A Computing Environment to Support Repeatable Scientific Big Data Experimentation of World-Wide Scientific Literature

Bob G. Schlicher[1], James J. Kulesz[2], Robert K. Abercrombie[3], and Kara L. Kruse[4]

[1] *schlicherbg@ornl.gov,* [2] *jim.kulesz@gmail.com,* [3] *abercrombier@ornl.gov,* [4] *krusekl@ornl.gov*
Oak Ridge National Laboratory, Computational Sciences and Engineering Division, 1 Bethel Valley Road, Oak Ridge, TN 37830-6085 (USA)

Abstract

A principal tenet of the scientific method is that experiments must be repeatable. This tenet relies on *ceteris paribus* (i.e., all other things being equal). As a scientific community, involved in data sciences, we must investigate ways to establish an environment where experiments can be repeated. We can no longer allude to where the data comes from, we must add rigor to the data collection and management process from which our analysis is conducted. This paper describes a computing environment to support repeatable scientific big data experimentation of world-wide scientific literature, and recommends a system that is housed at the Oak Ridge National Laboratory in order to provide value to investigators from government agencies, academic institutions, and industry entities. The described computing environment also adheres to the recently instituted digital data management plan, which involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation, as mandated by multiple United States government agencies. It particularly focuses on the sharing and preservation of digital research data. The details of this computing environment are explained within the context of cloud services by the three layer classification of "Software as a Service", "Platform as a Service", and "Infrastructure as a Service".

Conference Topic

Science policy and research assessment, Methods and techniques

Introduction[1]

The scientific policy and research assessment community is investigating methods and techniques to establish an environment where experiments can be repeated through the use of data management. This approach attempts to ensure the integrity of scientific findings and the processes from which scientific literature analysis is conducted.

Data Science is the study of the generalizable extraction of knowledge from data (Dhar, 2013). From this definition, scientific development thus becomes the piecemeal process by which these items have been added, singly and in combination, to the ever growing stockpile that constitutes scientific technique and knowledge (Kuhn, 1970). Scientific literature analysis, or Scientometrics, is the study of measuring and analysing science, technology and innovation. Organizations, such as Thomson Reuters, have long used these analyses to identify the most influential papers or researchers in a field. Recently, Foresight and Understanding from Scientific Exposition (Murdick, 2011) takes this further by mining millions of papers and patents in both English and Chinese, two of the most commonly used languages in scientific literature (Readron, 2014).

[1] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Scientometrics and its related research activities in today's world make extensive use of digital research data. The data management of this digital research data is, in essence, the quintessential requirement for repeatable scientific experimentation. This term, digital research data, encompasses a wide variety of information stored in digital form including: experimental, observational, and simulation data, codes, software and algorithms, text, numeric information, images, video, audio, and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analysed data, and published and archived data ("Statement on Digital Data Management," 2014). More specifically, research data are defined in regulation ("Intangible property - Code of Federal Regulations 2 CFR 200.315," 2014), continuing the definition in further statues and United States Government Directives ("2 CFR 215 - Uniform Administration Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular A-110) ", 2012) as follows:

- "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This 'recorded' material excludes physical objects (e.g., laboratory samples). Research data also do not include:
    - Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
    - Personnel and medical information and similar information, which the disclosure would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

Purpose of the Study

When addressing the reality of allocating the scarce resources of the current research budget constraints, the current institutions of science today operate, essentially the same, as from the time period just after the Second World War (Azoulay, 2012). Azoulay further argues it would be a fortuitous coincidence if the systems that served us so well in the twentieth century were equally adapted to twenty-first-century needs. Such is not the case. To leverage these finite resources and to adhere to the principle of the scientific method that all experiments must be repeatable, we, as a scientific community must investigate ways to establish environments where experiments can be repeated. We can no longer allude to from where the data come, we must add rigor to the data collection and data management process from which our analysis is conducted.

Data management involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation. The focus of this statement is the sharing and preservation of digital research data. The following principles apply to the effective management of digital research data ("Statement on Digital Data Management," 2014):

- Effective data management has the potential to increase the pace of scientific discovery and promote more efficient and effective use of government funding and resources. Data management planning should be an integral part of research planning.
- Sharing and preserving data are central to protecting the integrity of science by facilitating validation of results and to advancing science by broadening the value of research data to disciplines other than the originating one and to society at large. To the greatest extent and with the fewest constraints possible, and consistent with the requirements and other principles of this statement, data sharing should make digital

research data available to and useful for the scientific community, industry, and the public.

