Contemporary Challenges for Data-Intensive Scientific Workflow Management Systems

Ryan Mork
The University of Chicago
5801 S Ellis Ave
Chicago, IL 60637
mork@uchicago.edu

Paul Martin
University of Amsterdam
Science Park 904
1098XH Amsterdam,
Netherlands
P.W.Martin@uva.nl

Zhiming Zhao
University of Amsterdam
Science Park 904
1098XH Amsterdam,
Netherlands
z.zhao@uva.nl

ABSTRACT

Data-intensive sciences now represent the forefront of current scientific computing. To handle this 'Big Data' focus, scientists demand enabling technologies that can adapt to the increasingly distributed, collaborative, and exploratory scientific milieu. However, how these challenges have changed the design requirements of scientific workflow management systems (SWMSs) has not been assessed. First, how scientists currently use SWMSs was determined through a comprehensive usage survey examining 1455 research publications from 2013 to July 31st, 2015. To understand how data-intensive scientists are producing impactful research, we further examined usage of two major research clouds, the Open Science Data Cloud (OSDC) and Cornell's Red Cloud. Here, we present a road map for SWMS development for data-intensive sciences. SWMSs are now needed that interconnect diverse software packages while enabling data exploration and multi-user interaction across distributed software and hardware environments.

Keywords

Enabling Technology, Big Data, Cloud Computing, Scientific Workflow Management Systems, Workflow, Data-Intensive Science

1. INTRODUCTION

Data-intensive computing and Big Data currently play an increasingly important role in industry [1], scientific discovery [14], and public administration [37]. Data-intensive applications often have very high business value in e-commerce data mining [3] or social impact in climate simulation and disaster early warning [46]. However, these applications are also very difficult to implement and execute due to the huge quantities of data involved and their high storage and computing requirements [31].

The exponential growth of total information, around twofold per year, is already faster than the increase of computing and storage capacity, 1.5-fold pear year [8, 49]. Distilling meaning from this 'data deluge' requires concerted collaboration between data analysts, domain experts, and data engineers. Smart supporting technologies for handling such increasing data volumes and for utilizing the available technical resources for storage and computing are thus urgently demanded for exploiting data-intensive computing in various application fields.

In the past two decades, many data-oriented tools and technologies have been developed for data transfer, processing, cataloging, annotation and integration. Among these tools, scientific workflow management systems (SWMSs) are a primary enabling technology for integrating distributed resources and for handling complex computational tasks [47]—examples include Taverna [53], Kepler [12], Triana [19] and NS Pegasus (using Wings) [24, 27]. However, most of these tools were proposed and developed in the era of grid computing, wherein many data pipelines represented in workflows were pre-defined, handling static, small-size data volumes [56]. Computing models such as MapReduce [23] were next proposed that specifically supported data-intensive applications. Such models suited massive-scale data processing, although permit limited application logic complexity [35, 42].

Big Data has changed the market for scientific software. Changes in data size, technology, userbase, Quality of Service (QoS) requirements, and methodology have provided significant difficulties for grid-era SWMSs. Data-oriented research infrastructures, such as the European Strategy Forum on Research Infrastructures (ESFRI) [7], or data cloud based collaboration environments, such as the Open Science Data Cloud (OSDC) [32], need new workflow systems and other supporting technologies for their user communities. What demands do these communities present to SWMSs? What are the gaps between the requirements of data-intensive software and what is actually offered by SWMSs? These questions require a more comprehensive understanding of the problems SWMSs are currently used to solve as well as of the software usage of data scientists in practice.

Data-intensive scientific computing demand new design principles for SWMSs. In order to discover these principles, the disconnect between the usage and features of SWMSs and the software currently used by data-driven scientists must be determined. While there have been several recent reviews on SWMSs and their capabilities in current computing environments [45, 44, 54], we have chosen to understand these differences through a two-fold usage survey: determin-

SC '15, Austin, Texas USA

DOI: http://doi.dx.org/10.1145/2822332.2822336

ing how SWMSs are used by the scientific community, and capturing the properties of enabling technologies presently employed by data scientists conducting groundbreaking research on science clouds. This is accomplished through a comprehensive literature review, citation count, and accompanying analysis.

The main output of this paper includes:

1. An overview of the challenges posed by data-intensive science to scientific computing
2. Survey of SWMS usage in recent scientific publications (from 2013 to July 31st, 2015)
3. A software usage survey of two major scientific data clouds: the Open Science Data Cloud (OSDC) and Cornell's Red Cloud
4. Concrete design principles to guide future development of data-intensive SWMSs guided by the usage survey

In the following, the challenges posed by data-intensive science to SWMSs are first presented. Section 3 then provides both the survey of SWMS usage and the comparative overview of software usage on OSDC and Red Cloud. The suggested design principles are articulated and discussed in Section 4. To conclude, Section 5 provides a review of the contributions of the paper and its import.

2. CHALLENGES IN DATA-INTENSIVE SCIENCE

Classically, SWMSs have been envisioned to enable in software the 'analytic workflow' experienced by scientists—the process by which data is processed and analyzed to support exploration and understanding of an underlying model system [58, 38]. Since much of current scientific research necessitates understanding multidimensional datasets with requisite computational processing, SWMSs have been presumed to serve as an 'enabling technology' by which intellectual advancement can be accelerated. Taverna [53], Kepler [12], Triana [19], RapidMiner [2], VisTrails [13], Pegasus (using Wings) [24, 27], and Galaxy [28] are SWMSs intended to serve as enabling technologies that provide this analytic workflow environment. The software methodologies are similar: provide numerous pre-built tools for common workflows within a comprehensive platform that abstracts both middleware and programming using graphical user interfaces for the entire scientific workflow life cycle [45].

