"Environmental Engineering" 10th International Conference
Vilnius Gediminas Technical University
Lithuania, 27–28 April 2017
http://enviro.vgtu.lt

eISSN 2029-7092 / eISBN 978-609-476-044-0

Article ID: enviro.2017.216
DOI: https://doi.org/10.3846/enviro.2017.216

Tools for the Storage and Analysis of Spatial Big Data

Przemysław Lisowski[1], Adam Piórkowski[2], Andrzej Leśniak[3]

AGH University of Science and Technology
Department of Geoinformatics and Applied Computer Science
al. Mickiewicza 30, 30-059 Kraków, Poland
E-mails: [1]plis@agh.edu.pl (corresponding author); [2]pioro@agh.edu.pl; [3]lesniak@agh.edu.pl

Abstract. Storing large amounts of spatial data in GIS systems is problematic. This problem is growing due to ever-increasing data production from a variety of data sources. The phenomenon of collecting huge amounts of data is called Big Data. Existing solutions are capable of processing and storing large volumes of spatial data. These solutions also show new approaches to data processing. Conventional techniques work with ordinary data but are not suitable for large datasets. Their efficient action is possible only when connected to distributed file systems and algorithms able to reduce tasks. This review focuses on the characteristics of large spatial data and discusses opportunities offered by spatial big data systems. The work also draws attention to the problems of indexing and access to data, and proposed solutions in this area.

Keywords: Big Data, spatial data, data wharehouse.

Conference topic: Technologies of Geodesy and Cadastre.

Introduction

Spatial data systems are increasingly common because an increasing number of new IT systems use spatial data in algorithms in many fields such as logistics, administration, geology, geophysics, etc. Specifications under which European Union members are obliged to store and share spatial information (EU 2017) are defined as part of the INSPIRE directive.

Special database types have been developed for saving spatial information. These so-called spatial databases are able to store geometry as vectors (Cichociński, Dębińska 2010). The major advantages of spatial database relate to the processing and analysis of spatial objects. Many commercial and open-source spatial database systems have been developed. The most common open-source solutions are PostGIS (PostGIS 2017) spatial extension for PostgresSQL (PostgreSQL 2017), SQLLite with SpatiaLite (SpatiaLite 2017), and MySQL Spatial (MySQL 2017); the most common commercial commercial systems are Oracle Spatial (Oracle 2017) and SQL Server Spatial (Microsoft 2017).

The aforementioned systems use one of two storage standards: OGC (OpenGIS 2017) and SQL/MM-Spatial (ISO/IEC 13249-3:1999), both of which offer data storage in two and three dimensions. Vector notation makes it possible to save objects as points, lines, polygons, or a collection of these types. In addition, spatial databases have been developed that offer raster graphicstorage (Piórkowski 2011; Lisowski *et al*. 2014). Usage of this datatype has increased because it is valuable for analysis of spatial objects and their positions. For example, this method of storage is offered by PostGIS (PostGIS 2017) and Spatialite (SpatiaLite 2017), both of which provide functionality such as tiles and spatial analysis for rasters.

Currently, a lot of spatial analysis is performed; therefore, efficient algorithms and methods for accessing and finding information are required, such as data warehouses, which are an attractive method for storing spatial data as they offer superior features for accessing, organizing, and optimizing data.

Data warehouse

Data warehouses are used to optimize access to dataset archives, which is crucial for spatial data. The scheme of a warehouse should be adapted according to the type of data storage. Spatial datasets can be from heterogeneous sources, often resulting in different incompatible formats. A data warehouse has various schemas including OLTP (On-Line Transactional Processing), OLAP (On-Line Analytical Processing), ROLAP (Relational On-Line Analytical Processing), MOLAP (Multidimensional On-Line Analytical Processing), and HOLAP (Hybrid On-Line Analytical Processing). One of the most widely used is ROLAP (Ying *et al.* 2017). This scheme can be constructed with a variety of logic types: star, snowflake and constellation. Irrespective of the type of logic, all are built with elements such as fact and dimension tables. This enables faster access to data than the standard access methods of relational databases (Kimball, Ross 2011). Example of data warehouse for spatial mining data is shown in Figure 1.