- Not all data need to be shared or preserved. The costs and benefits of doing so should be considered in data management planning.

Procedure for a Computing Environment to Support Repeatable Scientific Big Data Experimentation

A data management plan is a formal document that outlines how a research institution and program will handle data both during research and after the project is completed ("Data management plan," 2014). The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins. This ensures that data are well-managed in the present and prepared for preservation in the future. Multiple United States government agencies now require proposals submitted to include a supplementary document labelled “Data Management Plan” (Collins, 2014; "Dissemination and Sharing of Research Results," 2010). These supplementary documents describe how the proposal will conform to scientific policy on the dissemination and sharing of research results.

FUSEnet is a data analytics cloud specializing in managing both data and computational processes for assessing technical knowledge for identifying emergent technologies and capabilities. Under a multi-year United States Government research effort sponsored by Intelligence Advanced Research Projects Activity (IARPA), the overall goal of the FUSE program is to produce a new capability to accelerate the process of identifying and prioritizing emerging technologies across the globe (Murdick, 2011). The FUSE Program was established to develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technical emergence using information found in published scientific, technical, and patent literature. A concise description is as follows (Murdick, 2011):

> A fundamental hypothesis of the FUSE Program is that real-world processes of technical emergence leave discernible traces in the public scientific, technical, and patent literature. FUSE envisions a system that can (1) process the massive, multi-discipline, growing, noisy, and multilingual body of full-text scientific, technical, and patent literature from around the world; (2) automatically generate and prioritize technical terms within emerging technical areas, nominate those that exhibit technical emergence, and provide compelling evidence for the emergence; and (3) provide this capability for literature in English and at least two non-English languages. Technology developed from the FUSE Program would automatically nominate both known and novel technical areas based on quantified indicators of technical emergence with sufficient supporting evidence and arguments for that nomination. The FUSE Program also addresses the vital challenge of validating such a system, using real world data.

FUSEnet is currently a government system hosted by ORNL that stores unclassified, copyright-protected scientific information and provides remote access for approved users to analyse the stored data within a cloud computing environment to satisfy the research objectives of the IARPA FUSE Program. A key tenet within FUSEnet is that data integrity and availability is maintained. An ORNL developed “data diode” embedded within FUSEnet gateways allows access to protected data, but prevents data removal by users. As necessary, a mechanism for approved data export is built into the system architecture. Also by design, the activities and work products of individual user teams are segregated from each other in the cloud computing virtual environment.

FUSEnet Capabilities

The FUSEnet computing environment is based on the Cloud service model. These models are usually described by a three layer classification called SPI for **S**aaS, **P**aaS, and **I**aaS (Tian & Zhao, 2015) and adapted as follows:

- SaaS – Software as a Service: applications that are available on-demand.
- PaaS – Platform as a Service: refers to a computing platform of software components and middleware that are used by end-users to develop and manage their cloud applications. Typically, cloud providers at this layer offer databases, web servers, development environments, and application monitoring tools.
- IaaS – Infrastructure as a Service: physical or virtual machines with access to data storage and other operating system services. The cloud user is typically expected to install and maintain operating-system images.

The unique processing capabilities of FUSEnet are in the SaaS and PaaS levels. The IaaS capabilities were established with off-the-shelf software and hardware solutions as a result of understanding the operational needs of FUSEnet users, big data analytics, and optimizing central processing unit (CPU) and input/output (I/O) performance. One of the major challenges with the computing environment is with moving large volumes of data (terabytes) to and from the disk storage to the CPUs for processing. This challenge is met with ever increasing improvements and replacements for the IaaS without having any operating impact on the SaaS or PaaS layers. FUSEnet demonstrated this with an improvement in the data I/O transfer by replacing the disk storage system over its earlier version. Further, FUSEnet SaaS and PaaS software can be hosted on commercial IaaS platforms that meet the requirements for its intended usage.