SWMSs have come to embody the presumed best enabling technology for scientific research that necessitates software. However, how the usage and design requirements of SWMSs have changed due to advancements data-intensive scientific computing have not been assessed. An overview of the significant challenges associated with big data science is presented below in order to understand the shortcomings of current realizations of SWMSs.

2.1 Data Handling and Processing

The 'Big Data' movement presents challenges to current computing at the processing, network and storage levels. As the volume, velocity, variety and veracity of data exponentially increases, new solutions are needed to comprehend and analyze information [20]. These systems must facilitate not only querying and accessing data from different catalogs with variable metadata, but also the efficient transfer of data within and between distributed infrastructure [51]. In many cases, the data is often produced in near real time, which requires the workflow system to rapidly process data and perform quality control [39].

2.2 Data Exploration

Data-intensive scientific discovery changes the nature of the questions asked by scientists. Scientists have often shifted from using simulations to understand data to using the data directly to develop generative models. Due to its size, scientists are now also able to develop models from the data *de novo*, rather than rely upon low-dimensional simulation to produce results [20]. Accordingly, there has been a significant rise in interest in data mining and analytics as well as associated techniques for data exploration [43]. As a result, machine learning techniques that solve problems such as feature selection, clustering and pattern recognition are becoming increasingly important [34]. Support is needed for dynamic data exploration and synthesis, whereby workflows reorient and redeploy themselves in response to upstream results.

2.3 Interdisciplinary Collaboration

Many current scientific problems are inherently interdisciplinary. Scientists are challenged to collaborate across conventional disciplinary boundaries; domain experts must also collaborate with data analysts and data engineers if they are to accomplish more complex data science. The study of climate change, for example, involves not only atmospheric science, but also earth sciences, ocean sciences, and ecology. This kind of interaction requires better tools and support environments, especially since remote collaboration is becoming increasingly common [33]. SWMSs are needed that provide not only the necessary tools for data discovery, access and manipulation, but also facilities to enhance collaboration between domain experts and computer scientists of different backgrounds.

2.4 Technological Evolution

Recent years have seen a dramatic evolution of scientific infrastructure into increasingly distributed and visualized environments: initially confined to local clusters, then grids, now scientific infrastructure is moving into the realm of cloud computing. Since workflow systems are generally tied to their underlying infrastructure, they too have changed to meet the demands of new systems. Accordingly, several SWMSs have started to support visualization and server solutions to accommodate these advances. Taverna 2 and Galaxy provide a server for executing developed workflows remotely in cloud environments [53, 11]. The rapid evolution of technology makes the execution of legacy workflows in new environments difficult, and also requires the re-engineering of individual software components in order to efficiently employ new infrastructure. As a counterpoint, Amazon's Simple Workflow Service (SWF) [5] and IBM's Bluemix [9] represent Platform as a Service (PaaS) cloud computing models for workflow solutions.

2.5 Quality of Service and Experience

Data-intensive applications experience a bottleneck not only when transferring and processing large volumes of data, but also when handling application control flow and responding to external events. Several solutions are currently be-

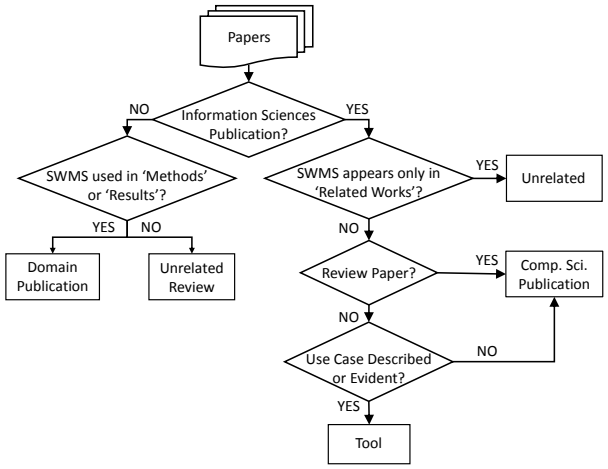

Figure 1: Workflow for the Manual Categorization for Citations of SWMSs. See Section 3.1 for a definition of all terms employed.

ing pioneered, generally by exploiting better research infrastructure or adopting better workflow patterns. Zhao et al. 2014 [55] provides a service framework for integrating the Swift SWMS into a cloud computing environment. CloudDragon uses OpenNebula and Swift to produce a scientific computing platform that provides a cloud workflow service with a static resource manager and a virtual cluster provisioner [57]. Possibly the most commonly used enabling software for cloud computing, MapReduce, provides a means for highly efficient parallel analysis of petabyte-scale datasets in cloud environments [23]. The quality of service of the system execution, in particular the experiences that users may have using data-intensive applications, significantly influences the realization of business value or research activities.

3. STATE OF THE ART SURVEY

To determine how SWMSs have been employed in light of the latest technological breakthroughs, a comprehensive survey of the usage of major SWMSs has been conducted and analyzed from 2013 to July 31st, 2015. Since the primary unit of scientific discovery is a publication, sorting and counting citations serves as one metric to measure the relative usage and efficacy of a SWMS. This survey was accomplished by gathering papers citing SWMSs using repositories such as Google Scholar, ScienceDirect, and Pubmed, collecting them with the Mendeley software environment, and manually sorting and curating these papers into usage categories. The methodology employed focused upon distilling how domain scientists have been employing SWMSs to accomplish research as well as the aspects of those SWMSs used that are most successful at galvanizing scientific discovery.

To complement this research, we also conducted a survey of the use of scientific software in general in two science data clouds, OSDC and Red Cloud, using a similar methodology. This supplementary survey serves to illustrate how applicable SWMSs are deemed by the scientific community in the context of advanced research infrastructure as well as better characterize the research practices of scientists using clouds.

