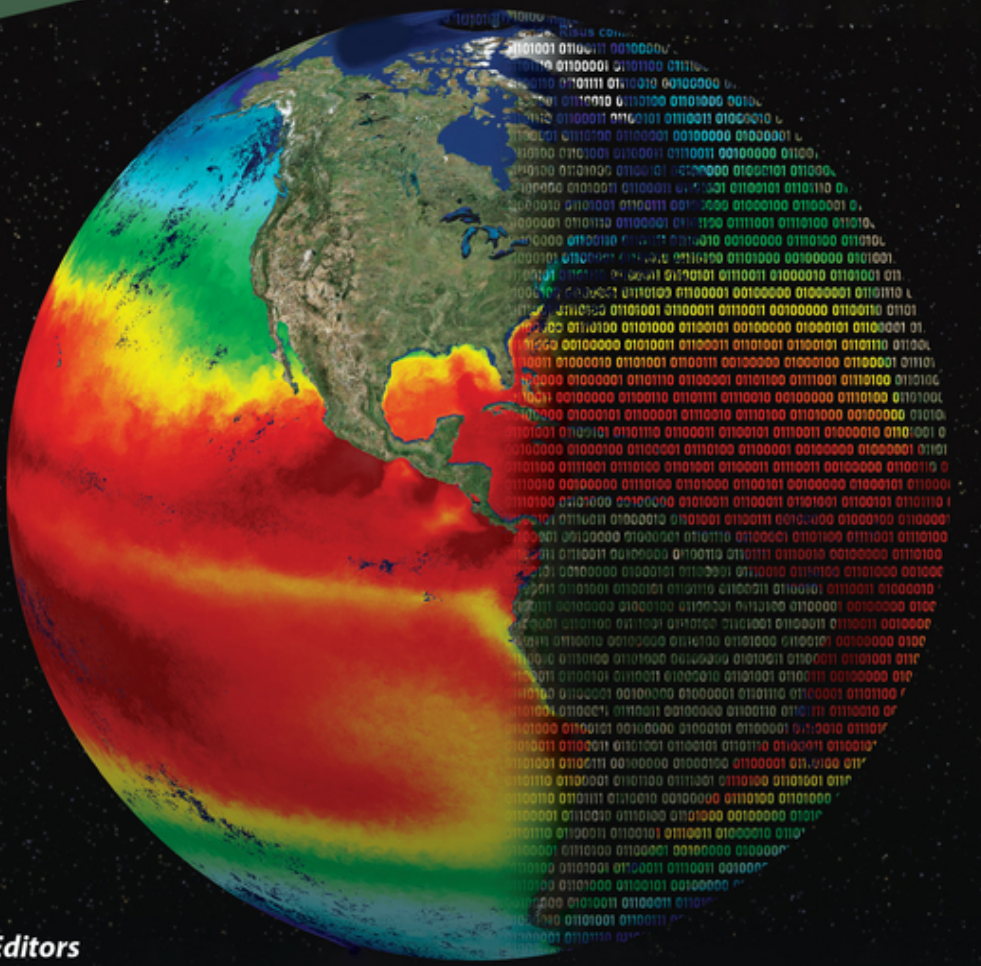
SPECIAL PUBLICATIONS SERIES

AGU
ADVANCING EARTH AND SPACE SCIENCE

Big Data Analytics in Earth, Atmospheric, and Ocean Sciences



Editors
Thomas Huang
Tiffany C. Vance
Christopher Lynnes

WILEY

Big Data Analytics in Earth, Atmospheric, and Ocean Sciences

Special Publications 77

BIG DATA ANALYTICS IN EARTH, ATMOSPHERIC, AND OCEAN SCIENCES

Thomas Huang
Tiffany C. Vance
Christopher Lynnes

Editors

This Work is a co-publication of
the American Geophysical Union and John Wiley and Sons, Inc.

