Big Data Europe

Hajira Jabeen, Phil Archer, Simon Scerri, Aad Versteden, Ivan Ermilov, Giannis Mouchakis, Jens Lehmann, Soeren Auer
The H2020 BigDataEurope Project Consortium
c/o Fraunhofer IAIS, Sankt Augustin, Germany
info@big-data-europe.eu

ABSTRACT

The BigDataEurope (BDE) project is developing exactly the kind of computing infrastructure that European stakeholders need when handling large volumes of data in a variety of formats; the results are open-source and their use is completely free. Coordinated by Fraunhofer IAIS, BDE is working directly with partners that represent the seven Societal Challenges identified by the European Commission (Health, Food, Energy, Transport, Climate, Social Sciences and Security). For each community, a pilot that makes use of BDEfis technology stack to address the Big Data needs identified by these challenges is well under way.

1 THE BIG DATA INTEGRATOR PLATFORM

BDE's Integrator Platform (BDI) makes the processing of big data simpler, cheaper and more flexible than ever before. It offers basic building blocks to get started with common big data technologies and makes integration of different technologies and applications easy. Components such as Apache Spark, Hadoop HDFS, Apache Flink, Apache Flume and Apache Kafka can be built into a pipeline through a simple graphical UI. Those components can help handle the velocity and volume dimensions, but BDI is also leading the way in tackling that third big data problem: variety. This is done through BDI's Semantic Data Lake and components like SANSA[1] which performs analytics on semantically structured RDF data by providing out-of-the-box scalable algorithms for massive datasets.

BDI is an open source platform based on Docker, today's virtualisation technique of choice. It works on a local machine or on hundreds of nodes using Docker Swarm, and can run in-house, or within an external cloud environment (not provided by BDE). BDE applications are provided as docker containers, making their installation and set-up a 10-minute job. With the help of latest Docker features, BDI offers:

- Swarm-based networking
- Load Balancing
- Service Discovery
- Multi-host networking with integrated KV-Store
- Fault tolerance

Docker Compose helps to create multiple containers on multiple nodes using a single command and a single compose file. Docker Compose V2 and Docker Swarm aim to implement full integration, which means that it is feasible to point a Compose app at a Swarm cluster and make its use possible in the same manner as if a single Docker host is being used. It is notable that the latest Docker components provide greater resemblance to Kubernetes in terms of orchestration features, and Swarm presents a better choice in terms of shifting from a local/development environment to a cluster.

The BDE Team provides baseline Docker images for Apache Hadoop, Spark, Flink and many others. Components were selected based on the requirements gathered from the seven Societal Challenges. Thus, the Platform makes it feasible to perform a variety of big data tasks, including message passing (Kafka, Flume), storage (Hive, Cassandra). The platform is able to handle RDF triples at scale using components like FOX, SemaGrow and 4Store; with particular emphasis on the triplification of geospatial data using GeoTriples, Sextant and Strabon.

BDI has enriched the Docker platform, a high-level depiction of which is shown in Figure 1, with a layer of supporting services, helping in the setup, maintenance and monitoring of the pipeline and workflows:

- The Init daemon allows to define workflows by monitoring the start-up status of inter-dependent Docker components.
- The Pipeline Service and Builder are developed to support the creation of workflows.
- The Pipeline Monitor front-end demonstrates the current status of the Docker components.
- The Integrator UI integrates the different official Web UIs of select pipeline components under one Integrated and personalised view. Furthermore, the Swarm UI visualises the status of a swarm cluster and allows to scale and monitor the cluster services.

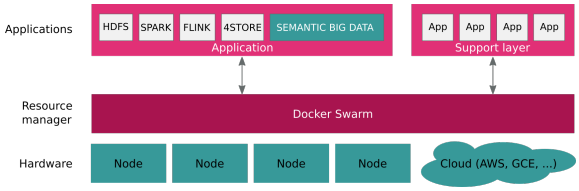

Figure 1: BDI platform's high-level modular architecture

For BDI platform progress updates please refer to the dedicated page[2]; or try it out or engage with our community[3].

ACKNOWLEDGMENTS

Supported by the BigDataEurope project (Empowering Communities with Data Technologies); EU-H2020-ICT-15-2014, Grant Agreement No.644564.

[1] http://sansa-stack.net/

[2] https://www.big-data-europe.eu/platform/

[3] https://github.com/big-data-europe