Big Geospatial Data — an OGC White Paper

Additional context, inspirational quote, etc. fits into this subheading area

Note
Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.
Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

Warning
This document is not an OGC Standard. This document is an OGC White Paper and is therefore not an official position of the OGC membership. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an OGC Standard.
Further, an OGC White Paper should not be referenced as required or mandatory technology in procurements.

OGC®
Making location count.
www.opengeospatial.org

License Agreement

Permission is hereby granted by the Open Geospatial Consortium, ("Licensor"), free of charge and subject to the terms set forth below, to any person obtaining a copy of this Intellectual Property and any associated documentation, to deal in the Intellectual Property without restriction (except as set forth below), including without limitation the rights to implement, use, copy, modify, merge, publish, distribute, and/or sublicense copies of the Intellectual Property, and to permit persons to whom the Intellectual Property is furnished to do so, provided that all copyright notices on the intellectual property are retained intact and that each person to whom the Intellectual Property is furnished agrees to the terms of this Agreement.

If you modify the Intellectual Property, all copies of the modified Intellectual Property must include, in addition to the above copyright notice, a notice that the Intellectual Property includes modifications that have not been approved or adopted by LICENSOR.

THIS LICENSE IS A COPYRIGHT LICENSE ONLY, AND DOES NOT CONVEY ANY RIGHTS UNDER ANY PATENTS THAT MAY BE IN FORCE ANYWHERE IN THE WORLD. THE INTELLECTUAL PROPERTY IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. THE COPYRIGHT HOLDER OR HOLDERS INCLUDED IN THIS NOTICE DO NOT WARRANT THAT THE FUNCTIONS CONTAINED IN THE INTELLECTUAL PROPERTY WILL MEET YOUR REQUIREMENTS OR THAT THE OPERATION OF THE INTELLECTUAL PROPERTY WILL BE UNINTERRUPTED OR ERROR FREE. ANY USE OF THE INTELLECTUAL PROPERTY SHALL BE MADE ENTIRELY AT THE USER'S OWN RISK. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR ANY CONTRIBUTOR OF INTELLECTUAL PROPERTY RIGHTS TO THE INTELLECTUAL PROPERTY BE LIABLE FOR ANY CLAIM, OR ANY DIRECT, SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM ANY ALLEGED INFRINGEMENT OR ANY LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR UNDER ANY OTHER LEGAL THEORY, ARISING OUT OF OR IN CONNECTION WITH THE IMPLEMENTATION, USE, COMMERCIALIZATION OR PERFORMANCE OF THIS INTELLECTUAL PROPERTY.

This license is effective until terminated. You may terminate it at any time by destroying the Intellectual Property together with all copies in any form. The license will also terminate if you fail to comply with any term or condition of this Agreement. Except as provided in the following sentence, no such termination of this license shall require the termination of any third party end-user sublicense to the Intellectual Property which is in force as of the date of notice of such termination. In addition, should the Intellectual Property, or the operation of the Intellectual Property, infringe, or in LICENSOR's sole opinion be likely to infringe, any patent, copyright, trademark or other right of a third party, you agree that LICENSOR, in its sole discretion, may terminate this license without any compensation or liability to you, your licensees or any other party. You agree upon termination of any kind to destroy or cause to be destroyed the Intellectual Property together with all copies in any form, whether held by you or by any third party.

Except as contained in this notice, the name of LICENSOR or of any other holder of a copyright in all or part of the Intellectual Property shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Intellectual Property without prior written authorization of LICENSOR or such copyright holder. LICENSOR is and shall at all times be the sole entity that may authorize you or any third party to use certification marks, trademarks or other special designations to indicate compliance with any LICENSOR standards or specifications. This Agreement is governed by the laws of the Commonwealth of Massachusetts. The application to this Agreement of the United Nations Convention on Contracts for the International Sale of Goods is hereby expressly excluded. In the event any provision of this Agreement shall be deemed unenforceable, void or invalid, such provision shall be modified so as to make it valid and enforceable, and as so modified the entire Agreement shall remain in full force and effect. No decision, action or inaction by LICENSOR shall be construed to be a waiver of any rights or remedies available to it.

None of the Intellectual Property or underlying information or technology may be downloaded or otherwise exported or reexported in violation of U.S. export laws and regulations. In addition, you are responsible for complying with any local laws in your jurisdiction which may impact your right to import, export or use the Intellectual Property, and you represent that you have complied with any regulations or registration procedures required by applicable law to make this license enforceable.

Table of Contents

I. Abstract....6
IV. Submitters....6
II. Keywords....6
III. Security Considerations....6
1. The Big Data Trend and Geospatial Information....1
2. The Value of Big Geo Data Applications....2
2.1. Earth Observations....2
2.2. Resource Management: Precision Agriculture....3
2.3. Mobile Location Services....5
2.4. Transportation and Moving Objects....6
2.5. Smart Cities....8
3. Use cases for Big Geo Data....9
3.1. Use Case commonality across applications....9
3.2. Collection & Ingest....11
3.3. Prepare and Structure....12
3.4. Analytics & Visualization....15
3.5. Modeling and Prediction....17
3.6. Use Cases applied to Agriculture....18
4. OGC Big Geo Data Opportunities....19
4.1. Objectives for open implementations of Big Geo Data....19
4.2. Cloud computing for EO data....19
4.3. Analysis Ready Data and Datacubes....21
4.4. Data Representation for Big Geo Data: Features, Coverages, DGGS....23
4.4.1. Simple Features applied to Big Data....23
4.4.2. Coverages applied to Big Data....23
4.4.3. Computing with Discrete Grids....25
4.5. Big Linked Geodata....26
4.6. Using Big Data Open Source....29
5. OGC activities on Big Geo Data....32
5.1. OGC Program Activities on Big Geo Data....32

5.2. External coordination: Standards..... 33
5.3. External coordination: R&D.....34
6. Acronyms..... 35
Annex A (informative) **Location Powers Emergent Themes**.....38

Table of Figures

Figure 1 — Population Distribution and Dynamics Modeling (Figure Source: J. Sanyal, ORNL) 3
Figure 2 — Example Application: Precision Farming (Figure Source: Manolis Koubarakis) 4
Figure 3 — Marketing Automation based on Location in Big Data (Figure source: Jon Spinney, Location Intelligence, Pitney Bowes) 6
Figure 4 — The coming flood of data in autonomous vehicles 7
Figure 5 — Collaboration of Big Data platforms to predict spatio-temporal features (Figure source: Akinori Asahara, Hitachi) 7
Figure 6 — Big Data in Smart Cities (Figure source: ESPRESSO Project) 9
Figure 7 — Big Geo Data Use Cases 10
Table 1 — Use Cases supporting Application Domains 10
Figure 8 — Use Case on Real Time streaming analytics (Figure Source: NIST Big Data Working Group) 11
Figure 9 — Six Star Model for Linked Data (Figure Source: L. van den Brink) 15
Figure 10 — Big Data Use Cases in Agriculture R&D Trials (Figure Source: K. Matson) 19
Figure 11 — OGC Testbed 13 Cloud Environment Overview 20
Figure 12 — DigitalGlobe Analysis Paradigm: Tile based (Figure Source: Dan Getman) 21
Figure 13 — USGS Analysis Ready Data structure (Figure Source: G. Guempel) 22
Figure 14 — Achieving the Big Geo vision with 4D Coverages (Figure Source: P. Baumann) 25
Figure 15 — Examples of DGGS mapping faces of Platonic solids to surface of the Earth. a) Rectilinear cells on rHealPIX projected hexahedron (rHealPIX DGGS see ref [41]); b) Hexagonal cells on ISEA projected icosahedron (ISEA3H DGGS — courtesy of PYXIS Inc.); c) Triangular cells on a Quaternary Triangular Mesh of an octahedron (QTM — courtesy of Geffrey Dutton). (Source: OGC Abstract Specification Topic 21: Discrete Global Grid Systems,OGC Document 15-045r5) 26
Figure 16 — Life Cycle of Linked Open EO Data (Figure Source: M. Koubarakis) 28
Figure 17 — Creating linked data in real time (Figure Source: G. Kepeklian) 28
Figure 18 — Pitney Bowes Big Data Spatial Components (Figure Source: Rose Winterton) 30

Figure 19 — Example Geospatially Enabled Apache Projects (Figure Source: Rob Emanuele) 31
Figure 20 — Distinctive Software/Hardware Architectures for Data Analytics (Figure Source: G. Fox) 32

I. Abstract

This white paper is a survey of Big Geospatial Data with these main themes:

— Geospatial data is increasing in volume and variety;
— New Big Data computing techniques are being applied to geospatial data;
— Geospatial Big Data techniques benefit many applications; and
— Open standards are needed for interoperability, efficiency, innovation and cost effectiveness.

The main purpose of this White Paper is to identify activities to be undertaken in OGC Programs that advance the Big Data capabilities as applied to geospatial information.

This white paper was developed based on two Location Powers events:

— Location Powers: Big Data, Orlando, September 20th, 2016; and
— Location Powers: Big Linked Data, Delft, March 22nd, 2017.

For information on Location Powers: http://www.locationpowers.net/pastevents/

IV. Submitters

Name	Affiliation
George Percivall, editor	OGC
Carl Reed	Carl Reed and Associates
Ingo Simonis	OGC
Josh Lieberman	Tumbling Walls
Steven Ramage	Group on Earth Observations

II. Keywords

The following are keywords to be used by search engines and document catalogues.

ogcdoc, OGC documents, Big Data, geospatial, location, open, standards, interoperability, cloud computing

III. Security Considerations

No security considerations have been made for this standard.

1. The Big Data Trend and Geospatial Information

Every second day the human race generates as much data as was generated from the dawn of humanity through the year 2003[1]. Big Data is both a challenge and an opportunity. Big Data is "extensive datasets — primarily in the characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis."[2]

Geospatial data has been Big Data for decades. New tools and technologies are now available to deal with Big Geo Data analytics and visualization. Geospatial information is advancing in all the dimensions of Big Data.

- Volume: The European Space Agency's Copernicus Missions archive is an ~8 PB archive and growing[3]. DigitalGlobe currently archives 70 PB of satellite imagery[4]. ECMWF currently has 180PB of weather data with plans to be archiving 1 PB/day.
- Variety: NASA distributed more than 3,500 distinct data products in 2015.[5] Geospatial attributes are being connected to data with an increasing diversity of structures and vocabularies.
- Velocity: For urban monitoring in Tokyo, the locations of one million people collected every minute adds up to 1.4 billion records per day[6]
- Veracity: Advances in Big Data processing based on machine learning and deep learning provide great predictive power. Understanding the algorithms and quantifying result uncertainties remains the subject of intense research. This white paper addresses Big Geo Data in the following sections.

Clause 2. **Value of Big Geo Data**	Applications of geospatial using Big Data techniques are described to show the value of these new capabilities.
Clause 3. **Use Cases for Big Geo Data**	Use cases are presented to demonstrate commonality across applications domains. This commonality allows best practices be defined through common standards and workflows. This helps manage the complexity in applying big data technology based on investments in

[1] http://www.pbs.org/show/human-face-big-data/

[2] ISO/IEC CD2 20546 Information Technology — Big Data — Overview and Vocabulary

[3] Cristiano Lopes, ESA, Location Powers Big Linked Geodata, March 2017

[4] Dan Getman, DigitalGlobe, Location Powers Big Data, September 2016

[5] https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20160008918.pdf

[6] Akinori Asahara, Hitachi, Location Powers Big Data, September 2016

Clause 4. **OGC Big Geo Data Opportunities**	Several high priority focus areas for advancing big geo data implementations based on open standards are presented as opportunities for OGC activities.
Clause 5. **OGC Activities on Big Geo Data**	Existing and potential new activities are listed for consideration to be undertaken in OGC Programs and in coordination with external alliances.

2. The Value of Big Geo Data Applications

2.1. Earth Observations

Observations of the Earth support global efforts to understand our shared physical environment. The environmental monitoring and modeling community generates Big Data to better understand Earth systems. High volumes (petabytes), at increasing velocity (distributed worldwide using high performance computing facilities) and variety (of data formats and resolutions) need to be handled and smoothly integrated to deal meet modern challenges such as global food security, effects and mitigation of climate change, or global logistics and infrastructures.

In a keynote presentation to the Location Powers: Big Geo Data workshop, Jibo Sanyal (ORNL) illustrated the value of Earth Observations as Big Earth data for estimating population (Figure 1). High-resolution population distribution data are critical for successfully addressing important issues ranging from socio-environmental research to public health to homeland security. Sanyal's keynote addressed how such data are of paramount importance for responding to policy topics, such as the UN 2030 Agenda and the sustainable development goals.

