STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Krzysztof CZAJKOWSKI, Łukasz CHODAK, Radosław KORKOSZ,
Michał URBAŃCZYK
Cracow University of Technology, Faculty of Physics, Mathematics and Computer Science,
Institute of Computer Networks

TOOLS AND TECHNOLOGIES IN THE BIG DATA ENVIRONMENT

Summary. Nowadays, the Big Data is a term frequently used in the literature, but still there is no consensus on the standards in implementations of such environments. The are many tools and technologies in this area. Designers and developers have to decide which solutions to implement and how to integrate many elements in one system. This paper is the review in the Big Data technologies.

Keywords: Big Data, large data set, software, review

NARZĘDZIA I TECHNOLOGIE W ŚRODOWISKU BIG DATA

Streszczenie. Termin Big Data coraz częściej pojawia się w literaturze, wciąż jednak nie uzgodniono standardów implementacji tego typu środowisk. Istnieje wiele narzędzi i technologii w tym zakresie, a projektanci i programiści muszą decydować, które rozwiązania implementować i jak integrować wiele elementów w jeden system. Artykuł stanowi przegląd kluczowych rozwiązań z tej tematyki.

Słowa kluczowe: Big Data, duży zbiór danych, oprogramowanie, przegląd

1. Introduction

The Big Data is a term, which became popular in recent years. It combines several aspects most important in nowadays data processing environment: amount of data (volume), constantly accumulating new data (velocity), and range of data types and sources (variety). Some publications define the 4V characteristics of Big Data by adding concept of variability, value (data quality), and veracity. The occurrence of such elements in one system generates a number of complex challenges. “Big Data” refers to datasets with size beyond the ability of

typical database software tools for capturing, storage, management, and analysis [1]. Many companies and foundations develop various approaches, projects and tools. Due to the short development time of the Big Data concept, there are no conventional approaches in this area.

To handle huge amount of data generated in many sources and to analyze them for making decision and prediction, it is necessary to use different solutions. In most companies, just a few percent of the data collected in different systems is properly processed. The most desirable aspects of the processing of large data sets are their integration, consistency, reliability, simplicity, timeliness [2]. Traditional technologies mostly suit structured and repeatable tasks. But for new challenges a set of specialized technics, that address speed and flexibility and skills in unstructured data analysis, is needed. In this way it may be possible to manage, explore and discover huge volume of data from heterogeneous sources.

The Big Data tools landscape is growing rapidly. Tools can be classified majorly into following area: data analysis, databases / data warehousing, operational, multi value database, business intelligence, data mining, key value, document store, graphs, grid solutions, multi model, XML databases and Big Data search [3]. In this paper selected programming model, tools and applications, which can be applicable in the Big Data infrastructure, are presented. We discuss the directions of development of such technologies and how the main software companies integrated them in their own products.

2. Hadoop

Apache Hadoop is an open source software framework that enables the distributed processing of large data sets across clusters of computers. The general concept is to divide the data into smaller fragments which are processed in certain nodes. Hadoop is not a replacement for existing infrastructure, but a tool with augmented data management and storage capabilities. It was created in 2005, first release 1.0.0 is available from 27 December 2011. Hadoop framework enables easy scaling by adding additional nodes without reconfiguration of an entire system. By MapReduce, Hadoop distributes large amounts of fragmented data to certain nodes. Currently, Hadoop is the most popular technology to store any type of data, structured or not, from any number of sources. This platform is licensed under the Apache Licensed 2.0.

Hadoop has been designed as a framework with support for large data collections (Fig. 1). An important advantage is flexibility and scalability of the system which is done by adding new nodes. These nodes are using standard computers and servers. This modification of architecture does not require reconfiguration of the entire system such as in the case of OLAP systems. Framework is independent of the operating systems and databases. Another ad-

vantage is reliability. In the case of failure of any node another computer does the task for him. There is no interruption in the delivery of the service.

