ActiveEon’s Workflows: from HPC to Data Analytics to Machine Learning

Denis Caromel
AE Vision

- Workflows are Now Strategic
- Convergence of Usage in Academia & Industry
- Convergence in usage: HPC to Data Analytics to Machine Learning

For Scientist, Application People, Data Scientists, ML + AI Experts
AE Architectural Vision

Studio: Easy to Use Workflow Designer, with Palettes and Presets + APIs

Automation Dashboard: PaaS, Job Planner, Cloud Watch, Analytics

Scheduling & Meta-Scheduling: on all Infra for Legacy, Cloud & Hybrid

Scalable & Elastic resources

Provides cloud computing power according to your needs. Minimize costs by deploying VMs only when needed (configurable load factor). Never exceed your budget (min/max VMs threshold).

100% resources usage, no waste

Incremental resource deployment

Smart scale down

Smart and fully configurable elastic policy. Shutdown unused VMs whenever it’s possible. Prevent time-consuming re-deployments by adjusting idle nodes’ release delay (avoid scale up/down cycles).
Space & Image Processing

Situation
Make Sentinel data available to the greatest number and encourage the development of applications using them (agriculture, maritime field...)

Solution
Proactive Solution provided by ActiveEon to execute on Azure in hybrid mode allows enhancing PEPS data and making them available to API providers:
• Multi-Cloud Ecosystem Platform
• Remove complexity for Data Scientists
• Provide Cloud performance

Benefits
• Faster execution, Optimisation of On-Prem ressources & Clouds,
• Easier to use by end-users
Key points to choose AE:

- OnPrem & Clouds with a unique Task definition
- Single Interface for Job Management
- Out-of-the-box effective error management
- Product integration with Docker
- Business tasks can trigger VM provisioning on Azure Cloud
Genomic Platform

by INRA
Hybrid Cloud Architecture Overview

Quantitative Metagenomics Platform for gene profiling and statistical analysis

Pre, Post Processing of Data Analysis
Flexibility, Speed of Analysis
Granular execution
Distribution for fast execution
Secure data transfer

REST / HTTPS

ProActive Workflows & Scheduling

Secure data transfer

1TB / Sequence Analysis

Total DNA → QC/Library preparation → Proton/Illumina Sequencing → 1.3PB Database

On-Prem

Windows Cluster 1
192 cores

Linux Cluster 2
366 cores

Microsoft Azure

CPU cores on demand
Cloud Bursting

Web Portal and Integration with Scientific tools

Quantitative Metagenomics Platform
for gene profiling and statistical analysis

Quantitative Metagenomics Platform
for gene profiling and statistical analysis

Quantitative Metagenomics Platform
for gene profiling and statistical analysis

Quantitative Metagenomics Platform
for gene profiling and statistical analysis
**CEA List Architecture**

- **On-Prem**
  - Dev. Cluster: 72 cores
  - Prod. Cluster: 192 cores
  - Partner Cluster: 144 cores

- **Azure Cloud**
  - CPU cores on demand
  - Cloud Bursting

**Features:**
- Secure data transfer
- REST / HTTPS
- Job monitoring
- Result visualization

**Techniques:**
- Multiple NDT Techniques
  - RT: weld inspection
  - CT: tomographic reconstruction of complex parts
  - ET: 2D map of a complex defect

**Software:**
- CIVA software to launch NDT
- CIVA tools
- ProActive Workflows & Scheduling
- Java
- Microsoft Azure
Finance (Solvency Risk) on Microsoft Azure

Profile Results without ProActive

Profile Results with ProActive

Batch optimization with ProActive:
- Enforce strong priorities
- Optimal compact execution
- Start tasks as early as possible
- Pipeline and co-allocate
- CPU-intensive with I/O intensive tasks

“Achievements & Benefits:
- High Priority Results in 2 H instead of 18 H!
- Compacting Executions for savings on the Cloud

What Legal & General said about Batch optimization with ProActive:
“It enforces strong priorities, optimally compacts execution, starts tasks as early as possible, pipelines and co-allocates CPU-intensive with I/O intensive tasks.”
AI – Machine Learning Customer Cases
Hospital - Human Brain Project

**Scientists**
- Provide new statistical models on demand
- Develop ML models and algorithms at scale
- Manage distributed ML model training
- Deploy solution everywhere

**Clinicians**
- Request additional diagnostic on demand
- Identify new patterns for prediction and advices
Machine Learning Analytics for Aerospace
Each campaign generates a large amount of data on more than 2000 sensors.