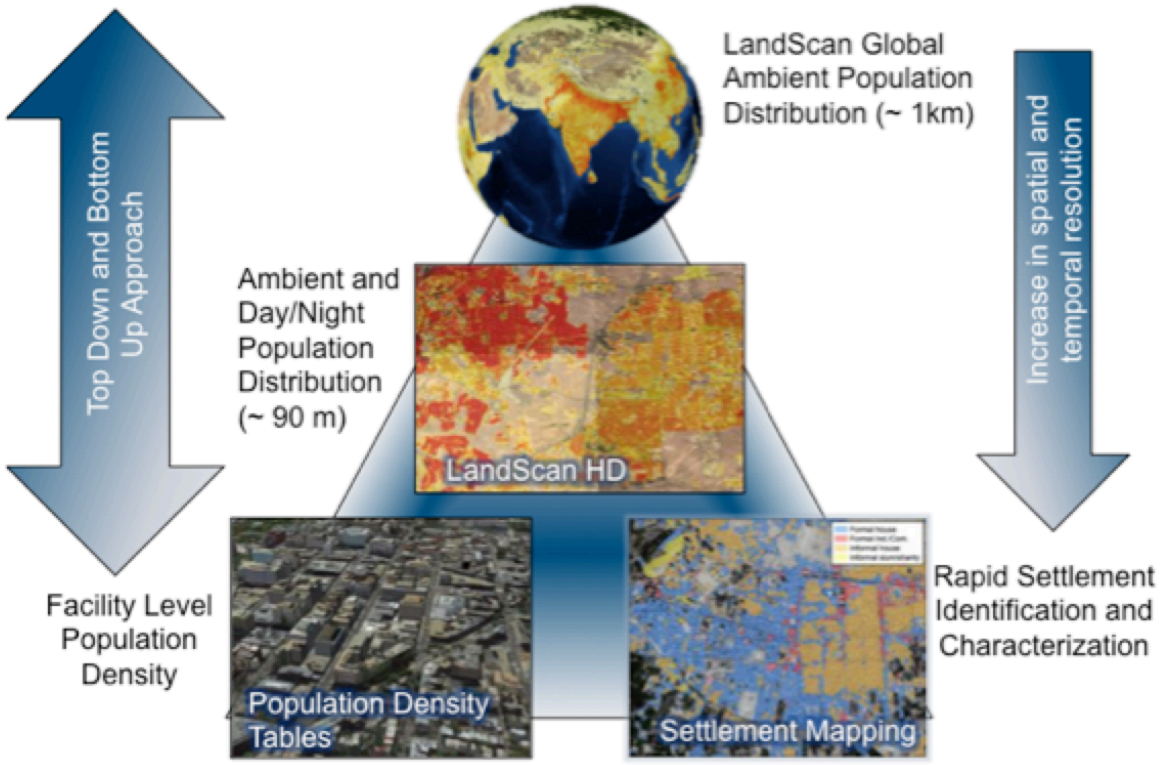

Figure 1 — Population Distribution and Dynamics Modeling (Figure Source: J. Sanyal, ORNL)

Satellite-based Earth Observations were an early driver to the Big Data explosion. Current emphasis on Big Data can be seen in the recent Big Data from Space 2106 conference in Europe and in the Big Earth Data Initiative (BEDI) in the United States. Ground-based Earth Observations, such as in-stream flow monitoring and air particulates monitoring, have traditionally been lower volume outputs than space based sensors but this situation is changing. NOAA's Big Data Project is engaged with several cloud providers with one of the most innovative being the hosting of data from ground based NEXRAD high-resolution Doppler radar. Non-traditional ground based sensors coming from IoT and Smart City applications will also drive new applications of ground based Earth Observations.

2.2. Resource Management: Precision Agriculture

Due to growing world population and changing climatic norms, the sustainment of crop quality as well as quantity from existing agricultural land is an important challenge today in reducing global food insecurity. The goal of precision agriculture research is to define a decision support system (DSS) for whole farm management with the goal of optimizing returns on inputs while preserving resources[7]. Precision can only improve decision-making and farm management if agriculture farmers have access to the necessary small scale, detailed information to make informed choices. The creation of field or even plant-level information can support farmers to improve their crop production and attract long-term investment. The more comprehensive and up-to-date picture that farmers have about their crops (e.g., through remote sensing and GPS technologies), the better decisions they can make as to where and when to apply seed, how much to fertilize, when to irrigate and so forth. Longer-term records of agricultural processes from precision farming data allow farmers to use their cropland more

[7]McBratney, A., Whelan, B., Ancev, T., 2005. Future Directions of Precision Agriculture. Precision Agriculture, 6, 7-23

efficiently, increase crop size and quality, and respond more effectively to climatic challenges such as drought. Precision farming has the potential to make a worthwhile difference in farmers' income, crop yields, and resilience while mitigating negative environmental impacts farming.[8]

Activities such as GEO GLAM provide regular Earth observations to feed into crop monitoring for early warning and production systems. www.cropmonitor.org The Group on Earth Observations (GEO) Global Agricultural Monitoring (GLAM) flagship follows the GEO data sharing and data management principles. Realizing this potential depends greatly on the cost and difficulty for the farmer of collecting and working with big geospatial standards.

The LEO Horizon2020 (H2020) Project developed software tools that support the whole lifecycle of reuse of EO data and related linked geospatial data. To demonstrate the benefits of linked open EO data and its combination with linked geospatial data to the European economy, a precision farming application was developed (Figure 2).

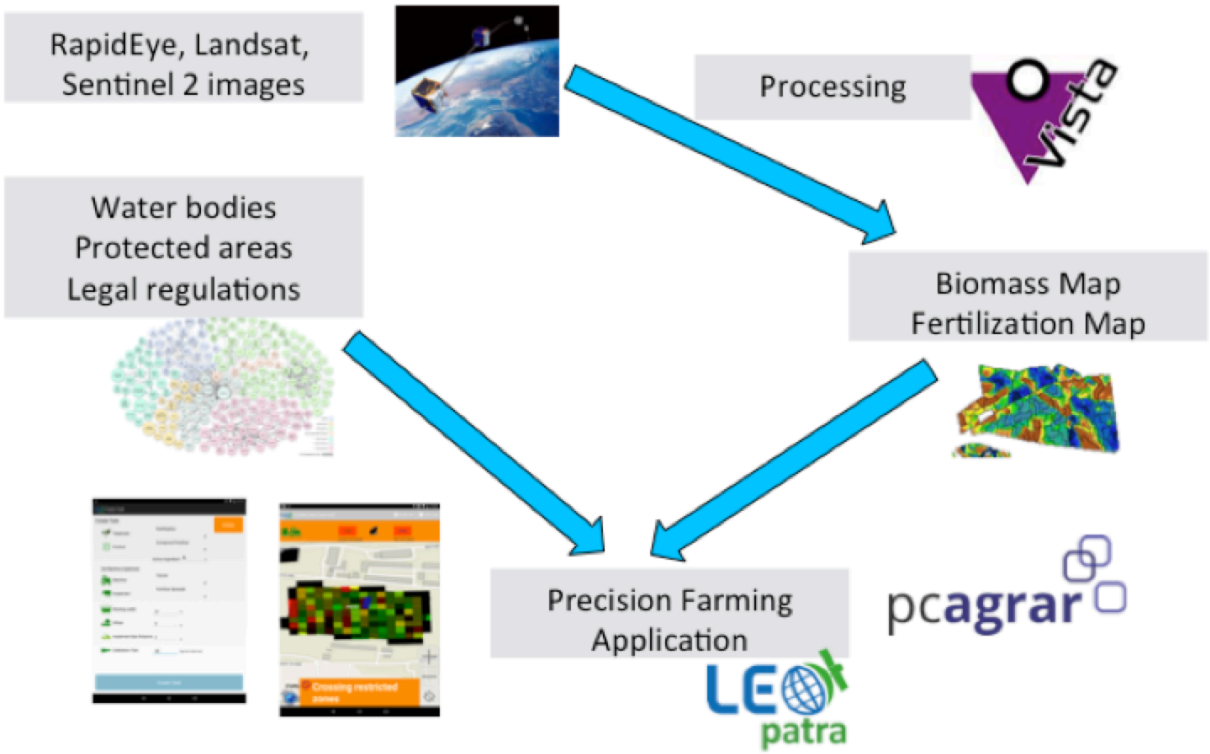

Figure 2 — Example Application: Precision Farming (Figure Source: Manolis Koubarakis)

Another example of Big Data for Agriculture includes satellite image processing to calculate the available area of arable land. Fritz et al. have shown that land suitable for cultivating biofuel crops has been vastly overestimated. They have reduced the estimate by almost 80 percent and expressed a growing concern about how production of biofuels will impact food security. Based on Big Data

[8] http://www.linkedeodata.eu/Precision_Farming

analytics, Fritz's et al. study showed that previous studies had overestimated the amount of arable land and had underestimated the amount of land already being cultivated[9].

Other initiatives are currently exploring the reusability of Big Data concepts, technologies, and architectures across domains to leverage synergy effects. The OGC participates in the research and development project DATABIO, co-funded by the European Commission. DATABIO focuses on the data intensive target sector Data-Driven Bioeconomy. More specifically, DATABIO explores the potential of Big Data integration and analytics in the domains agriculture, forestry, and fishery/aquaculture including taking into account interoperability and sustainability aspects in the heterogeneous European bioeconomy landscape.

DATABIO proposes to deploy a state of the art big data platform on top of existing partners' infrastructure and solutions, the Big DATABIO Platform. DATABIO features continuous cooperation of experts from end user and technology provider companies, from bioeconomy and technology research institutes, standardization organizations such as OGC, and of other partners, mainly of the public administration sector. A series of pilots allows associated partners and other stakeholders to get actively involved in the project.

2.3. Mobile Location Services

Location-enabled mobile devices are a major source of Big Data. Location data coming from the mobile devices and their associated networks enables many Big Data applications. The Ways Big Geospatial Data Is Driving Analytics In the Real World begins with this observation:

"Amid the flood of data we collect and contend with on a daily basis, geospatial data occupies a unique place. Thanks to the networks of GPS satellites and cell towers and the emerging Internet of Things, we're able to track and correlate the location of people and objects in very precise ways that were not possible until recently".

Recent studies of mobile devices identified the predictability of human mobility. A study reported in Science found that "by measuring entropy of individual's trajectory, we find 93% potential predictability in user mobility" as determined based on a study of ~10 million anonymous mobile phone users. Cardiff University Researchers have shown the effectiveness of detecting real-world events using Twitter based on location detection and disambiguation[10]. The power of location data was highlighted by Sir Martin Sorrell, CEO WPP, during his speech at Mobile World Congress in his comment that "Location targeting is holy grail for marketers."

Location based contextual awareness is relevant to location based marketing, first responders, urban planners and many other applications. Creating useful local context requires Big Data analytics platforms. Big data processing and high velocity streaming of location-based data creates the richest contextual awareness. Data from many sources including IoT devices, sensor webs, social media and crowd-sourcing are combined with semantically rich urban and indoor spatial data. The resulting context information is delivered to and shared by mobile devices in connected and disconnected operations. Open standards play a key role in establishing context platforms and marketplaces. Successful approaches will consolidate data from ubiquitous sensing technologies on to enabled

[9] http://pubs.acs.org/doi/abs/10.1021/es103338e

[10] http://dl.acm.org/citation.cfm?doid=3068849.2996183

context-aware analysis of environmental and social dynamics. For example Pitney Bowes is applying big data developments of automating marketing based on location (Figure 3).

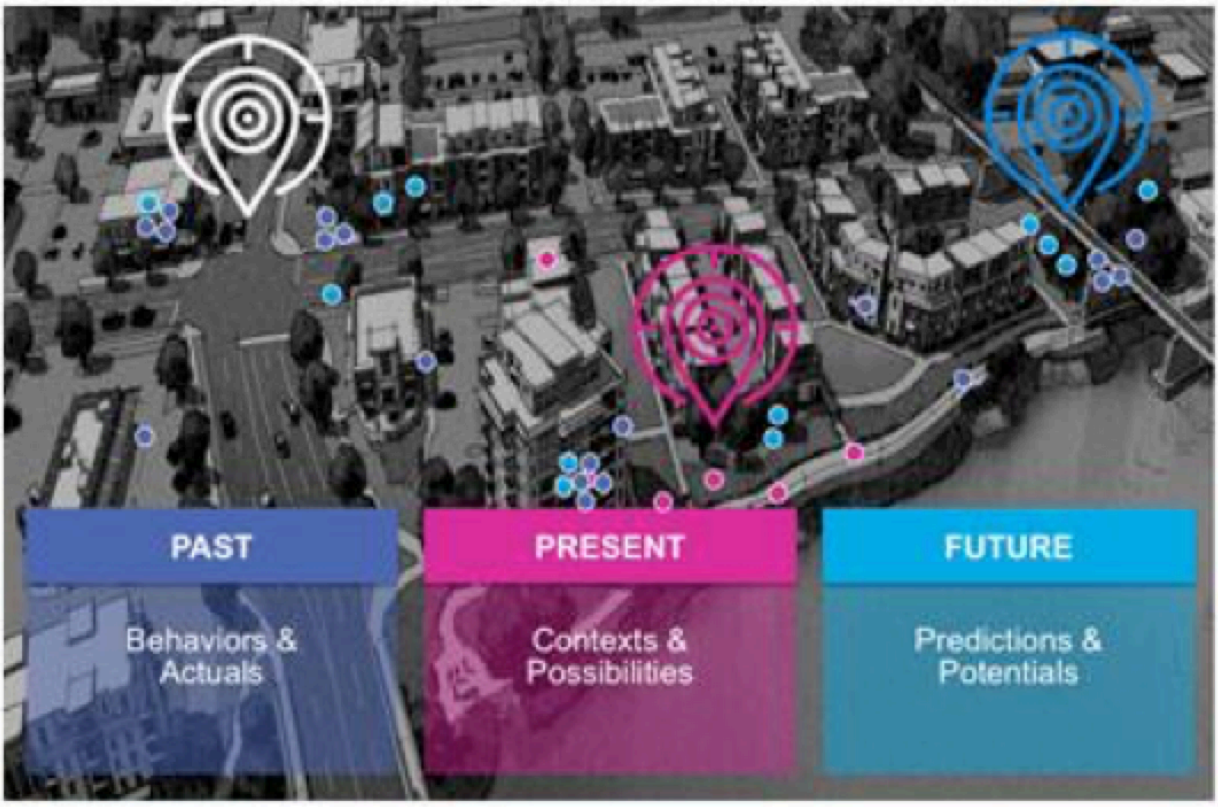

Figure 3 — Marketing Automation based on Location in Big Data (Figure source: Jon Spinney, Location Intelligence, Pitney Bowes)

2.4. Transportation and Moving Objects

Management and optimization of transportation systems benefits from the Big Data platforms to monitor, visualize and perform predictive analytics of objects moving in space and time. Traffic congestion is reduced as trip demand data collected using transportation surveys is integrated with real time or projected traffic data. Combined, optimal schedules and routes can be calculated. Thanks to the availability of real time reports from location enable devices, these schedules and routes can even be optimized at runtime. Automobiles will continue to increase as generators of location based big data. Intel has predicted that autonomous vehicles will generate 4 TB of observation and measurement data per day (Figure 4).

