Data curation, ingestion and processing layer

Omer Boehm - June 06, 2019

In the era of the Mobility and IoT, the generation of new data by the many applications in every aspect of life, work, industry, consumerism, entertainment, and play has drastically increased. Huge volumes of data coming from various sources in different formats, such as sensors, logs, structured data from an RDBMS, etc. are becoming available at a fast velocity which brings the need for an efficient Analytics Systems.

Big data analytics platforms goal is to help in extracting the desired knowledge and insights required for smarter decision making, better products and solutions tailored for customer needs, better models of future behaviour and outcomes in Business, Government, Security, Science, Healthcare, Education, and more.

I-BiDaaS is becoming a unified Big Data as-a-service solution which will address the needs of both non-IT and IT professionals by enabling easy interaction with Big Data technologies.

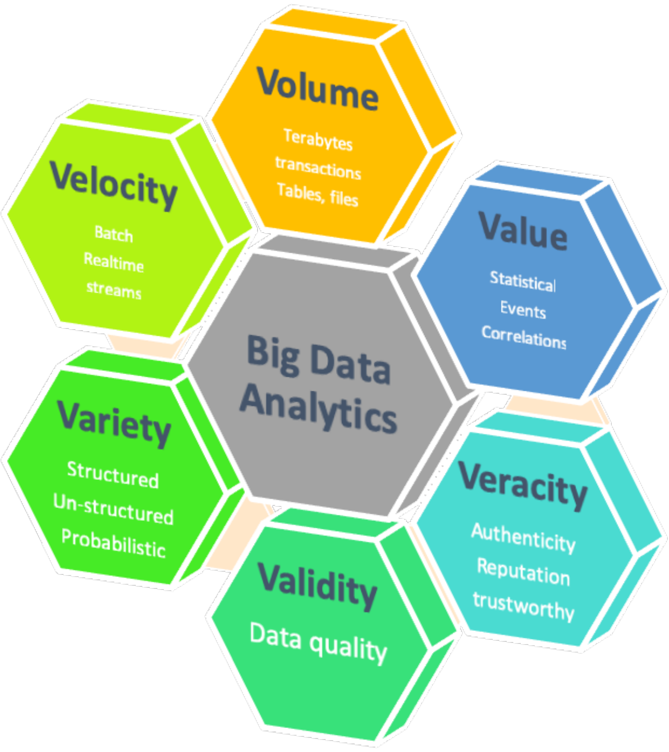

Figure 1. Big Data and Analytics most important V's

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.

The data curation, ingestion and processing layer is the first step for the nine datasets of the generic and domain representative use cases derived from the Finance, Manufacturing, and Telecommunications sectors (reported in D1.1).

The processes and activities related to the organization and integration of the data collected for these use cases and the definition of the data needed to be complemented synthetically are now defined. For each dataset we provide a high-level overview, it's objectives and goals and a description of it's nature and formats and most importantly, a concrete definition of the data structure and relationships it contains. These definitions are critical for the analytics components, but also for the test data fabrication platform which complement and augment datasets quickly and efficiently while avoiding the risks related to using sensitive production data. The real data provided for each of the use cases are anonymized, masked, tokenized and manipulated in various ways to avoid privacy regulations violations as well as to avoid leakage of certain sensitive business details.

Data at this layer are ingested into the batch processing and the streaming analytics modules via the Universal Message broker. The broker uses a scalable, publish-subscribe mechanism for distributing the data to the relevant components. It collects, aggregates, pre-processes and transforms real-time streaming raw data (structured, unstructured, noisy and possibly incomplete) into structured messages of a common, unified, format. The data exchange uses open standards like MQTT to allow for a flexible and easily extensible component architecture.

I-BiDaaS, as mentioned above, isn't designed for expert users alone. The Visualization layer follows the User Centric Design methodology to design user interfaces of increased usability. The main characteristics, technologies and integration points of the tools comprising the visualization and monitoring framework have been carefully designed, considering system requirements and the basic end-users workflows, to maximize the usability of the platforms' components. Wireframes sketches have been designed to demonstrate how the groups of end users interact with the platforms' environment in order to manage their datasets and perform the offered analyses so as to extract useful insights.

Find & Follow us

Website | **Twitter** | **LinkedIn**

Zenodo | **OpenAIRE** | **GitHub**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780787.