3.1 SWMSs and Present Use

Since researchers can cite a SWMS paper for a number of reasons, including as a background document or as an example in a technical report, it was necessary to develop a sorting framework for publications that distinguishes genuine use-cases in domain science from extraneous references. This framework is presented below and summarized in Figure 1:

Domain Publications: papers describing the use of a SWMS to conduct original scientific research in a scientific discipline outside of the information sciences; for example, determining new properties of a star, conducting a microbial ecology experiment, or measuring how Twitter could be used to determine natural disaster patterns.

Computer Science Publications: papers that describe the functionality of a SWMS, query its efficacy, enhance its capacity with regard to a certain task, or elaborate upon possible use cases. This also includes publications that provide an in-depth review of the functionality of a SWMS.

Tools: papers that document an extension of a SWMS; for example by adding additional support for new kinds of infrastructure, such as cloud computing, or adding specific functions and workflows to the SWMS that could be employed by researchers.

Primary Tools: domain publications that cite and employ a tool developed for the SWMS. These do not include information science primary publications.

Unrelated Mentions: papers that cite the SWMS but do not describe nor mention in any detail how the SWMS was employed. An example is citing an SWMS in the related work section of a paper. These papers were excluded from further analysis.

Table 1 records the number of papers in each category for the surveyed SWMSs. A caveat to this table is that some scientific disciplines generate far more publications than others. While Galaxy has generated significantly more papers than any other SWMS surveyed, it is also the most closely tied with a specific domain (biological science), which produces significantly more papers than other disciplines.

Next, we asked specifically how the researcher in each publication chose to employ the SWMS. Usage was divided into four categories based upon the ENVRI Reference Model for research infrastructure [18]:

Data Access: these papers employed the SWMS to fetch data from a web service or database or to upload the raw data or data products generated from experiments or analysis.

Data Processing: these papers employed the SWMS for data reduction, data pre-processing, or data analysis.

Data Curation: these papers employed the SWMS to log how the data was used, develop metadata for the raw data, or annotate the metadata in a consistent manner.

Community Support: these papers employed the SWMS to record the developed analysis pipeline, preserve the state of the workflow, upload the pipeline to a repository, or employ the logging features of the SWMS.

SWMS Software	Domain Publications	CompSci Publications	Tools and Applications	Primary Publications from Tools	Total Examined
Galaxy	264	63	72	26	465
Kepler	3	128	16	12	283
Pegasus/Wings	0	90	23	2	239
RapidMiner	72	16	7	4	178
Taverna 1 & 2	5	33	23	12	131
Triana	0	9	0	0	29
VisTrails	3	55	5	6	133

Table 1: Publication Counts of Commmon SWMSs 2013 to July 31st 2015. See Section 3.1 for the column definitions. 'Total Examined' represents the non-repeated number of papers for each SWMS found.

SWMS Software	Access	Processing	Curation	Community Support
Galaxy Domain	**14**	**247**	**3**	**16**
Galaxy Tools	18	62	29	52
Kepler Domain	**1**	**3**	**2**	**1**
Kepler Tools	0	12	12	12
Pegasus/Wings Tools	0	2	0	0
RapidMiner Domain	**0**	**72**	**1**	**3**
RapidMiner Tools	0	4	2	0
Taverna 2 Domain	**3**	**2**	**0**	**4**
Taverna 2 Tools	7	4	4	9
VisTrails Domain	**0**	**3**	**0**	**2**
VisTrails Tools	6	6	6	6

Table 2: How SWMSs are used in Domain Research 2013 to July 31st, 2015. Usage categories are defined by the ENVRI Reference Model (Section 3.1). 'Domain' refers to domain science publications that directly employed the SWMS, while 'Tools' refers to publications using a developed Tool through the SWMS.

Table 2 records these usages for each of the SWMSs surveyed. A given publication can employ a SWMS for multiple reasons. Further, there is a distinction drawn in these categories between employing a SWMS for its intended purpose, to aid in reproducibility of both the data and the workflow, or its minimal necessary purpose, to retrieve or process data. Table 2 clearly indicates that while Taverna, Triana, and Kepler are rarely used, when employed, these SWMSs are fully utilized to provide a comprehensive workflow environment. Conversely, RapidMiner and Galaxy are often used but as little more than toolboxes—one specific function is transiently employed as part of a larger multi-platform data pipeline.

3.2 SWMSs and Developed Tools

Next, we wanted to understand what properties of the tools, or developments within the SWMSs, in turn led to the highest use. Further, we wanted to understand what tool attributes, independent of specific SWMSs, led to the largest number of publications. This was accomplished by searching for papers using Google Scholar that directly cited the tool paper and counting the number of these that fell into the 'domain publication' category. Table 2 documents the number of primary papers generated by these tools and places them into the previously described use categories by the type of extension provided by the tool. For example, a tool that adds a specific function, such as implementing a new algorithm, would be considered in the 'data processing' category. Developing a new workflow in the SWMS would fall under both 'data processing' and 'community support'. A similar pattern appears compared to the application of the SWMS directly—tools that generated the most citations primarily did so as a result of providing a data processing or analysis method.

Several of the tools and applications that generated the most citations were then examined in-depth to understand which design principles led to their success. Table 3 describes several of the most successful tools by citation count. While some tools provide access to web services or databases, all provide pre-made workflows for scientists to employ in their research. From this, it appears that domain scientists tend not to directly employ SWMSs such as Kepler and Taverna, but will use pipelines developed using this software if these provide a novel analysis method or data processing step. While these SWMS provide a comprehensive set of solutions for building workflows, it is possible that the barrier of entry, either through learning a new software system, or the computational knowledge required to build a workflow in these systems, is too high to be employed by a domain scientist.