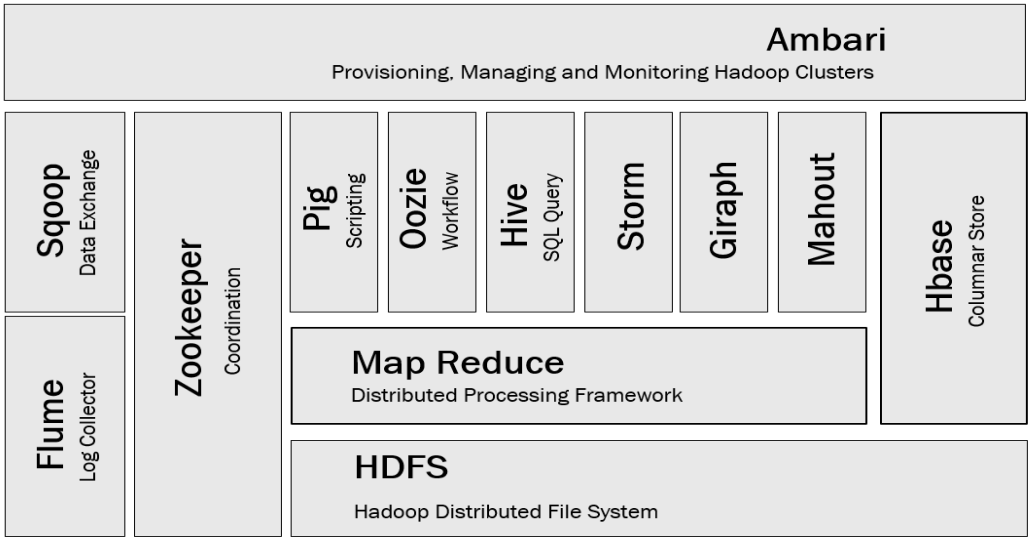

Fig. 1. Hadoop Ecosystem
Rys. 1. Środowisko Hadoop

The framework is used by large corporations such as Yahoo, Facebook, Amazon, Ebay, American Airlines, The New York Times, Federal Reserve Board, Chevron, IBM.

The basic elements of the Hadoop framework are: Apache Hive, Apache HBase, Apache ZooKeeper, Apache Pig, Apache Sqoop, Apache Oozie, Apache Ambari, Apache Flume, Apache Mahout (Fig. 1) – will be described in detail section below.

3. Tools included in the Hadoop

3.1. MapReduce

MapReduce is a programming model used to process enormous amounts of data, i.e. financial data or data collected from meteorological stations. This approach relays on dividing set of data into independent portions, which are processed by the mapping algorithms in the form of <key, value> and then sorted. Sorted output data from mapping process is used as input data to the process of reducing not very essential values. Prepared data are used in application, which sends a request for the needed information, e.g. to make a report. Figure 2 shows the information reading from the database, then using MapReduce in order to process data, and then processing this data to another source of information.

Having to improve efficiency of systems working on, 4000 or more nodes, clusters, MapReduce 2 project have been issued. Its name is YARN (Yet Another Resource Negotiator).

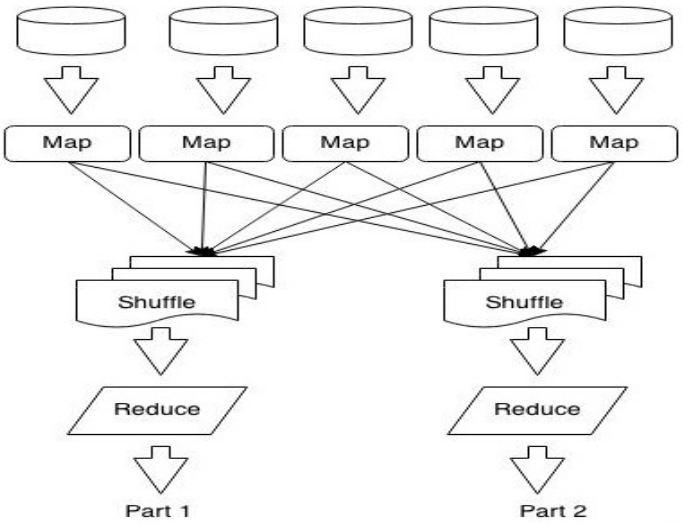

Fig. 2. Data flow in MapReduce
Rys. 2. Przepływ danych w MapReduce

MapReduce on YARN involves more entities than classic MapReduce. They are [4]:

- The client, which submits the MapReduce job.
- The YARN resource manager, which manages compute resources on the cluster.
- The YARN node managers, which are responsible for starting and monitoring the compute containers on nodes in the cluster.
- The MapReduce application master, which is responsible for coordinating the tasks running the MapReduce job.
- The distributed file system, which allows to share job files between the other entities.