ProActive allows to parallelize data analytics and scale the infrastructure to optimize compute times.

2 ML Models: a Streaming Version during the Test and a Post-Mortem for fine grain detections. It also allows to **Industrialize**, **Manage** and **continuously** improve the Models.
Anomaly Detection on Satellite Sensors

**FEATURES**

- Data analysis: data cleaning and merging
- Signal processing
- Big data workflows for automation of test scenarios
- Automatic detection of the different categories of anomalies (out of order, noisy, axis inversion)
- Data visualization in-browser

**BENEFITS**

- Data fetching from many sources
- Extraction of relevant features
- Faster results with **parallel parsing and processing**
- **Automatic anomaly detection** with machine learning
- Extensible & Evolving solution using Workflow tasks
FEATURES

- Introduce Artificial Intelligence technologies in the aerial/satellite image chain
- Give them autonomy, opening up many opportunities for remote sensing with a capacity for reaction
- Creation of large database of labeled aerial images dedicated to learning
- Evaluate performances on concrete settings (use cases, data, hardware)

BENEFITS

- A collaborative platform to design, evaluate and share ML algorithms
- Direct access to diverse hardware (different GPU, FPGA, Clouds, On-premise)
- On-demand scalability for diverse experiments
Live Demo on https://try.activeeon.com/
Technology

Processing & Automation Workflows

Workflow Studio

Scheduling & Orchestration
+ Meta-Scheduling

Job Console

Resource Management & Monitoring

Resource Manager
Workflows stored in **buckets** in the Catalog

**RBAC** support for each bucket / Users can share workflows and templates

Keep track of the revisions with a **versioning** feature integrated
Cloud Automation: On-demand PaaS

On-Demand PaaS Services with full Life-Cycle Management
UK Ministry of Interior is using ActiveEon for 2 critical applications:

- Visa Delivery Process, and
- Big Data & Analytics platform for Crime Reduction (HODAC).

**Situation**

25 different sources of Data.

Build a consolidated Data Lake and analytics platform to be used for many Home Land security applications.

**Solution**

ActiveEon used as the central Orchestrator to Schedule and Meta-Schedule all the Big Data, ETL, Analytics, Machine Learnings software appliance of the platform (Hadoop, SAS, TIBCO Spotfire, Python, Anaconda, GreenPlum, ElasticSearch, ...).
Meta-Scheduler, Big Data ETL + new ELT

Main Benefits
- Central Orchestration Tool
- Workflow Expressiveness: universal & comprehensive
- Management of Security for highly sensitive environments
- Management of Resources for all appliances (SAS, GREENPLUM, TIBCO, …)

Virtualized Infrastructure using Docker

4 Thousands of physical cores

« The only solution capable to Schedule any Big Data Analytics, mono-threaded, multi-threaded, multi-core, parallel and distributed » Cap Gemini Lead Engineer for Home Office
IoT

Komatsu is a Japanese multinational corporation. It manufactures construction, mining, industrial and military equipment.

Situation
ActiveEon Orchestrates on Cloud execution over hot and cold storage for Streaming and Batch Analytics ➔ 1,200 tasks executed per hour

Solution
ActiveEon has enabled control over scheduling and execution:
- Error Management – Notification, Automated Recovery
- Job Planner
- Distribution & Parallelization

Benefits
- Reliable execution to orchestrate multiple services and resources
- Provide consistent results and KPIs to end users and BI Tools

ActiveEon allowed to migrate from AWS to Azure
Fault-Tolerance & Error Management

[Table and Diagram as shown in the image]
Supporting Specialized Processing

- Cuda
- OpenCL
- FPGA (e.g. Altera)
- ASIC (if needed one day)

- Edges for IoT (Raspberry Pi, …)
Conclusion

- Both Portals and APIs, CLI (Java, Python, R, …)
- Error Management and Notifications
- Machine Learning Open Studio
- Data Connectors
- Data Analytics
- ML + DL
- Job Analytics
Building Blocks in Studio

Synchro between Tasks & even Jobs.