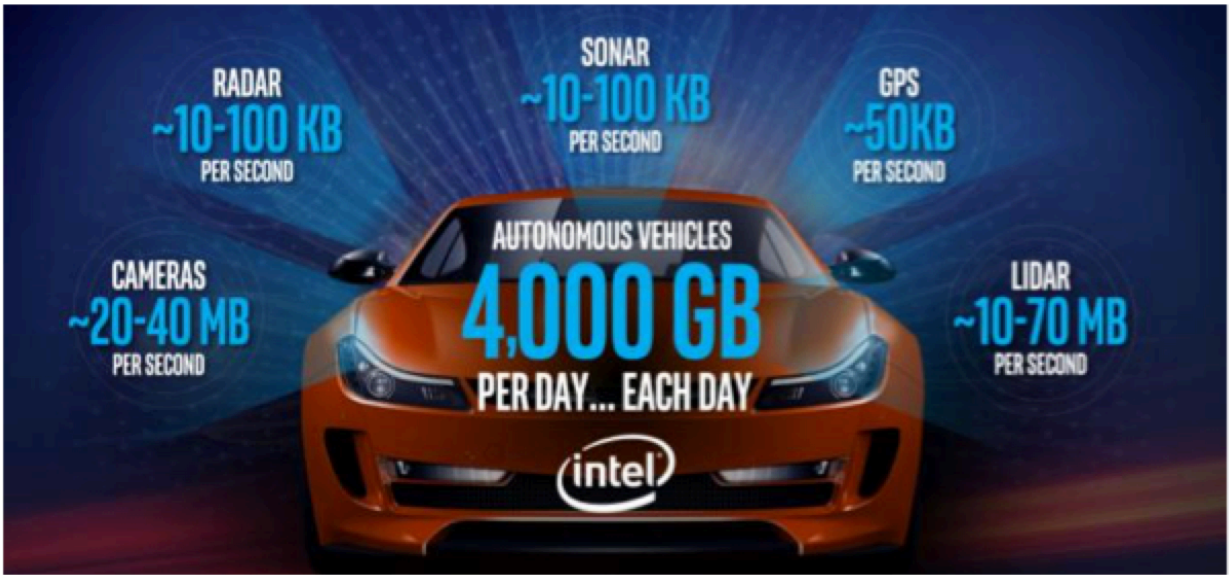

Figure 4 — The coming flood of data in autonomous vehicles[11]

The more real-time information is made available, the better the optimization algorithms work defining requirements on Big Data handling and processing for, standards can leverage these capabilities. OGC Moving Features standard[12] allows seamless integration of mobile objects and predictions based on mobile objects across systems.

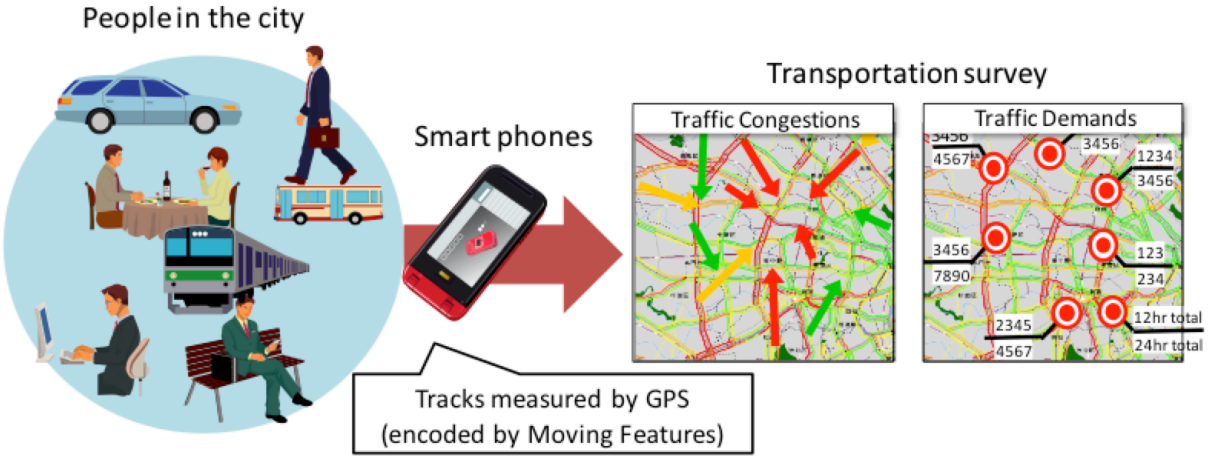

Figure 5 — Collaboration of Big Data platforms to predict spatio-temporal features (Figure source: Akinori Asahara, Hitachi)

[11] http://www.networkworld.com/article/3147892/internet/one-autonomous-car-will-use-4000-gb-of-dataday.html

[12] http://www.opengeospatial.org/standards/movingfeatures

2.5. Smart Cities

Imagine knowing practically every detail about a city: The state of the infrastructure, inhabitants, and the environment are all known to you, at high resolutions in time and in space. You are able to fuse physical data streams with socio-economic data. Transport data tells you where people are going. Sales and transaction data tells you what they are going to see or do or buy. Social media tells you how groups feel about events. And of course high-quality weather data is already built into your system. Suppose that practically every movement and action within the city's systems and infrastructure created location enabled data that could be used to enhance livability, provision of services, and more. Think of the data streams that would exist or could be created; the rates at which those data streams would flow; the technology and skills that would be necessary to acquire, store, and analyze such massive data. Think also of the theories and models that social scientists could generate and test, the problems that system operators and policy-makers could solve if they had access to those models and applications; and of the speed at which those problems could be addressed. Therein is the potential of big data in Smart Cities.[13]

A Smart City provides effective integration of physical, digital and human systems in the built environment to deliver a sustainable, prosperous and inclusive future for its citizens[14]. International Standards organizations are working to advance open standards to meet the needs of the widespread deployment of information technology to cities. Of particular note is the recently initiated ISO/IEC JTC 1/WG 11 for Smart Cities. The OGC's contributions to the JTC 1/WG 11 are based on the OGC Smart Cities Spatial Information Framework white paper.

[13]The entire paragraph is an edited version from: Ann Keller, S., Koonin, S. E. and Shipp, S. (2012), Big data and city living — what can it do for us? Significance, 9: 4—7. doi:10.1111/j.1740-9713.2012.00583.x

[14]Smart city definition from BSI PAS 180

Figure 6 — Big Data in Smart Cities (Figure source: ESPRESSO Project)

OGC explores Smart City aspects as part of its Smart Cities Domain Working Group, and as project coordinator of the ESPRESSO project, a development project co-funded by the European Commission. In an effort to leverage the promise of a system approach, ESPRESSO focuses on development of a conceptual Smart City Information Framework based on open standards. This framework will consist of a Smart City platform (the “Smart City enterprise application”) and a number of data provision and processing services to integrate relevant data, workflows, and processes. The project will build this framework by identifying relevant open standards, technologies, and information models that are currently in use or in development in various sectors. The project will analyse potential gaps and overlaps among standards developed by the various standardisation organizations and will provide guidelines on how to effectively address those shortcomings.

3. Use cases for Big Geo Data

3.1. Use Case commonality across applications

The benefits of Big Data described in Clause 2 vary across the domain of application and in the types of benefits. Fortunately, from a technology development perspective there is high commonality of use cases and associated requirements across the application domains. This section provides a set of Big Geo Data Use Cases that can be used to help manage the complexity of Big Geo Data development. The Use Cases are organized into four groups as shown in Figure 7. Three of the four use case groups are inspired by and similar to the roles of a big data application provider as defined in a reference

architecture under development by ISO/IEC JTC 1/WG 9 — Information technology — Big Data[15]. The Models and Simulation category shown in Figure 7 is not in the WG9 architecture, but is included in the geospatial use cases as geospatial models and simulations meet the definition of big data.

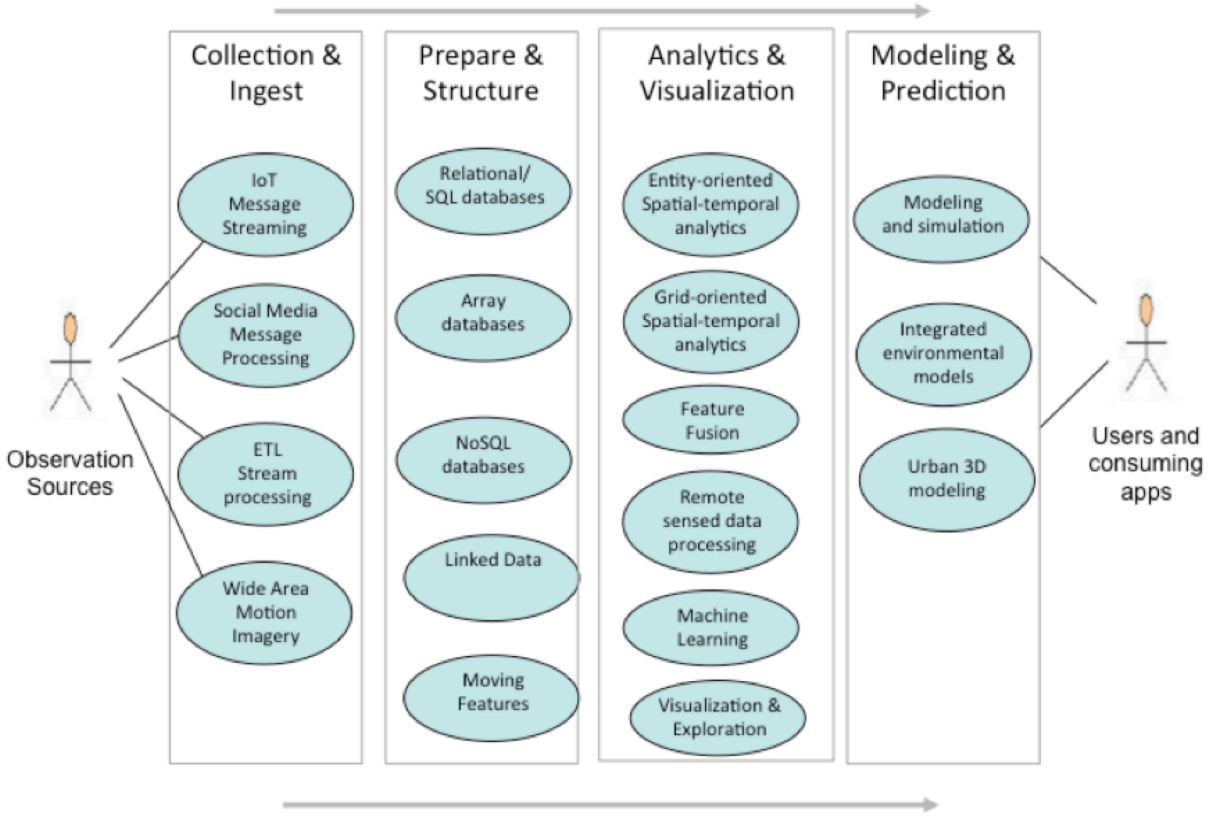

Figure 7 — Big Geo Data Use Cases

Table 1 shows how the use cases support the application domains described in Clause 2.

	Collection and Ingest	Prepare & Structure	Analytics & Visualization	Models & Prediction
Earth Observations	Observations from sensors to processing centers	Met/Ocean 4D queries on array databases	Processing of observations in data clusters and exascale processing	Integrated Environmental processing modeling and predictions
Resource Management: Precision Agriculture	Observations from remote, in-situ, and on-vehicle sensors	Processing, normalization, cross-scale integration of raw observations	Identification of trends, correlations, scale of resource evolution	Prediction on resource quantities in the future, generation of treatment prescriptions

[15] https://www.iso.org/committee/45020.html

	Collection and Ingest	Prepare & Structure	Analytics & Visualization	Models & Prediction
Mobile Location Services	Mobile devices location tracking	Linked data structuring	Sentiment analysis from social media	Prediction of future state
Transportation	Real-time Tracking of moving features	Detect and monitor real-time event detection	Pattern recognition in moving features	Predicting future states for decision making
Smart Cities	Observations of urban environment. Social media from citizens	Queries on quality of life in cities	Prediction of urban needs	Prediction for urban planning using spatial models

Table 1 — Use Cases supporting Application Domains

3.2. Collection & Ingest

Collection and ingest of much Big Data requires the abilities of high velocity data streaming. Geoffrey Fox, Indiana University, provided an excellent overview of high velocity streaming to the Location Powers, Big Geo Data workshop. Professor Fox's presentation was based in part of two previous workshops on streaming systems. Additionally Fox described use cases coming from the US National Institute of Standards and Technology (NIST) Public Data Working Group, e.g., Figure 8

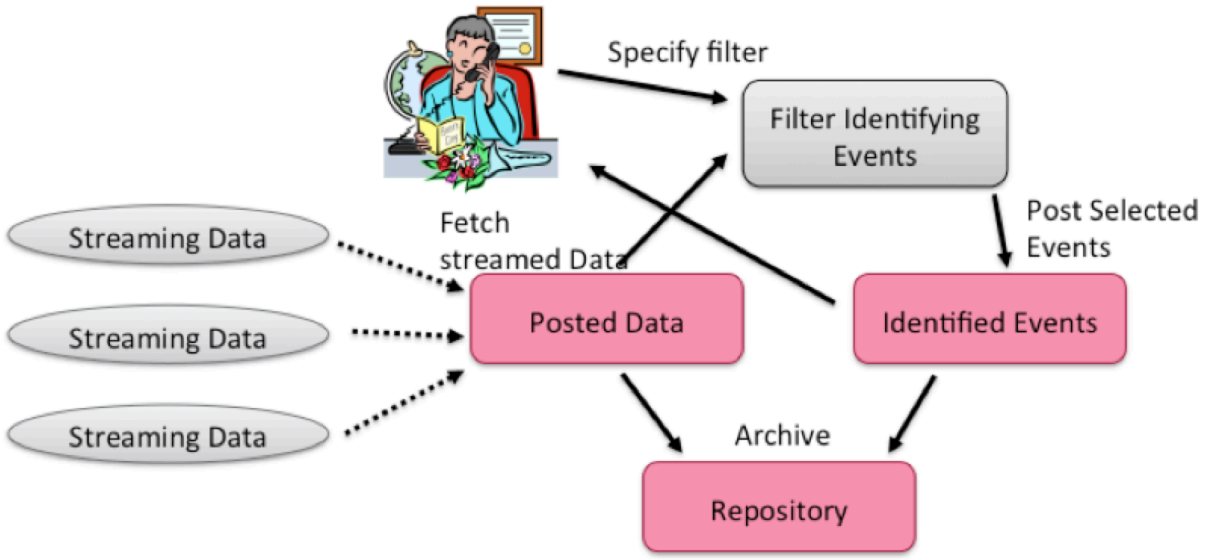

Figure 8 — Use Case on Real Time streaming analytics (Figure Source: NIST Big Data Working Group)

Use Cases in the High Velocity category include the following.