3.2. HDFS

HDFS (Hadoop Distributed File System) is a cluster distributed file system designed to hold large data (terabytes or petabytes), and to provide wide bandwidth to access to this information. To ensure enough durability and high availability in parallel applications, files in HDFS are replicated within whole cluster [5]. HDFS is split into blocks, and each of them stores data. Minimum block size is 64MB.

HDFS implements master/slave architecture with NameNode and DataNodes mechanism. NameNode runs as master server that manages the file system namespace, regulates files permissions and executes operations, such as opening, closing and renaming files and directories. NameNode is a place, where metadata about whole cluster are stored. DataNodes, however, provide mechanisms, which split data set (externally exposed as single file) into many internal blocks. One external file may be stored within different physical machines. DataNodes are responsible for serving read/write request from clients and perform block operations, such as creation, deletion and replication, upon instructions from NameNode.

Application can communicate with HDFS in many ways. Naturally, HDFS provides Java API to use in application but also there is a C language wrapper for his Java API along with Pydoop (Python HDFS and MapReduce API for Hadoop). WebDAV protocol is the way to expose and interact with HDFS data via web browsers.

3.3. HBase

HBase is a non-relational database management system that runs on top of HDFS. It is written in Java. It is used to store and analyze large amounts of data (billions of rows, millions of columns). It is not an ACID (Atomicity, Consistency, Isolation, and Durability) compliant database but it does guarantee certain specific properties. Tables can serve as the input and output for MapReduce jobs run in Hadoop. It can be accessed through the API.

3.4. Ambari

The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters [6]. Ambari provides management through web UI.

3.5. Storm

Storm is a free and open source distributed real-time computation system [7]. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Storm can be used with any programming language**.** Storm has many use cases: real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more.

3.6. Giraph

Apache Giraph is an iterative graph processing system built for high scalability. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections [8].

3.7. Sqoop

Sqoop is a tool designed for data transfer between Hadoop and relational databases [9]. Sqoop might be used to import data from a relational database management system (RDBMS) such as SQL Server or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS –

Fig. 3. Data can be imported also to Hive or HBase. Tasks in Sqoop can be repeated using jobs. Saved jobs remember the parameters used to specify a job, so they can be re-executed by invoking the job by its handle. There is a possibility to run Sqoop job using Oozie. Apache Oozie is a workload scheduler for Hadoop. It has been designed to manage the executions of connected workflows with data dependencies between them.

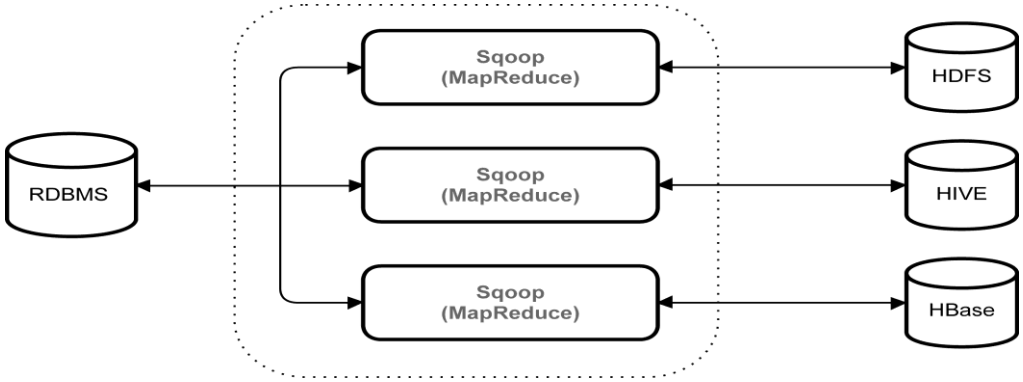

Fig. 3. Example of flow between RDBMS and Hadoop
Rys. 3. Przykład przepływu pomiędzy RDBMS i Hadoop

3.8. Mahout

Mahout is a machine learning, open source library written in Java [10]. The main feature of Mahout is scalability and collaboration with Hadoop. It uses the data stored in HDFS. Mahout focuses on three key areas of machine learning at the moment. These are recommender engines (collaborative filtering), clustering, and classification. Mahout incubates a number of techniques and algorithms, many still in development or in an experimental phase.