© 2017 Przemysław Lisowski, Adam Piórkowski, Andrzej Leśniak. Published by VGTU Press. This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY-NC 4.0) License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

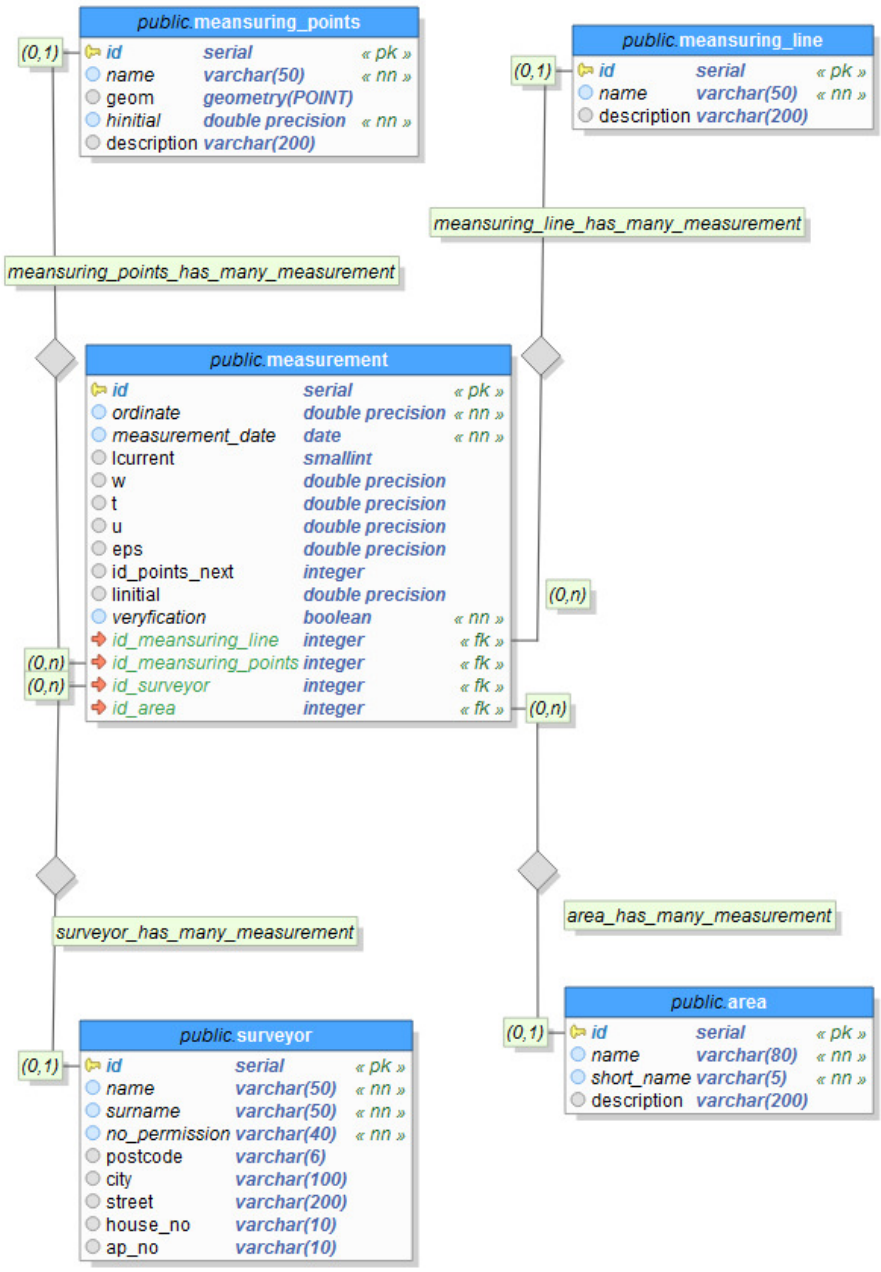

Fig. 1. Example spatial data warehouse (source: Lisowski *et al*. 2015)

Loading data into data warehouses is often a problematic task as it requires extensive preparation, conversion, and transfer of data from the source. In the literature, this is referred to as the ETL (Extract, transform, load) process (Gorawski 2009). As is described in the literature, commercial ETL solutions such as Oracle Warehouse Builder, SQL Server Integration for MS SQL Server, and IBM InfoSphere Warehouse for IBM DB2 are popular and open-source applications are also available. A solution for loading spatial data into data warehouse in the literature describes features offered in Talend Open Studio and GeoKattelle (Klisiewicz *et al.* 2011). Data warehouse structures can be used to store spatial data and to decrease time taken to execute spatial analysis.

Big data

Increasing amounts of data are being generated, a phenomenon referred to Big Data. When, traditional solutions do not work efficiently enough and a data warehouse cannot cope with the load, a Big Data solution should be used.