A summary of the FUSEnet benefits and capabilities that support repeatability of big data experiments includes:

- An organized repository of 100 million published scientific and patent documents,
- Technical in-house expertise for maintenance of data pertaining to integrity and availability, pedigree, and version control,
- Reliable data sources including data provided by, Thomson Reuters, Lexis-Nexis, Elsevier, Institute of Electrical and Electronics Engineers (IEEE), Nature Publishing Group, PubMed Central, and others,
- Technical expertise with the format and details of the data, and
- Four analytical software applications with evidentiary traceability and indicators for assessing repeatability:
    - Assess and forecast technical research and technology developments,
    - Reverse-search the events contributing to a technology or development,
    - Drill down the evidence supporting the assessment and forecast,
    - Remote end-user workspaces ready-to-run the applications and the analytics platform,
    - Multiple analytics capabilities including Natural Language Processing (NLP), Parts-of-Speech (PoS) detectors, deduplication, belief network modelling, and machine learning,
    - Operation of the system with 24/7 and 99.8% availability within domain-specific expertise with the current ORNL technical staff,
    - Rapid custom development to meet unique end-user analytics requirements, and
    - Immediate data protection for the repository and custom end-user data.

The FUSEnet SaaS Level

At the SaaS level, four unique software applications perform automated technical assessments for supporting the detection and forecasting. Each of these applications process and analyse published scientific and engineering papers that are made available in the FUSEnet data repository. Unlike previous approaches to detecting emergence, which are based on the citation analysis of papers and patents (Bettencourt et al., 2008; Huang et al., 2014), the following application systems extract information from the text of publications and patents, identifying authors, their affiliations, addresses, as well as classifying types of organizations and publications. Although these applications have the same objectives, their analytical techniques are uniquely different and hence provide different insights into the organization and search of the data (Babko-Malaya et al., 2013). These analysis techniques include: feature extraction (Michaelis et al., 2012), time series analysis, sentiment and network analysis (Fürstenau & Rambow, 2012), and emergent detection and prediction (Brock et al., 2012), among others. The four main applications developed within the FUSEnet system are ARBITER from BAE Systems, Copernicus from SRI International, Emerge from BBN, and DETAiLS from Columbia University.

The FUSEnet PaaS Level

The aforementioned SaaS applications use a variety of tools and libraries at the PaaS level. While the SaaS level in FUSEnet is the automated assessment, the FUSEnet PaaS computing platform can best be described as a "Network Analysis" (Otto & Rousseau, 2002) and text analytics platform. Text analysis uses statistical pattern learning to find patterns and trends from text data (in our case, scientific literature and patents). A summary of several key tools that FUSEnet provides are in Table 1. A selection of software libraries for network analysis and text analysis in FUSEnet, available for ensuing that experiments can be repeated, is shown in Table 2.

The FUSE Program licensed and installed a large number of scientific papers and patents from several suppliers in multiple languages including English and Chinese. The data sets include bibliographic citations of journal articles (108+ million), full text journal articles (5+ million), patent backfile records (14+ million at beginning of 2013 for the US and China), and updates to the patent backfile records, (51+ million for the US and China). A backfile is a single file containing the original patent application data plus all updates to the patent (both by the originator and by the patent office) up to the time the backfile was created.

Fig. 1 illustrates the large increase in scientific journal articles and patent applications as included in the FUSE research system during the past two decades. The number of Chinese patent applications is increasing dramatically and has now surpassed the number of US patent applications. Also, the number of Chinese journal articles is increasing at a rate faster than the rest of the world.

Table 1. FUSEnet PaaS support software packages.