4. Hadoop in commercial applications

Table 1
Commercially supported Hadoop-related products

Company	Solution
SAP	SAP Hana
IBM	InfoSphere BigInsights
Oracle	Oracle Big Data Appliance
HP	HAVEn
Teradata	Teradata Appliance for Hadoop
EMC	Pivotal HD
Microsoft	HDInsight

There are a number of companies offering commercial implementations and/or providing support for Hadoop. The major players and their solutions were put in table 1 [11]. Some of them were further described in next paragraphs.

4.1. Microsoft

Microsoft HDInsight is an enterprise-ready distribution of Hadoop that runs on Windows servers and as cloud service [12]. HDInsight was developed in partnership with Hortonworks and Microsoft.

Service is offered in PaaS (Platform as a Service) model. The service is integrated with other known analytical tools offered by Microsoft. End users can use Power Query function from Excel 2013 to get data from different sources, including Hadoop. Data are visualized using Power View or Power Map in Excel or Office 365. Map and reduce code can be developed using .NET platform. Another feature is the ability to integrate with Active Directory which makes Hadoop reliable and secure.

Polybase is part of an overall Microsoft "Big Data" solution [13], built into SQL Server 2012 Parallel Data Warehouse. Designed for an easy way to connect data stored in relational and non-relational way. The advantage of this solution is no need to supply the data warehouse and no need for training in MapReduce. Users can execute T-SQL queries that join tables containing a relation source with tables in Hadoop Fig. 4. Further, data from Hadoop can be fetched with Select queries that contain JOINs and GROUP BY clauses.

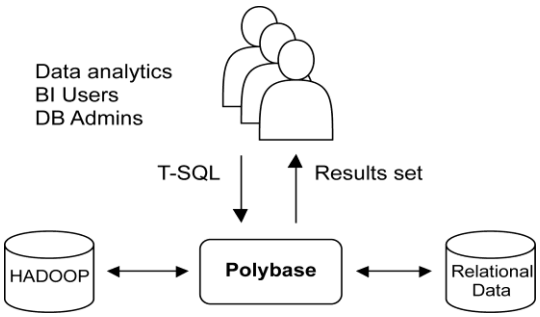

Fig. 4. Interaction in Polybase
Rys. 4. Interakcja w Polybase

4.2. SAP HANA

SAP HANA is an in-memory, data platform which operates on columnar database to increase data processing speed. It supports R language spatial processing, natural language processing, and text analytics libraries [14]. SAP HANA provides smart data access and access to information from remote sources – such as a SAP IQ or Hive data warehouse – using virtualization techniques. Query is executed on the remote server, hence only a minimal amount of data is returned to SAP HANA.

Sybase IQ is a column database. It is designed for efficient analysis of large amounts of data, which find usages in the area of Business Intelligence. SAP IQ offers a native MapReduce application programming interface (API). Hence, programmers can develop and deploy MapReduce programs natively in SAP IQ [15].

SAP Data Services delivers a Hadoop connector that provides high-performance reading from and loading into Hadoop. With SAP Data Services, unstructured data files can be pushed into Hadoop to extract relevant text from files. After getting relevant text, in next step data can be loaded in to SAP HANA or SAP IQ for real-time analysis with structured data.

The SAP BusinessObjects BI platform offers an analytic modeling environment and BI tools. Data models (called universes) can be deployed against data in SAP HANA, SAP IQ, or non-SAP data sources, including Hive and HDFS. SAP supports integration with Hadoop in several ways, including the SAP HANA smart data access capability, SAP Sybase IQ native MapReduce API, SAP Data Services Hadoop connector, and SAP BusinessObjects BI universes. [16], Fig. 5.