AGU
ADVANCING EARTH
AND SPACE SCIENCE

WILEY

This edition first published 2023

Published under the aegis of the AGU Publications Committee

Matthew Giampoala, Vice President, Publications
Carol Frost, Chair, Publications Committee
For details about the American Geophysical Union visit us at www.agu.org.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Huang, Thomas (Technologist), editor. | Vance, C. Tiffany, editor. | Lynnes, Christopher, editor.
Title: Big data analytics in earth, atmospheric, and ocean sciences / Thomas Huang, Vance, C. Tiffany, Christopher Lynnes, editors.
Description: Hoboken, NJ : Wiley-American Geophysical Union, 2023. | Includes bibliographical references and index.
Identifiers: LCCN 2022020168 (print) | LCCN 2022020169 (ebook) | ISBN 9781119467571 (cloth) | ISBN 9781119467564 (adobe pdf) | ISBN 9781119467533 (epub)
Subjects: LCSH: Earth sciences–Data processing. | Atmospheric science–Data processing. | Marine sciences–Data processing. | Big data.
Classification: LCC QE48.8 .B54 2022 (print) | LCC QE48.8 (ebook) | DDC 550.0285/57–dc23/eng20220722
LC record available at https://lccn.loc.gov/2022020168
LC ebook record available at https://lccn.loc.gov/2022020169

Cover Design: Wiley
Cover Images: Courtesy of Kate Culpepper with design elements provided by Esri, HERE, Garmin, FAO, NOAA, USGS, EPA | Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Set in 10/12pt TimesNewRomanMTStd by Straive, Chennai, India

CONTENTS

LIST OF CONTRIBUTORS

Edward M. Armstrong
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Alberto Arribas
Microsoft
Reading, UK; and
Department of Meteorology
University of Reading
Reading, UK

Jessica Austin
Axiom Data Science, LLC
Anchorage, Alaska, USA

Rob Bochenek
Axiom Data Science, LLC
Anchorage, Alaska, USA

Mark A. Bourassa
Center for Ocean-Atmospheric Prediction Studies, and
Department of Earth, Ocean, and Atmospheric Science
Florida State University
Tallahassee, Florida, USA

Jonathan Brannock
North Carolina Institute for Climate Studies
NOAA Cooperative Institute for Satellite Earth System Studies
North Carolina State University
Asheville, North Carolina, USA

Otis Brown
North Carolina Institute for Climate Studies
NOAA Cooperative Institute for Satellite Earth System Studies
North Carolina State University
Asheville, North Carolina, USA

Kevin A. Butler
Environmental Systems Research Institute
Redlands, California, USA

Nga T. Chung
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Thomas Cram
National Center for Atmospheric Research
Boulder, Colorado, USA

Jenny Dissen
North Carolina Institute for Climate Studies
NOAA Cooperative Institute for Satellite Earth System Studies
North Carolina State University
Asheville, North Carolina, USA

Kyle Dumas
ARM Research Facility
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

John-Marc Dunaway
Axiom Data Science, LLC
Anchorage, Alaska, USA

Jocelyn Elya
Center for Ocean-Atmospheric Prediction Studies
Florida State University
Tallahassee, Florida, USA

Eamon Ford
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

David W. Gallaher
National Snow and Ice Data Center
Boulder, Colorado, USA

Kevin Michael Gill
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Glenn E. Grant
National Snow and Ice Data Center
Boulder, Colorado, USA

Frank R. Greguska III
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Mahabaleshwara Hegde
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA

Alex Held
CSIRO Centre for Earth Observation
Canberra, ACT, Australia

Erik Hoel
Environmental Systems Research Institute
Redlands, California, USA

Benjamin Holt
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Hook Hua
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Thomas Huang
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Joseph C. Jacob
NASA Jet Propulsion Laboratory
Pasadena, California, USA

Zaihua Ji
National Center for Atmospheric Research
Boulder, Colorado, USA

Karthik Kashinath
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Edward J. Kearns
First Street Foundation
Brooklyn, New York, USA