Tolerance aux Fautes, Versionning Management

30 mn:
- Intro: 5 mn
- Archi: Automation Dashboard / Studio / Scheduler / RM + API
- Features: Node Request, Data Space, Streaming/Interactive
- Automation: Call Graphs,
- Use Cases:
- Demo:
- Message: Easy to Use – New Usages in Scientific Computing, Hybrid, Portable, Open
UK Ministry of Interior is using ActiveEon for 2 critical applications:

- Visa Delivery Process,
- Big Data & Analytics platform for Crime Reduction (HODAC).

**Situation**

25 different sources of Data.

Build a consolidated Data Lake and analytics platform to be used for many Home Land security applications.

**Solution**

ActiveEon used as the central Orchestrator to Schedule and Meta-Schedule all the Big Data, ETL, Analytics, Machine Learnigs software appliance of the platform (Hadoop, SAS, TIBCO Spotfire, Python, Anaconda, GreenPlum, ElasticSearch, ...).
Meta-Scheduler, Big Data ETL + new ELT

Main Benefits
- Central Orchestration Tool
- Workflow Expressiveness: universal & comprehensive
- Management of Security for highly sensitive environments
- Management of Resources for all appliances (SAS, GREENPLUM, TIBCO, …)

Virtualized Infrastructure using Docker
4 Thousands of physical cores

« The only solution capable to Schedule any Big Data Analytics, mono-threaded, multi-threaded, multi-core, parallel and distributed »
Cap Gemini Lead Engineer for Home Office
Activeeon Solution

**Workflow Studio**
Intuitive, multi-language, open

**Schedule & Control**
Error management, Job Planner, Event Base

**Monitor**
Real-time job and resource monitoring

**Resource Management**
Hybrid, multi-cloud, auto-scaling

Packaged Solution
Open Source
ProActive Workflows & Scheduling Technology

Processing & Automation Workflows

Workflow Studio

Scheduling & Orchestration + Meta-Scheduling
Job Console

Resource Management & Monitoring
Resource Manager
Workflows: All kinds & All Resources

ProActive Workflows

Scheduler & Meta-Scheduler

Resource Manager

Fault Tolerance
Cloud bursting
Resource agnostic

API

Clusters

LSF

PBS Works

GRID ENGINE

slurm

GREENPLUM DATABASE

ANACONDA

SAS

MATLAB

SAP

kubernetes
docker
Azure
amazon web services
Google
OPENSHIFT
Scalable & Elastic resources

Provides cloud computing power according to your needs. Minimize costs by deploying VMs only when needed (configurable load factor). Never exceed your budget (min/max VMs threshold).

Incremental resource deployment

100% resources usage, no waste

Smart scale down

Smart and fully configurable elastic policy. Shutdown unused VMs whenever it’s possible. Prevent time-consuming re-deployments by adjusting idle nodes’ release delay (avoid scale up/down cycles).
ProActive Process Flow

1. Design
   - PROACTIVE STUDIO
   - IT department, Data scientist, Business Lines, Activeeon Consulting

2. Schedule, Monitor & Connect the resources
   - PROACTIVE SCHEDULER
   - Business owner

3. Execute
   - PROACTIVE CLOUD AUTOMATION
   - Operational team
### Some Supported Languages and Connectors

#### Machine Learning & Artificial Intelligence

- Cognitive Services
- CNTK
- Keras
- PyTorch
- TensorFlow
- Scikit-Learn
- MXNet
- YOLO
- H2O
- Caffe
- Spark MLlib
- Pandas
- JupyterLab
- DL4J
- Deep Learning
- G4J
- DLlib
- BigDL

#### Big Data

- Zookeeper
- Kafka
- Databricks
- Spark
- Hadoop HDFS
- Hadoop
- Swarm
- Storm
- Clearwater
- Twitter
- Cuda
- Open/CL
- FPGA
- Visdom
- Grafana
- Kibana