	FUSEnet PaaS Analytics Tool	*Technical Usage*	*SaaS application that uses it*
1	MySQL[2]	SQL[3] database typically used to store document, term, and author data.	Emerge, ARBITER
2	MongoDB[4]	Document-oriented, NoSQL database used to store extracted entities and indicator-specific data.	Emerge, Copernicus
3	MALLET	Machine Learning and NLP[5] Toolkit for Java. Provides topic modelling for document clustering.	Emerge
4	Sofia-ml	Fast incremental machine learning algorithm. Provides clustering of documents from topic models generated by MALLET.	Emerge
5	Lucene IR system	Used for its indexing engine.	Emerge
6	Scikit-learn	Machine learning models.	Emerge
7	Tomcat/Solr Web Server	Used for Term indexing.	ARBITER
8	Apache ActiveMQ[6]	Messaging and integration patterns.	ARBITER
9	Cassandra	NoSQL database.	ARBITER
10	Virtuoso	RDF[7] triple storage.	ARBITER
11	OpenRDF/Sesame	RDF processing including parsing, storing, reasoning and querying.	ARBITER
12	Spring Framework	Used for Integration using JMS.	ARBITER
13	Lucene/Solr	Document level information search, retrieval and storage engine.	ARBITER, DETAiLS
14	Open NLP	Machine learning based toolkit for processing natural language text.	ARBITER
15	Netica	Used for working with belief networks and influence diagrams.	ARBITER
16	Elasticsearch	Extension on Lucene that provides search and analytics.	Copernicus
17	Hadoop 2+	Used for extract, transform, and load (ETL) and de-duplication processing.	Copernicus
18	Berkeley Parser	Sorts and assigns words in sentences into subjects, verbs, and objects.	DETAiLS
19	Duke	Deduplication engine written in Java operating with Lucene.	DETAiLS
20	Stanford Chinese Word Segmenter	Split Chinese text into a sequence of words.	DETAiLS
21	Stanford Part-of-Speech (POS) Tagger	Reads text and assigns parts of speech to each word (noun, verb, adjective, etc.).	DETAiLS
22	UIMA	Unstructured Information Management Architecture (UIMA) is a general framework for analysis of unstructured information and its integration with search technologies.	DETAiLS
23	Weka	Machine learning software written in Java for data analysis and predictive modelling.	DETAiLS

[2] MySQL is a well-known relational database manager used in a wide variety of systems, including Twitter, Wikipedia, Facebook, Google, Wordpress, and countless more websites and other applications.
[3] SQL (Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS),
[4] MongoDB is a document-oriented, NoSQL database.
[5] NLP is Natural Language Processing where algorithms are used to derive meaning from human language.
[6] Apache ActiveMQ is an open source message broker written in Java together with a full Java Message Service (JMS) client.
[7] RDF is Resource Description Framework and is used to express data in subject-predicate-object expressions.

Table 2. Subset of FUSEnet software libraries for social network and text analysis.

	Library/Package	*Description*	*SaaS application that uses it*
1	Arpack	Linear algebra routines for Java	Emerge
2	JDOM	XML processing library for Java	Emerge
3	Jwnl	Java WordNet library	Emerge, ARBITER
4	Matrix-toolkits-java	Linear algebra data structures for Java	Emerge
5	BLAS	Linear algebra subroutines	Emerge
6	LAPACK	Linear algebra data structures and subroutines	Emerge
7	Libquadmath	High-precision math libraries	Emerge
8	Beanshell	Scripting for Java	Emerge
9	Trove4j	High-performance data structures for Java	Emerge
10	JGrapht	Graphical data structures and algorithms for Java	Emerge
11	JUNG	Java Universal Network/Graph Framework	ARBITER
12	R	Development environment for statistical computing and graphics	ARBITER