IoT Message Streaming

— Description: Data from IoT devices is made available using high velocity transaction messaging. The objective is rapid and diverse distribution of IoT data with limited semantics sufficient for more downstream processing.

— OGC Standards: OGC Sensor Observation Service (SOS) and SensorThings API
— References:
  — OGC SensorThings API Standard; and
  — OGC Sensor Observation Service.

Social Media Message Processing

— Description: Processing of social media messages with relatively basic processing to make it suitable for downstream processing. Social media is a key source of Big Data for understanding local sentiment and decisions. Examples including geotagging messages and basic cluster identification based on message content from: Twitter, Instagram, Flickr, etc.
— Standards: W3C Activity Streams 2; OGC SOS, GeoSPARQL, WPS
— Reference: OGC Testbed-11 Incorporating Social Media in Emergency Response Engineering Report, OGC Document 15-057

ETL Stream processing

— Description: Extract, Transform, Load (ETL) processing of messages based on structuring the messages with RDF to support inference, e.g., identify events of interest. For example, process sensor messages into filed-named tuples using a semantic ontology relevant to the observations.
— Standards: RDF, OWL, SPARQL, SWE, SSN
— References:
  — Integrating Big Data: A Semantic Extract- Transform-Load Framework
  — Scalability in RDF Stream Processing Systems

Wide Area Motion Imagery

— Description: Motion Imagery is a video stream where each image in the stream is spatio-temporally related to the next image. Motion Imagery contains metadata to provide context, e.g. sensor information, position, time, image quality, etc. Wide Area Motion Imagery (WAMI)[16] has a large footprint, typically tens of square kilometers per image with an image capture rate of 2 to 24Hz
— Standards: WAMI, WebSockets (RFC 6455) and/or UDP for streaming WAMI data, in addition to HTTP(S)
— Reference: OGC Best Practice — WAMI Services: Dissemination Services for Wide Area Motion Imagery

3.3. Prepare and Structure

Data preparation and structuring involves processes that convert raw data into organized information. These processes and resulting structured data stores enhance the value of the data for search, analytics, and visualization. Unstructured data does not have a pre-defined data model or is not organized in a pre-defined way. Data structuring is performed mainly to support analytics (See Clause 3.4).

The activities of preparing and structuring data lead to Data Representation. An excellent survey of data representation for Big Data is provided in "Frontiers in Massive Data Analysis" by the US National Academies. They emphasize the goals of data representation as:

[16] http://www.opengeospatial.org/pressroom/pressreleases/1759

"Although a picture may be worth a thousand words, a good representation of data is priceless: a single data representation, or sometimes multiple ones, allows one to carry out a large number of data processing and analysis tasks in a manner that is both algorithmically efficient and statistically meaningful."

Tasks performed by this activity could include data validation (e.g., checksums/hashes, format checks), cleansing (e.g., eliminating bad records/fields), outlier removal, standardization, reformatting, or encapsulating. This activity is also where source data will frequently be persisted to archive storage and provenance data will be verified or attached/associated. Verification or attachment may include optimization of data through manipulations (e.g., de-duplication) and indexing to optimize the analytics process. This activity may also aggregate data from different Data Providers, leveraging metadata keys to create an expanded and enhanced data set.

Geography is often described as the "glue that binds conceptually linked data". The links between spatial things — and between other resources and spatial things — describe how the world around us is structured and interrelated and form an important facet of the Web of Data[17].

Key to data structuring big data are new data types. During the Location Powers Big Geo Data workshop Keith W. Hare (JCC Consulting) provided a comparison of traditional data types vs. "Big Data" data types. Data structuring using these new data types enables new use cases.

Use Cases in the Geospatial Databases category include the following.

Relational Databases/SQL

- Description: Relational databases have been a dominant data structure and technology for many years. SQL (Structured Query Language) is a language used in programming and designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS).
- Standards: SQL
- References: (SQL for large databases)

Array databases

- Description: Geospatial Coverages when structured as arrays are not well suited to traditional RDBMSs. Query optimization for array data is difficult, and the relational model is based on sets, not ordered data. Several efforts to incorporate array data into the relational model have appeared in the research literature, but without lasting effect.[18].
- Standards: ISO/IEC JTC 1/SC 32 is creating a new part to the SQL standards for arrays: WD 9075-15 Multi-dimensional arrays (SQL/MDA). OGC WCS provides access to arrays with the OGC WCPS standard as an input to SQL/MDA development.
- References: EarthServer is an example implementation of an array database.

No-SQL/Non-Relational databases

[17]Source: https://www.w3.org/TR/sdw-bp/

[18]3 Scaling the Infrastructure for Data Management." National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, DC: The National Academies Press. doi: 10.17226/18374

— Description: Non-relational model database paradigms are sometimes referred to as NoSQL (Not Only or No Structured Query Language [SQL]) systems). Since NoSQL is in such common usage it will continue to refer to the new data models beyond the relational model. However, the term refers to databases that do not follow a relational model. Examples of non-relational database models include the column, sparse table, key-value, key-document, and graphical models[19].
— Standards: NoSQL data management systems, which are intended to provide support for non-tabular structured data, as well as unstructured and semi-structured data, have not yet settled on a common access language.

Moving Features

— Description: Mobile devices are providing increasing Big Data sets of features moving in space and time. Several use cases based on spatial-temporal analysis motivate the access methods for databases storing moving feature data. For example, these operations retrieve positions, trajectories, and velocities of a moving feature such as a car, a person, a vessel, an aircraft, and a hurricane.
— Standards: OGC Moving Features Access Standard.
— References:

Linked Data

— The term 'Linked Data' refers to an approach to publishing data that puts linking at the heart of the notion of data, and uses the linking technologies provided by the Web to enable the weaving of a global distributed database. By identifying real world entities — be they Web resources, physical objects such as the Eiffel Tower, or even more abstract things such as relations or concepts — with URLs data can be published and linked in the same way Web pages can.
— Standards: HTTP, URIs, RDF, SPARQL, JSON LD
— References: https://www.w3.org/TR/sdw-bp/#linked-data

[19]The text for this use case derives from a NIST Big Data Working Group draft document.

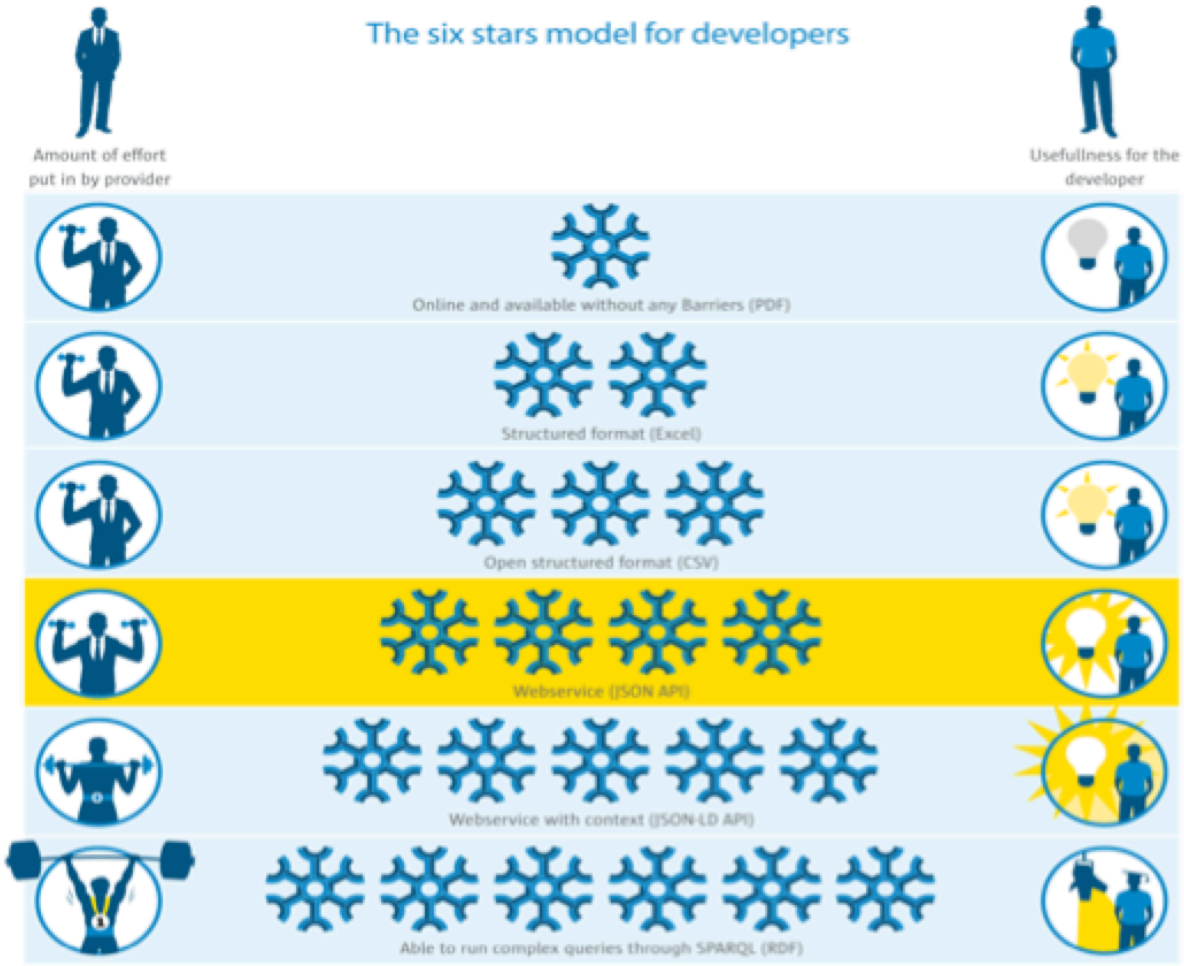

Figure 9 — Six Star Model for Linked Data (Figure Source: L. van den Brink[20])

3.4. Analytics & Visualization

"Big data analytics is the process of examining large and varied data sets — i.e., big data — to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.[21]" The requirements specify the data processing algorithms to produce new insights that will address the technical goal.

The ESIP Federation defines Earth Science Data Analytics as "Process of examining, preparing, reducing, and analyzing large amounts of spatial (multi-dimensional), temporal, or spectral data encompassing a variety of data types to uncover patterns, correlations and other information, to better understand our Earth". ESIP goes on to say that analytics encompasses:

— Data Preparation — Preparing heterogeneous data so that they can be jointly analyzed;

[20] https://www.linkedin.com/pulse/why-apis-missing-link-linked-open-data-dimitri-van-hees

[21] http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

— Data Reduction — Correcting, ordering and simplifying data in support of analytic objectives; and
— Data Analysis — Applying techniques/methods to derive results.

In this white paper, Data Preparation and Data Reduction were discussed in the previous section on Data Structuring. Data Analysis is the synthesis of knowledge from information.

Many of the topics addressed in this set of use cases are considered in Data Science. There are many definitions and activities labeled as Data Science. A good perspective on this topic is "50 Years of Data Science" by David Donoho. "The Fourth Paradigm" by Tony Hey, et.al, provides an excellent perspective on this topic by considering data-intensive scientific discovery.

Use Cases in the Analytics and Visualization category include the following.

Entity-oriented Spatial-temporal analytics

— Description: Fusing structured and unstructured data for geospatial analysis in support of activity-based intelligence (ABI), object-based production (OBP), and human geography (HG) analysis. Spatiotemporal analytics on high velocity streaming data-in-motion and high volume batch data-at-rest. Geospatial examples include GeoWave, GeoTrellis, GeoMesa, GeoJinni; JTS Topology Suite; Esri geometry-api-java
— Standards: OGC Simple Features
— References:
  — Large-scale Analysis of Event Data, Hagedon, Sattler, Gertz
  — Applying Geospatial Analytics Using Apache Spark Running on Apache Mesos — Apache Big Data conference presentation
  — Activity-Based Intelligence: Revolutionizing Military Intelligence Analysis

Grid-oriented Spatial-temporal Analytics

— Description: Data is organized in tile structures with attributes associated with each tile regarding the physical and human geography of the geographic space of the tile. Analysis of moving entities across the tile structure.
— Standards: OGC Discrete Global Grid System (DGGS) Abstract Model

Feature Fusion

— Description: Feature fusion is a type of data fusion where the data elements being associated are features. Data Fusion is the act or process of combining or associating data or information regarding one or more entities considered in an explicit or implicit knowledge framework to improve one's capability (or provide a new capability) for detection, identification, or characterization of that entity. A primary example of Feature Fusion is Conflation
— Standards:
— References:
  — OGC Fusion Standards Study, Phase 2 Engineering Report

Remote-sensed data processing

— Description: Processing of remote sensed data has traditionally used purpose built algorithms depending on the specific sensor. In particular for lower level processing to Level 1 (Data Processing Levels). Level 2 and 3 processing has recently become the subject of generic compute platforms for hosting remote sensed processing algorithms on distributed clusters.

- Standards: NetCDF, HDF; WCS, WPS
- References:
  - Cloud Computing Enabled Web Processing Service for Earth Observation Data Processing; Z. Chen, N. Chen, C. Yang, L. Di — IEEE JSTARS
  - Exploitation Platforms Open Architecture, S. Pinto at BiDS'16 Conference

Machine Learning

- Description: Machine learning is a general class of algorithms that learn from and make predictions on data. Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation. Machine learning can be classified into three broad categories: Supervised learning, Unsupervised learning, Reinforcement learning. Example Implementations: Apache Mahout, Spark/MLlib
- Use cases:
- Standards:
- References:
  - "What Led Computer Vision to Deep Learning?" Jitendra Malik
  - "Architecture and Implementation of a Scalable Sensor Data Storage and Analysis System Using Cloud Computing and Big Data Technologies", Galip Aydin, Ibrahim Riza Hallac, and Betul Karakus

Data Visualization and Exploration

- Description: The visualization activity prepares elements of the processed data and the output of the analytic activity for presentation. The objective of this activity is to format and present data in such a way as to optimally communicate meaning and knowledge. The visualization preparation may involve producing a text-based report or rendering the analytic results as some form of graphic. The visualization activity frequently interacts the analytics activity to provide interactive visualization of the data.