3.3 Enabling Technologies Employed in Scientific Clouds

Researchers at the forefront of scientific discovery are facing significant infrastructure challenges as datasets and the requisite analysis of this data now require significant computational resources. Several science data clouds have arisen to provide tailored, discovery-driven ecosystems for scientists to accomplish novel research. OSDC embodies a dramatic new computational community: by linking a data commons with cloud computing resources, researchers are now able to share large datasets while simultaneously running their own analyses combining complementary datasets [32]. Red Cloud is another example of a large science cloud deployed

Tool	SWMS Software	Primary Citations	Acquisition	Curation	Access	Processing	Community Support
ESO Reflex Environment	Kepler	12	X	✓	X	✓	X
RenSeq	Galaxy	12	X	X	X	✓	X
deepTools	Galaxy	10	X	X	✓	✓	✓
Orione	Galaxy	6	X	X	✓	✓	✓
Banana Genome Hub	Galaxy	5	X	✓	✓	✓	✓
Peroxibase	Galaxy	5	X	✓	✓	✓	✓
K-mer SVM	Galaxy	5	X	X	✓	✓	✓
Genetics Software Suite [25]	Taverna 2	4	X	X	✓	✓	✓

Table 3: Overview of Most-Cited Tools using SWMSs 2013 to July 31st, 2015. Examples of SWMS Tools, either enhancements to the software or generated workflows using the software, that galvanized the largest number of publications, along with the provided features as defined by the ENVRI Reference Model. This only includes tools and generated publications from 2013 to July 31st 2015.

by Cornell University that offers computational resources for scientists that are a part of Cornell or are collaborating with Cornell scientists [6]. These novel scientific computing services have helped generate a significant number of high impact publications [14, 22, 26, 50, 52, 29, 17] as a result of leveraging the level of collaboration possible between researchers utilizing these linked data commons to produce field-defining results. We hoped to determine how scientists are using software in these data clouds as well as the properties of this software to elucidate the necessary requirements and possible gaps in the enabling technologies used in data-intensive sciences. We accomplished this through a literature review of papers published that employed the OSDC and Red Cloud.

Publications employing OSDC were harvested through the documented publications on their website from 2013 to July 31st, 2015 [10]. As domain scientists do not directly cite the OSDC papers, nor do they tend to mention employing cloud resources, these are the only papers that can be directly associated with OSDC. Red Cloud publications were harvested manually by reviewing the publication history of the professors known to employ these services as listed by the website from 2013 to July 31st, 2015 [6]. As with OSDC papers, papers employing Red Cloud do not directly cite the employment of cloud services. Thus, relying upon listed users is the only means to directly link publications to cloud computing resources. For this reason, public data clouds that also host a significant amount of scientific research, such as Amazon, were unable to be surveyed. Amazon does not list specific users, and the users generally do not specifically mention employing Amazon cloud services.

3.3.1 Open Science Data Cloud

Table 4 lists the properties of these publications by primary investigator and scientific discipline. Note that every major paper employed the OSDC not only for data storage and retrieval but also for computation. There is a clear trend between the domain and the degree of software documentation as well as the enabling technologies employed. Biology papers tended to employ a described list of software, both proprietary and custom-made, to accomplish data processing and analysis. Typically, several software packages were stitched together via custom scripts to produce a workflow pipeline for analysis. While links to the software are provided, parameters used and scripts are not provided for these tools. Thus, the workflows are not easily reproducible. The two linguistics papers surveyed are less reproducible, as the algorithm for analysis is listed but the software—even the programming language—to accomplish such analysis is not listed. Conversely, Mandl et al. [46], an earth sciences paper, describes a comprehensive, custom-built and openly available workflow pipeline for satellite image analysis to conduct flood prediction and prevention.

3.3.2 Red Cloud

Table 5 lists which software was used and how by primary investigator and discipline for Red Cloud resources. As with the OSDC, Red Cloud users employed these resources for both data and processing services. Red Cloud software employment, compared to the OSDC, tends to be slightly more workflow-oriented, although pipeline reproducibility remains a problem. The biology papers primarily employed QIIME [16], a software pipeline for microbial ecology research, to accomplish their cloud-based data processing followed by a variety of downstream scientific software and custom scripts to conduct analysis. The Astronomy papers show a similar pattern: thier pipelines first consist of data reduction through SpexTool [21] followed by analysis conducted through scientific software packages and scripts.

Conclusions

Enabling technologies employed on the surveyed scientific data clouds, OSDC and Red Cloud, tend to either be developed by the users or by domain scientists and used in several publications. This is exemplified with Spextool, the Namibia Early Flood Warning Sensorweb, modENCODE, and Ohmage. Another common trend is integrating previously developed scientific software packages produced by domain scientists, such as QIIME or Accelrys [4], with a custom software pipeline.

While data sources tended to be well-documented, data infrastructure and processing were poorly documented. Data sources used, either generated during the publication or used from previous publications, were documented. Data processing pipelines, however, were usually a patchwork of custom scripts and a variety of previously developed scientific software. While workflow-oriented software is employed in the initial, static, data processing stages, the analysis stage itself is conducted by a diverse set of software tools tailored for answering specific questions. Data exploration is a key step in this scientific workflow during the analysis stage and is handled by a variety of separate software packages.

Principle Investigator	Scientific Domain	Papers Using Cloud	Collaborative	Usage Description
Kevin White	Genetics and Systems Biology	5	2/5	Data housing of biological samples; Data analysis (UU)
Tyler Schnoebelen (Idibon)	Linguistics	1	1/1	Twitter Data Set and Analysis using Custom Scripts (UU)
Eric Xing	Linguistics	1	1/1	Twitter Data Set and Analysis using Custom Scripts (UU)
Andrey Rzhetsky	Bioinformatics	1	0/1	Data housing; usage of Wikipedia data set; Data Analysis using Custom Scripts
W. Rathmell and Chad Creighton	Genetics and Cancer Biology	1	1/1	Data processing; Data analysis (UU)
Pedro Galante	Genetics and Cancer Biology	1	0/1	Data usage of housed data, 1000 Genomes
Marsha Rich Rosner	Genetics and Cancer Biology	1	1/1	Biological Data Housing; Data processing; Data Access (UU)
Jorg Szarzynski	Earth and Weather Sciences	1	1/1	Satellite data housing; Data processing and analysis for the Namibia Early Flood Warning SensorWeb
Theodore Karrison	Genetics and Cancer Biology	1	1/1	Data analysis using R statistics packages (UU)
Lincoln Stein	Software Engineering; Genetics and Cancer Biology	1	1/1	Software Development, modENCODE, for the Bionimbus Protected Data Cloud (part of OSDC) and Galaxy

Table 4: OSDC Software Usage from 2013 to July 31st 2015. '(UU)' indicates 'Undocumented Use': the scientist does not describe how the cloud was used, and was inferred given the applications employed. 'Collaborative' represents the number of papers with more than two authors from different academic institutions.