Bhargavi Krishna
ARM Research Facility
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Vitaliy Kurlin
Department of Computer Science
University of Liverpool
Liverpool, UK

Michael M. Little
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA

Qin Lv
Department of Computer Science
University of Colorado
Boulder, Colorado, USA

Christopher Lynnes
NASA Goddard Space Flight Center (retd.)
Greenbelt, Maryland, USA

Gerald Manipon
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Theo McCaie
Met Office
Exeter, UK

Dmitriy Morozov
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Grzegorz Muszynski
Lawrence Berkeley National Laboratory
Berkeley, California, USA; and
Department of Computer Science
University of Liverpool
Liverpool, UK

Matt Paget
CSIRO Centre for Earth Observation
Canberra, ACT, Australia

Tom Powell
Met Office
Exeter, UK

Giri Prakash
ARM Research Facility
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Prabhat Ram
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Niall Robinson
Met Office
Exeter, UK; and
University of Exeter
Exeter, UK

Sujen Shah
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Adrienne Simonson
Office of the Chief Information Officer
National Oceanic and Atmospheric Administration
Asheville, North Carolina, USA

Shawn R. Smith
Center for Ocean-Atmospheric Prediction Studies
Florida State University
Tallahassee, Florida, USA

Kate Szura
Interactions LLC
Franklin, Massachusetts, USA

Ronnie Taib
CSIRO Data61
Sydney, NSW, Australia

Jacob Tomlinson
NVIDIA
Reading, UK

Vardis Tsontos
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Tiffany C. Vance
U.S. Integrated Ocean Observing System
National Oceanic and Atmospheric Administration
Silver Spring, Maryland, USA

Peter Wang
CSIRO Data61
Sydney, NSW, Australia

Michael Wehner
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Brian D. Wilson
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Robert Woodcock
CSIRO Centre for Earth Observation
Canberra, ACT, Australia

Elizabeth Yam
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Chaowei Phil Yang
George Mason University
Fairfax, Virginia, USA

Alice Yepremyan
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Hailiang Zhang
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA

PREFACE

The seeds for this book were sown in sessions on Big Data Analytics, held at the 2016 Fall Meeting of the American Geophysical Union. At the time, Earth Science data were projected to rise by orders of magnitude in the coming decade, and the community was investigating a variety of emergent technologies and techniques to make the best use of the coming deluge. The chapters of this book are a representative, but by no means exhaustive, collection of those and similar investigations.

Big Earth Data Analytics can be defined as the application of increasingly sophisticated tools for data analysis and display to the rapidly increasing volume of Earth science data to obtain information, and eventually insight. This combines two concepts: Big Earth Data and Data Analytics. Big Earth Data refers both to the volume of data sets and the combination of data from a variety of sources, in a variety of formats, and from a variety of disciplines. To get a sense of the volume, NOAA generates tens of terabytes of data a day from satellites, radars, ships, weather models, and other sources. The National Aeronautics and Space Administration (NASA) Earth Observation archives were growing by more than 30 TB per day in 2020 with daily growth expected to increase to 130 TB/day by 2024 as new satellites launch; and the European Centre for Medium-Range Weather Forecasts (ECMWF) meteorological data archive adds 200 terabytes of new data daily. However, the data are "big" not only in their volume but in their varied formats, disciplines, structures, and formats. As such, they are disruptors to traditional analysis methods, and to the kinds of questions that can be asked by researchers. Data analytics are increasingly driven by the availability of high-volume and heterogeneous data sets. Data size and complexity affect all aspects of data management and usage, requiring new approaches and tools. Despite the challenges to acquire, use, and analyze Big Earth Data, they are already being utilized extensively in climate, oceanographic, and biology related works. Easily available data lead to the ability to analyze longer scale records and patterns over large spatial domains.

