B. Mitschang et al. (Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2017),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 523

Big Data is no longer equivalent to Hadoop in the industry

Andreas Tönne[1]

Abstract: For a long time, industry projects solved big data problems with Hadoop. The massive scalability of MapReduce algorithms and the HBase database brought solutions to an unanticipated level of computing. But this obstructs the view for the need of change. Business goals that emerge from Industry 4.0 or IoT have long been addressed with a suboptimal architecture. New business goals require a rethinking of the big data architecture instead of being driven by the known Hadoop ecosphere. We discuss the transformation of a Hadoop-centric middleware solution to a streaming architecture from a business value perspective. The new architecture also replaces a single NoSQL database by polyglot persistence that allows to focus on best performance and quality of each data processing step. We also discuss alternative architecture approaches like Lambda that were evaluated in the course of the transformation. We show that a single technology choice likely leads to a solution that is suboptimal.

1 Extended Abstract

Depending on whether you are a researcher or practitioner, Hadoop was created in 2002 as Yahoo's research solution to a better (scaling) search engine, called project Nutch. Or it really came to life when it was promoted top-level project of the Apache Software Foundation in 2008. Also in 2008 Cloudera was founded to offer a commercial version of Hadoop. This was one critical first step towards maturity that made Hadoop usable for CIOs around the world. Hadoop offered an unanticipated level of computing power to corporates and a promise of endless, massive scalability both for storage in HDFS/HBase and computing using the map-reduce concept of Google (2004). Many companies started out with Hadoop, investigating its usefulness for business problems from a purely technological perspective. But the concepts of Hadoop were too "alien" for the traditionally conservative corporate world to immediately relate them to business value. And although the term "big data" has been around for more than two decades, the big data hype peaked around 2012/2013. At this time, big data = Hadoop was firmly rooted in the heads of corporate IT.

Many companies (57%) that have invested in the Hadoop technology suffered from pains growing the needed skills or recruiting the needed experts, and a similar percentage of companies (49%) is still trying to figure out how to get value out of their Hadoop investment (Gartner Hadoop Adoption Study 2015), and this matches very well with our experience. This situation leaves corporate IT stuck with a Hadoop installation and the related promises, desperately trying to monetize the investment. Since Hadoop is simply a massively scalable technology to process big data amounts in an asynchronous way in a distributed data lake, new hypes like data science (analytics), complex event processing or deep learning, while not matching the processing paradigm of Hadoop, are nonetheless sought to yield

[1] dibuco GmbH, Franz-Schubert Str. 75, 70195 Stuttgart, andreas.toenne@dibuco.de

the promised business value. But these new promises expose a mismatch between the architecture underlying the Hadoop technology and many business goals that are to be solved with analytics and machine learning concepts. These new business goals may emerge not only from hypes like Industry 4.0 or IoT, or also as a response strategy to digital disruptors, which are threatening many businesses. Common to these business values is the aspect of time (i.e., batch versus stream processing). Often it is the case that the business value of big data deteriorates over time, requiring near realtime analysis or data pattern recognition. For example, machine data, online payment processing or business statistics of an online platform cannot wait for a long running Hadoop batch to complete. These goals require a rethinking of the big data strategy and the adoption of more modern big data solutions. The current technology trends are streaming platforms like Apache Storm or Apache Spark Streaming and inmemory processing. Also the Lambda architecture by Nathan Marz that combines batch style processing (data lake) and event processing (data stream) into a common view is growing in popularity. Yet in business meetings we still observe the identification of big data storage with HBase as the only choice. We also see resistance of IT to admit further big data technologies to their technology stack because "we already have Hadoop in operation".

In order to break this assumed equality of big data and Hadoop, IT needs to stop wagging the dog by the tail. It is unfortunately common to drive a big data project from the technology end. This is wrong and ultimately harms the customers! We need to start with the business goals and work our way down to technology and will discover that we have an exponentially growing number of big data technologies to choose from. In our talk we demonstrate this way of evolving a Hadoop centric big data solution to a more capable and even more scalable streaming solution on a customer product development project that we were involved with in the last three years. We highlight the importance of polyglot persistence to choose both the right data model and the right database technology to achieve the business goals. Very important was also the question whether more data is actually better from a business value perspective. Finally, if one tries to address every problem related to concurrency and consistency with Hadoop batches, one may discover that one simply runs out of time to schedule the batches. The lesson learned is that a premature single big data storage and processing technology choice might lead to a solution that is suboptimal for many business goals.