	5	
On	example_trigger_controller_dag	@once	me	1	
On	example_trigger_target_dag	None	airflow	2	
On	example_upload	@daily	airflow	5	
On	example_xcom	@once	airflow	3	

https://docs.qubole.com/en/latest/user-guide/data-engineering/airflow/monitor-airflow-cluster.html

Qubole | Sparklens

binder

Turn a Git repo into a collection of interactive notebooks

ONNX

mlflow

DASK

Continuous Technology Scouting

Best Practice Ambassador Network @ PMI

- Supporting technical assessment of tools, methods, etc.
- Preparing & conducting internal trainings
- Providing on-site support to fellow data scientists
- Facilitating retrospectives of other teams
- Actively participating in local meetups & conferences
- Designing & implementing data science specific solutions for improving the data science work effectiveness

In Conclusion

- Engineering smart systems around a machine-learned core is **difficult**
- It requires teams of exceptionally talented individuals to work together.
- What makes data scientists special is their ability to work with **both** business leaders and technology experts.
- We must acknowledge that we are a part of something much bigger and learn to play well with each other and with all parties involved.

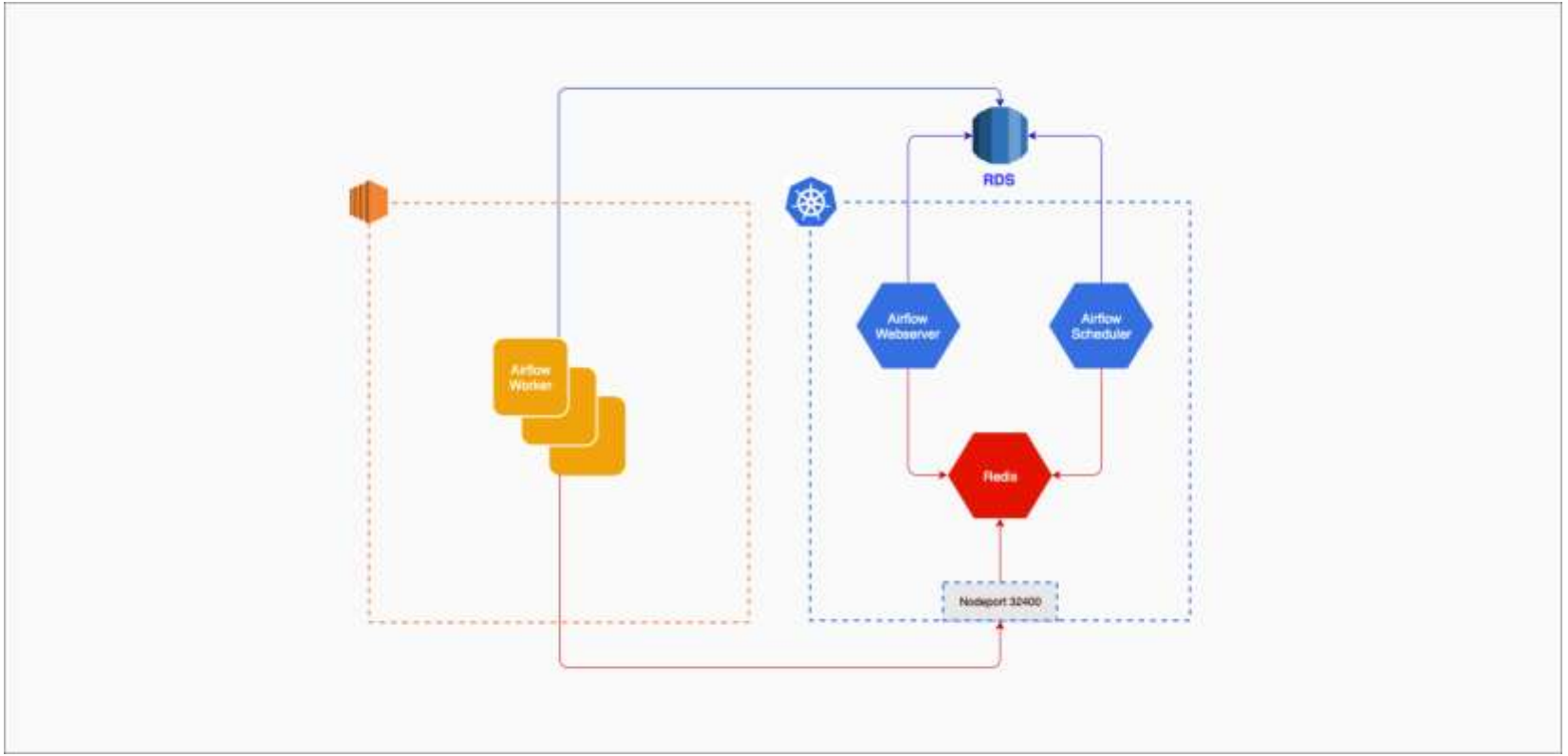
Our hope is that these systems, principles and best practices will help you take the first steps in that direction