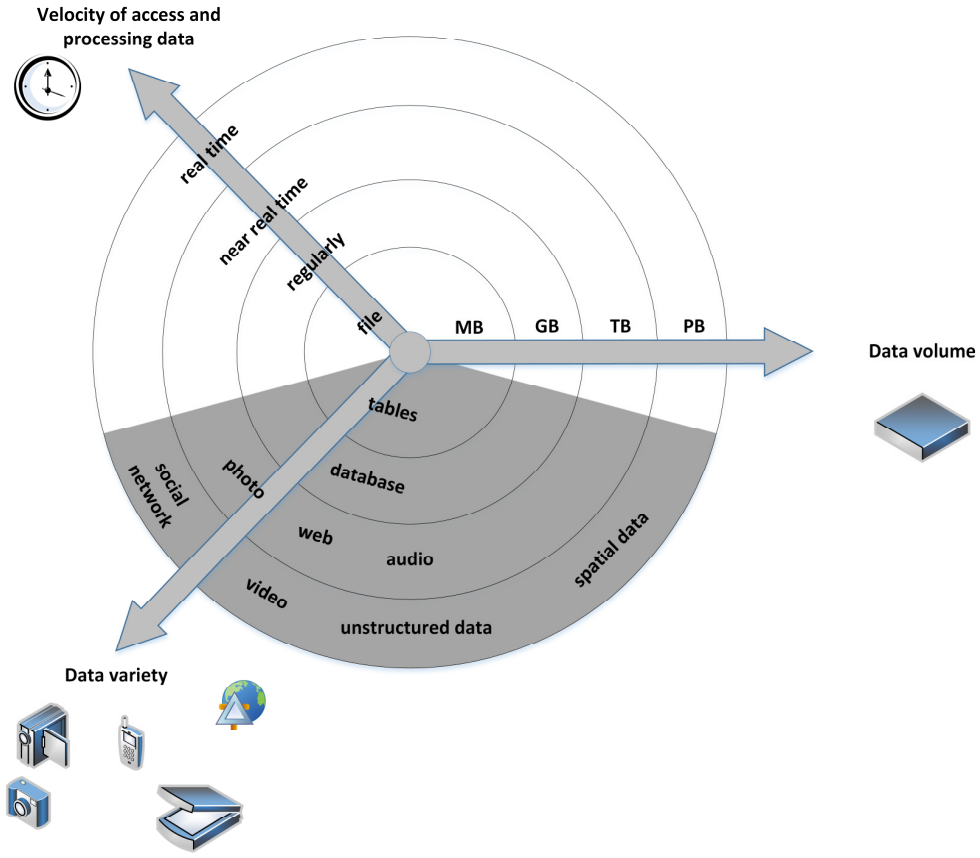

Fig. 2. Model 3-V (Klein *et al.* 2013)

A significant issue associated with Big Data is the storage of large volumes of collected data. The volume of data being produced is increasing faster than the ability to process and store it. Big data is defined by three concepts: large data volumes, the variety of information types, and the varying velocities at which data is collected (Krishnan 2013).

A few large dataset models are defined in literature (see Rutkowski 2014). For example, a 3-V model is proposed (Fig. 2). The dimensions of this model are described as: quantity of data (*volume*), speed of data (*velocity*), and diversity of data (*variety*) (Klein *et al.* 2013).

Data can be collected from various sources including websites, online community projects, and measuring devices. Spatial information can also be saved. Data is usually collected in various structures. This data cannot be processed by traditional systems as the data sets are too large, varying from terabytes to petabytes. 4-V and 5-V models are also described in the literature. These models contain additional definitions such as the veracity of data and the value in collecting it (see Rutkowski 2014). Another important problem is the real-time processing of collected data (*velocity*)

The models listed above precisely characterize the properties of large datasets. Datasets often include spatial information, a fact which affects the approach taken to analysis of Big Data sets. However, many methods have been developed for working with large spatial datasets. One such solution involves statistical methods such as PCA analysis, data dictionaries, data clustering, and data sampling. These methods make it possible to filter data and access to critical data more rapidly (Slavakis *et al.* 2014). Moreover, these methods optimally organize datasets and enable classification and reduction of data dimensions to the minimum required for processing. Also, statistical methods for Big Data can be used for forecasting. Currently, these operations for large datasets are performed in distributed systems that store data in the cloud or distributed data file systems. These Big Data solutions are developed using MapReduce/Hadoop (Slavakis *et al.* 2014; Apache 2017). MapReduce was developed for use in automated parallel processing (Katal *et al.* 2013). The MapReduce operating diagram is shown in Figure 3.