Analyses of these data borrow both from traditional scientific analyses and from tools developed for business applications. These types of data

analytics are developed by university and other research teams. They are increasingly becoming an area of interest to cloud providers and analytics companies. From Google's Earth Engine for analyzing Earth science data at scale, to the National Oceanic and Atmospheric Administration's (NOAA's) Big Data Program, big data about the Earth and their analysis are increasingly common. Amazon's Elastic MapReduce and SageMaker are common building blocks for cloud-based analysis and Galileo (a.k.a. Service Workbench) is Amazon's latest Web application for interactive analysis. Microsoft Azure ML Studio is another popular cloud-based data analysis solution. Big Earth Data analyses increasingly rely on cloud-based storage and processing capabilities as the volume of the data and the computing resources needed go beyond local resources.

This book is organized into three parts. It starts with the big picture, covering Big Data Analytics Architecture. This part begins with a chapter addressing the geospatial aspect of Big Earth Data from a variety of perspectives. This is followed by a chapter discussing the data management challenges posed by data at scale, particularly in the context of making them available for analysis. This is complemented by a chapter discussing the challenges of scaling up the analysis itself. The following chapters cover large-scale projects such as NASA's Earth Exchange, which enables large scale data analysis in a supercomputing environment and the NOAA Big Data Project, which makes data sets available to end users via several cloud providers. Part I also includes chapters on architectures and fully realized systems, such as Data Cube, NEXUS and the Apache Science Data Analytics Platform, and a NoSQL based platform for exploring and analyzing in situ data.

The second part of the book, Analysis Methods for Big Earth Data, addresses some specific techniques to derive information and/or insight from big data, emphasizing the unique aspects of Earth Observations. Part II begins with two chapters on the use of geospatial statistics for analysis, followed by a chapter melding machine learning with geophysical constraints, and finally a chapter benchmarking different analytical methods for spatiotemporal analysis.

The third part of the book, Big Earth Data Applications, describes a few specific applications of big analysis techniques and platforms: weather and climate model analysis, atmospheric river patterns, Antarctic land surface temperatures extremes, satellite in situ match-ups of oceanographic data, and vessel tracking. This is clearly a small sample of existing applications; rather, the sample shows how some very different analysis methods can find diverse applications in the Earth sciences.

While the application of big Earth data analytics covers a range of applications, a number of common themes in the chapters of this book

include (1) the role of the cloud, especially with ever increasing data sizes; (2) limitations and costs of using the cloud, including the unpredictability of costs and the high cost of data egress from the cloud; (3) techniques to maintain data integrity during file transfers; (4) efficiencies via partial reads from Web object storage; (5) the use of data/object stores; (6) serverless and other intrinsic functions to standardize computations; (7) data pipelines and the use of Docker to encapsulate analyses; (8) development of application programming interfaces; (9) GeoTIFFs, Zarr, and Parquet as cloud file formats for satellite and in situ data; and (10) hard limits on data sizes in the cloud, which is especially important with satellite data.

While the chapters in this book provide a broad introduction to the subject, there are still many opportunities to address challenges posed by big data analytics, such as incorporating new data sources, implementing data standards, optimizing the use of cloud and supercomputing resources, and incorporating artificial intelligence and machine learning. As these challenges are surmounted, the computing power and agile infrastructure of the cloud will support the emergence of important new analyses and insights, in turn supporting new policy making. At the same time, new policy challenges are raised by the solutions. The use of cloud resources for data storage and analysis has the potential to both enable and complicate the accessibility of both the data and the analysis methods by the wider community, particularly as the community broadens to new application, education, and citizen scientist users. On the other hand, data egress fees or cloud provider-specific tools may impair long-term data preservation, scientific reproducibility, and basic equity.

Thomas Huang
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA
Tiffany C. Vance
U.S. Integrated Ocean Observing System
National Oceanic and Atmospheric Administration
Silver Spring, Maryland, USA
Christopher Lynnes
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA (retd.)

