APACHE HADOOP

Ein Framework für die Implementierung von Big Data Analytics

Christian Anlauf

10. May 2015
COREtechmonitor | Blogpost

Key Facts

- The growing complexity of exploitable data in the business world requires a future oriented list of companies in the area of Big Data Analytics
- The high infrastructural requirements for traditional technologies to support Big Data Paradigms can only be implemented at great expense
- Apache Hadoop framework addresses current and future technological challenges of Big Data and enables the integration of big data analytics in businesses

Report

The raw data and the capabilities of the analysis to move are becomingan important competitive factor for companies. In the banking and finance industry, massive amounts of data are generated by multiple services and service performances in the areas of mobile banking, credit, investment, and insurance. The purpose of these collected data extends from the operation of services through the improvement of the offer to the customer to fraud detection and risk analysis. The volume of exploitabledata amounts is today increasing steadily, stands out by its most diverse data types, and requires flexibility in the analysis capability. All in all, companies must position themselves as future-oriented entities, because the trend “Big Data” is driven further by the braid of future technological innovations.

Contrary to what the name indicates, the challenges of Big Data does not relate solely to massive amounts of data in the petabyte or hexabyte area. In particular, it also refers to data of high variety with different data formats in structured and unstructured form. There is also a high and continuous lasting change frequency of the stored data and the indispensable requirement to carry out valid and fault-tolerant data analysis. This high complexity of the "Big Data Analytics" places high demands on hardware resources such as memory, RAM and CPU, where a solid infrastructure becomes - even more so - a critical factor for performance and availability. Thus, in this context, the high cost of implementing a high-performance, highly available, and highly scalable infrastructure increasingly comes into consideration. In this area cloud computing offers solutions.

In the sphere of Big Data, cluster and cloud computing Apache Hadoop has been successfully established as an open source solution for distributed systems. The master/slave architecture promises a high degree of scalability, availability, and fault tolerance. Additional memory, increasing I/O capacity, and enhanced performance can be realized cost-effectively by the simple addition of standard hardware.

Map-Reduce is a framework for parallel computing of large data sets on different nodes within large clusters. Since the processing and managing of data are two things naturally in direct connection, Map-Reduce at the level of "Distributed data processing" along with the database "HBase" can be considered. Apache HBase realizes in this context, read and write access to the HDFS in real

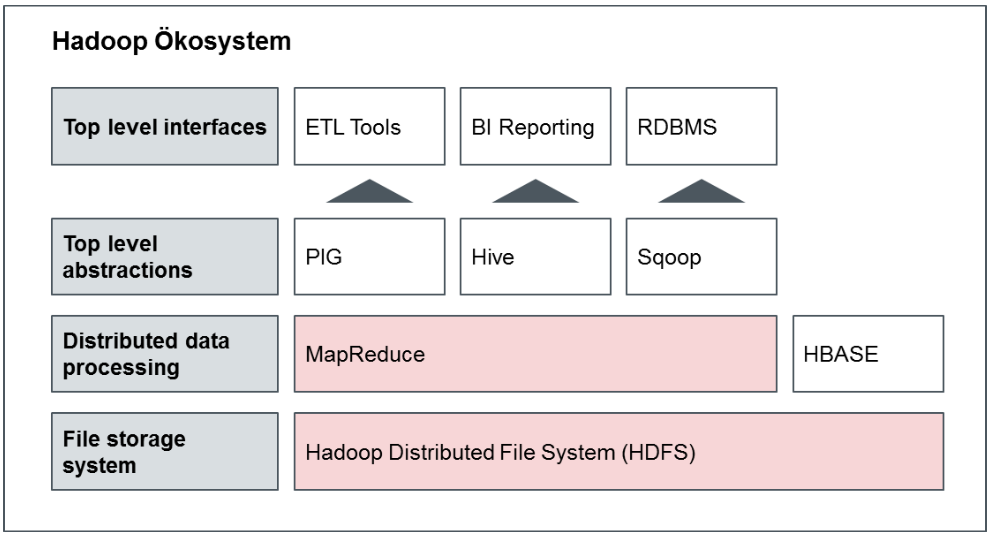

Figure 1: Illustration of a possible Apache Hadoop ecosystem

time and forms very large tables with billions of rows and millions of columns in a Hadoop cluster.

There are other tools part of the Hadoop HBase family that can be operated on the HDFS- and Map-Reduce-Framework. The in Figure 1 illustrated "top level abstractions" show an exemplary set of tools to perform the distributed stored "Big Data" analyzes. This abstraction layer makes the link between the distributed systems in Hadoop and existing analysis environments in companies (see "Top level interfaces" in Figure 1). It is in this fashion that Apache Pig as platform for the implementation of data flows in a Hadoop allows an integration of ETL tools (Extract, Transform, Load). For the operation of business intelligence solutions, Apache Hive provides a distributed data warehouse infrastructure and Apache Sqoop allows the data transfer between relational databases (Relational Database Management Systems) and the HDFS.

In addition to the availability of additional tools, such as Ambari, Avro, and Mahout, and additional modules such as Hadoop YARN, thedevelopment of Hadoop frameworks is promoted in the context of the Apache project. Concerning Big Data Analysis Hadoop with HDFS and Map-Reduce offer not only a response to todays challenges of scalability, availability, and performance, but also to the necessary flexibility in the integration of big data analytics in business.

However, the implementation of Big Data solutions in large companies is confronted with obstacles while facing a variety of legacy systems. Thus, the existing architectures, technologies, and applications of different origin and culture technology are joined together by various compliance policies and security policies. This results in the introduction of Big Data technologies in necessary implications for the organization:

Introduction of procedures, methods and skills for the implementation and operation of Big Data Analytics

Development of new policies to address the driven by Big Data requirements for data quality and data security

Sponsorship of top management to ensure a holistic Big Data Governance

This topic that concerns Big Data technologies as a whole must be considered in the corresponding calculations for the development of Big Data technologies. We must increase awareness in the corporate worldin order to know which challenges, all together, must be addressedwith a modern data platform and to what extent the organization itself must take responsibility over data management.

Christian Anlauf is Transformation Associate at CORE. He studied at Brandenburgischen Technischen Universität Cottbus-Senftenberg Business Administration. At CORE, the focus points of Christians activities are in the field of IT-strategy and data management where his expertise is used in different project situations.

Mail: christian.anlauf@coretransform.com

COREinstitute
Am Sandwerder 21-23
14109 Berlin | Germany

www.coreinstitute.org

Phone: +49 30 16344 020

office@coreinstitute.org

COREtransform GmbH
Am Sandwerder 21-23
14109 Berlin | Germany

www.coretransform.de

Phone: +49 30 26344 020

office@coretransform.de

COREtransform GmbH
Limmatquai 1
8001 Zürich | Helvetia

www.coretransform.ch

Phone: +41 442 610 143

office@coretransform.ch

COREtransform Ltd.
One Canada Square
London E14 5DY | Great Britain

www.coretransform.co.uk

Phone: +44 203 319 0356

office@coretransform.co.uk