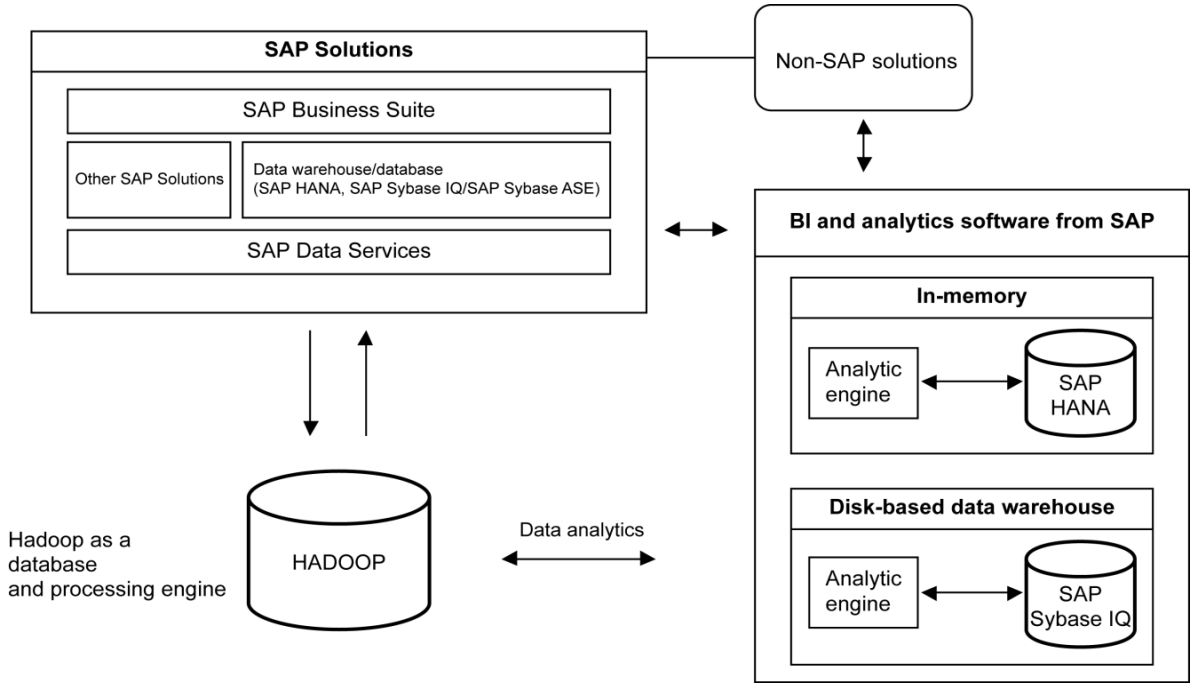

Fig. 5. SAP Hana
Rys. 5. SAP Hana

4.3. Oracle Big Data Appliance

Oracle Big Data Appliance is a system comprising not only software but also hardware elements. The hardware is optimized so as to run the enhanced Big Data software components. Oracle Big Data Appliance, to achieve maximum performance, can be connected with Oracle Exadata Database Machine, which ensures great performance in data warehouses hosting and transaction processing databases. To make this environment complete, Oracle Exadata Database Machine can be connected with Oracle Exalytics In-Memory Machine, which provide high performance of business intelligence and planning applications [17].

Individual tools are dedicated for different steps of data processing as well as levels of infrastructure – Fig. 6.

Oracle NoSQL Database ("Not Only SQL") is a distributed key-value database, built on Berkeley DB Java Edition. This database system indexes the data and supports transactions in the opposite to HDFS, which stores unstructured data in very large files. Oracle NoSQL Database has less strict consistency rules, no schema structure, and only modest support for joins, particularly across storage nodes (in contrast to Oracle Database, which stores highly structured data). Such database is typically used to store customer profiles and similar data for identifying and analyzing Big Data.

Cloudera's Distribution, including Apache Hadoop (CDH), uses the Hadoop Distributed File System (HDFS), while Cloudera Manager provides a single administrative interface to all Oracle Big Data Appliance servers configured (as a part of the Hadoop cluster).

