Big Data – Real Time Processing of Data Streams

Abstract

Since the Internet digitally connected the world, the amount of data, having high business value for many organizations, has been increasing tremendously. Existing technologies could, mostly, handle the increasing volume of data until the emergence of social media, search engines and e-commerce, which, briefly saying, caused so-called Data Boom. In order to meet customers' demand, to be innovative or to have a competitive advantage, businesses needed to gather the data from many sources, ingest, transform, analyze it and make quick decisions based on it. It was not an easy task, considering the limited capacities of available technologies. The volume of this data was not the only problem here. Along with this factor, the main challenge with processing of the large datasets was related to their various, not necessarily structured nature and velocity.

The term 'Big Data' was officially launched in 2005. Hadoop – the open-source heart of big data universe was created the same year. Since then the open source community has been actively working on and contributing to the Big Data. Numerous technologies and tools have been developed to process, store, manage, analyze large sets of data. As a result, today the Big Data ecosystem is very diverse and its architecture – quite complex.

The following paper overviews the nature of Big Data, its relevance, main concepts and principles along with its key architectural characteristics. The key part of the work has been devoted to one of the most popular topics in the Big Data world – (near) real-time processing of data streams. There are discussed two important architectures of it. As for the practical side of the work, a software solution to a specific task, related to real time data streams processing, has been presented.

Ana Japaridze