ZDMP Blog

AI Analytics Runtime – What is it and how it works

By Ferrán Bou, Senior Software Engineer & Francisco Javier Barrena, Senior Software Architect & Security Advocate, Instituto Tecnológico de Informática (ITI)

Some questions for you

- Why would an AI Analytics Runtime is needed?
- What is the position of the component in the overall architecture of ZDMP?
- What kind of features does it provide?
- What are their technical foundations?

Data Scientists vs Software Engineers

Artificial Intelligence (AI) pipelines are complex and involve several technical profiles usually specialized in specific areas. In fact, it is recommended to have at least one of each of the following profiles to appreciate and implement AI: Software Architect, Software Engineer, DevOps Engineer, Data Scientist, IT systems administrator and Big Data Engineer.

If it is thought about it carefully, it makes sense. Artificial Intelligence requires significant data which means there is a need to store and retrieve that data efficiently and for this a Big Data Engineer is needed. Usually, the machines in which the Big Data tools are installed need to be provisioned by someone, here is where the IT systems administrator is relevant. On top of this, the data will be consumed, processed, and transformed into AI's models; these tasks are usually performed by a Data Scientists. But, at the end, these models need to be deployed and executed in a specific infrastructure and needs to be secured and orchestrated in an overall application. This is where the roles of DevOps Engineer and Software Engineer can contribute. Of course, how the systems involved are interconnected one to each other, the selection of technologies and the architectural decisions are made by the Software Architect.

It is not likely to have a one-man-army who performs well all the skills. Thus, for that reason, usually, the Data Scientists are not the best role to develop APIs or backends, and Software Engineers are not the best role to develop AI models.

For this reason, the AI Analytics Runtime was concepted and it has been conceived to provide an infrastructure to run AI models automatically. The aim is that everything, from the component deployment to the AI model deployments, is performed without human intervention. To achieve this behavior, the component uses automation technologies in every layer. In addition, integrates with the other components in the ZDMP architecture, providing functionalities such as: Securization, authentication and authorization, monitoring and alerting, data acquisition and marketplace integration.

AI Analytics Runtime supports multiple AI based technologies:

- Python (PKL, ONNX)
- H20.ai models (Java based)
- Generic integration using Docker Layers

Technologies involved

AI Analytics Runtime uses several technologies:

- **Docker in Swarm mode:** This is to provide container orchestration. Docker Swarm creates a cluster using multiple machines and allows to orchestrate Docker containers in an easy way. Every machine belonging to the cluster can be defined in one of two available flavors: Manager or worker. Swarm managers oversee the control panel, route the traffic between nodes, maintain the state of the cluster, etc. On the other hand the worker nodes execute the tasks received from the managers
- **Portainer:** This is a Docker management tool. It provides a comprehensive UI to manage Docker endpoints. These endpoints can be the local Docker daemon or remote Docker daemons. In ZDMPs case, it manages the Docker Swarm cluster. It also offers a REST API to allow interaction from other parts of the component.

Portainer is deployed in a manager node inside the Docker Swarm cluster, as it requires access to the Swarm low-level API. The main purpose of the component is to allow AI model deployments. These models are wrapped inside a Docker image and deployed as a Docker container.

- **Container registry:** Later, the models are exposed to the users via an URL. To store the Docker images, the component includes a container registry. The registry allows image versioning and it also offers a web-UI to manage the registry itself. This component part is deployed inside the Docker Swarm cluster
- **MinIO:** This is an object storage server designed for performance and to be cloud-native. It can be deployed as a single server or in cluster mode. By using MinIO, the AI Analytics Runtime stores every model training result in a well-defined path to allow model reutilization. It also provides a web-UI to navigate through the stored objects. MinIO is deployed inside the Docker Swarm cluster
- **Træfik:** Træfik is a HTTP reverse proxy and load balancer designed to be natively compliant with multiple cluster technologies: Kubernetes, Swarm, Docker, etc. The routing configuration can be performed in real-time without the need of restart. The ease-of-use is its main goal. Træfik is deployed in a manager node inside the Docker Swarm cluster. It requires a domain name pointing to the manager node's IP
- **Custom backend**. This exposes a REST API to allow interaction with other ZDMP components. The backend receives the AI code and a manifest containing all the needed information to deploy and operate the AI model, create the Docker images, upload them to the registry and deploy them in the Swarm cluster. During the build process, the backend also wraps the AI model inside a REST API, allowing users to use the model. It also stores the manifest in a MongoDB database to allow users to find which manifests has been deployed (version, parameters, name, etc.)

What will ZDMP achieve

The AI Analytics Runtime, provided as one of the component services on the ZDMP platform, can help to reduce the number of technical roles needed to deploy AI models into production, leaving data scientists focused on what they do best, abstracting them of the technical details usually required to retrieve data efficiently, envelope the models in APIs to be consumed and deploy them in an elastic cluster.

ZDMP Links

• Architecture Component(s)	AI Analytics Runtime
• Work Package	WP5 – ZDMP Platform Building
• Tasks	T5.6 – AI Analytics Runtime

References

None