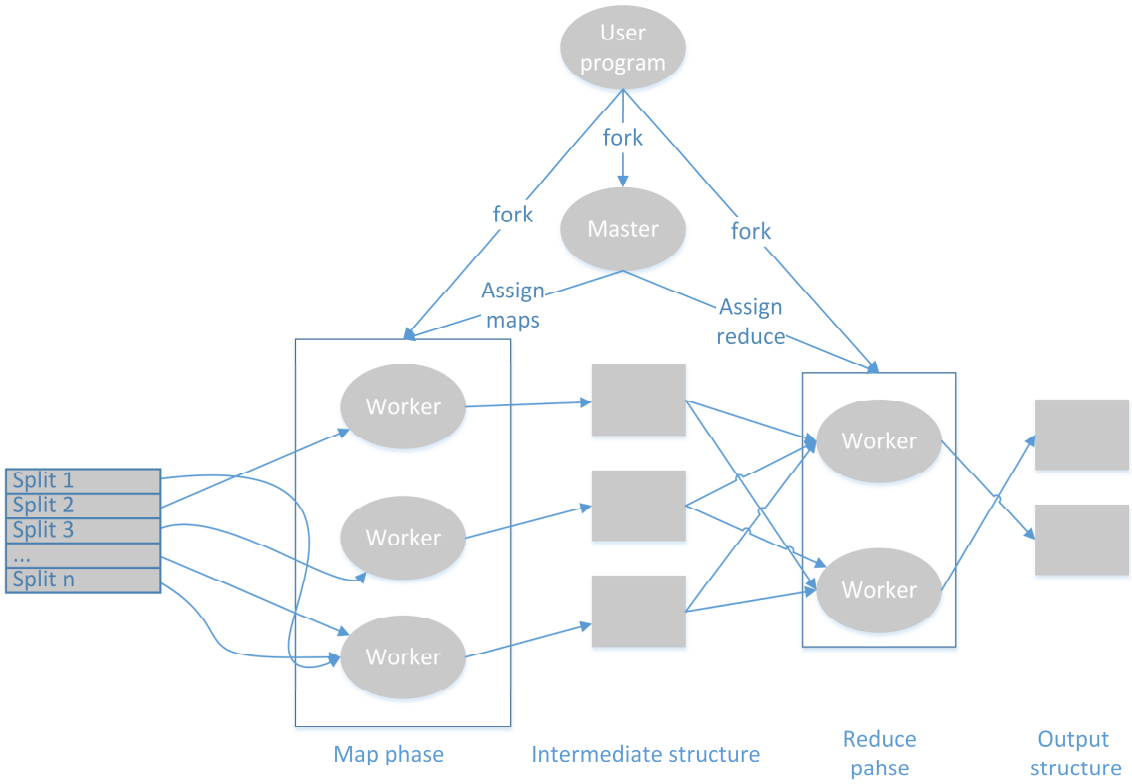

Fig. 3. MapReduce operating diagram (Dean, Ghemawat 2008)

Spatial data sets

A spatial dataset is classified according to the source data. The first type is data from online communities such as Wikimaps (Wikimap 2017) or OpenStreetMap (OpenStreetMap 2017). Examples of this data are shown in Figure 4. These sources contain geographic information with text or number attributes. The second type is data from social networking services such as Facebook or Twitter. Spatial location, such as addresses and the location of outdoor events are published by users (Gao *et al.* 2014). Another type is datasets from measurement instruments. One example is data from sensors that monitor flood embankments (Pięta *et al*. 2014). The examples listed above can be classified as Big Data because they contain a variety of information which can be processed in real time. Additionally, this data can be large in size (terabytes or more).

Fig. 4. Illustration of example spatial data from OpenStreetMap

Working with Big Data sets requires innovative methods. One example is the way in which public information is used in the construction of gazetteers. Nowadays, a lot of geographic objects are created, but it is not possible create meaningful new names for all of them. The automatic creation of new object names was proposed using datasets

from social networks and maps. This would create POI points with an implementation of MapReduce, using continuously collected data. The authors of one paper developed a solution using Apache Hadoop with GPHadoop (Gao *et al.* 2014) combined with a cluster containing nodes for processing spatial data. In this study, a classification algorithm is proposed which makes it possible to add to an object the most commonly used name. Thus, geographic objects are automatically named (Gao *et al.* 2014).