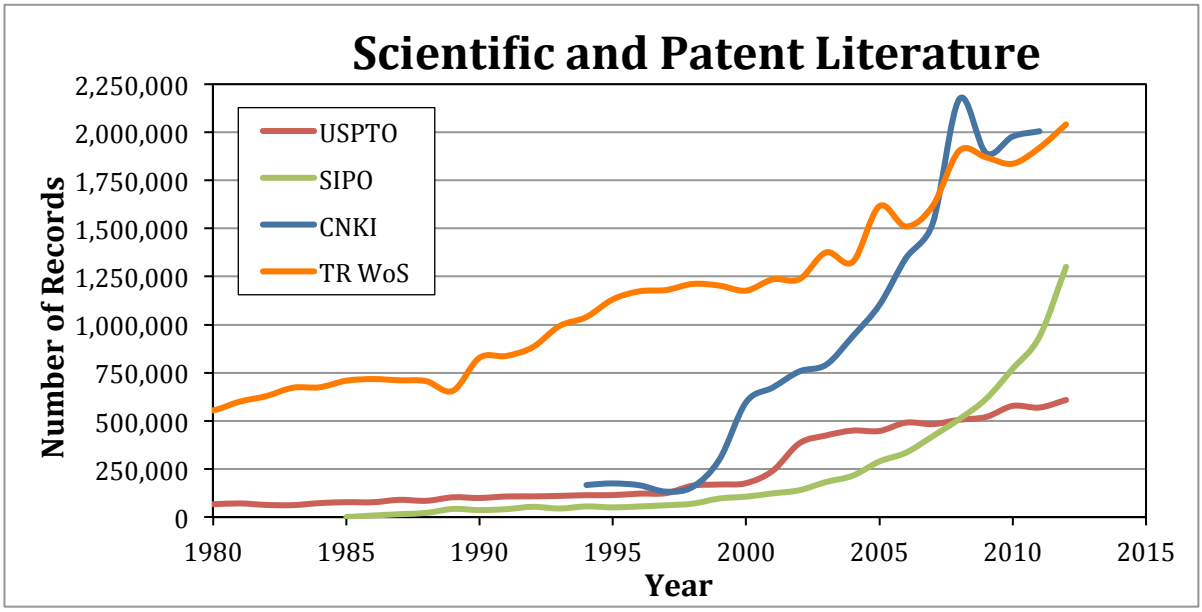

Figure 1. Number of records per year for the four largest datasets in the FUSEnet collection including patent records from the US (USPTO) and Chinese (SIPO) patent offices (i.e. number of backfile records at the beginning of 2013) and journal article citations from China (CNKI) and Thomson Reuters' Web of Science (TR WoS).

The FUSEnet IaaS Level

The deployed second generation FUSEnet at ORNL has the following summary specifications:

- 770 gigaFLOPS[8] of maximum performance,
- 16 blade servers (plus 2 support blades), each with 2 CPUs, each with 6 cores, totalling 192 cores, or processors; additional blade with USB 3.0 for dedicated data transfer/export,
- 3.07 TB of RAM w/ 192 GB per node,
- Disks:
    - EMC Isilon: 340 TB (useable; includes 6.4 TB SSD) running NFS over 10 Gb/s Ethernet,
    - HP LeftHand: 260 TB of effective disk storage; will be reconfigured for backup and
    - Isilon disk I/O up to 1 gigabyte/sec per blade,
- Networking: Flex-10 modules totalling 160 Gbits/sec bandwidth per enclosure x 2 enclosures (theoretical maximum),
- Virtualized computing space through VMware[9],
- Access and control policies enforced by ORNL Computing Data Center, and
- Call Center and metrics for service quality.

Table 3. Characteristics of cloud providers and applicability to FUSEnet requirements.

	Vendors	*Cloud Offering Overview*	*Applicability to FUSEnet*
1	Amazon Web Services	Overall market leader offering virtual servers, MapReduce (Hadoop) for search engine, large data storage, SQL databases, NoSQL databases, mobile integration, business applications including email, payment systems, and workflow.	PaaS (databases), IaaS
2	Google Cloud Platform	App Engine web application platform (PaaS), virtual machines, file storage, SQL databases, NoSQL, big dataset support, mobile integration.	PaaS (databases, web apps), IaaS
3	IBM SmartCloud	SaaS including data warehousing and analytics, business analytics engine, business process management, financial modelling, payment systems, medical analysis, social media analysis, transportation management, medical analytics, SQL databases, NoSQL databases, mobile integration.	SaaS (social media analysis), PaaS (databases, web apps), IaaS
4	Microsoft Azure	Windows or Linux virtual machines, messaging, scheduling, SQL databases, NoSQL databases, mobile integration.	PaaS (databases), IaaS
5	Rackspace Cloud	High bandwidth networking, virtual machines, data storage, process load balancing.	IaaS

Analysis of Technical Requirements and Alternatives versus Commercial Cloud Providers

Representative current cloud solution offerings from commercial vendors include but are not limited to the following: Amazon Web Services (AWS), IBM SmartCloud, Microsoft Azure, Google Cloud Platform, and Rackspace Cloud Servers. Considering the data management, experimentation requirements and the strategic issues, the question arises, "Are the IaaS and

[8] In computing, FLOPS (for FLoating-point Operations per Second) is a measure of computer performance, useful in fields of scientific calculations that make heavy use of floating-point calculations. For such cases, it is a more accurate measure than the generic instructions per second. Computers capable of performing greater than 1 Giga FLOPS are termed as supercomputers.

[9] VMware, Inc. is a software company that provides cloud and virtualization software and service.

PaaS from these selected vendors sufficient for hosting and maintaining the FUSEnet SaaS and PaaS?" A summary of the cloud providers and the offering are described in Table 3.

Analysis of SaaS Technical Alternatives

FUSEnet consists of four unique technical emergence software applications. Current cloud providers are not in the business of providing this niche capability. Cloud providers offer more general SaaS services such as Enterprise Resource Planning (ERP), general accounting, medical, and financial applications for managing business administration operations. If FUSEnet were to be employed on a 3rd party cloud, unique, domain-specific expertise would be required to operate and manage the FUSEnet software applications.