1

AN INTRODUCTION TO BIG DATA ANALYTICS

Erik Hoel

Environmental Systems Research Institute, Redlands, California, USA

Big data analytics, in the context of geospatial data, employs distributed computing using advanced tools that support spatiotemporal analysis, spatial statistics, and machine learning algorithms and techniques (e.g., classification, clustering, and prediction) on very large spatiotemporal data sets to visualize, detect patterns, gain deeper understandings, and answer questions. In this chapter, the key definitions, domain specific problems, analysis concepts, current technologies and tools, and remaining challenges are discussed.

1.1. Overview

Big data analytics involves analyzing large volumes of varied data, or big data, to identify and understand patterns, correlations, and trends that ordinarily are invisible due to the volumes involved in order to allow users and organizations to make better decisions. These analytics, in the context of geospatial data, commonly involve spatial processing, sophisticated spatial statistical algorithms, and predictive modeling. Big data can be obtained from a wide variety of sources; this includes sensors (both

Big Data Analytics in Earth, Atmospheric, and Ocean Sciences, Special Publications 77, First Edition. Edited by Thomas Huang, Tiffany C. Vance, and Christopher Lynnes.

DOI: 10.1002/9781119467557.ch1

stationary and moving), aerial and satellite imagery, Lidar, videos, social networks, website activity, sales transaction records, and real-time stock trading transactions. Users and data scientists apply big data analytics to evaluate these large collections of data, data with volumes that traditional analytical systems are unable to accommodate (Miller & Goodchild, 2014). This is particularly the case with unstructured or semistructured data (such data types are problematic with data warehouses, which often utilize relational database concepts and work with structured data).

To address these complex demands, many new analytic environments and technologies have been developed. This includes distributed processing infrastructures such as Spark and MapReduce (Dean & Ghemawat, 2008; Garillot &Maas, 2018; Zaharia et al., 2010), distributed file stores, and NoSQL databases (Alexander & Copeland, 1988; DeWitt & Gray, 1992; Klein et al., 2016; NoSQL, 2022; Pavlo & Aslett, 2016). Many of these technologies are available in open-source software frameworks, such as Apache Hadoop (2018), that can be used to process huge data sets with clustered systems.

When working with big data, there is a collection of objectives that users have when performing big data analytics (Marz & Warren, 2013; Mysore et al., 2013). These include

1. *Discovering value from big data*. Visualize and analyze big data in a way that reveals patterns, trends, and relationships that traditional reports and spatial processing do not. Data may exist in many disparate places, streams, or web logs.
2. *Exploiting streaming data*. Filter and convert raw streaming data from various sources, which contain geographical elements, into geographic layers of information. The geographical layers can then be used to create new, more useful maps and dashboards for decision making.
3. *Exposing geographic patterns*. Use maps and visualization to see the story behind the data. Examples of identifying geographical patterns include retailers seeing where promotions are most effective and where the competition is, banks understanding why loans are defaulting and where there is an underserved market, climate-change scientists determining the impact of shifting weather patterns.
4. *Finding spatial relationships*. Seeing spatially enabled big data on a map allows you to answer questions and ask new ones. Where are disease outbreaks occurring? Where is insurance risk greatest given recently updated population shifts? Geographic thinking adds a new dimension to big data problem solving and helps you make sense of big data.

5. *Performing predictive modeling*. Predictive modeling using spatially enabled big data helps you develop strategies from if/then scenarios. Governments can use it to design disaster response plans. Natural resource managers can analyze recovery of wetlands after a disaster. Health service organizations can identify the spread of disease and ways to contain it.