3.5. Modeling and Prediction

The US National Strategic Computing Initiative noted the need to address both Data Analytic Computing along with Modeling and Simulation:

Historically, there has been a separation between data analytic computing and modeling and simulation. Data analytics focuses on inferring new information from what is already known to enable action on that information. Modeling and simulation focuses on insights into the interaction of the parts of a system, and the system as a whole, to advance understanding in science and engineering and inform policy and economic decision-making. While these systems have traditionally relied on different hardware and software stacks, many of the current challenges facing the two disciplines are similar. A coherent platform for modeling, simulation, and data analytics would benefit both disciplines while maximizing returns on R&D investments[22].

Use Cases in the Modeling and Predication category include the following.

Modeling and simulation

[22]National Strategic Computing Initiative Strategic Plan, July 2016. Box 3.

— Description: Modeling and simulation applications support applications in which inter-connected simulators share a common view of the simulated environment.
— Standards: OGC CDB
— Reference: “Cloud Terrain Generation and Visualization Using Open Geospatial Standards”, Chambers and Freeman, 2014. From proceedings of the Interservice/Industry Training, Simulation, and Education Conference.

Integrated environmental models

— Description: Integrated environmental modeling provides a science-based structure to develop and organize relevant knowledge and information and apply it to explain, explore, and predict the behavior of environmental systems in response to human and natural sources of stress.
— Standards: OpenMI
— References: Environmental model access and interoperability: The GEO Model Web initiative

Urban 3D modeling

— Description: A 3D city model is a representation of an urban environment with a three-dimensional geometry of common urban objects and structures, with buildings as the most prominent feature. 3D city models have become valuable for purposes beyond visualization, and are utilized in a large number of domains and applications.
— Standards: CityGML
— References: Applications of 3D City Models: State of the Art Review

3.6. Use Cases applied to Agriculture

This section shows how the Big Data use cases support the agriculture application presented in Clause 2. Figure 10 is an adaptation of an earlier version the use cases shown in Figure 7. Figure 10 shows Big Data uses in an agriculture research and development trials. The use cases are organized in the categories as shown in Figure 7. The names of the use cases reflect the specifics of agriculture. (Figure 10 was presented by Kris Matson, Bayer to the OGC Agriculture DWG, Sept 22nd 2017, Orlando, FL, USA).

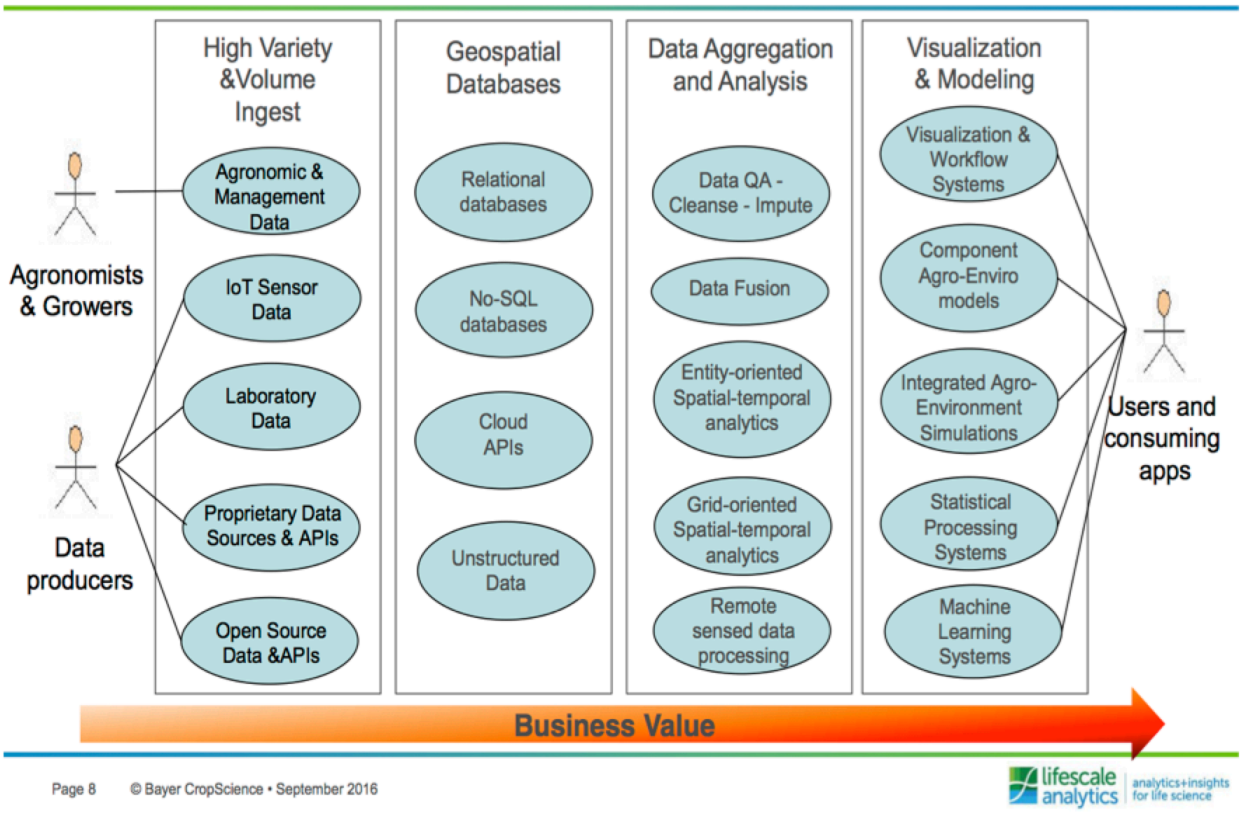

Figure 10 — Big Data Use Cases in Agriculture R&D Trials (Figure Source: K. Matson)

4. OGC Big Geo Data Opportunities

4.1. Objectives for open implementations of Big Geo Data.

OGC standards enable the development of open geospatial processing frameworks. The broader IT community has recognized the need for open standards in Big Data. For example, “Use of standards and related issues in predictive analytics” a presentation by Paco Nathan, O’Reilly Media at the KDD conference, 2016-08-16 identified a Lesson from the success of Apache Spark is “lack of interchange for analytics represents a serious technical debt and potential liability.”

The implementations listed in the remainder of this section are based on the discussions of the two Location Powers events as well as in the OGC Big Data Domain Working Group (DWG).

4.2. Cloud computing for EO data

OGC Testbed 13 has a focus on “Cloud Computing Environment for Earth Observation Data”. Testbed-13 addresses key elements in cloud computing research, such as loosely-coupled PB-sized archives for rapid geospatial information product creation at any scale based on open standards. The topics of interoperability and portability are significant considerations in relation to the use of cloud services. Testbed-13 will help clarifying the specific interoperability and portability concerns that

arise in the large cloud eco-system with its variety of cloud services offered. From the interoperability perspective, in particular two aspects are of primary interest.

a) Cloud API interoperability is a major issue at the moment, with all cloud providers offering a dedicated API to interact with their specific cloud. The APIs provide multiple interfaces to cover all types of interactions, such as e.g. for the functional administration of cloud services, authentication and authorization, billing, or invoicing. Ideally, these APIs would be standardized, so interaction with different clouds would have minimal impact on the customer's components. Changing cloud services across providers would be a smooth experience.
b) Application portability is the ability to easily transfer an application or application components from one cloud service to a comparable cloud service and run the application in the target cloud service. The ease of moving the application or application components is the key here. The application may require recompiling or relinking for the target cloud service, but it should not be necessary to make significant changes to the application code.

Testbed-13 is developing a cloud-computing environment for Earth observation data (Big Data) that is integrated with OGC web services. The environment as shown in Figure 11 will support hosting of data processing tools including all necessary deployment and management steps.

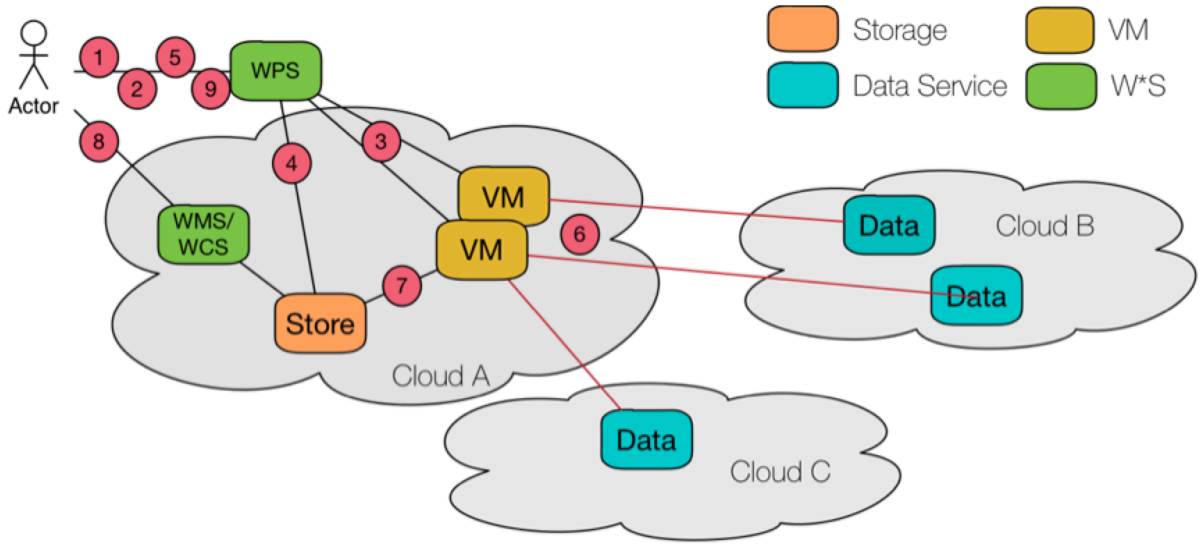

Figure 11 — OGC Testbed 13 Cloud Environment Overview

OGC Testbed-13 supports the development of the European Space Agency's (ESA) Thematic Exploitation Platforms (TEP) by exercising envisioned workflows for data integration, processing, and analytics based on algorithms developed by users. Algorithms are initially developed by users in their local environments and afterwards tested on the TEP. The goal is to put an already developed application into an Application Package, upload this package to the TEP, and deploy it on infrastructure that is provided as a service (IaaS) for testing and execution. The entire workflow should support federated user management (Identity provider and security token service) and makes use of already available catalog services and catalog interfaces as part of the cloud platforms.

During the Location Powers: Big Geo Data workshop, Dan Getman presented several approaches for processing satellite imagery that are being used by DigitalGlobe. Figure 12 shows how the DigitalGlobe approach of pre-computing and storing image chips in an object store. This defers processing that

supports analysis ready paradigm. With this approach the data is smaller and more targeted, enabling "map and reduce" analysis. This treats imagery as a service rather than determining which imagery strips to order and calling a sales rep, etc.

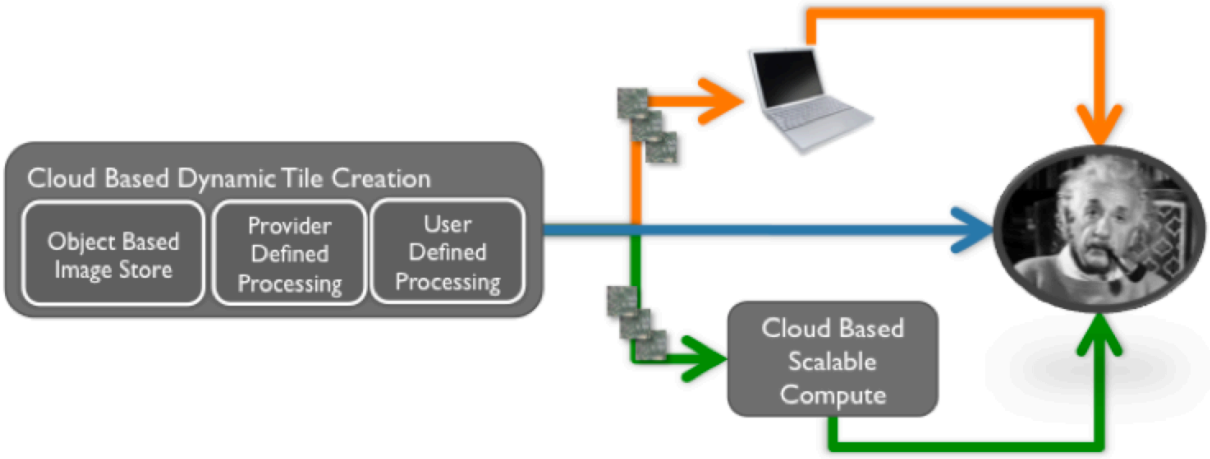

Figure 12 — DigitalGlobe Analysis Paradigm: Tile based (Figure Source: Dan Getman)

An example of the application of Big Data computing techniques to Earth Observation data is the SciSpark project by JPL that marries Apache Spark with climate science. SciSpark is a scalable system for interactive atmospheric analysis. SciSpark performs scientific data ingestion, visual interaction and metrics generation using the Spark engine.

A commercial project leading in this area is the hosting of Exelis ENVI Analytics on DigitalGlobe's Geospatial Big Data Platform. This approach allows the 15-year catalog of high-resolution satellite imagery from DigitalGlobe to be processed using ENVI's analytic tools in a cloud-based environment. This will allow users to extract incredibly valuable information and insight about our changing planet — at scale.

Planning for OGC Testbed 14 is underway including potential cloud processing topics[23].

4.3. Analysis Ready Data and Datacubes

A main theme emerging from the Location Powers Big Geo Data workshop was "Analysis Ready Data (ARD)." Several additional themes that support ARD were also discussed such as "Download as Last Resort Mentality," "Analytics as a service," and "Datacubes."

As presented by Glenn Guempel at the Location Powers Big Geo Data workshop, the USGS Land Change Monitoring Assessment and Projection (LCMAP) information system aims to provide interactive access to the Analysis Ready Data (Figure 13). To achieve that perspective, USGS, identified three approaches:

a) Store data in unzipped, optimal formats ready for direct processing by standard services or custom processes;

[23] http://www.opengeospatial.org/projects/initiatives/testbed14

b) Provide basic visualization, analysis and extraction functions through services on an open platform; and
c) Provide the potential processing capacity for building unforeseen custom workflows and processes against Big Data.