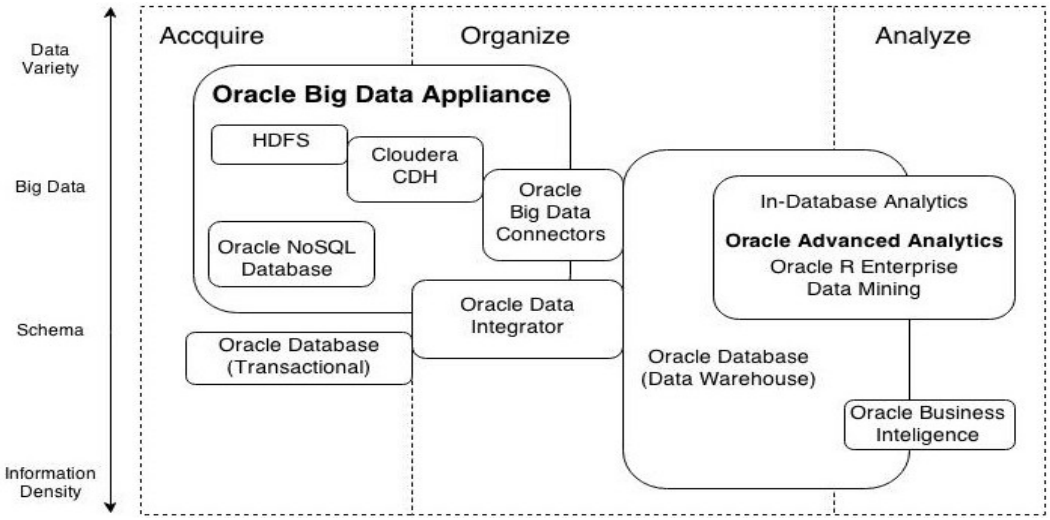

Fig. 6. Oracle Big Data Appliance Software Overview
Rys. 6. Oprogramowania Oracle Big Data Appliance

Oracle Big Data Appliance provides support for R language – an open source language and environment for statistical analysis and graphing (linear and nonlinear modeling, standard statistical methods, time-series analysis, classification, clustering, and graphical data displays). In the Comprehensive R Archive Network (CRAN) there are available open-source packages for bioinformatics, spatial statistics, and financial and marketing analysis [17].

4.4. IBM InfoSphere BigInsights

IBM BigInsights provides all features of Hadoop and additionally analytical capabilities from IBM Research. The software supports structured, semi-structured and unstructured data. It delivers management, security and reliability features to support large-scale deployments and reduces access time to the information. IBM provides both software and hardware solutions for management of Big Data [18].

Stream Computing allows to analyze large volumes of streaming data with sub-millisecond times. It can execute MapReduce models for advanced analytics. Information

Integration and Governance provides information to business initiatives. Accelerators are software components that accelerate data read with pre-packed analytical and industry specific content. Data Warehousing delivers deep operational insight with advanced in-database analytics.

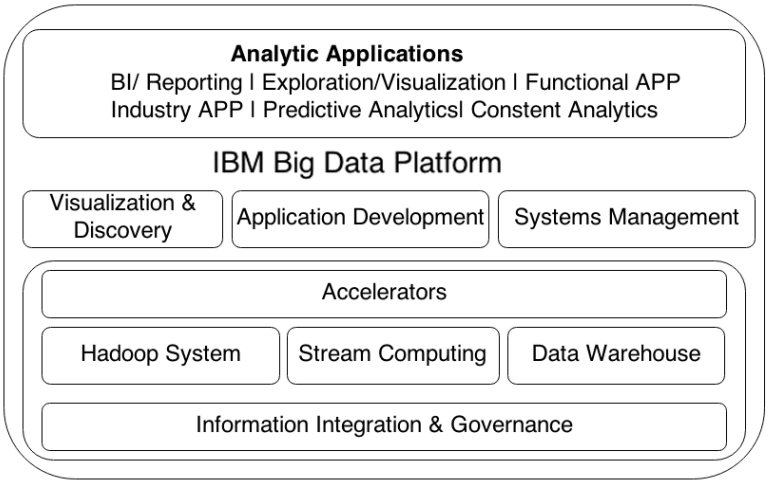

Fig. 7. IBM Big Data
Rys. 7. IBM Big Data

5. Conclusions

Concepts of Big Data, such as information volume, velocity, and variety, focus on new challenges in nowadays information environments. Data acquisition, integration, mining and analysis of huge amount of data generate specific problems and their solutions require new approaches in the design of modern software. The paper discusses selected tools and technologies available to cope with Big Data. At this moment, Hadoop can be treated as a main part of Big Data ecosystem. For users, a very important issue is, that Hadoop is an open platform. Different vendors offer their own custom versions, support as well as additional functionalities. Selected tools were developed by large companies and available for free. Hadoop is elastic solution which can cooperate with existing system which use RDBMS and can be used in commercial environments to simplify processes of integration, manipulation and analysis of the information in the Big Data environments. Hadoop is a new technology; many tools are still being intensively developed. It is probable that, in near future Hadoop would be strongly integrated with other solutions which use RDBMS, OLAP and BI technologies to process Big Data.