MapReduce can quickly prepare data sets for analysis. A platform based on MapReduce for exporting OpenStreetMap data was developed. This contains algorithms for spatial indexing data and algorithms make it possible to make queries using MapReduce. This platform is mentioned in the literature (Alarabi *et al.* 2014). Spatial indexation in this solution is based on SpatialHadoop (SpatialHadoop 2017), which can create indexation using structures such as grid, R-tree, and R+-tree. This platform is also capable of partitioning data using Hadoop algorithms.

The aforementioned solutions are developed using existing algorithms and – with some adjustments – can be used in other applications. Moreover, Apache Hadoop can create data warehouses using MapReduce. Hadoop-GIS is solution for spatial data sets using the Apache Hadoop platform. Hadoop-GIS (Aji *et al.* 2013) should be used for spatial data sets. It was developed for Hive (part of Hadoop): a queries system for data sets stored in an HDFS file system. An alternative option for Hadoop is PigLatin with the Rout extension. This extension can perform typical operations for spatial big data using MapReduce (Jayalath, Eugster 2013). Irrespective of the chosen technology, the aforementioned solutions make it possible to create scalable and effective systems for processing spatial big data.

Table 1. Comparison of available solutions for storing spatial datasets

	Relational database	Data warehouse	GPHadoop	Hadoop-GIS	PigLatin with Rout
Hardware Requirements	One computer	One computer	Computer cluster	Computer cluster	Computer cluster
Licence	Commercial or open source	Commercial or open source	Open source	Open source	Open Source
MapReduce	No	No	Yes	Yes	Yes
Communication	SQL language	SQL language	Processing with ESRI GIS tools	QL^{SP}	PigLatin language

Conclusions

Available solutions are capable of processing and storing spatial big data. Existing spatial big data systems have various requirements, which can limit using these systems (Table 1). This study describes solutions for storing spatial data sets and focuses on storing spatial big data such as GPHadoop, Hadoop-GIS, and PigLatin with Rout. Examples for processing big data are described. The presented studies show new ways in which data can be processed. Statistical methods such as clustering and filtration of data sets are proposed in many studies. Also, MapReduce is used by many authors for optimization of access to Big Data sets. MapReduce is often implemented using the Apache Hadoop system.

The aforementioned studies show the advantages of solutions that use Hadoop for processing and storing spatial big data. Moreover, a lot of research focuses on issues such as indexation and data access times. The cited studies contain proposed solutions for these issues. A universal solution cannot be used for analysis and storage of spatial big data as each issue requires an individual algorithm.

References

Aji, A.; Wang, F.; Vo, H.; Lee, R.; Liu, Q.; Zhang, X.; Saltz, J. 2013. Hadoop-GIS: a high performance spatial data warehousing system over MapReduce, *Proceedings VLDB Endowment* 6(11): 1009–1020. https://doi.org/10.14778/2536222.2536227

Alarabi, L.; Eldawy, A.; Alghamdi, R.; Mokbel, M. 2014. TAREEG: A MapReduce-Based Web Service for Extracting Spatial Data from OpenStreetMap, in *SIGMOD '14 Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, New York, 897–900.

Apache Hadoop Homepage [online], [cited 13 January 2017]. Available from Internet: http://hadoop.apache.org/

Cichociński, P.; Dębińska, E. 2010. Spatial database supporting local covernments in implementing entrepreneurship development policy. ZN Pol. Sl. *Studia Informatica* 31(90), Gliwice.

Dean, J.; Ghemawat, S. 2008. MapReduce: simplified data processing on large clusters [online], *Communications of the ACM* 51(1): 107 [cited 13 January 2017]. Available from Internet: http://portal.acm.org/citation.cfm?doid=1327452.1327492

EU. 2007. Directive 2007/2/EC of the European Parliament and of the council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) [online], *Official Journal of the European Union* 50: 1–14

[cited 13 January 2017]. Available from Internet: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:EN:PDF

Gao, S.; Li, L.; Li, W.; Janowicz, K.; Zhang, Y. 2014. Constructing gazetteers from volunteered big geo-data based on Hadoop, *Comput. Environ. Urban Syst.*

Gorawski, M. 2009. Advanced data warehouses ZN Pol. SI *Studia Informatica* 30(3B). Gliwice

ISO/IEC 13249-3:1999. *Information technology – Database languages – SQL Multi-media and Application Packages* – Part 3: Spatial International Organization for Standardization, 2000.

