July 16, 2020

Autonomous Driving with Airflow

Where big-data meets high performance computing

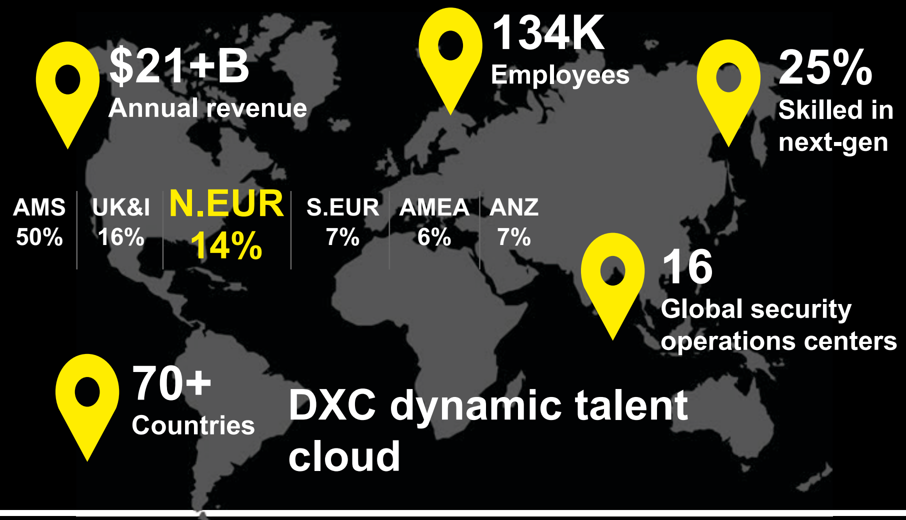
Amr Noureldin – Solution Architect
Michal Dura – Big Data Engineer

DXC.technology

luxoft
A DXC Technology Company

The world's leading independent, end-to-end IT services company

Scale & Skills



$21+B
Annual revenue

134K
Employees

25%
Skilled in next-gen

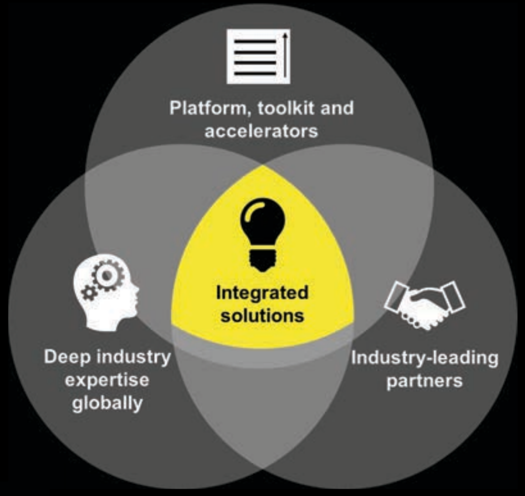
AMS	UK&I	N.EUR	S.EUR	AMEA	ANZ
50%	16%	14%	7%	6%	7%

16
Global security operations centers

70+
Countries

DXC dynamic talent cloud

DXC Value for AD



Platform, toolkit and accelerators
Integrated solutions
Deep industry expertise globally
Industry-leading partners