Thank you!

Maciej.Marek@pmi.com

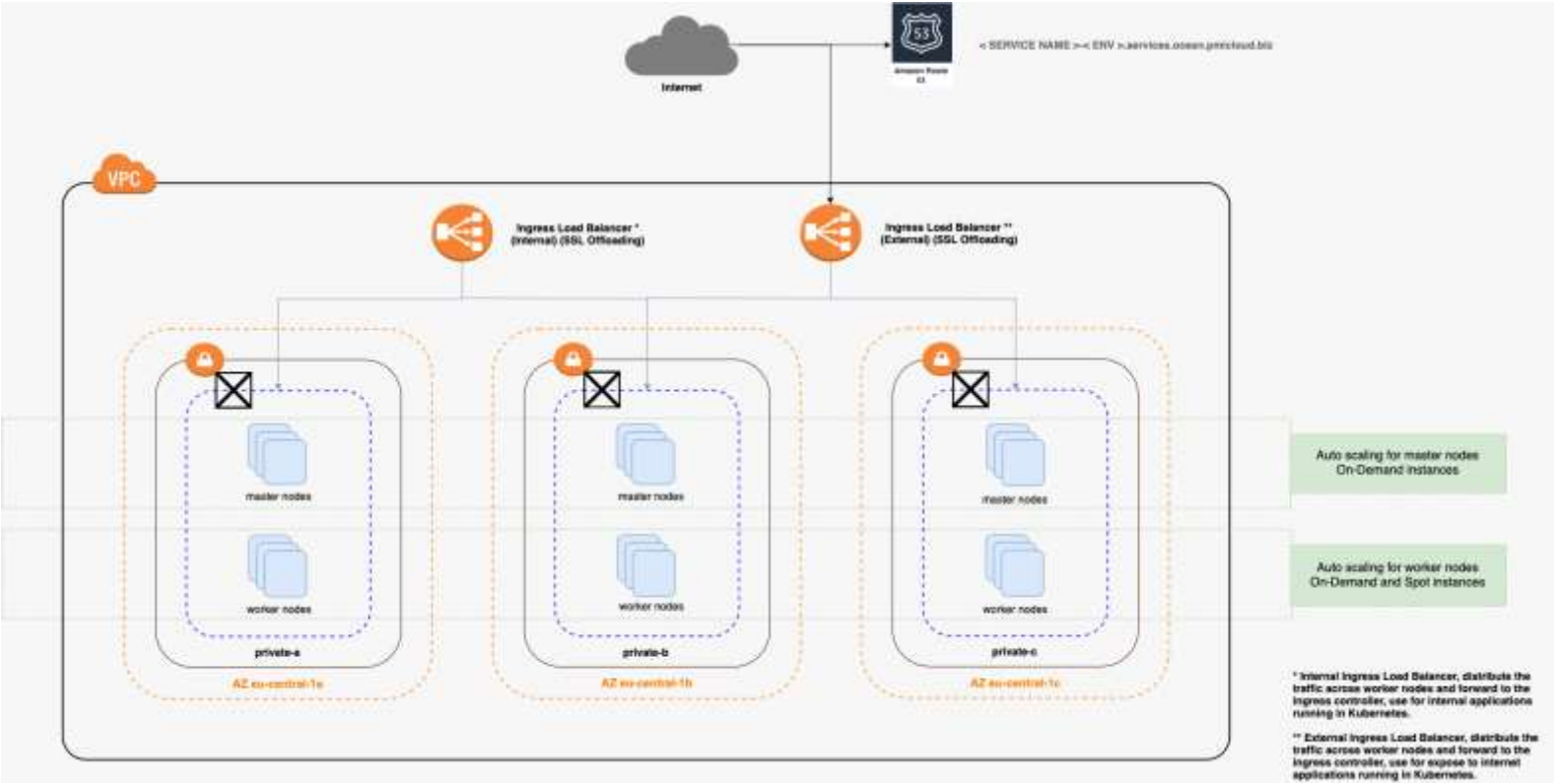
Airflow

Architecture