Analysis of PaaS Technical Alternatives

FUSEnet consists of several framework and middleware solutions combined with math-based libraries that are unique to network and text analysis. With the exception of IBM SmartCloud, current cloud providers are not in the business of exclusively providing this niche capability. Cloud providers offer more general PaaS software such as databases, email, and web servers. The features of the network and social analytics tools in SmartCloud should be further evaluated.

Analysis of IaaS Technical Alternatives

FUSEnet is operated in a secured, cloud environment at the Data Computing Center at ORNL. It currently operates on the hardware infrastructure described above. This FUSEnet hardware was performance tested to determine its disk I/O (input/output) throughput under various load conditions. Software programs were used to perform these tests at a low level or 'raw' I/O set of read and write tests and at the application layer with tests that simulated application disk usage. From these initial test results and further repeated testing, the FUSEnet disk I/O was optimized for handling the volume and type of data used in the system. Further tests were performed to compare FUSEnet with another commercial cloud offering, which demonstrated similar or better performance for FUSEnet depending on the operating conditions selected. Currently, the FUSEnet storage system is in its second generation as a result of these performance tests and evaluations. The FUSEnet software and data can be operated on 3rd party (IaaS) environments that can meet the overall system requirements as follows:

- Handle big data that is mixed structured and unstructured and continuously growing.
- Protect selected data and apps (commercial, proprietary) that remain in the cloud.
- Rapidly deploy software solutions to the data.
- Provide virtualization for operating systems including common Linux distributions, Windows and Mac OS.
- Rapidly ingest data into the system.
- Provide the computing performance involving big data analytics software services.
- Provide an easy-to-use big data analytics platform.
- Provide high-performance big data storage and retrieval up to 500 TBs and continue to scale.
- Provide robust, state-of-the-practice cyber security.

In general, commercial firms are advised to consider strategic issues with regards to cloud scope, service levels, and deployment needs. For the FUSEnet environment, Table 4 summarizes these strategic concerns.

The overall need for a secured FUSEnet environment involves the capability to employ software services, such as the analytics described earlier, that uses the data within the FUSEnet cloud, but cannot copy the data out of the cloud. FUSEnet is equipped with custom

middleware software within the PaaS called a Data Diode that monitors activities and prevents the exfiltration of data. Thus, the commercial and proprietary data is protected from being taken outside the FUSEnet enclave (Abercrombie, MacIntyre, & Schlicher, 2011). The Data Diode involves a change to the Linux distro (distribution)[10] so that an IaaS provider must approve the customer to host their own virtualized and configurable operating system (MacIntyre, Paul, & Schlicher, 2011).

Table 4. Strategic issues for the FUSEnet environment.

	Strategic Issue	*Description*	*Assessment for FUSEnet*
1	Cloud Scope – what is the design to meet the need?	Identifies the availability, performance, and security needs; sufficient and planned computing power, storage, and bandwidth.	FUSEnet is monitored daily and reported monthly with the current operational stats: Availability: 99.8%; CPU usage: 12-18%; Memory usage: 56-65%; Storage usage: 69%. FUSEnet is installed with a Data Diode that protects against data exfiltration of its repository. FUSEnet is a virtual environment with separated computing enclaves. Each user or user group within an enclave has the freedom to compose and perform their needed computational research without directly impacting other users.
2	Service Levels	Identifies the expected workload, admin support, service delivery needs, timing and I/O response.	FUSEnet Test and Evaluation (T&E) simulates heavy end-user loading. This is measured to be an increase of 5-10% of the daily load. For its initial usage, FUSEnet could simultaneously host 3-4 heavy end-users loading. The Admin support is at two levels: operating system and the virtual layer through VMware.
3	Deployment Needs	Identifies the integration needs with infrastructure services.	FUSEnet operates on VMware that isolates the PaaS from dependencies on the hardware and the Operating System. The current FUSEnet system, including the number of cores, performance of the cores, memory, and the Isilon storage, is a proven baseline for simultaneously hosting 3-4 heavy end-user loading.