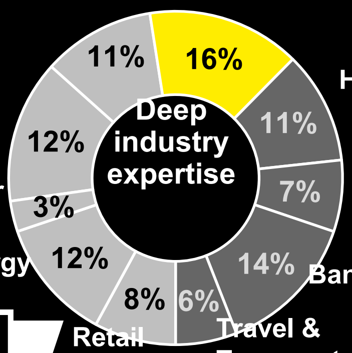
Accelerate time to market

Reduce cost and risk

Improve market leadership

DXC.technology

Customer Intimacy



Global PS

Manufacturing & Automotive

U.S. Public

Healthcare

Other

Insurance

Energy

Banking

Retail

Travel & Transportation

Excellent client coverage across the globe …

~6,000	200+	36
Clients	F500 clients	NPS

… enhanced through world-class partner network

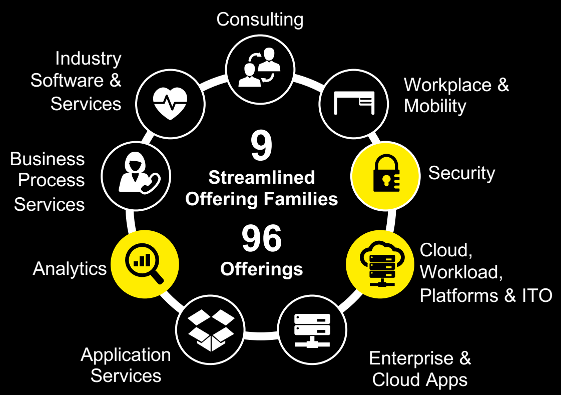
Technology-Driven Innovation

$4B
Digital revenue

250+
global partners

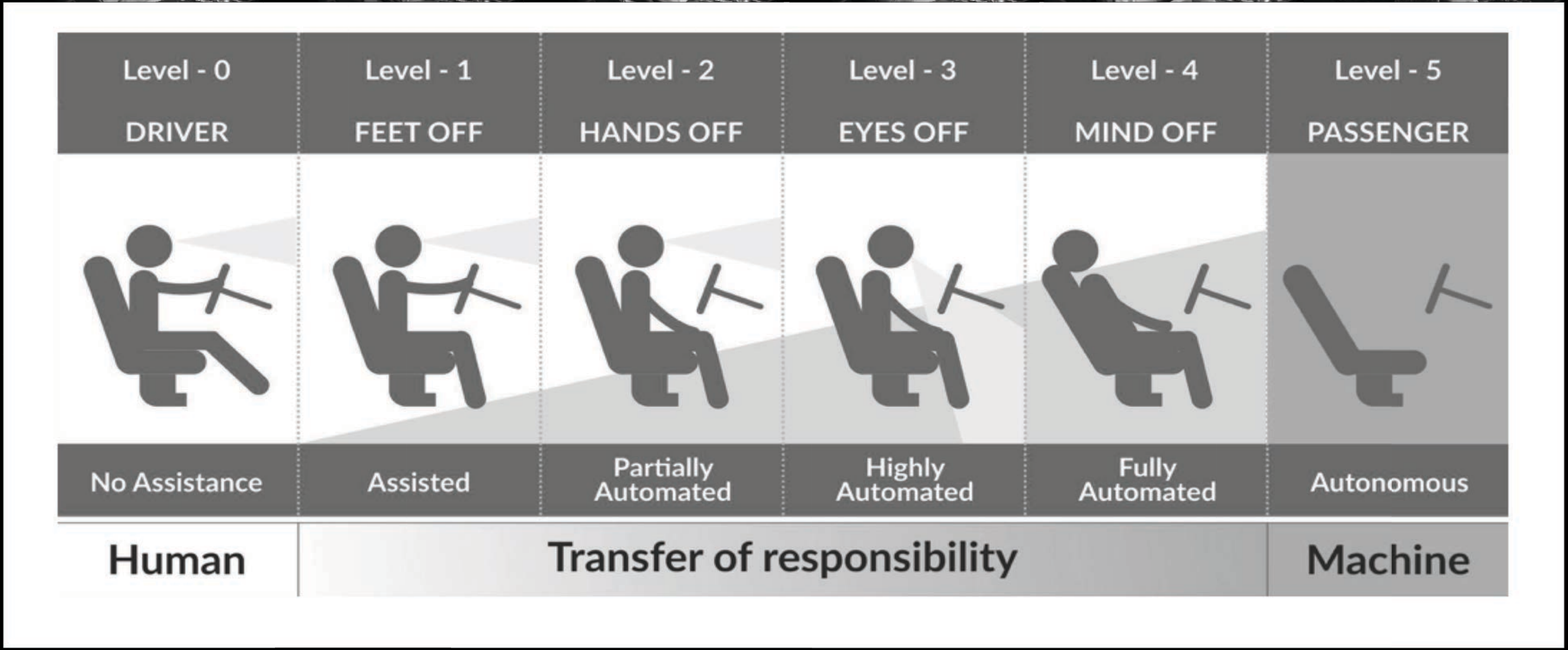
14 strategic co-investing partners

Streamlined offerings



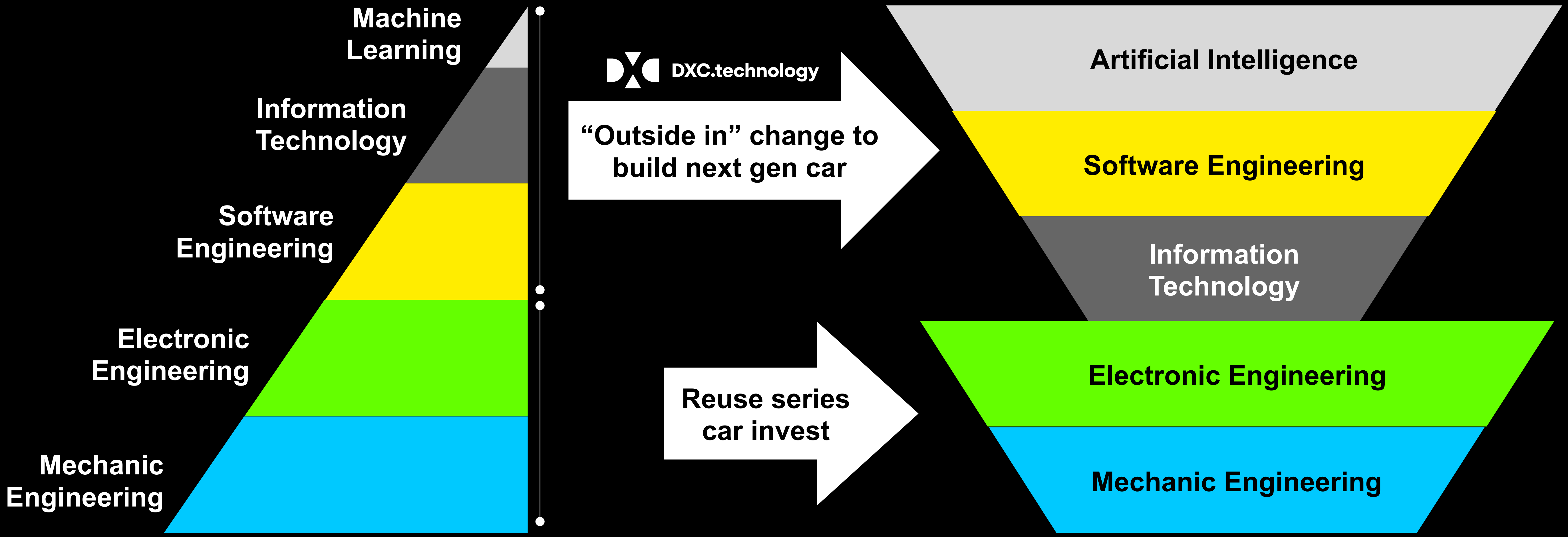
Consulting
Workplace & Mobility
Security
Cloud, Workload, Platforms & ITO
Enterprise & Cloud Apps
Application Services
Analytics
Business Process Services
Industry Software & Services
9
Streamlined Offering Families
96
Offerings