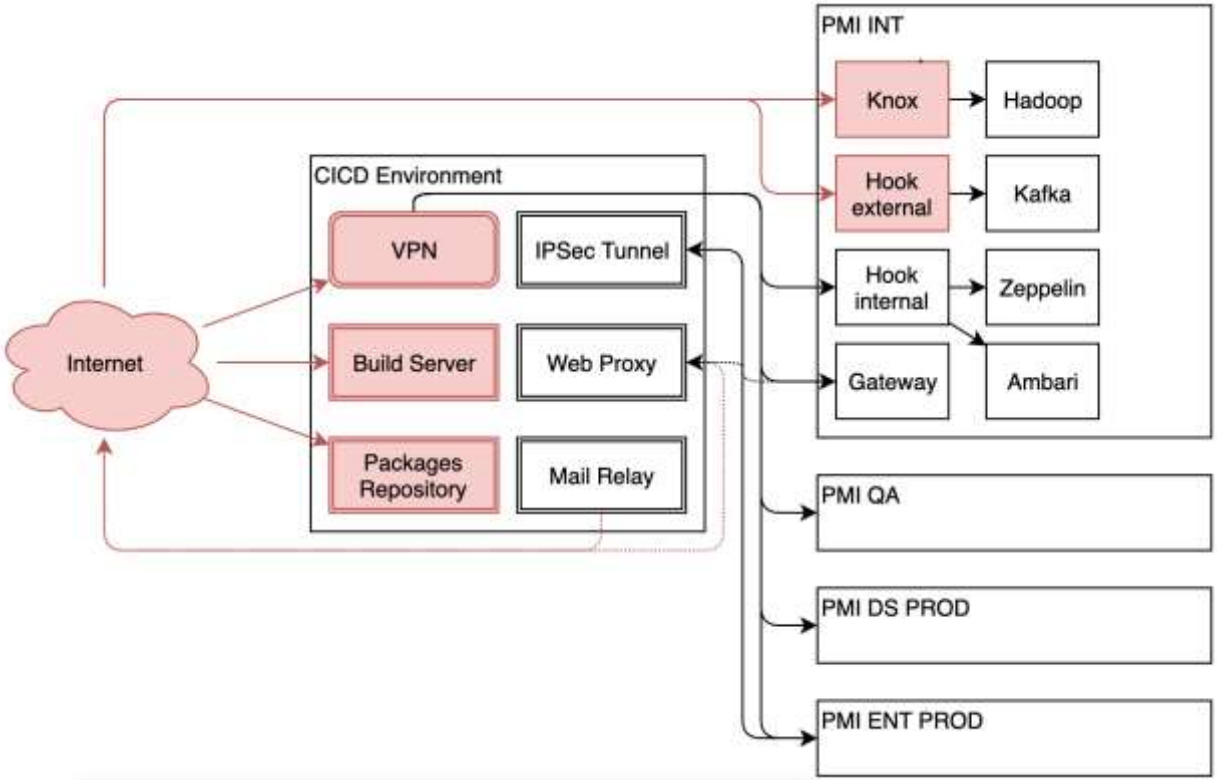
Kubernetes

Infrastructure



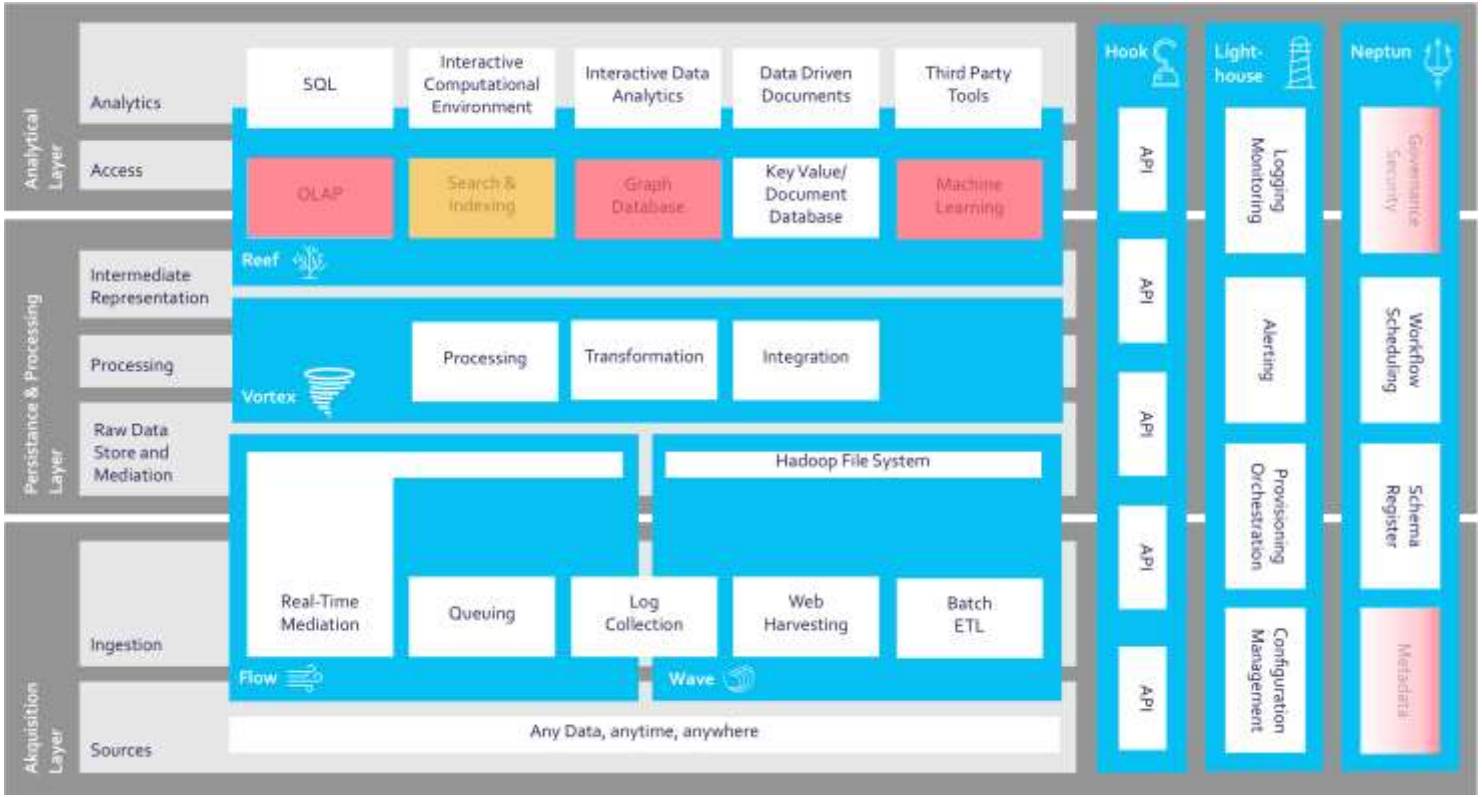
Ocean platform

Infrastructure



Ocean platform

Capability view



Ocean platform

Instantiation view