1.1.1. What Differentiates Spatial Big Data

Spatial big data are differentiated from standard (nonspatial) big data by the presence of spatial relationships, geostatistical correlations, and spatial semantic relations (this can be generalized to include the temporal domain (Hägerstrand, 1970). Spatial big data offer additional challenges beyond what is encountered with more traditional big data. Spatial big data are characterized by the following (Barwick, 2011):

- *Volume*. The quantity of data. Spatial big data also include global satellite imagery, mobile sensors (smart phones, GPS trackers, and fitness monitors), and georeferenced digital camera imagery.
- *Variety*. Spatial data are composed of 2D or 3D vector or raster imagery. Spatial data are more complex and subsume the types found with conventional big data.
- *Velocity*. Velocity of spatial data is significant given the rapid collection of satellite imagery in addition to mobile sensors.
- *Veracity*. For vector data (points, lines, and polygons), the quality and accuracy vary. Quality is dependent upon whether the points have been GPS determined, determined by unknown origins, or determined manually. Resolution and projection issues can also alter veracity. For geocoded points, there may be errors in the address tables and in the point location algorithms associated with addresses. For raster data, veracity depends on accuracy of recording instruments in satellites or aerial devices, and on timeliness.
- *Value*. For real-time spatial big data, decisions can be enhanced through visualization of dynamic change in such spatial phenomena as climate, traffic, social-media-based attitudes, and massive inventory locations. Exploration of data trends can include spatial proximities and relationships.

Once spatial big data are structured, formal spatial analytics can be applied, such as spatial autocorrelation, overlays, buffering, spatial cluster techniques, and location quotients.

1.2. Definitions

The terms in Table 1.1 are referenced in this chapter and are included here to facilitate a more rapid understanding of the general concepts discussed later.

Table 1.1 Terms for understanding general concepts

Amazon Web Services	(AWS) A secure, on-demand, cloud computing platform where users pay for the computing resources that they consume (e.g., computing, database storage, and content delivery).
Artificial Intelligence	Computer systems or machines that are able to perform tasks and mimic behavior that normally requires human intelligence, such as visual perception, speech recognition, and language translation.
Big Data as a Service (BDaaS)	Cloud-based hardware and software services that support the analysis of large or complex data sets. These services can provide data, analytical tools, event-driven processing, visualization, and management capabilities.
Cloudera	A software company that provides a software platform that can run either in the cloud or on-prem, supporting data warehousing, machine learning, and big data analytics. The company is a major contributor to the Apache Hadoop platform (e.g., Avro, HBase, Hive, and Spark).
Computer Vision	A scientific discipline that focuses on the acquisition, extraction, analysis, and understanding of information obtained from either single or multidimensional image or video data.
Data as a Service (DaaS)	Built on top of software as a service, data are provided to users on demand for further processing and analysis. The centralization of the data enables higher quality curated data at a lower cost to the client.
Databricks	A company that provides a cloud-based platform for working with Apache Spark. Databricks traces it origins to the AMPLab project at Berkeley that evolved into an open-source distributed computing framework for working with big data.
Data Mining	The process of discovering and extracting hidden patterns and knowledge found in big data using methods and techniques that are commonly associated with database management, machine learning, and statistics.

Table 1.1 (*continued*)

Deep Learning	A subfield of machine learning that focuses on algorithms and computational architectures that mimic the structure of the brain (commonly termed artificial neural networks). Recent advances in large-scale distributed processing have enabled the development and use of very large neural networks.
Elastic Compute Cloud (EC2)	Infrastructure within Amazon Web Services (AWS) that provides scalable computing capacity; clients can develop, deploy, and run their own applications. EC2 is elastic and allows clients to scale their compute and storage up or down as necessary.
Hadoop	An open-source framework and set of software modules that enable users to solve problems on big data sets using a distributed cluster of hardware resources. This includes distributed data storage and computation using the MapReduce programming model. Apache Hadoop was originally inspired by Google's work in the distributed processing domain.
HDFS	A distributed and scalable file system and data store that is part of Apache Hadoop. HDFS stores big data files across a cluster of machines and supports high reliability by replication of the data across different nodes in the cluster.
Hive	Data warehouse software module in Apache Hadoop that facilitates querying and analyzing big data stored in HDFS in a distributed and replicated manner using a SQL-like language termed HiveQL.
IBM Cloud	A set of cloud computing capabilities and services that provides capabilities including Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).
Infrastructure as a Service (IaaS)	A type of cloud computing infrastructure that virtualizes computing resources, storage, data partitioning, scaling, and networking. Unlike Software as a Service (SaaS) or Platform as a Service (PaaS), IaaS clients must maintain the applications, data, middleware, and operating system.
Machine Learning	A subset of artificial intelligence where software systems can automatically learn and improve without any explicit programming, relying upon statistical methods for pattern detection and inference. Machine learning software creates statistical models using sample data in order to make decisions or predictions.