Discussion and Conclusions

This paper addresses science policy with a method and a technique to assess research, increasing its value to the US national scientific community by making available a computing environment to support repeatable scientific big data experimentation of world-wide scientific literature. The computational capability ensures the integrity, availability and confidentiality of new technologies and new technical knowledge. This will position scientific investigators (academic, commercial, and government) with an advantage to address the technical and political challenges all three entities face. FUSEnet offers this unique capability and this paper describes a computing environment necessary to support repeatable experimentation, and recommends a system that is housed at the ORNL Data Center in order to provide value to investigators from a variety of sources while adhering to recently mandated Data Management Planning.

[10] A Linux distribution (often called a distro for short) is an operating system made as a collection of software based around the Linux kernel and often around a package management system

Acknowledgments

We thank colleagues and other reviewers for their assistance and helpful comments. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Energy (DOE). This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOE, ORNL, or the U.S. Government.

References

2 CFR 215 - Uniform Administration Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular A-110) (2012).

Abercrombie, R. K., MacIntyre, L. P., & Schlicher, B. G. (2011). *Protection of Data in Virtual and Physical Computing Environments (Invention Disclosure Number: 201102659, DOE S-Number: S-124,217)*. Oak Ridge: Oak Ridge National Laboratory.

Azoulay, P. (2012). Research efficiency: Turn the scientific method on ourselves. *Nature, 484*(7392), 31-32.

Babko-Malaya, O., Hunter, D., Amis, G., Meyers, A., Thomas, P., Pustejovsky, J., et al. (2013, May 8-10). *Characterizing Communities of Practice in Emerging Science and Technology Fields.* Paper presented at the 2013 International Conference on Social Intelligence and Technology (SOCIETY).

Bettencourt, L. A., Kaiser, D., Kaur, J., Castillo-Chávez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics, 75*(3), 495-518.

Brock, D. C., Babko-Malaya, O., Pustejovsky, J., Thomas, P., Stromsten, S., & Barlos, F. (2012, November 2-4). *Applied Actant-Network Theory: Toward the Automated Detection of Technoscientific Emergence from Full-Text Publications and Patents.* Paper presented at the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposiurm Social Networks and Social Contagion, Arlington, VA.

Collins, D. (2014). *Knowledge Article: Data Management*. Oak Ridge: Oak Ridge National Laboratory.

Data management plan. (2014). Retrieved June 21, 2015 from: http://en.wikipedia.org/wiki/Data_management_plan

Dhar, V. (2013). Data Science and Prediction. *Commununications of the ACM, 56*(12), 64-73.

Dissemination and Sharing of Research Results. (2010). Retrieved June 21, 2015 from:http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Fürstenau, H., & Rambow, O. (2012). *Unsupervised induction of a syntax-semantics lexicon using iterative refinement*. Paper presented at the First Joint Conference on Lexical and Computational Semantics.

Huang, M.-H., Huang, W.-T., Chang, C.-C., Chen, D.-Z., & Lin, C.-P. (2014). The Greater Scattering Phenomenon Beyond Bradford's Law in Patent Citation. *Journal of the Association for Information Science and Technology, 65*(9), 1917-1928.

Intangible property - Code of Federal Regulations 2 CFR 200.315(2014).

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: University of Chicago Press.

MacIntyre, L. P., Paul, N. R., & Schlicher, B. G. (2011). *Data Diode (Copyright Document Number 90000008)*. Oak Ridge: Oak Ridge National Laboratory.

Michaelis, J. R., McGuinness, D. L., Chang, C., Luciano, J. S., & Hendler, J. (2012). *Applying Multidimensional Navigation and Explanation in Semantic Dataset Summarization*. Paper presented at the 11th International SemanticWeb Conference (ISWC 2012).

Murdick, D. A. (2011). Foresight and Understanding from Scientific Exposition (FUSE). Retrieved June 21, 2015 from: http://www.iarpa.gov/index.php/research-programs/fuse

Otto, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441-453.

Readron, S. (2014). Text-mining offers clues to success. *Nature, 509*, 410.

Statement on Digital Data Management. (2014). *DOE Office of Science Funding Opportunities* Retrieved June 21, 2015 from: http://science.energy.gov/funding-opportunities/digital-data-management/

Tian, W., & Zhao, Y. (2015). Chapter 1 - An Introduction to Cloud Computing. In W. Tian & Y. Zhao (Eds.), *Optimized Cloud Resource Management and Scheduling* (pp. 1-15). Boston: Morgan Kaufmann.