Jayalath, C.; Eugster, P. 2013. Efficient geo-distributed data processing with rout, in *IEEE 33rd International Conference on Distributed Computing Systems*, 470–480.

Katal, A.; Wazid, M.; Goudar, R. H. 2013. Big data: Issues, challenges, tools and Good practices, in *Sixth International Conference Contemporary Computing (IC3)*, Noida, 404–409.

Kimball, R.; Ross, M. 2011. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

Klein, D.; Tran-Gia, P.; Hartmann, M. 2013. Big Data, *Informatik-Spektrum* 36(3). Berlin Heidelberg.

Klisiewicz, J.; Piórkowski, A.; Porzycka, S. 2011. Construction of spatial data ETL process. ZN Pol. Sl. *Studia Informatica* 32(97). Gliwice.

Krishnan, K. 2013. *Data Warehousing in the age of Big Data*. Elsevier, Walthman.

Lisowski, P.; Krawczyk, A.; Leśniak, A. 2015. A method of data storage for case of deformations in mining area, in *IAMG 2015: The 17th Annual Conference of the International Association for Mathematical Geosciences*, 5–13 September 2015, H. Schaeben (Eds.), Freiberg, Germany, 1273–1278. ISBN 978-3-00-050337-5.

Lisowski, P.; Krawczyk, A.; Porzycka-Strzelczyk, S. 2014. Possibilities of 3D data storage in spatial databases. ZN Pol. Sl. *Studia Informatica* 35(116). Gliwice.

Microsoft SQL Server Hompage [online], [cited 13 January 2017]. Available from Internet: http://www.microsoft.com/sql/

MySQL Homepage [online], [cited 13 January 2017]. Available from Internet: http://www.mysql.com/

OpenGIS Implementation Specification for Geographic information – Simple feature access [online], [cited 13 January 2017]. Available from Internet: http://www.opengeospatial.org/standards/sfs/

OpenStreetMap Homepage [online], [cited 13 January 2017]. Available from Internet: http://www.openstreetmap.org

Oracle Spatial and Graph Homepage [online], [cited 13 January 2017]. Available from Internet: http://www.oracle.com/technetwork/database/options/spatialandgraph/

Pięta, A.; Lupa, M.; Chuchro, M.; Piórkowski, A.; Leśniak, A. 2014. A Model of a system for stream data storage and analysis dedicated to sensor networks of embankment monitoring, in *International Conference on Computer Information Systems and Industrial Management*, Berlin, Heidelberg, 514–525. https://doi.org/10.1007/978-3-662-45237-0_47

Piórkowski, A. 2011. Mysql Spatial and Postgis – Implementations of Spatial Data Standards, *EJPAU* 14(1), #03, Wrocław.

PostGIS Homepage [online], [cited 13 January 2017]. Available from Internet: http://postgis.refractions.net/

PostgreSQL Homepage [online], [cited 13 January 2017]. Available from Internet: http://www.postgresql.org/

Rutkowski, L. 2014. *Big Data – Nowe Wyzwania Informatyki*. Laudacja w przewodzie dotyczącym nadania tytułu Doktora Honoris Causa AGH dla członka Polskiej Akademii Nauk prof. dr. hab. inż. Leszka Rutkowskiego, Dyrektora Instytutu inteligencji Systemów Informatycznych Politechniki Częstochowskiej. Thesis AGH University of Science and Technology. Kraków, 25–39.

Slavakis, K.; Giannakis, G. B.; Mateos, G. 2014. Modeling and Optimization for Big Data Analytics: (Statistical) learning tools for our era of data deluge, *IEEE Signal Processsing Magazine* 31(5). New York.

SpatialHadoop Homepage [online], [cited 13 January 2017]. Available from Internet: http://spatialhadoop.cs.umn.edu/

SPATIALITE Home page [online], [cited 13 January 2017]. Available from Internet: http://www.gaia-gis.it/gaia-sins/

Wikimap Homepage [online], [cited 13 January 2017]. Available from Internet: http://www.wikimapia.org

Ying Chen; Dehne, F.; Eavis, T.; Rau-Chaplin, A. 2017. Parallel ROLAP data cube construction on shared-nothing multiprocessors, in *Proceedings International Parallel and Distributed Processing Symposium*, 10. IEEE Comput. Soc. Udostępniono styczeń 4.