(Continued)

Table 1.1 (*continued*)

MapReduce	A programming model, originally developed at Google, that is often used when processing big data sets in a distributed manner. MapReduce programs contain a map procedure where data can be sorted and filtered, and a reduce procedure where summary operations are performed. MapReduce systems, such as Apache Hadoop, are responsible for managing communications and data transfer among the collection of distributed processing nodes.
Microsoft Azure	A cloud computing service from Microsoft for creating, deploying, and managing applications using data centers managed by Microsoft. Hundreds of services are available that provide functionality related to compute, data management, messaging, mobile, and storage capabilities.
Natural Language Processing (NLP)	A portion of artificial intelligence that focuses on enabling computers to understand and communicate (including language translation) through human language, both written and spoken.
NoSQL data stores	A non-SQL or non-relational database that provides a mechanism for storage and retrieval of data. NoSQL data stores often trade consistency in favor of availability, speed, horizontal scalability, and partitionability.
Oracle Cloud	A collection of cloud computing services from Oracle providing servers, storage, network, applications, and services using Oracle-managed data centers. The Oracle Clouse provides Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS), and Data as a Service (DaaS).
Pig	An Apache platform to develop programs for analyzing big data sets that run on Apache Hadoop using a high-level language (Pig Latin). Pig can be used to develop functionality that runs as MapReduce, Tez, or Spark jobs.
Platform as a Service (PaaS)	A category of cloud computing service that allows clients to develop, deploy, run, and manage applications without needing to build or maintain the cloud computing infrastructure. Unlike software as a service (SaaS), the client is responsible for maintaining the applications and data.

Table 1.1 (*continued*)

Predictive Analytics	A group of statistical and machine learning algorithms that are used to predict the likelihood of future or other unknown events based upon existing historical data.
Real-time Data Processing	A collection of software and hardware that processes data on-the-fly and is subjected to a constraint where responses must be provided within a short interval of time (e.g., fractions of a second), independent of system or event data load.
Redshift	A column-oriented, fully managed, data warehouse for big data. Redshift is similar to other columnar NoSQL databases as it is intended to scale out with distributed clusters of low-cost hardware.
Simple Storage Service (S3)	An object storage service offered by Amazon Web Services (AWS); it is intended to store any type of data (objects) that can later be used for big data analytic processing.
Software as a Service (SaaS)	A category of cloud computing service that allows clients to license applications, web-based software, on-demand software, and hosted software. The delivery model is on a subscription basis and is centrally hosted. Differing from Platform as a Service (PaaS), SaaS does not require client to manage either data or software.
Spark	An analytic engine and cluster-computing framework, part of Apache Hadoop, that supports applications that run across a distributed cluster. Originally developed at Berkeley in 2009, it provides a framework for programming clusters of machines with data parallelism.
Speech Recognition	A collection of methodologies and techniques that enables the recognition and transformation of spoken language into text for further computational processing.
Storm	A real-time, distributed, high-volume, stream-processing framework for big data. It is part of the Apache Hadoop open-source framework.
Stream Processing	A computer programming paradigm (similar to dataflow programming), where given a sequence of data (a stream), a series of pipelined operations (or kernel functions) is applied to each element in the stream.

1.3. Example Problems

There are a significant number of industries and application domains that benefit from spatiotemporal big data analytics (Hey et al., 2009). As the sheer number of processes and technologies that are collecting spatial data grows, the ubiquity and significance of the data have grown. Spatial big data analytics has wide applicability and value across numerous domains; a few of these are the following.