#### Data Connectors

- URL
- FTP
- SFTP
- PostgreSQL
- Greenplum
- MySQL
- Oracle
- SQL Server
- MongoDB
- Cassandra
- Azure Storage
- Azure Data Lake
- AWS-S3
- SAP
- Elasticsearch
- Logstash

#### Languages and Predefined Tasks

- Linux Bash
- Cmd
- Java
- Scala
- Javascript
- Groovy
- Ruby
- Jython
- Python
- Perle
- PHP
- PowerShell
- C++/C#
- R
- Cron
- LDAP

#### Infrastructure

- Linux
- Windows
- Solaris
- AIX
- AS/400
- VMware
- Openstack

#### Clouds

- Azure
- AWS
- GCP
- Docker
- Kubernetes
- OpenShift
- Slurm
- PBS
- LSF
- SGE

#### HPC Schedulers

- SGE
- LSF
- PBS
- OpenShift
- Kubernetes
- Docker
- AWS
- Google Cloud Platform
- VMware
- Openstack
- Slurm
- PBS
- LSF
- SGE
Machine Learning Open Studio

https://www.youtube.com/watch?v=mbrQxCf4lqM
Some Major Customers

- Telco & IT
- Engineering
- Aeronautics & Space
- Energy
- Bio Tech & Health
- Media Distribution
- Finance

Some Partners:
- orange™
- AMADEUS
- Government
- Home Office
- IBM
- Microsoft Azure
- Capgemini
- Amazon Web Services
- Atos
- PA
- Mastek
- Legal & General
- IoT
- Cosmetics
- Orange
- Safran
- Sncema
- CNES
- CEA
- Daimler
- INRA
- L’Oréal
- Joy Global
- Komatsu
IoT

Komatsu is a Japanese multinational corporation. It manufactures construction, mining, industrial and military equipment.

**Situation**

ActiveEon Orchestrates on Cloud execution over hot and cold storage for Streaming and Batch Analytics ➞ 1,200 tasks executed per hour

**Solution**

ActiveEon has enabled control over scheduling and execution:

- Error Management – Notification, Automated Recovery
- Job Planner
- Distribution & Parallelization

**Benefits**

- Reliable execution to orchestrate multiple services and resources
- Provide consistent results and KPIs to end users and BI Tools
UK Ministry of Interior is using ActiveEon for 2 critical applications:

- Visa Delivery Process,
- Big Data & Analytics platform for Crime Reduction (HODAC).

**Situation**

25 different sources of Data.

Build a consolidated Data Lake and analytics platform to be used for many Home Land security applications.

**Solution**

ActiveEon used as the central Orchestrator to Schedule and Meta-Schedule all the Big Data, ETL, Analytics, Machine Learnings software appliance of the platform (Hadoop, SAS, TIBCO Spotfire, Python, Anaconda, GreenPlum, ElasticSearch, ...).
Meta-Scheduler, Big Data ETL + new ELT

Main Benefits
- Central Orchestration Tool
- Workflow Expressiveness: universal & comprehensive
- Management of Security for highly sensitive environments
- Management of Resources for all appliances (SAS, GREENPLUM, TIBCO, …)

Virtualized Infrastructure using Docker
4 Thousands of physical cores

« The only solution capable to Schedule any Big Data Analytics, mono-threaded, multi-threaded, multi-core, parallel and distributed »
Cap Gemini Lead Engineer for Home Office
Positioning - Differentiation

High-Level Business & Application Owner:

- Important Savings of Cloud Expenses
- Time To Value, Speed of Development
- Speed of Execution
- Multi-System Multi-Data-Source Aggregation
- Complex and Large Scale Applications

STRONG TECHNICAL DIFFERENTIATORS

Cloud-Ready: Integrated Resource Manager with one-click elastic Multi-Cloud Scheduling and Meta-Scheduling of any application

Muli-Language, comprehensive, resilient Workflows at large Scale
Automate Accelerate & Scale
10K Nodes, 20K Tasks, 1M Jobs