Autonomous Driving



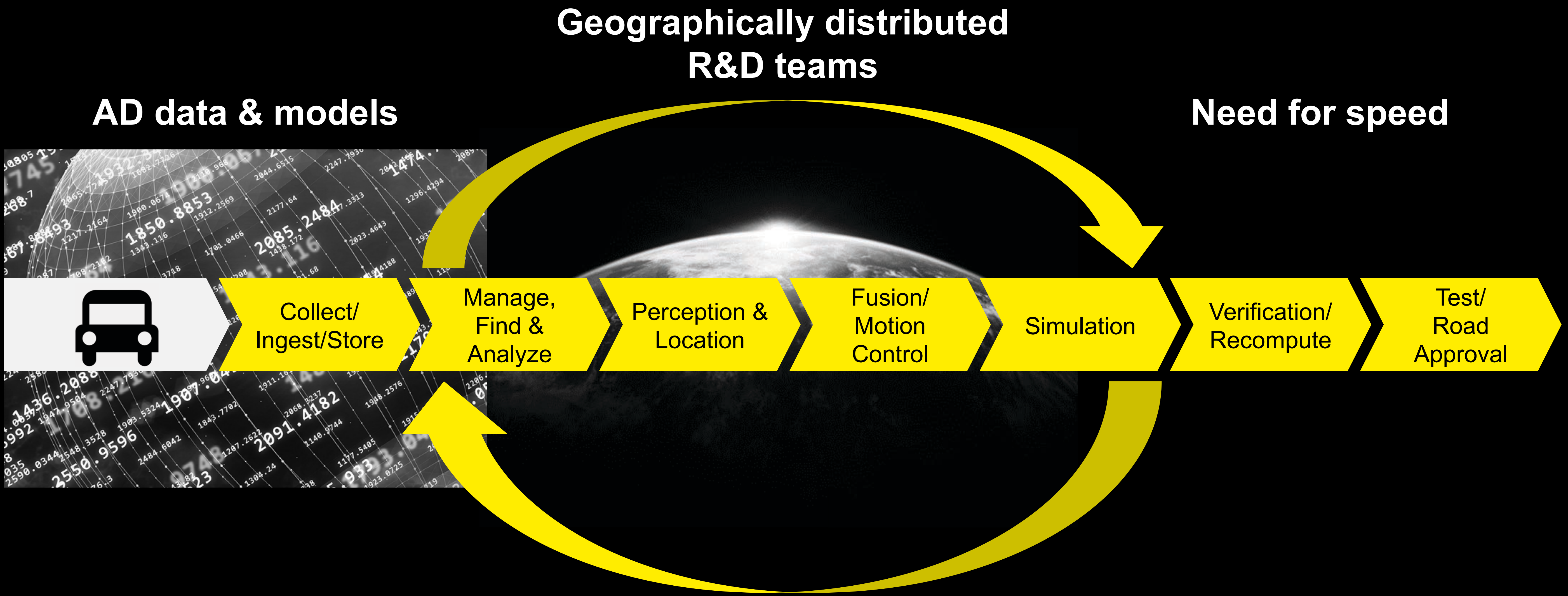
Level - 0	Level - 1	Level - 2	Level - 3	Level - 4	Level - 5
DRIVER	FEET OFF	HANDS OFF	EYES OFF	MIND OFF	PASSENGER
No Assistance	Assisted	Partially Automated	Highly Automated	Fully Automated	Autonomous
Human	Transfer of responsibility				Machine