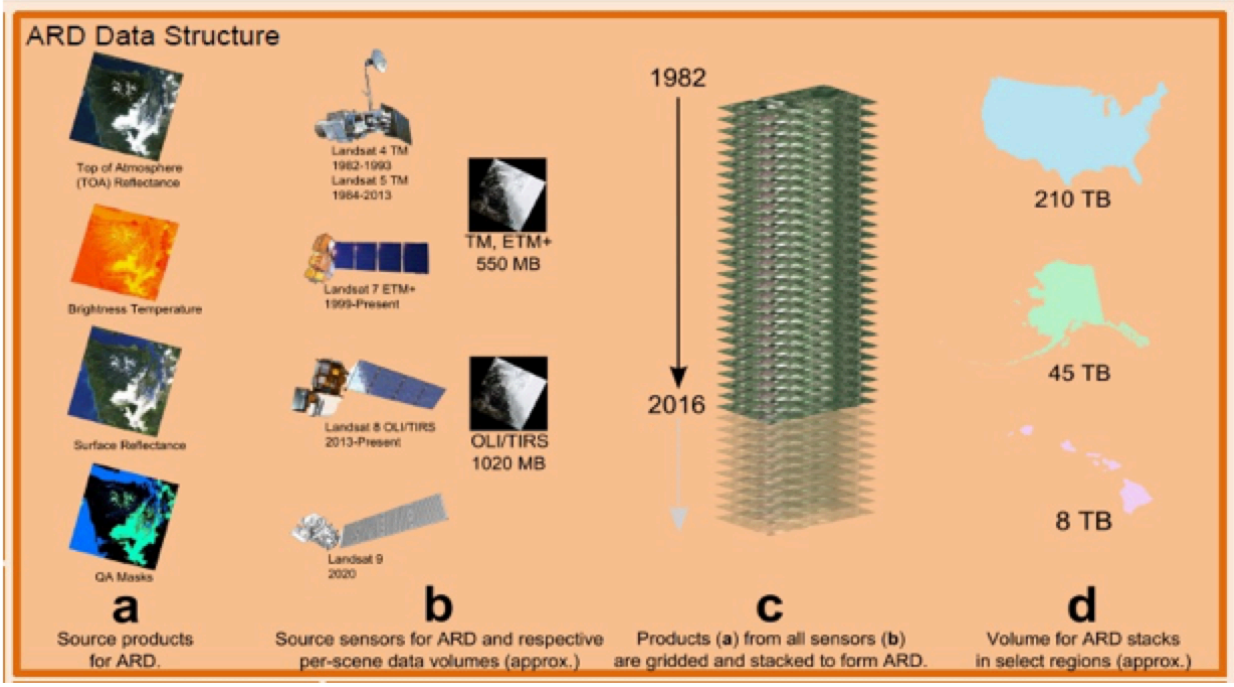

Figure 13 — USGS Analysis Ready Data structure (Figure Source: G. Guempel)

The "Datacubes" term is used in several similar but differing developments.

— The Australian Geoscience Data Cube is an collaborative approach for storing, organizing and analyzing the vast quantities of satellite imagery and other Earth Observations. The Data Cube is a series of structures and tools that calibrate and standardize datasets, enabling the application of time series and the rapid development of quantitative information products.
— At the Location Powers: Big Data Workshop, Peter Baumann presented the Earthserver project as an example of a data cube approach. EarthServer makes Agile Analytics on Big Earth Data Cubes of sensor, image, simulation, and statistics data a commodity for non-experts and experts. EarthServer provides an array databases as an interface for analysis ready data. EarthServer datacube services are operating with greater than 500 TB of data. Intercontinental datacube fusion has been demonstrated live between ECMWF and NCI Australia, both running the European datacube tool, rasdaman.
— The CEOS Open Data Cube (ODC) is a common analytical framework based on analysis ready data from current CEOS satellite systems. The ODC is a technological solution that removes the burden of data preparation, yields rapid results, and utilizes an international global community of contributors.
— The QB4ST activity addresses a related but slightly different topic of building an RDF Data Cube extensions for spatio-temporal dimensions, components, and profiles

Progress toward coordination of Datacube developments could be based on discussion of common requirements. The Earthserver Project provides this list of requirements for consideration[24].

a) Datacubes shall support gridded data of at least one through four spatial, temporal, or other dimensions.
b) Datacubes shall treat all axes alike, irrespective of an axis having a spatial, temporal, or other semantics.
c) Datacubes shall allow efficient trimming and slicing along any number of axes from a datacube in a single request.
d) Datacubes shall convey similar extraction performance along any datacube axis.
e) Datacubes shall allow adaptive partitioning, invisible to the user when performing access and analysis.
f) Datacubes shall support a language allowing clients to submit simple as well as composite extraction, processing, filtering, and fusion tasks in an ad-hoc fashion.

4.4. Data Representation for Big Geo Data: Features, Coverages, DGGS

Recall this comment from Clause 3.3 by the National Academies "A good representation of data is priceless: a single data representation, or sometimes multiple ones, allows one to carry out a large number of data processing and analysis tasks." For Big Geo Data this means placing the existing data representation methods for geospatial information into a big data context. Priceless previous geospatial analytics are based on Features, Coverages and Coordinate Reference Systems (CRS). These data representation methods along with the newer DGGS-based analytics are proposed to be a basis of OGC big data activities.

4.4.1. Simple Features applied to Big Data

OGC Simple Features has been used as a fundamental geospatial data representation for two decades. Simple Features — also published as an ISO standard — provides geometries and feature model that is used in many OGC Compliant and other implementations.

Recently, Raj Singh of IBM commented that an impactful activity for OGC would be to bring Simple Features to the Big Data world's "DataFrame" object types. Spark, Python Pandas, and R all have DataFrame objects as their primary data structure[25]. An excellent OGC big data activity would be to define how DataFrame object types support a spatial data type.

4.4.2. Coverages applied to Big Data

Geospatial Coverages have been a radical and effective method to bring the big data of remote sensing in accord with the GIS-oriented data concepts. The Coverages standard — which began in OGC and subsequently was also published as an ISO standard — defines a conceptual schema that maps from a spatiotemporal domain to feature attribute values where feature attribute types are common to all geographic positions within the domain.

[24] http://earthserver.eu/tech/datacube-manifesto

[25] https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html

Chris Lynnes, NASA, commented there is an opportunity for OGC to lend unique value to the Big Data problem by addressing the challenges of handling the spatial alignment / co-location problem that occurs both between datasets, and within datasets where we are looking at them over time. The Datacubes discussion began this theme. Here the role of Coverages data representation is highlighted.

Coverages and the associated standards for Web Coverage Service (WCS) and Coverage Implementation Schema (CIS) were initially applied to the 2D imagery case. More recently the standards and implementations of WCS, CIS and the EO profile of WCS have been extended to 3D and 4D. Also the organization of coverage collections has been extended to handle the high number of layers in applications such as Meteorology. These new advances point the way to achieve the vision of big geo data based on coverages. These advances provide consumers the ability to singularly request data in 3D/4D domains and receive N-Dimensional range/feature data about geography, time, altitude, & ensembles, etc.

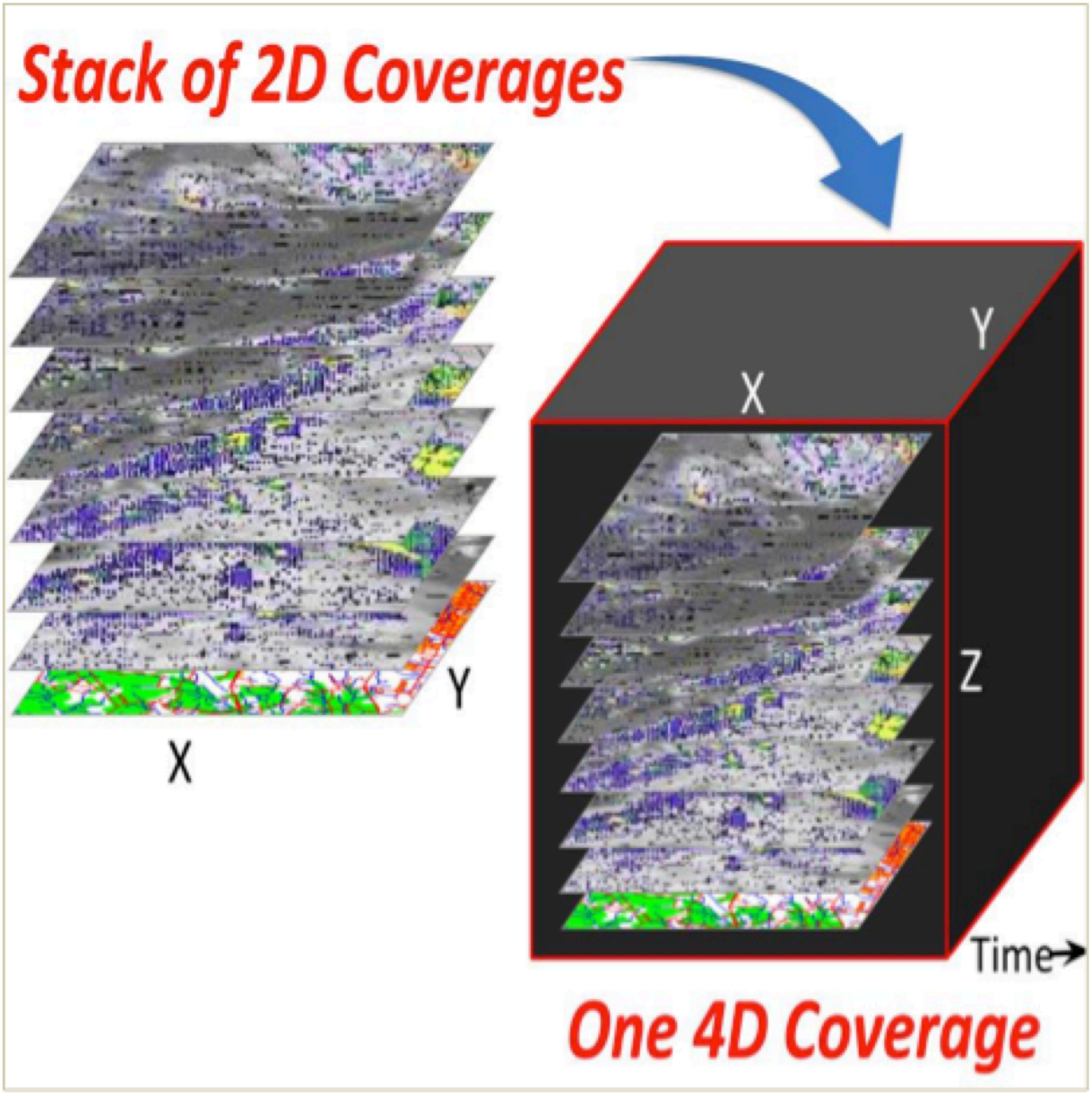

Figure 14 — Achieving the Big Geo vision with 4D Coverages (Figure Source: P. Baumann)

4.4.3. Computing with Discrete Grids

Rose Winterton, Pitney Bowes, presentation during the Location Powers Big Geo Data Workshop identified a key analytical capability of "Reduce the complexity of billions of transactional records by assigning data to geographic bins and aggregating results." By binning data into discrete grids, analytics can efficiently answer questions like:

— Is the average 4G network coverage in this area better than a competitor?
— Is the accumulated exposure at risk of hurricane damage too high?
— Is this data point inside or outside of a geofence?

To support grid-based analysis OGC has recently approved the OGC Discrete Global Grid System (DGGS) Core Standard [15-104r5] as a new OGC Abstract Specification Topic. This document specifies the core standard and extension mechanisms for Discrete Global Grid Systems (DGGS). A DGGS is a spatial reference system that uses a hierarchical tessellation of cells to partition and address the globe. DGGS are characterized by the properties of their cell structure, geo-encoding, quantization strategy and associated mathematical functions. The OGC DGGS standard supports the specification of standardized DGGS infrastructures that enable the integrated analysis of very large, multi-source, multi-resolution, multi-dimensional, distributed geospatial data. Interoperability between OGC DGGS implementations is anticipated through extension interface encodings of OGC Web Services.

Publication of the DGGS Abstract Specification can provide a structure to big data analysis activities using various grids. Results of a coordinated open development would be consensus agreement on specific grids as well as improved data quality and analysis based on increased understanding of the grid-based analysis.

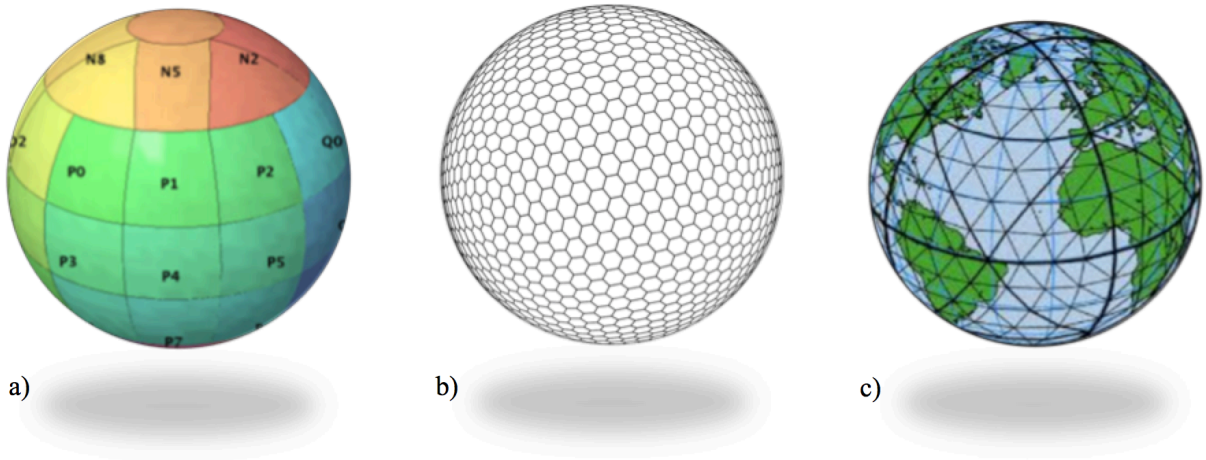

Figure 15 — Examples of DGGS mapping faces of Platonic solids to surface of the Earth. a) Rectilinear cells on rHealPIX projected hexahedron (rHealPIX DGGS see ref [41]); b) Hexagonal cells on ISEA projected icosahedron (ISEA3H DGGS — courtesy of PYXIS Inc.); c) Triangular cells on a Quaternary Triangular Mesh of an octahedron (QTM — courtesy of Geffrey Dutton). (Source: OGC Abstract Specification Topic 21: Discrete Global Grid Systems,OGC Document 15-045r5)

4.5. Big Linked Geodata

How can we make sense of big data? Developments in the Semantic Web make it possible to link data based on geographic information in a way that provides more insight. The Location Powers: Big Linked Geodata workshop investigated scaling effective exploitation of linked geodata by using big data approaches. Here, two approaches need to be differentiated. First, links between Big Data entities, and second, links between metadata for Big Data. Both enable more holistic views, but approach it from a different angle.