Principle Investigator	Scientific Domain	Papers Using Cloud	Collaborative	Usage Description
Ruth Ley	Microbial Ecology	5	4/5	Data housing of biological sequence data; Processing of data using QIIME; Local Analysis
James Lloyd	Astronomy	3	3/3	Data housing for the TripleSpec Spectrograph; Data analysis through SpexTools
Ludmilla Aristilde	Biological and Environmental Engineering	1	0/1	Data Analysis through Molecular Dynamics simulations using Accelrys
Denina Hospodsky	Biological and Environmental Engineering	1	1/1	Data processing through QIIME with local, custom analysis tools
Josh Selsky	Medical Software	1	1/1	Data housing and processing for Ohmage, a web client for mobile data

Table 5: Red Cloud Software Usage from 2013 to July 31st, 2015. 'Collaborative' represents the number of papers with more than two authors from different academic institutions.

4. DISCUSSION

From the conducted survey, it is clear that the generalist SWMSs developed in the grid era represents a minority solution for scientific computing (Section 3.1). Furthermore, scientists conducting groundbreaking scientific research using state-of-the-art science clouds are not employing SWMSs and usually develop their own software solutions to solve their problems (Section 3.3). While the computer science field has spent considerable effort in developing usable, comprehensive software solutions for enabling research, these are not being commonly employed by domain scientists. They are, however, used by other computer scientists to develop specific workflows which domain scientists then use (Section 3.2).

Usage rates for tools developed for these SWMSs can provide insights into why these enabling technologies have been discarded by current researchers. The common thread of usage by the most successful tools are those that provide specific solutions for common problems in a particular field (Table 3). This software is typically produced either by a domain scientist or by computer scientists working closely with a domain scientist. However this case applied only for a few of the tools developed; the majority of new tools extended the general usability of a given SWMS, by (for example) including better data provenance, execution support,

etc., that were not cited. There may be bias in that these papers either do not receive proper exposure and so are not actually employed by researchers, or are integrated into the SWMS and thus are not generally directly cited even when they *are* used.

We now discuss the conclusions that we can draw from our survey with respect to the challenges we identified earlier.

4.1 Data Processing and Analysis

Since current scientific pipelines utilize numerous software packages across distributed systems, SWMSs must in turn adopt a framework-based architecture that allows the assembly and logging of workflows produced by various software sources. This shift from employing SWMSs to develop software pipelines *de novo* to integrating other software is first visible in the dominant usage of Galaxy and RapidMiner as tools for executing a smaller part of a larger data pipeline (Section 3.1). The minimal use of Taverna, Kepler, and Triana reinforce this observation: data-intensive scientists utilize pre-built analysis tools and tend not to re-develop these tools within an overarching SWMS. The data cloud survey reveals a similar trend, with custom-built, poorly documented data processing pipelines representing standard practice for workflows (Tables 4, 5). Scientists already favor particular software packages for data processing; reimplementing these packages in a SWMS expends time and effort that is likely not considered worthwhile.

Goodrich et al. 2014 [30] presents a key use case that would significantly benefit from these new design requirements. In this paper, three levels of software over two computational infrastructures, cloud and local lab, were used to first process microbial samples in Red Cloud using QIIME, followed by PICRUSt to determine further genetic information, and culminating with a heritability estimate with OpenMx. A SWMS capable of wrapping these software components together along with logging use and important algorithm parameters would provide an invaluable means of reproducing experiments in addition to allowing other researchers to employ a similar pipeline with their own datasets. SWMS architectures that allows easy plug-and-play of other software resources, while maintaining invaluable workflow provenance and reproducibility tools represent invaluable, and highly useful, enabling technologies for researchers.

4.2 Supporting Data-Intensive Computing Workflows

Given the huge complexity and size of current datasets, a key focus of analysis is determining the remarkable features of any dataset. This explains the massive rise in the employment of machine learning to accomplish data reduction and feature selection. The significant usage of RapidMiner as a data processing environment illustrates such a trend (Table 2). Several of the more successful tools developed as measured by generated publications, such as K-mer SVM, represent implementations of machine learning techniques in SWMSs (Table 3). There is significant demand for open source workflow systems employing machine learning in data-intensive science. SWMSs that directly provide access to existing machine learning libraries, such as R or Python's Scikit Learn, would be highly appealing to researchers.

Data exploration is another significant aspect of the data-intensive approach. The increasing demand for visualization in software for data analtyics serves as an indirect measure of this trend that is difficult to directly detect from publications. As datasets become increasingly high-dimensional, visualizations of lower-dimensional embeddings serve as a proxy for exploration [36]. Some of the most popular tools (Table 3) from the survey, including DeepTools and the Banana Genome Hub, focus on providing visualizations to better understand complex analytics from genomics datasets. Moreover, as shown in Section 3.3, data-driven scientists are constantly updating and shifting analysis methods to better understand the underlying data—this is highlighted in the popularity of more toolbox-style enabling software such as Galaxy and RapidMiner. Workflow systems must therefore ensure that data exploration through visualization is readily supported, both by integrating various software packages and providing intermediate visualizations and feature selection methods within the data processing pipeline.