R&D in Automotive Industry – Capabilities are Changing



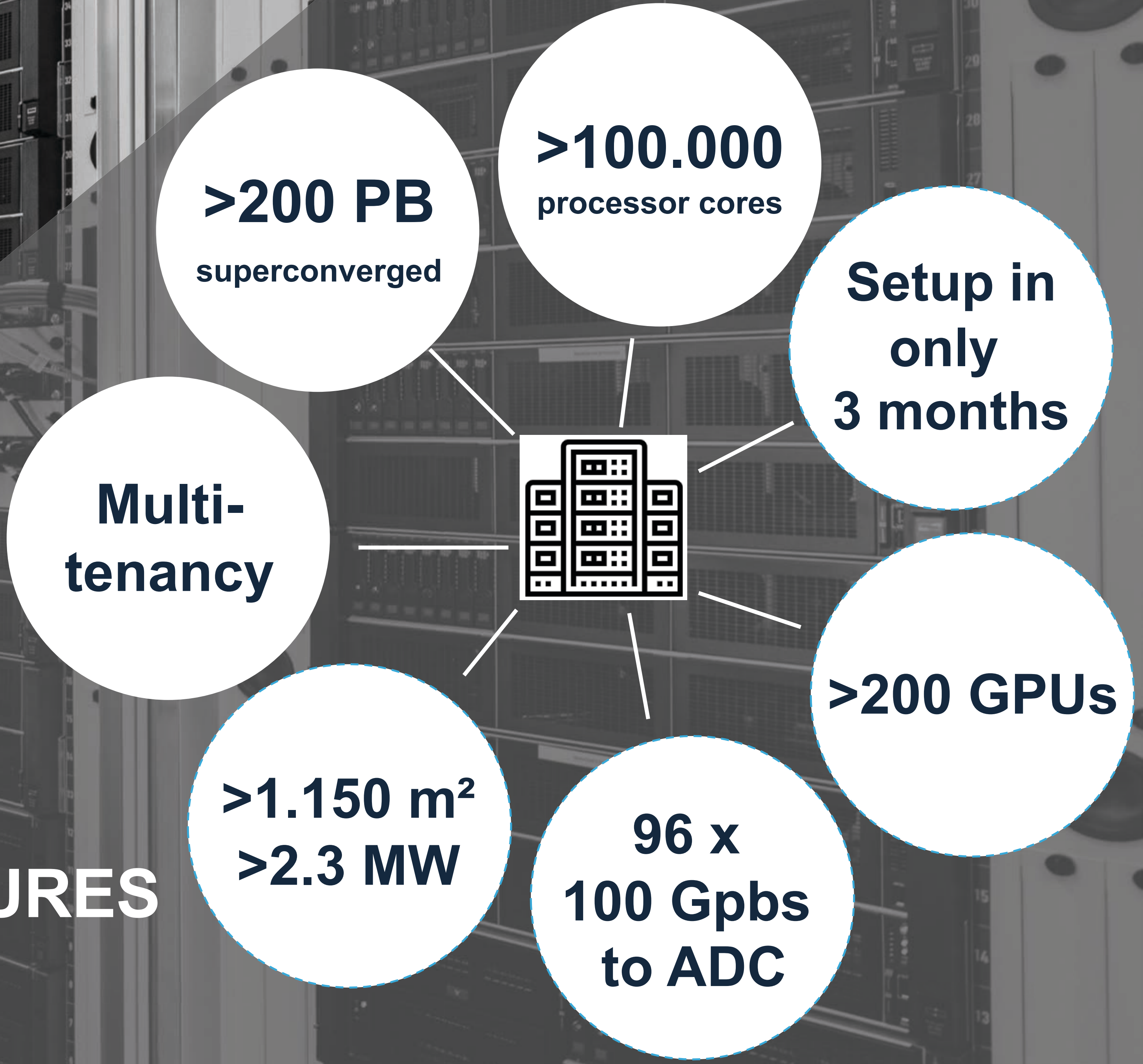
Extend & reuse current mechanic engineering capability combined with artificial intelligence, software building and IT is important to survive

