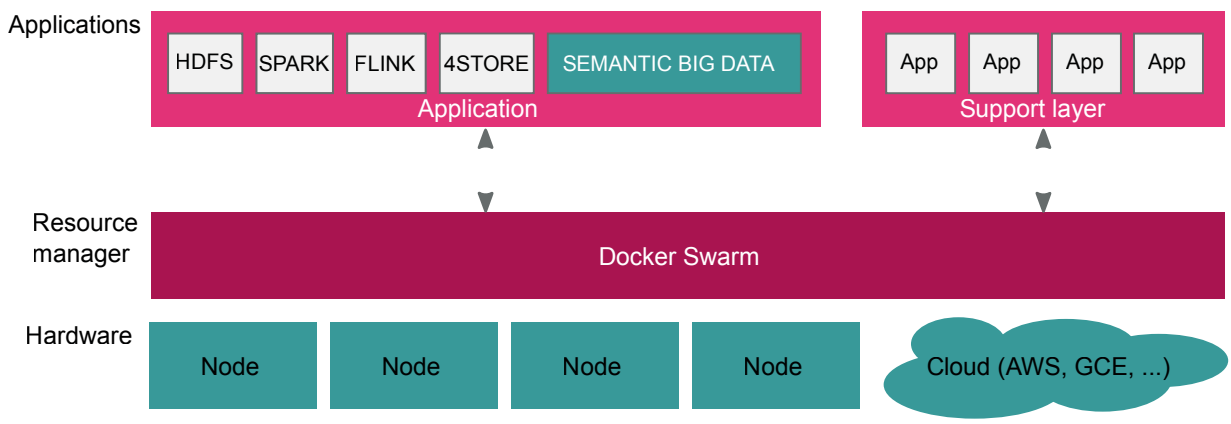
BIG DATA EUROPE
General platform description

The Big Data Europe (BDE) platform makes Big Data simpler, cheaper and more flexible than ever before. We offer basic building blocks to get started with common Big Data technologies and make integration with other technologies or applications easy.

Examples of available blocks are Apache Spark, Hadoop HDFS, Apache Flink and many others. Research efforts are conducted with Smart Big Data, by adding semantics to a data lake and performing structured machine learning on semantically structured data.



The Big Data Integrator (BDI) is an Open Source platform based on Docker, today's virtualization technique of choice. The base Docker platform is enriched with a layer of services, which support the workflows' setup, creation and maintenance. BDI can work on your local development machine, or scale up to hundreds of nodes connected in a swarm. It is proven to work on the differing requirements of each one of Europe's 7 Societal Challenges. The platform can be run in house, or can be hosted by vendors like Amazon Web Services or Microsoft Azure.

BDE Wiki

BDE Github

For any questions related to the BDI platform, please contact us at **platform@big-data-europe.eu**

@BigData_Europe
www.big-data-europe.eu

BIG DATA EUROPE
Pilots descriptions

Health, Demographic Change and Wellbeing (SC1)

This pilot demonstrates reproducing the functionality of the Open PHACTS Discovery Platform on the Big Data Integrator. The pilot explores replacing the distributed Virtuoso triple store used by the current platform with open-source solution offering equivalent functionality. The pilot deployment is based on the 4store component of the Big Data Integrator, which will be usedtogether with the Scientific Lens query term expansion service of the Open PHACTS Discovery Platform.

Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy (SC2)

This pilot demonstrates the automatic annotation of scientific publications in the viticulture domain by extracting named entities (locations, domain terms) and the captions of images, figures and tables. The pilot deployment is based on the Kafka, Flume, Spark, HDFS and 4store components of the Big Data Integrator. In the pilot, the SWC Pool Party Semantic Suite will be used to consolidate and link the terms in the ingested documents, which are stored in HDFS, and to write the consolidated metadata to a 4store repository. The 4store repository will be exposed to viticultural researchers via a search interface.

Secure, Clean and Efficient Energy (SC3)

The pilot demonstrates the ingestion and processing of data streams from wind farms to enhance condition monitoring. This is achieved by pooling together data from multiple units to consider the cluster's operation as a whole. The pilot deployment is based on the HDFS and Spark components of the Big Data Integrator. In the pilot, existing analysis tools will be orchestrated by a Spark program to operate in parallel over data ingested in HDFS from multiple units.

Smart, Green and Integrated Transport (SC4)

The pilot demonstrates the ingestion, processing and storing of data to classify traffic conditions and to make predictions. The pilot deployment is based on the Kafka, Flink, PostGIS, and Elastic Search components of the Big Data Integrator. In the pilot, the algorithms for map matching and classification will be implemented in R and accessed by a Java/Flink application. Kafka will be tested as the middleware upon which different modules implemented in different languages will be integrated.

Climate Action, Environment, Resource Efficiency and Raw Materials (SC5)

The pilot demonstrates the ingestion transformation, cropping, and combining of datasets from the climate domain, in order to prepare input for a climate modelling experiment. The pilot deployment is based on the HDFS, Cassandra, Strabon, Spark and Semagrow components of the Big Data Integrator. In the pilot, the ability to manage and transform large-scale datasets and their derivatives will be tested, including the maintenance of lineage annotations pertaining to the datasets and modelling parameters used to produce each derivative dataset.

Europe in a changing world - inclusive, innovative and reflective societies (SC6)

The pilot demonstrates the ingestion of open financial data (i.e. budget and budget execution data) from three different cities, and the homogenizing of their structures and formats so that they can be compared, analyzed and visualized in a comprehensible way. The pilot deployment is based on the Kafka, Flume, Spark, HDFS and 4store components of the Big Data Integrator. In the pilot, the SWC PoolParty Semantic Suite will be used to consolidate and link the terms in the ingested documents, which are stored in HDFS, and to write the consolidated metadata to a 4store repository. The 4store repository will be exposed via a dashboard that provides search/discovery, aggregation, analysis, correlation, and visualization functionality.

Secure societies – Protecting freedom and security of Europe and its citizens (SC7)

The pilot demonstrates the detection of geo-located events in news sites and social media, and the correlation of such events with changes detected by comparing Earth Observation satellite images (such as construction/destruction of settlements). The pilot deployment is based on the HDFS, Cassandra, Strabon, Spark and Semagrow components of the Big Data Integrator. In the pilot, the integration of Spark-based processing (change detection) and external processing (event extraction) by writing results to common data stored is tested. The results are exposed via Sextant, a visualization tool for viewing images, changes, and event information.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644564.