4.3 Supporting Collaboration

Data-intensive scientific research requires collaboration between both domain experts from varying fields as well as interactions with data analysts and engineers. A critical question is which party ought to, or will, use a SWMS to develop a scientific workflow. The most sensible answer is the analyst: these users tend to have the most computational knowledge and act as an interface between the domain scientists executing their questions within the constraints provided by the data engineers. This is already the case with most SWMSs: computer scientists or analysts develop a domain-specific workflow which is then employed by domain scientists for a specific application (Section 3.2). SWMSs must then handle two major use cases: first, how data analysts use the software to construct a new workflow, and second, how domain scientists can easily and reproducibly execute these constructs in various infrastructures.

A second dimension of collaborative support often neglected is collaboration between domain scientists. The bulk of papers surveyed from OSDC and Red Cloud exhibited a high degree of collaborativity—most papers were the result of interactions between different academic departments in addition to different institutions (Tables 4 and 5). Even those papers without collaborators had at least three authors or more, indicating that current scientific computing projects are overwhelmingly the product of multiple researchers developing close ties with information scientists.

For workflow systems, three levels of support must be facilitated:

Laboratory level: At this level, scientists are in the same lab and working on a problem concurrently. Workflow versioning and provenance are key features to ensure that multiple users can fluidly interact with the same workflow and reproduce results. An example is Blair et al. 2014 [15], in which three authors all worked on a medical word corpus pipeline.

Domain level: Scientists often collaborate with researchers from other fields of study. Since scientific jargon plays a key role in most experimental pipelines that dramatically shifts from field to field, data provenance becomes a critical issue. Fields may also differ in their statistical and processing methodologies, so facilitating multiple analyses, or branches, for a given dataset is also desirable. Kittler et al. 2013 [40] serves as a prime example,

in which researchers from genetics, computer science, ecology and evolution, and biophysics collaborated to provide a multi-level view of a breast cancer regulatory network.

Institution level: Large temporal, spatial, and cultural barriers may separate scientists collaborating with different institutions. Computation may also be distributed over multiple sites. Workflow systems need to seamlessly bridge these resources while maintaining network awareness of distributed data sources. In Muirhead et al. 2014 [48], for example, Cornell researchers shared datasets with researchers from California to New York nationally and Portugal internationally to study Kepler satellite planet observations.

The growing use of dedicated science clouds for collaborative research indicates a desire for virtual research environments. SWMSs need to be able to either provide that environment, or plug into such environments if they are to see use in high-impact collaborative research.

4.4 Performance, QoS Optimization, and Reproducibility

In order to provide performance evaluation metrics and optimization information, software infrastructures must now record system usage to report and ensure future QoS. Due to a lack of adequate documentation of software usage in surveyed papers in general, it is difficult to comprehensively assess the usage of performance optimization in data-intensive research environments. Accordingly, quality optimization analyses cannot be assessed from the domain science perspective, and must be determined through the infrastructures they use. Comprehensive usage logs from data clouds would provide invaluable information about the properties of current scientific datasets and the requisite quality needed for efficient runtime execution. Furthermore, this information would allow for the development of customized algorithms for the size and computational intensity of data-intensive analyses. Another solution is to have SWMSs directly log task size and completion times to help the developer better understand their user base. Determining the desired attributes of performance and infrastructure utilization necessitates direct software usage records, as such details are not typically recorded by domain scientists.

Reproducibility is essential in both assessing performance optimization and maintaining consistency within ongoing research efforts. One of the primary advantages a SWMS provides is logging, documenting, and providing a contained software product. These properties are necessary to ensure comparability and reproducibility of scientific research, and are lacking in data-intensive science at present (Tables 2, 4, 5). Since most users tend not to bother with data provenance, minimal, consensus-based standards must be developed and employed for SWMSs that can be applied with negligible user interaction. These services must exist below the software and databases employed, as fine-grained interaction with such technologies may be beyond the expertise of the user.

4.5 Utilization of Advanced Infrastructures

Data-driven scientists employing scientific data clouds demand enabling technologies that solve specific scientific problems capable of being easily deployed on virtualized environments. Interfacing with localized computation is also essential for downstream data exploration and analysis. Geneticists, which represent the majority of scientists using the surveyed research clouds, specifically demonstrate such software demands. These users employ a science cloud to handle initial data processing using QIIME or Bowtie [41] followed by local data exploration using a mix of data integration, machine learning, and visualization software (Tables 4, 5). Intermediate data products are presumed to be manually passed between these distinct software solutions. SWMSs should accommodate data processing pipelines spanning multiple computational levels, from large-scale analysis in cloud or grid environments to medium and small-scale processing of data intermediates on local computers. This is a distinct solution from such cloud workflow systems as SWF or Bluemix, as it would allow scientists to exploit advanced infrastructure without locking them in to that infrastructure and preventing the use of their work in other contexts.

The use of sophisticated virtualized infrastructure is not usually well-documented in research papers and may lead to long-term sustainability problems for science clouds. Domain-specific scientific papers do not cite the papers referencing cloud resources or the software and middleware behind these computing resources. The resources used to store and curate data are also not mentioned in these papers. Usage of cloud resources is not directly mentioned aside from occasional thanks to technical staff in the acknowledgments section of papers. In addition to software reproducibility as previously mentioned, hardware reproducibility also becomes a serious issue since both are critical in ensuring other scientists are capable of implementing a given data pipeline in their own research. A lack of hardware citation further diminishes the profile and perceived importance of scientific research infrastructure, which may have implications for the future sustainability of advanced infrastructures dedicated to data-intensive science.