BIBLIOGRAPHY

1. Manyika J. et al.: Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, http://www.mckinsey.com, 2011.
2. Chwin M.: Big Data at work: Transform your business with analytics, http://www.oracle. com/in/corporate/events/bigdata-at-work-2041503-en-in.html, 10.10.2013.
3. Sharma S.: Big Data landscape. International Journal of Scientific and Research Publications, Volume 3, Issue 6, 2013, p. 1÷8.
4. White T.: Hadoop: The Definitive Guide. O'Reilly Media / Yahoo Press 2012.
5. Gigaom: Because Hadoop isn't perfect, http://gigaom.com (2013.12.22).
6. The Apache Software Foundation: Ambari, http://ambari.apache.org/ (2013.12.22).
7. Storm Community, http://storm-project.net/ (2013.12.22).
8. Main site of Giraph Project, http://giraph.apache.org/ (2013.12.22).
9. The Apache Software Foundation: Sqoop User Guide, http://sqoop.apache.org/docs/1.4.3/SqoopUserGuide.html (2013.12.22).
10. The Apache Software Foundation: Apache Mahout – An algorithm library for scalable machine learning on Hadoop, http://hortonworks.com/hadoop/mahout/ (2013.12.22).
11. Farley B.: Big data and the analytic race. SAS Institute, 2012.
12. Nadipalli R.: HDInsight essentials. Packt Publishing, 2013.
13. Microsoft: PolyBase, http://www.microsoft.com/en-us/sqlserver/solutions-technologies/data-warehousing/polybase.aspx (2013.12.22).
14. SAP Hana In-Memory, http://www.sapbigdata.com/platform/in_memory/ (2013.12.22).
15. Moore T.: The Sybase IQ Survival Guide. TDM Computing Limited, 2010.
16. SAP Hana Platform/Hadoop, http://www.sapbigdata.com/platform/hadoop/ (2013.12.22).
17. Oracle Corporation: Oracle Big Data Appliance Software Users Guide, http://docs.-oracle.com (2013.12.22).
18. IBM Software: Using IBM InfoSphere BigInsights to accelerate big data time-to-value Thought Leadership. White Paper, 2013.

Wpłynęło do Redakcji 7 lutego 2014 r.

Omówienie

Z uwagi na popularyzowanie się pojęcia Big Data można zaobserwować zwiększone zainteresowanie różnych podmiotów rozwiązaniami tego typu. Big Data obejmuje wiele aspektów związanych z szeroko pojętymi zagadnieniami systemów przetwarzania danych. Pojęcie to jest często definiowane jako 3V i wyróżnia się w nim następujące elementy: Volume – rozumiane jako ilość danych tak duża, że stawia nowe wyzwania dla infrastruktury informatycznej, Velocity – duża prędkość pozyskiwania (generowania) nowych danych, Variety – różnorodność danych (ich źródeł, formatów, sposobu pozyskiwania). Nierzadko dodawane są kolejne pojęcia (a skrót modyfikowany jest do 4V): Variability (zmienność danych), Value (ich jakość), Veracity (prawdziwość). Występowanie tak wielu różnych aspektów w rozproszonym systemie generuje nowe, specyficzne wyzwania.

W odpowiedzi na te wyzwania tworzone są różne technologie, środowiska oraz aplikacje. Ich mnogość, przy jednoczesnej nowości samych koncepcji, powoduje, że brak jest wciąż powszechnie przyjętych standardów projektowania oraz implementacji tego typu środowisk. Można zaobserwować gwałtowny rozkwit różnorodnych rozwiązań w tej dziedzinie. W niniejszym artykule zaprezentowano przegląd głównych kierunków rozwoju oprogramowania przeznaczonego do pracy w środowiskach typu Big Data.

Addresses

Krzysztof CZAJKOWSKI: Cracow University of Technology, Faculty of Physics, Mathematics and Computer Science, Institute of Computer Networks, ul. Warszawska 24, 31-155 Kraków, Polska, kc@pk.edu.pl.

Łukasz CHODAK: Cracow University of Technology, Faculty of Physics, Mathematics and Computer Science, ul. Warszawska 24, 31-155 Kraków, Polska, lukasz.chodak89@gmail.com.

Radosław KORKOSZ: Cracow University of Technology, Faculty of Physics, Mathematics and Computer Science, ul. Warszawska 24, 31-155 Kraków, Polska, radoslaw.korkosz@gmail.com.

Michał URBAŃCZYK: Cracow University of Technology, Faculty of Physics, Mathematics and Computer Science, ul. Warszawska 24, 31-155 Kraków, Polska, michal@urbanczyk.it.