Requires an end-to-end data and AI capability ecosystem for AD development



High Performance Data Driven Development Platform

Facts & Figures



>200 PB superconverged

>100.000 processor cores

Setup in only 3 months

Multi-tenancy

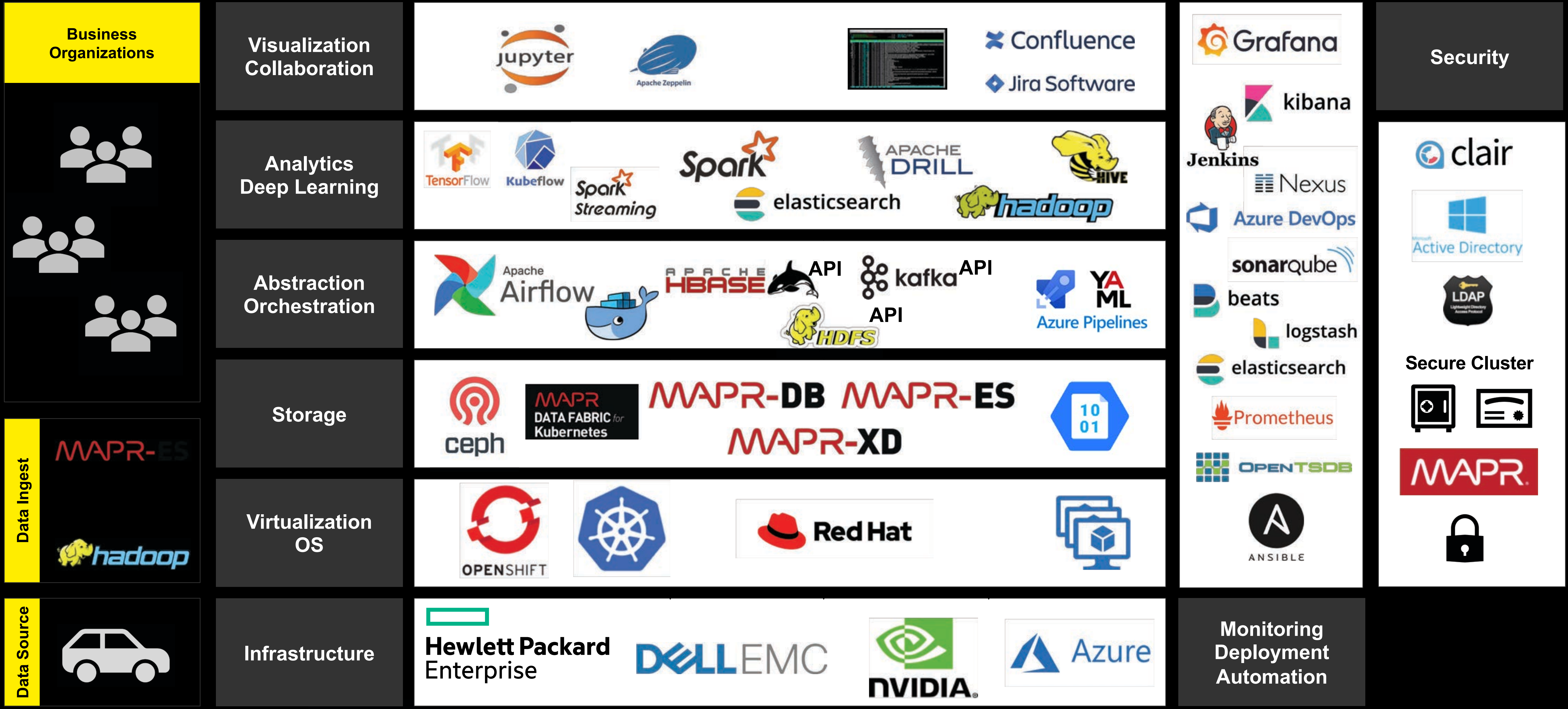
>200 GPUs

>1.150 m² >2.3 MW

96 x 100 Gpbs to ADC

DXC.technology

Technology Stack



Business Organizations

Data Ingest

Data Source

Visualization Collaboration

Analytics Deep Learning

Abstraction Orchestration

Storage

Virtualization OS

Infrastructure

Monitoring Deployment Automation

Security

Secure Cluster

Autonomous Driving with Airflow – In a Nutshell (1/3)

- **Airflow on OpenShift**
  - 1 scheduler instance – big risk
  - Multiple webservers – load balanced (Rest API Calls)
  - Numerous workers – multiple queues

- **Automated deployment via Helm charts**
  - Various configurations for different instances (also different Airflow versions)
  - History tracking via version control

- **On-demand Airflow instances (ex: development purposes, isolated testing – on a service level)**

Autonomous Driving with Airflow – In a Nutshell (2/3)

- **Integration with the Data and Storage Platform – MapR**
  - Loading DAGs from different locations
  - Loading job configurations, used by the different operators

- **Integration with the Compute Platform – OpenShift**
  - KubernetesPodOperator
  - KubernetesExecutor

Autonomous Driving with Airflow – In a Nutshell (3/3)

- **Metrics Collection & Monitoring**
  - StatsD → Prometheus → Grafana

- **Log collection and aggregation: ElasticSearch + Kibana**

- **Large scale orchestration: aiming at orchestrating jobs at the scale of 100,000's / month**
  - Ingestion, simulation, reprocesssing, machine learning, …etc
  - Complex DAG dependencies

July 16, 2020

Apache Airflow - Robotic Drive orchestrator

DXC.technology

Platform Orchestration Requirements

- Open Source
- Scalability
- Easy to adapt / extend
- Active community

What do we Orchestrate?

- Data Ingestion
- Machine Learning
- Reprocessing
- Simulation Jobs

July 16, 2020

Journey from PoC to Production

DXC.technology

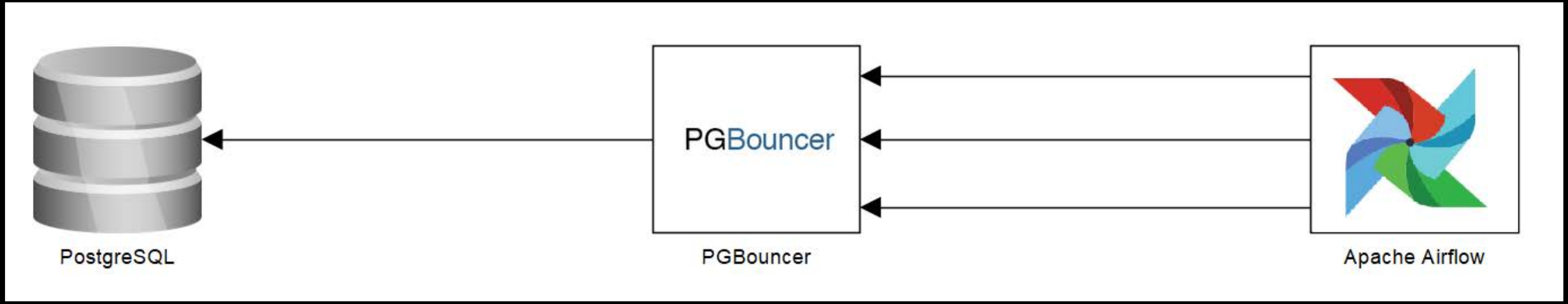
Airflow at Robotic Drive – the beginnings