5. SUMMARY

The primary focus of the paper is to provide usage information for SWMSs in the cloud computing era, as well as information about the other software employed in scientific cloud environments. By surveying 1455 citations of SWMSs, we provide a snapshot of the scientific impact of leading SWMSs, as well as applications and tools developed for these systems. We also document how these workflows are being used by scientists. A software usage survey of two major science clouds, OSDC and Red Cloud is also presented. From this information, we provide a set of design requirements necessary if SWMSs are to be effectively used by researchers in data-intensive science.

5.1 Future Directions

Software usage surveys are critical in understanding how individuals are using software, and thus how to produce effective tools to enable new research. The primary unit of usage employed here is the publication, which may exhibit significant bias depending on the scientific domain and the type of software used, especially considering that software is generally very poorly documented in primary research publications. Another facet of this problem is that generalist SWMS's could be considered akin to programming languages. Much as a researcher does not consider it worthwhile to cite Python due to its perception as a ubiquitous,

open resource, they might also not consider Taverna worth mentioning.

A significantly more comprehensive usage survey could be conducted by interviewing data-intensive science laboratories and determining how enabling software is employed, how software solutions are developed, and how software is recorded for future use. A weakness of studying publications is an inability to see in detail the data-intensive scientific exploration process, which is invaluable in practice to design effective exploratory tools. Such a survey would provide high resolution use-cases for scientific software and significantly improve the applicability of scientific enabling software like SWMSs.

5.2 Conclusions

Data-intensive sciences have dramatically shifted the computational landscape in the areas of data size, technology, methodology, collaboration, and quality demands. Grid era SWMSs are not being significantly used as enabling technologies to generate scientific advancement due to these altered design requirements. Instead, data-intensive scientific clouds tend to use specialized, domain-specific software packages in lieu of enabling technologies developed by the computer science community. However, there is still a case to be made for SWMSs that focus on providing a highly interoperable framework for connecting diverse software packages that consider multi-user interaction, data exploration and low-effort software reproducibility.

6. ACKNOWLEDGMENTS

This research was made possible by the OSDC Partnerships for International Research and Education (PIRE) Program, funded by NSF Award #1129076. The work is also partially supported by European Union's Horizon 2020 research and innovation program under grant agreements No. 654182 (ENVRI$^{\text{PLUS}}$ project), 643963 (SWITCH project), and 676247 (VRE4EIC project).

7. REFERENCES

[1] Big data, analytics and the path from insights to value. http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/, 2010. Accessed: 2015-07-27.

[2] Rapidminer from Rapid-I at CeBIT 2010. http://www.data-mining-blog.com/cloud-mining/rapidminer-cebit-2010/, 2010. Accessed: 2015-07-22.

[3] How companies learn your secrets. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1\&_r=2\&hp, 2012. Accessed: 2015-07-20.

[4] Accelrys software. http://accelrys.com/, 2015. Accessed: 2015-7-27.

[5] Amazon simple workflow service (SWF). http://aws.amazon.com/swf/, 2015. Accessed: 2015-7-29.

[6] Cornell faculty, staff, and students. https://www.cac.cornell.edu/clients/cu.aspx, 2015. Accessed: 2015-07-5.

[7] ESFRI. http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri, 2015. Accessed: 2015-7-27.

[8] Executive summary data growth, business opportunities, and the IT imperatives. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm, 2015. Accessed: 2015-7-31.

[9] IBM Bluemix. http://www.ibm.com/cloud-computing/bluemix/, 2015. Accessed: 2015-7-29.

[10] OSDC publications. https://www.opensciencedatacloud.org/publications, 2015. Accessed: 2015-07-5.

[11] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, and J. Taylor. Harnessing cloud computing with Galaxy Cloud. *Nature biotechnology*, 29(11):972–4, Nov. 2011.

[12] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: an extensible system for design and execution of scientific workflows. *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, 2004.

[13] L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo. VisTrails: Enabling Interactive Multiple-View Visualizations. In *VIS 05. IEEE Visualization, 2005.*, pages 135–142. IEEE, 2005.

[14] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.

[15] D. R. Blair, K. Wang, S. Nestorov, J. A. Evans, and A. Rzhetsky. Quantifying the impact and extent of undocumented biomedical synonymy. *PLoS computational biology*, 10(9):e1003799, Sept. 2014.

[16] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–6, May 2010.

[17] B. Chassaing, O. Koren, J. K. Goodrich, A. C. Poole, S. Srinivasan, R. E. Ley, and A. T. Gewirtz. Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. *Nature*, 519(7541):92–96, 2015.

[18] Y. Chen, P. Martin, B. Magagna, H. Schentz, Z. Zhao, A. Hardisty, A. Preece, M. Atkinson, R. Huber, and Y. Legré. A Common Reference Model for Environmental Science Research Infrastructures. *Proceedings of the 27th Conference on Environmental Informatics 2013*, pages 665–673, 2013.

[19] D. Churches, G. Gombas, A. Harrison, J. Maassen, C. Robinson, M. Shields, I. Taylor, and I. Wang. Programming scientific and distributed workflow with Triana services. *Concurrency Computation Practice and Experience*, 18(10):1021–1037, 2006.

[20] J. Collins. *The fourth paradigm.* 2014.

[21] M. C. Cushing, J. T. Rayner, and W. D. Vacca. An Infrared Spectroscopic Sequence of M, L, and T Dwarfs. *The Astrophysical Journal*, 623(2):1115–1140, Apr. 2005.

[22] C. F. Davis, C. J. Ricketts, M. Wang, L. Yang, A. D. Cherniack, H. Shen, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer cell*, 26(3):319–30, Sept. 2014.

[23] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107, Jan. 2008.

[24] E. Deelman, J. Blythe, Y. Gil, and C. Kesselman. Pegasus: Mapping scientific workflows onto the grid. *Grid Computing*, 3165/2004:131–140, 2004.