Josh Lieberman at Location Powers: Big Linked Data characterized Linked Open Data as both one of the best thing that happened to semantics and also one of the worst things:

— Best — because it solves the island problem;
— Worst — because of missing link semantics;

— HOW things are related is important to make sense of Big Data!

Several presentations at the Location Powers: Big Linked Geodata workshop shows the opportunity for coordinated open developments.

— Linda van den Brink presented “5 years of linking spatial data in the Netherlands.” Over the last five years, a group led by Geonovum developed a wealth of knowledge and practice on Linked Data for both spatial and non-spatial data.
— Manoulis Koubarakis presented “Scaling linked geodata to cross-border and cross-sector public services.” Including a Life Cycle of Linked Open EO Data (Figure 16).
— Gabriel Kepeklian presented “From Linked Datasets to Linked Data Streams” about the Datalift project that developed a platform to publish and interlink datasets on the web of data. In Datalift, the input data are raw data coming from multiple heterogeneous formats (databases, CSV, XML, RDF, RDFa, GML, Shapefile, …). The output data produced are « Linked Data », they are also named semantic and interconnected data. Progress is made now on WAVES as Big Data Platform for Real-time Semantic Stream Management (Figure 17).
— Oracle has demonstrated scalability of its platform to RDF trillion triple store.
— Wouter Beeks presented on “How to Query Cadastral Big Data Using GeoSPARQL?” including the pointers to the LOD Laundromat tools to improve performance and perhaps change approach.
— INSPIRE is developing RDF encoding guidelines.
— Chuck Heazel presented OGC Testbed 13 results on the integration of the General Feature Model and Linked Data principles. The results suggest how Bayesian techniques can be used with link objects to represent and manage uncertainty in a General Feature Model-based data store.
— Open implementations of Linked Geo Data can build on the work of the W3C/OGC Spatial Data on the Web builds on many lessons learned.

The Location Powers: Big Linked Data workshop triggered Rein van ‘t Veer to blog: Is it time to drop the Linked Data fixation on SPARQL and move on to more stable options? Rein’s blog highlighted advances being made with databases like ElasticSearch and MongoDB as well as embracing JSON-LD as a native RDF serialization. JSON document stores with a huge user base offer a scalable, performant, cheap and highly available.

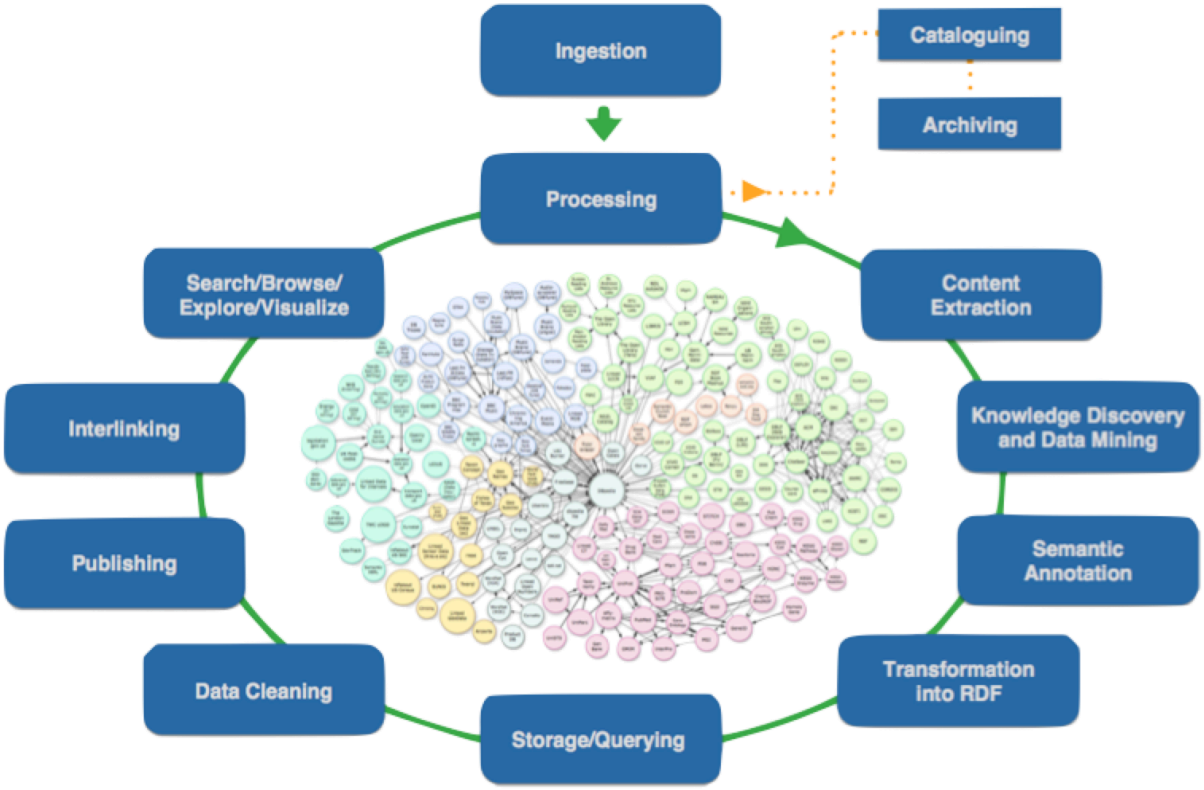

Figure 16 — Life Cycle of Linked Open EO Data (Figure Source: M. Koubarakis)

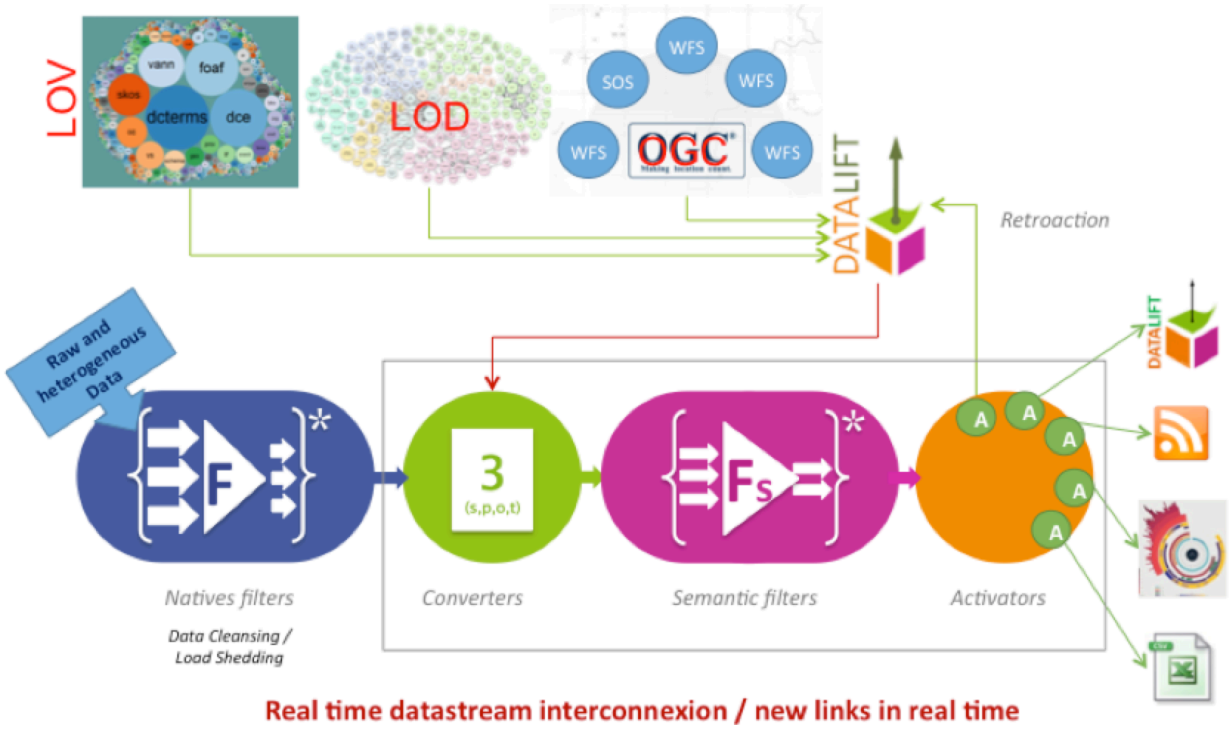

Figure 17 — Creating linked data in real time (Figure Source: G. Kepeklian)

4.6. Using Big Data Open Source

The Apache Software Foundation and Location Tech is developing open source projects applicable to the Big Geo Data.

- Apache:
  - General: Spark, Hadoop, Marmotta, NiFi, Kafka, Accumulo, Storm, Lucene, Jena, Mahout, Cassandra
  - Geospatial: Spatial Information System (SIS), Magellan
- LocationTech:
  - GeoWave, GeoTrellis, GeoMesa, GeoJinni

Multiple members of OGC and other organizations are using those open source projects to on big data applications.

- Adam Mollenkopf (Esri) presented “Applying Geospatial Analytics Using Apache Spark Running on Apache Mesos” during the Geospatial track of Apache Big Data conference.
- The Mission Exploitation Platform PROBA-V as presented as BiDS’14 and BiDS’16 has developed scalable processing and data analytics platform based on a Hadoop Cluster.
- Rob Emanuele (Azavea) during the Location Powers: Big Geo Data workshop presented “Enabling access to big geospatial data with LocationTech and Apache projects” (Figure 19).
- Rose Winterton (Pitney Bowes) during the Location Powers: Big Geo Data workshop presented “Transforming Insurance, Financial Services and Telecommunications with Big Data technology“. Showing an architecture for the Pitney Bowes Big Data Spatial Components Figure 18.
- The UK Met Office is working on cloud based suite of technology to work with huge data sets in a way that’s user friendly but powerful. The Informatics Lab developed a prototype using an Infrastructure as Code approach based on Docker, Jupyter, Dask and more.

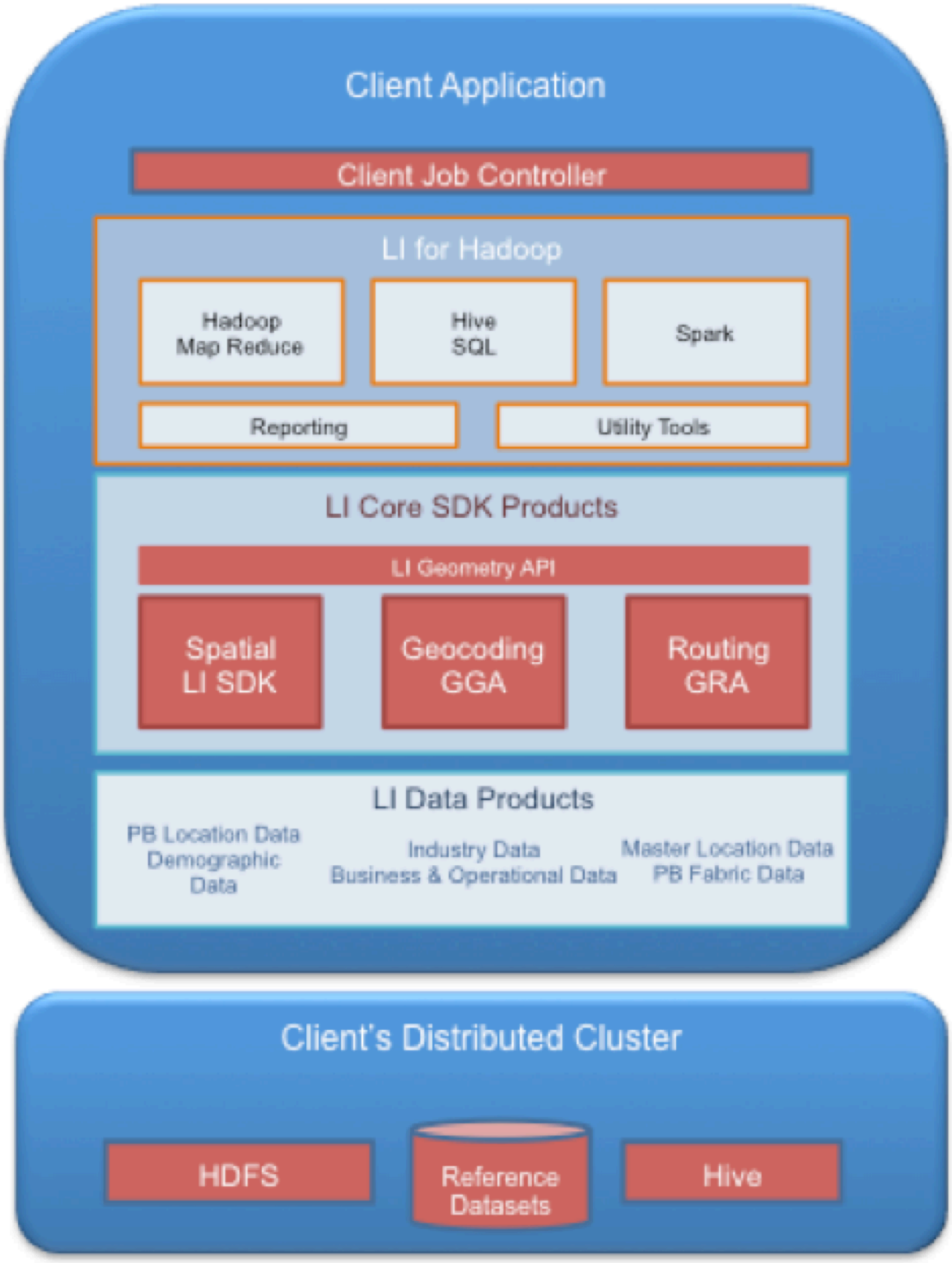

Figure 18 — Pitney Bowes Big Data Spatial Components (Figure Source: Rose Winterton)

Figure 19 — Example Geospatially Enabled Apache Projects (Figure Source: Rob Emanuele)

Many of the popular open source projects for big data focus on the pleasing parallel or embarrassing parallel data problems. Some analyses of geospatial data are not well suited to these parallelization methods. Geospatial data with multiple dimensions for space and time along with dimensionality of attribute values, e.g., vector domains in coverages, require different approaches to parallelization. Professor Fox presented Figure 20 at the Location Powers: Big Geo Data workshop. The figure is described in a research paper[26].