- Initial work started at Q2 2019
- Airflow 1.10.2 with CeleryExecutor
- PostgreSQL 9.4
- RabbitMQ

Technical Challenges and Lessons Learned

- **Airflow stress and scalability tests**

- **Bottlenecks in the Architecture:**
  - PostgreSQL connection scalability: directly proportional relationship between number of running tasks and database connections
  - Scheduler configuration & performance



July 16, 2020

Tailor-made Solutions

DXC.technology

Operators / Hooks Customizations (1/3)

Customization and standardization of SparkSubmitOperator

Included „properties_file” in operator constructor

- Spark application configuration can be provided via separate properties file
- It allows submitting jobs via Airflow, in the same fashion as submitting them standalone from the Hadoop cluster

Extend list of parameters where templating is supported

- Better DAGs reusability
- Reduced code duplication

Operators / Hooks Customizations (2/3)

Enrich Airflow logs by adding YARN application details for Spark jobs

- Correlation between Spark application (triggered by Airflow) and submitted YARN job is challenging to discover when using YARN cluster mode
- Out of the box: YARN Application ID logged in the task logs only when a YARN job fails
- Extension: Extract and log the following for all Spark tasks:

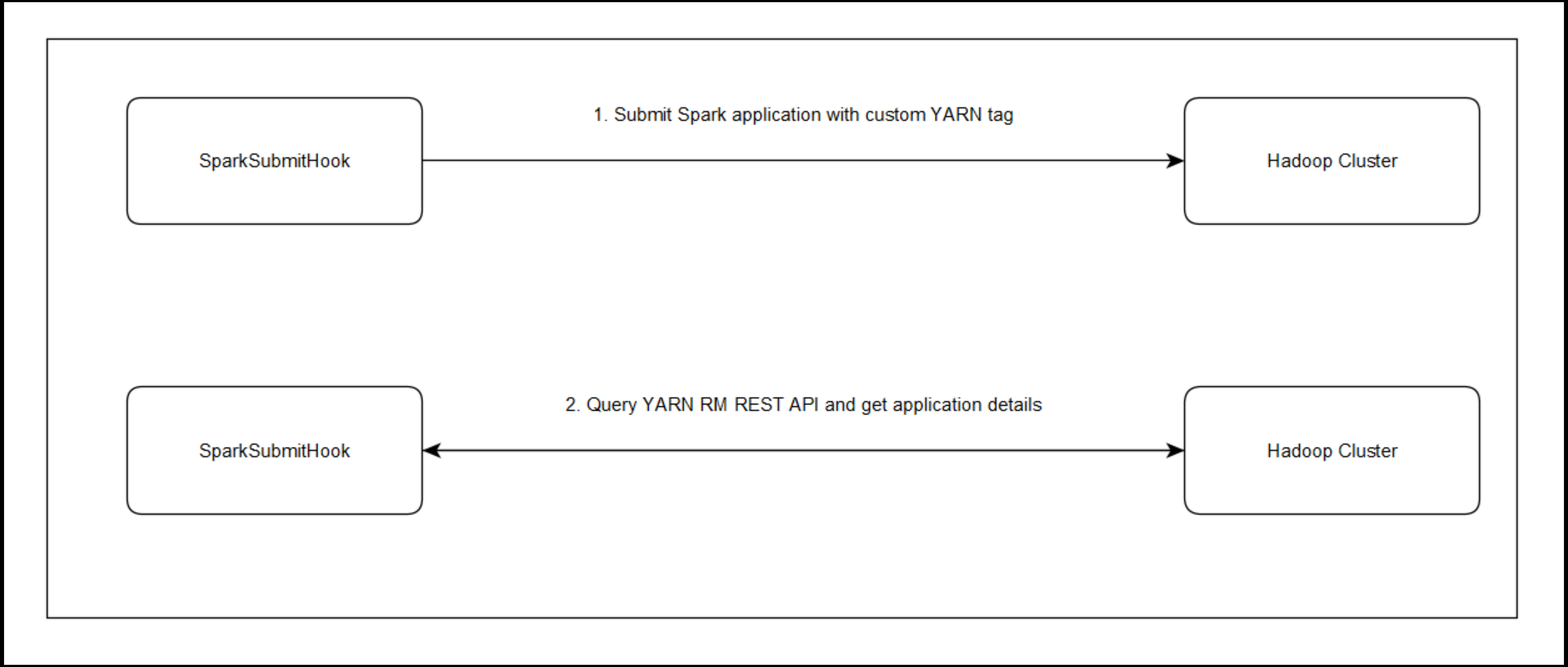
  1) YARN Application ID

  2) YARN Tracking URL

  3) Diagnostics (Failure root cause)

Operators / Hooks Customizations (3/3)

YARN application details visible in Airflow logs for Spark jobs



1. Submit Spark application with custom YARN tag
SparkSubmitHook
Hadoop Cluster
2. Query YARN RM REST API and get application details
SparkSubmitHook
Hadoop Cluster

Custom Authentication Methods

LDAP Secured REST API

- REST API usage allowed only for dedicated AD role
- Complete integration with LDAP
- Only one role specified for REST API – no separation between endpoints so far

July 16, 2020

Airflow in Production

DXC.technology

Production-ready Airflow instances

Robotic Drive supports by default 3 main instances used in the platform:

- **Development** (used mostly to test new Airflow deployments / features)
- **Staging** (testing new DAGs)
- **Production** (for full production usage)

Current stable setup is created using **Airflow 1.10.10** and **Celery Executor**

User-based Airflow instances

- Deployment automated via **Helm Charts**

- Airflow created as a Kubernetes project

These 2 points makes it possible to fully parametrize Airflow deployment and create many Airflow instances on-demand. In the Robotic Drive Platform each user is able to create their own, separate Airflow instance.

