Simplifying AI Applications with containers and Kubernetes

Glyn Bowden, Chief Architect, AI & Data Science Practice

SUSECON 2019
The world is replacing programming with training

4 in 10 companies mention **lack of analytical skills** as a key challenge

By 2020, **50%** of organizations will **lack sufficient AI and data literacy skills** to achieve business value

23 million developers worldwide

Handful of data scientists
What you need to do defines what you have to do

- Edge
  - Where is the data generated?
  - What does that data consist of?
  - How do you prepare and integrate data for advanced analytics?

- Cloud
  - How long do you have to take action?
  - What are your business goals?
  - What governance and security regulations do you need to comply with?
  - What do you have to do to put that data in a form you can use?
AI ecosystem

A working solution requires every layer of this stack

- **Expertise**
  - Advisory / consulting services
  - Vertical SMEs

- **Data**
  - Public data sources
  - Proprietary data collections
  - Data labeling providers

- **Software**
  - Custom applications
  - As-a-service offerings
    - Machine learning and deep learning frameworks and libraries (TensorFlow, Caffe, Scikit-learn, ...), data platforms
    - Systems software and libraries (CUBLAS, MKL, cuDNN, MKL-DNN)

- **Hardware**
  - Memory
  - Storage
  - Network
  - Hardware accelerators
Challenges to gaining insight

Existing AI process

**Individual** – AI solutions tend to be bespoke. Designed and deployed as part of the project

**Time consuming** – AI solution cycle is long due to bespoke nature of design

**Risk** – Bespoke solutions and iterative platform development introduces risk of inconsistent results and even project failure

**Inefficient** – Skilled and costly resources used on tasks which could easily be done by others
The Four Pillars of Data Science
“simplify AI consumption bringing together the AI ecosystem”

Data Sources
- Multiple formats
- Multiple Sources
- Multiple Standards
- Security

Data Management
- Data Platform
- Data Lake
- On Premise
- Hybrid Cloud
- Integration

Analytics
- Multiple Machine Learning Libraries
- Multiple AI trained models
- Multiple AI languages

Insight
- Business Use Cases
- Integrated Analysis
- 360 view of data
- CDO Dashboards
AI Ecosystem Summary

- **14** Distinct ML Services
- **49** ML Model Candidates
- **15** Distinct ML Services
- **20+** AI / ML Candidate Partners
- **50+** Language Frameworks
- **40+** Model Categories
AI – Vertical Solution Use Case Examples

Finance
- Algorithmic Trading
  - Market data
  - Trade Execution
- Risk Analysis
  - Simulation
  - Decision Execution
- Fraud Detection
  - Financial data
  - Notification

Manufacturing
- Material Analysis
  - Simulation
  - Conclusion
- Predictive Maintenance
  - Sensor Analysis
  - Notification
- Self-Functioning Devices
  - Environment Analysis
  - Task execution

Healthcare
- Genome Sequencing
  - Simulation
  - Conclusion
- Resource Management
  - Resource data
  - Proposal
- Connected Health
  - Patient sensors
  - Medical opinion

Telco / Media
- Customer Profiling
  - Activity & Social analysis
  - Directed advertising
- Network Performance Analysis
  - Sensor analysis
  - Config execution
- SmartCity
  - Log analysis
  - Notification

Enterprise
- Supply Chain optimisation
  - Simulation
  - Order execution
- Image / Video Recognition
  - Image data
  - Classification
- Natural Language comms.
  - Text analysis
  - Chat comms
Layered Architecture

- Task Execution Engine
- Visualisation Engine
- Models and Algorithm (Content)
- Big Data Frameworks
- AI Frameworks
- Developer Frameworks SUSE CAP
- Unified Data Warehouse
- On-Site DWH SUSE Enterprise Storage
- Container Platform SUSE CaaSP
- Cloud DWH
- Infrastructure
- Cloud

Hybrid IT Practice

Cloud Technology Partners

Business Users

Data Scientists

IT Operators
Accelerate time-to-value in your AI journey with OneAI

Rapid deployment of solution components during PoV leading to, smooth production

Conventional steps for a PoV / Production implementation

- Provisioning of Infrastructure
- Setup AI s/w stack
- Containerization process
- Messaging and transformation components
- Monitoring tools

Elapsed Time: Weeks to Months

Faster PoV → Getting proof points early and fail fast!

Automated setup with OneAI

- Automated Provisioning
- Containerize
- Easy Monitoring

Elapsed Time: Hours

Automation deployment of infrastructure, Monitoring tools and Management: simple
How Project OneAI works – an illustrative view

**Templates**

Use Case Parameters
- Use Case parameters
- Data sources
- Dashboard parameters
- Accounts / Roles
- Monitoring parameters

**OneAI**

- Automated Provisioning of Container based ecosystem
- Orchestrated with Kubernetes
- Unified monitoring on open frameworks (Prometheus)

**Use case based output**
### Addressing the challenges to Insights

#### Existing AI process

- **Individual** – AI solutions tend to be bespoke. Designed and deployed as part of the project
- **Time consuming** – AI Solution cycle is long due to bespoke nature of design
- **Risk** – Bespoke solutions and iterative platform development introduces risk of inconsistent results and even project failure
- **Inefficient** – Skilled and costly resources used on tasks which could easily be done by others

#### HPE Pointnext Framework

- **Reuse** – Leverage HPE globally industrialized templates and project IP
- **Rapid** – Deploy an AI solution, so it is ready to start receiving data
- **Learn & Evolve** – Solution gallery will grow over time as more AI projects are successfully delivered
- **Efficient** – Consultants are driving insight from the customer data not configuring and experimenting with underlying technologies
Project OneAI Functional Architecture v2.0

Kubernetes Cluster
Istio Service Mesh
OneAI Microservices
- catalog
- environment
- datasource
- lifecycle
- operations

Ambassador API Gateway
Keycloak OAUTH2/OIDC

{{api_client}}
oai (cli)

Web browser

hub.docker.com
HPE Pointnext Docker Registry
Private Docker Registry
NVIDIA NGC
NVIDIA NGC

Grommet UI

Content
Service Scaling
Cluster Scaling

Diagram showing the setup of Docker containers within a Kubernetes (K8S) environment. The diagram illustrates the interaction between OneAI service, K8S service, and the containers within a pod.
Component Hierarchy Design
Use Case Categories

- Manufacturing: 2 cases
- IT Operations: 1 case
- Healthcare: 1 case
- Smart City: 1 case
- HPE Discover: 1 case
- Transportation: 0 cases
Kafka
Apache
Kafka® is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands of companies.

Published on: 11 Jan 2019, 11:55 GMT-8

Nifi
Apache
Apache Nifi supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic.

Published on: 11 Jan 2019, 11:55 GMT-8

Zookeeper
Apache
ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications. Each time

Published on: 11 Jan 2019, 11:55 GMT-8

GitLab
Apache
GitLab is a single application for the entire software development lifecycle. From project planning and source code management to CI/CD, monitoring, and security.

Published on: 11 Jan 2019, 11:55 GMT-8

TensorRT
NVIDIA
NVIDIA TensorRT™ is a platform for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime that delivers low latency and high-throughput for deep learning inference applications. TensorRT-

Published on: 11 Jan 2019, 11:55 GMT-8

TensorFlow
Google
TensorFlow™ is an open source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of

Published on: 11 Jan 2019, 11:55 GMT-8
Thank You

gjb@hpe.com