[26] http://grids.ucs.indiana.edu/ptliupages/publications/nistHPC-ABDS.pdf

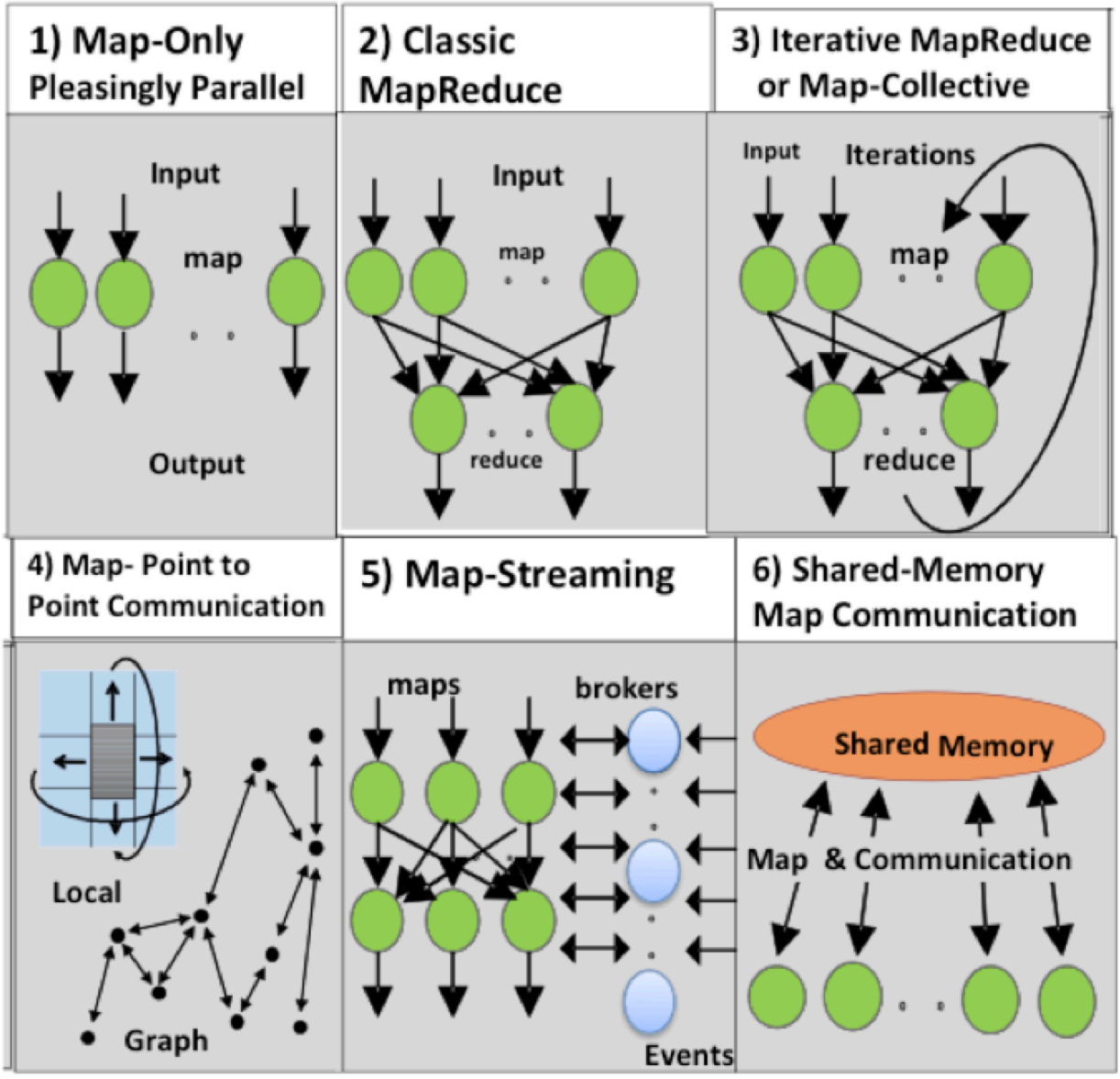

Figure 20 — Distinctive Software/Hardware Architectures for Data Analytics (Figure Source: G. Fox)

5. OGC activities on Big Geo Data

5.1. OGC Program Activities on Big Geo Data

OGC Technology Strategy

— The OGC Technology Strategy provides coordination of technology across the OGC Programs. The OGC Technology is managed by the CTO in support of the OGC Board of Directors.

— As part of the strategy a Geospatial Technology Trends tracking activity is maintained with review and input by the OGC Architecture Board (OAB). Emerging geospatial technologies opportunities for OGC activity are listed as a “Ripe Trends.”
— Big Geo Data is currently listed as a Ripe Trend. Each of the OGC Programs is addressing Big Geo Data.

OGC Standards Program

— The OGC Big Data Domain Working Group (DWG) serves as a focal point in the OGC Standards program for discussion of Big Data topics.
— The OGC Technical Committee Chair regularly conducts “Future Directions” sessions as part of the TC meetings. Big Data topics are routinely discussed in the Future Directions sessions.
— Multiple OGC Standards development are related to Big Data including: WPS, WCPS, WCS, DGGS, SensorThings, Moving Features, more.

OGC Innovation Program (IP)

— The OGC IP conducts studies, testbeds, pilots and other projects to advance geospatial technology innovation. Engineering Reports from OGC IP are provided to the OGC Standards Program for consideration of new and refined standards.
— OGC Testbed 13 currently under way includes the Earth Observation Clouds (EOC) thread that focuses on cloud computing for Earth Observation data.
— OGC Testbed 14 is currently being planned and is anticipated to include a variety of Big Data Topics.

OGC Communication and Outreach Program

— The Location Powers events led by COP have been key to identifying the topics in this white paper.

5.2. External coordination: Standards

ISO/IEC JTC 1/WG 9 Big Data

— WG 9 serves as the focus of and proponent for JTC 1’s Big Data standardization program; Develop foundational standards for Big Data; and Engage with the community outside of JTC 1 to grow the awareness of and encourage engagement in JTC 1 Big Data standardization efforts within JTC 1, forming liaisons as is needed.
— Current projects under development:
  — ISO/IEC 20546 Information technology — Big data — Overview and vocabulary
  — ISO/IEC 20547 Information technology — Big data reference architecture
  — Part 1: Framework and application process
  — Part 2: Use cases and derived requirements
  — Part 3: Reference architecture
  — Part 4: Security and privacy fabric (transferred to JTC 1/SC 27)
  — Part 5: Standards roadmap
— OGC has formed an alliance relationship with JTC 1/WG 9. JTC 1 identifies this as a C-liaison

US NIST Big Data Public Working Group (NBD-PWG)

— The focus of NBD-PWG is to form a community of interest from industry, academia, and government, with the goal of developing a consensus definitions, taxonomies, reference

architectures, and technology roadmaps. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable Big Data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.
— NBD-PWG documents have been provided to JTC 1/WG 9
— OGC members and staff participate in NBD-PWG activities.

5.3. External coordination: R&D

Apache Software Foundation (ASF)

— ASF oversees more than 350 leading Open Source projects, including Apache HTTP Server. ASF provides an established framework for intellectual property and financial contributions. ASF has 38 projects in the category of Big Data.
— OGC organized a geospatial track at the Apache Big Data Conference 2016. Discussion in that session led creation of a geospatial@apache.org mailing list.
— OGC may organize geospatial sessions at future ASF events.

LocationTech

— LocationTech is the Eclipse Foundation's industry working group focusing on location aware technologies. LocationTech has several projects that support Big Data including GeoJinni (formerly SpatialHadoop), GeoTrellis, GeoMesa, GeoWave
— OGC is a Participating Member of Location Tech
— LocationTech participated in the Location Powers: Big Data workshop.

Open Source Geospatial Foundation (OSGeo)

— OSGeo is a not-for-profit organization whose mission is to foster global adoption of open geospatial technology by being an inclusive software foundation devoted to an open philosophy and participatory community driven development. OSGeo organizes the annual FOSS4G conference.
— The FOSS4G conference 2017 has multiple papers regarding Big Data and OSGeo projects. OGC Members and Staff will be attending FOSS4G.

NSF-funded activities

— The US National Science Foundation funds several activities that are advancing geospatial Big Data in the geospatial.
— The CyberGIS Center at the University of Illinois at Urbana-Champaign, advances cyberinfrastructure, geographic information science and technologies, and various geospatial knowledge domains. OGC Staff have served on the external review board for CyberGIS
— Big Data Hubs including aim to stimulate an agile and sustainable national Big Data (including Big Geo Data) innovation ecosystem.

European Commission Horizon 2020

— The H2020 DATABIO project explores the potential of Big Data integration and analytics in the domains agriculture, forestry, and fishery/aquaculture; taking into account interoperability and

sustainability aspects in the heterogeneous European bioeconomy landscape. OGC Innovation Program is a member of the DATABIO Project.

BiDS Conference Series

— The Big Data from Space (BiDS) conferences of 2014 and 2016 were relevant to the topics addressed in this paper. Participation in BiDS 2017 would again be relevant.

6. Acronyms

ABI	Activity-Based Intelligence
ASF	Apache Software Foundation
BEDI	Big Earth Data Initiative
BiDS	Big Data from Space
DGGS	Discrete Global Grid System
DWG	Domain Working Group
ESA	European Space Agency
ESIP	Earth System Information Partners
ETL	Extract, Transform, Load
EU	European Union
EUSC	European Union Satellite Centre
FOSS	Free and Open Source Software
GEO	Group on Earth Observations
GIBS	Global Imagery Browse Services (NASA)
GPS	Global Positioning System
HDF	Hierarchical Data Format
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
IEC	International Electrotechnical Commission
IoT	Internet of Things
ISO	International Standards Organization
JSON LD	JavaScript Object Notation for Linked Data
LCMAP	Land Change Monitoring Assessment and Projection (USGS)
LEO	Linked Earth Observation

LOD	Linked Open Data
MDA	Multi-Dimensional Arrays
NASA	National Aeronautics and Space Administration
NBD-PWG	NIST Big Data — Public Working Group
NetCDF	Network Common Data Form
NEXRAD	Next Generation Weather Radar
NIST	National Institute of Standards and Technology
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
OAB	OGC Architecture Board
OBP	Object-Based production (OBP),
OGC	Open Geospatial Consortium
OPeNDAP	Open-source Project for a Network Data Access Protocol
ORNL	Oak Ridge National Laboratory
OSGeo	Open Source Geospatial Foundation
OWL	Web Ontology Language (W3C)
PB	Peta-Byte
PLDN	Platform implementatie Linked Open Data
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RDSMS	Relational Data Stream Management System
RSA	Raster Storage Archive
SOS	Sensor Observation Service
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SSN	Semantic Sensor Network
SWE	Sensor Web Enablement
TB	Terra-Byte
TEP	Thematic Exploitation Platforms (ESA)
THREDDS	Thematic Real-time Environmental Distributed Data Services
URI	Uniform Resource Identifier
USGS	US Geological Survey

W3C	World Wide Web Consortium
WAMI	Wide Area Motion Imagery
WCPS	Web Coverage Processing Service
WCS	Web Coverage Service
WMS	Web Map Service
WPS	Web Processing Service

Annex A
(informative)
Location Powers Emergent Themes

Location Powers is a place to share what we understand about Location and what more we want to know. The Location Powers Summits are provided by the OGC, whose members have been making the world's location standards for over 20 years. As well as being a place of learning for everyone about the power of understanding location, the summits are also designed to help inform the future direction and work within the consortium to continue developing relevant and useful standards.

Location Powers: Big Data in September 2016 Emergent Themes:

- We live in a download mentality. How do we move to answering questions? Analytics on the fly
- Focus shifting from understanding what happened last week to being able to predict what will happen next week
- Geo Big Data could take better advantage of developments in Big Data Proper, which is only tangentially interested in Geo Big Data
- Challenges
  - Moving from the disorganized attic model to a standard
  - Standardizing rest interface
  - Sharing algorithms among users
- Multiple Applications: Telecommunications, property casualty insurance, financial services, Energy Monitoring and prediction, Population Dynamics, Settlement Mapping
- Input to OGC Testbed 13

Location Powers Big Linked Geodata, March 2017 Emergent Themes:

- Overarching questions
  - How can we use location linked data to make sense of Big Data?
  - What are the challenges when scaling linked data to Big Data spatial analytics?
  - Is Linked Data a way to achieve Analysis Ready geospatial Data?
- Emergent themes about Linked Geodata
  - Semantics: Which relationships and classes need to be standardized for location?
  - The ETL challenge of keeping the data up to date on a daily basis (or faster)
  - How to model uncertainty > ties in with statistics

- Emergent Themes about Big Linked Geodata
  - Big linked data is challenging
    - but there are still issues (performance)
    - commercial vs. open source
  - Big data strategies
    - Using linked metadata to make Big Data more manageable… … and use other data crunching techniques, tools and (REST?) APIs to then deal with it
    - Bring the user to the data
  - Also consider the “variety” V
    - different solutions and formats for different user groups
    - standards and governance
    - think “supply chains”