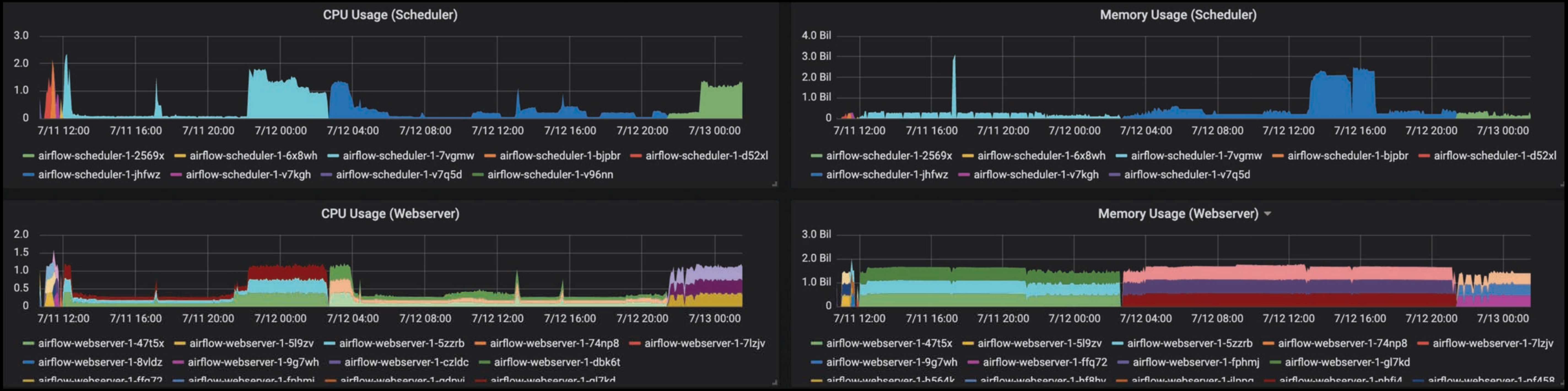
It helps to eliminate problems related to testing new DAGs, Operators and other features and changes that might impact other users, especially when multiple developers are working on the same component.