[25] H. Dharuri and P. Henneman. Automated workflow-based exploitation of pathway databases provides new insights into genetic associations of metabolite profiles. *BMC ...*, 2013.

[26] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science (New York, N.Y.)*, 330(6012):1775–87, Dec. 2010.

[27] Y. Gil, V. Ratnakar, and E. Deelman. Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows. *IAAI'07, the 19th national conference on Innovative applications of artificial intelligence, Volume 2*, pages 1767–1774, 2007.

[28] J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, 2010.

[29] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, et al. Human Genetics Shape the Gut Microbiome. *Cell*, 159(4):789–799, Nov. 2014.

[30] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, et al. Human Genetics Shape the Gut Microbiome. *Cell*, 159(4):789–799, 2014.

[31] I. Gorton, P. Greenfield, A. Szalay, and R. Williams. Data-intensive computing in the 21st century. *Computer*, 41(4):30–32, 2008.

[32] R. L. Grossman, Y. Gu, J. Mambretti, M. Sabala, A. Szalay, and K. White. An overview of the Open Science Data Cloud. *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing - HPDC '10*, page 377, 2010.

[33] J. Hoekman, K. Frenken, and R. J. Tijssen. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, 39(5):662–673, June 2010.

[34] H. H. Huang and H. Liu. Big Data Machine Learning and Graph Analytics : Current State and Future Challenges. *2014 IEEE International Conference on Big Data*, pages 31–32, 2014.

[35] V. Kalavri and V. Vlassov. MapReduce: Limitations, Optimizations and Open Issues. In *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pages 1031–1038. IEEE, July 2013.

[36] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[37] G.-H. Kim, S. Trimi, and J.-H. Chung. Big-Data Applications in the Government Sector. *Communications of the ACM*, 57(3):78–85, 2014.

[38] H. S. Kim, S. C. In, and H. Y. Yeom. A task pipelining framework for e-science workflow management systems. *Proceedings CCGRID 2008 - 8th IEEE International Symposium on Cluster Computing and the Grid*, pages 657–662, 2008.

[39] R. Kitchin. The real-time city? Big data and smart urbanism. *GeoJournal*, 79(1):1–14, Nov. 2013.

[40] R. Kittler, J. Zhou, S. Hua, L. Ma, Y. Liu, E. Pendleton, C. Cheng, M. Gerstein, and K. P. White. A comprehensive nuclear receptor network for breast cancer cells. *Cell reports*, 3(2):538–51, Feb. 2013.

[41] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, Apr. 2012.

[42] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, and B. Moon. Parallel data processing with MapReduce: A survey. *ACM SIGMOD Record*, 40(4):11, Jan. 2012.

[43] C. K.-s. Leung. Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data. 2014.

[44] J. Liu, E. Pacitti, P. Valduriez, and M. Mattoso. Parallelization of Scientific Workflows in the Cloud. 2014.

[45] J. Liu, E. Pacitti, P. Valduriez, and M. Mattoso. A Survey of Data-Intensive Scientific Workflow Management. *Journal of Grid Computing*, 2015.

[46] D. Mandl, S. Frye, P. Cappelaere, M. Handy, F. Policelli, M. Katjizeu, et al. Use of the earth observing one (EO-1) satellite for the namibia sensorweb flood early warning pilot. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2):298–308, 2013.

[47] S. H. I. Meilin, Y. Guangxin, X. Yong, and W. U. Shangguang. Workflow management systems: a survey. *Communication Technology Proceedings, International Conference on*, 2:1 – 6, 1998.

[48] P. S. Muirhead, J. Becker, G. a. Feiden, B. Rojas-Ayala, A. Vanderburg, E. M. Price, et al. CHARACTERIZING THE COOL KOIs. VI. H - AND K -BAND SPECTRA OF KEPLER M DWARF PLANET-CANDIDATE HOSTS. *The Astrophysical Journal Supplement Series*, 213(1):5, 2014.

[49] M. Roser. Technological progress. http://ourworldindata.org/data/technology-and-infrastructure/moores-law-other-laws-of-exponential-technological-progress/, 2015. Accessed: 2015-7-31.

[50] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science (New York, N.Y.)*, 330(6012):1787–97, Dec. 2010.

[51] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. Computational solutions to large-scale data management and analysis. *Nature reviews. Genetics*, 11(9):647–57, Sept. 2010.

[52] Z. Su, P. P. Labaj, S. Li, J. Thierry-Mieg, D. Thierry-Mieg, W. Shi, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, Aug. 2014.

[53] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, et al. The Taverna workflow suite: designing and executing workflows of Web

Services on the desktop, web or in the cloud. *Nucleic acids research*, 41(Web Server issue):W557–61, July 2013.

[54] J. Y. and R. Buyya. A Taxonomy of Workflow Management Systems for Grid Computing. *Journal of Grid Computing*, 3(3-4):171–200, 2005.

[55] Y. Zhao, Y. Li, S. Lu, I. Raicu, and C. Lin. Devising a cloud scientific workflow platform for big data. *Services (SERVICES), 2014 ...*, 2014.

[56] Y. Zhao, I. Raicu, and I. Foster. Scientific Workflow Systems for 21st Century, New Bottle or New Wine? In *2008 IEEE Congress on Services - Part I*, pages 467–471. IEEE, July 2008.

[57] Y. Zhao, Y. Zhang, W. Tian, R. Xue, and C. Lin. Designing and deploying a scientific computing Cloud platform. *Proceedings - IEEE/ACM International Workshop on Grid Computing*, pages 104–113, 2012.

[58] F. Zulkernine, P. Martin, Y. Zou, M. Bauer, F. Gwadry-Sridhar, and A. Aboulnaga. Towards cloud-based analytics-as-a-service (CLAaaS) for big data analytics in the cloud. *Proceedings - 2013 IEEE International Congress on Big Data, BigData 2013*, pages 62–69, 2013.