1.3.1. Agriculture

Farmers can use spatial big data analytics to detect and analyze patterns in weather data, correlated with historical crop yields, surface topography, and soil characteristics. This helps farmers determine the best seed varieties to use and times and places to plant crops in order to maximize yields. In addition, the distribution of fertilizer can be optimized based upon historical information. Tractor and heavy equipment movement can also be tracked via GPS and incorporated into the logistic optimization analytics, and the areas of usable and productive land within a field can be identified.

1.3.2. Commerce

Commercial retailers have always used local shopping patterns and demographics to drive marketing strategies and site selection. However, retailers can now use spatial big data analytics to analyze the locations and characteristics of customers along with social media conversations and browsing behavior in order to better understand customers' needs. Retailers can essentially build a richer and more useful understanding and relationship with their customer base. New store site selection on regional or national levels can be optimized based on the locations of customers, competitors, and other nontraditional data.

1.3.3. Connected Cars

Developers of systems for connected cars and autonomous vehicles can use spatial big data analytics to provide accurate situational awareness to drivers and vehicles about their surrounding environment.

Systems can apply analytics capabilities such as road snapping, predictive road snapping, change detection of objects sensed by the vehicle but not on the map, and accident prediction. This is all under the topic of improved vehicle reliability and passenger safety.

1.3.4. Environment

Environmental organizations can employ spatial big data analytics to answer a number of important questions including whether there are spatiotemporal correlations between species observations (this can be by geographic area or species).

1.3.5. Financial Services

In the financial services/insurance industry, spatial big data analytics are used to overlay weather data with claim data to assist companies in detecting possible instances of fraud. In other contexts, non-traditional data sources like satellite imagery are combined with traditional topographic data sources to identify the potential risk of offering flood insurance. Insurers can also assess spatial relationships between their insurance portfolios and past hazards to balance risk exposure. Finally, banks can use spatiotemporal historical transaction data to help them detect evidence of fraud.

1.3.6. Government Agencies

National and regional government agencies would like to use spatial big data analytics to process and overlay nationwide data sets containing land use; parcels; planning information; geological informational, and environmental data in order to create information products that can be used by analysts, scientists, and policy makers to make better policy decisions.

1.3.7. Health Care

Public health agencies can use spatial big data analytics to see how far patients are from health facilities helping them evaluate access to care. Hospital networks can determine the density of hospitals in certain areas

to identify gaps and opportunities. They can also measure the prevalence of certain habits and illnesses in the community using demographic data. Public health agencies can also utilize tracking data to perform contact tracing of infected individuals to identify who they have been in contact with in the past. The contact information can then be utilized to help reduce the infections in the general population. Proximity tracing is a variant in which contact is specified using a proximity-based filtering criteria (e.g., spatial and temporal range) in order to identify potential contact events.

1.3.8. Marketing

Geospatial big data analytics is frequently used in corporate marketing for prospect and customer segmentation. Data from body sensors (e.g., smart phones, smart watches, fitness monitors) can be used to segment the customer base according to physical activity or behavioral patterns and deliver advertising in a targeted manner. Companies also want to be able to identify where their customers are in relation to their competitors' customers. This allows them to identify areas where they are losing the market and help determine where they need to focus their marketing efforts.

1.3.9. Mining

Mining companies can apply spatial big data analytics to perform complex vehicle tracking analysis to find ways to better manage equipment moves. For example, they can analyze patterns of equipment locations when braking, and they can review shock absorption, RPM changes, and other telematics information. They can also analyze geochemical sample results.

1.3.10. Petroleum

Spatial big data analytics enable petroleum companies to identify suitable areas for exploration based upon historical production, geographic composition, and competitor activity (including leasing activity). Spatial big data analytics can also be used to review historical production data to assess reservoir production over time. Vehicle tracking data can be analyzed to determine time spent on both commercial and noncommercial roads. They can also review vessel tracks over offshore blocks using AIS vessel tracking information.