July 16, 2020

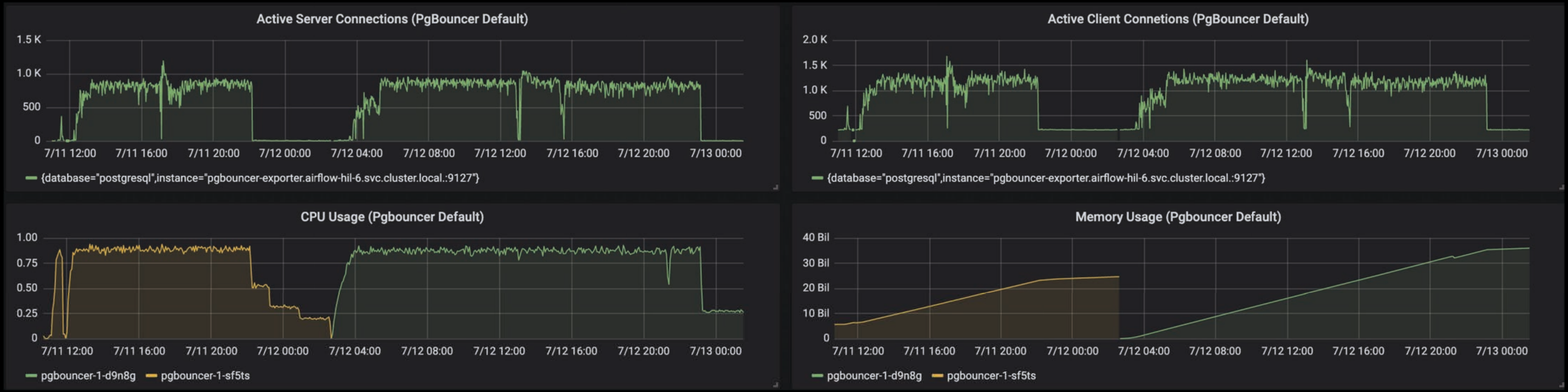
Monitoring Airflow

DXC.technology

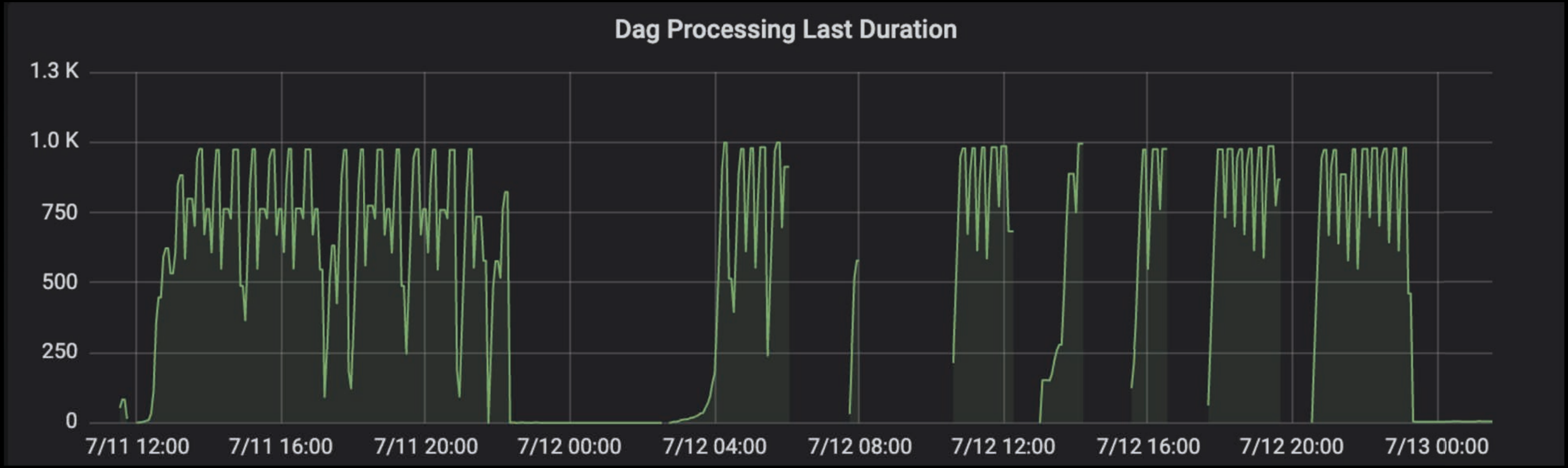
Monitoring Airflow


CPU Usage (Scheduler)
3.0
2.0
1.0
0
7/11 12:00
7/11 16:00
7/11 20:00
7/12 00:00
7/12 04:00
7/12 08:00
7/12 12:00
7/12 16:00
7/12 20:00
7/13 00:00
airflow-scheduler-1-2569x
airflow-scheduler-1-6x8wh
airflow-scheduler-1-7vgmw
airflow-scheduler-1-bjpbr
airflow-scheduler-1-d52xl
airflow-scheduler-1-jhfwz
airflow-scheduler-1-v7kgh
airflow-scheduler-1-v7q5d
airflow-scheduler-1-v96nn
Memory Usage (Scheduler)
4.0 Bil
3.0 Bil
2.0 Bil
1.0 Bil
0
7/11 12:00
7/11 16:00
7/11 20:00
7/12 00:00
7/12 04:00
7/12 08:00
7/12 12:00
7/12 16:00
7/12 20:00
7/13 00:00
airflow-scheduler-1-2569x
airflow-scheduler-1-6x8wh
airflow-scheduler-1-7vgmw
airflow-scheduler-1-bjpbr
airflow-scheduler-1-d52xl
airflow-scheduler-1-jhfwz
airflow-scheduler-1-v7kgh
airflow-scheduler-1-v7q5d
CPU Usage (Webserver)
2.0
1.5
1.0
0.5
0
7/11 12:00
7/11 16:00
7/11 20:00
7/12 00:00
7/12 04:00
7/12 08:00
7/12 12:00
7/12 16:00
7/12 20:00
7/13 00:00
airflow-webserver-1-47t5x
airflow-webserver-1-5l9zv
airflow-webserver-1-5zzrb
airflow-webserver-1-74np8
airflow-webserver-1-7lzjv
airflow-webserver-1-8vldz
airflow-webserver-1-9g7wh
airflow-webserver-1-czldc
airflow-webserver-1-dbk6t
airflow-webserver-1-ffq72
airflow-webserver-1-fphmj
airflow-webserver-1-gdpvj
airflow-webserver-1-gl7kd
Memory Usage (Webserver)
3.0 Bil
2.0 Bil
1.0 Bil
0
7/11 12:00
7/11 16:00
7/11 20:00
7/12 00:00
7/12 04:00
7/12 08:00
7/12 12:00
7/12 16:00
7/12 20:00
7/13 00:00
airflow-webserver-1-47t5x
airflow-webserver-1-5l9zv
airflow-webserver-1-5zzrb
airflow-webserver-1-74np8
airflow-webserver-1-7lzjv
airflow-webserver-1-9g7wh
airflow-webserver-1-ffq72
airflow-webserver-1-fphmj
airflow-webserver-1-gl7kd
airflow-webserver-1-h564k
airflow-webserver-1-hf8hv
airflow-webserver-1-jlpng
airflow-webserver-1-nhfj4
airflow-webserver-1-nf458

Monitoring Airflow



Monitoring Airflow



July 16, 2020

What’s next?

DXC.technology

Looking forward to…

- **Airflow 2.0: HA Scheduler + Performance Optimizations**
- **Advanced Authentication + Authorization**
- **Extend and stabilize monitoring metrics**
- **Stable API vs Experimental API**

Q&A

July 16, 2020

Amr Noureldin <amr.noureldin@dxc.com>
Michal Dura <mdura2@dxc.com>