Carey *et al. Integrating Materials and Manufacturing Innovation* (2016) 5:7
DOI 10.1186/s40192-016-0049-0

Integrating Materials and Manufacturing Innovation

CASE STUDY Open Access

Data integration for materials research

Nicholas S. Carey[1*], Tamás Budavári[2,1,3], Nitin Daphalapurkar[3,4] and K. T. Ramesh[3,4]

*Correspondence: ncarey4@jhu.edu
[1] Department of Computer Science, Johns Hopkins University, 3400 N Charles St, 21218 Baltimore, MD, USA
Full list of author information is available at the end of the article

Abstract

Introduction: A new data science initiative in materials research has been launched at The Johns Hopkins University within the Materials in Extreme Dynamic Environments (MEDE) Collaborative Research Alliance (CRA). Our first goal is to build a solution that facilitates seamless data sharing among MEDE scientists. We expect to shorten the design and development cycle of new materials by providing integrated storage, database, and analysis services, building on proven components of the SciServer project developed at the Institute for Data Intensive Engineering and Science (IDIES).

Case description: Here we present our system design and demonstrate the power of our approach through a use-case that enables easy comparison of simulations and measurements. This prototype effort, focusing on boron carbide (BC), brings together multiple materials research elements in the Ceramics group within the MEDE CRA.

Discussion and evaluation: The SciServer platform offers single-sign on access to various general purpose data analysis tools familiar to materials scientists in MEDE. During the case study deployment, users appreciated the simple data file upload process, automated database ingestion, and platform applicability to both students of the art and power users.

Conclusions: From our case study experience in aggregating data from both simulations and physical experiments, we developed a template workflow from which a user may run a common data comparison task outright or customize to another purpose. Next, we turn to acquiring data from more MEDE groups and expanding the user base to the Metals group.

Keywords: Materials research, Data science, Infrastructure

Background

The Materials in Extreme Dynamic Environments (MEDE) Collaborative Research Alliance (CRA) is a multiscale materials research organization consisting of a consortium of major research universities, a national lab, translational institutions, and the Army Research Laboratory and industry, which was established to implement a strategically driven and fundamental science-based paradigm for the development of lightweight protection materials for extreme dynamic environments. The MEDE CRA approach uses advanced experimental techniques, advanced validated modeling and simulation tools, and advanced synthesis, processing, and fabrication capabilities to develop a materials-by-design capability. Our approach is founded on the development of an understanding of the fundamental mechanisms active within extreme dynamic environments at all relevant length and time scales. Using this foundation, our collaborative approach develops

a validated modeling and simulation paradigm for the design, optimization, and fabrication of materials for extreme dynamic environments. A critical part of this approach, of course, is the effective and efficient use of data science. We view this program as one of the major testbeds for the integration of data science into materials research, following the Materials Genome Initiative paradigm. This manuscript presents our system design and demonstrates the power of our approach through a specific use-case.

The MEDE research effort is divided by materials class, with groups of faculty, scientists, and students working together on a model metal system, a model ceramic system, a model polymer system, and a model composite system. Our approach is to develop our methods and protocols first for one of these groups (we chose the Ceramics group) and then extend the approach across to the other materials classes once the essentials of the approach have been worked out and demonstrated within the first group. The Ceramics group in MEDE is working on the boron carbide system, using experiments, materials characterization, modeling, large-scale simulations, and synthesis and processing. The core science approach ensures that all of the scientists in the group are working on the same baseline material and that they all work simultaneously on each new generation of materials as they are developed. Thus the timely and efficient sharing of data and metadata is critical, but this is also challenging given the variety of disciplines that are involved in this materials research effort (a sense of the range of disciplines can be obtained from the MEDE website at hemi.jhu.edu/cmede).

An example of one of the major challenges is provided by the magnitude and variety of the experimental data developed just within the ceramics effort in MEDE. Modern technology provides a dizzying array of high-resolution (in time and in 3D space) experimental information during an impact event. Imaging using high-resolution and ultra-high-speed cameras, computer-aided tomography, and other three-dimensional characterization methods such as serial sectioning provides large datasets on every material sample. Analysis of such datasets provides insights into the complex processes associated with materials undergoing rapid configurational changes (e.g., deformation, failure, phase change) across a wide range of scales. For the most part, these insights are length-scale dependent, but the failure mechanisms are typically multiscale, and information passing and scale bridging are not well-defined within the community in terms of the datasets (as compared to the equations).

As a specific example, the dynamic compressive strength of a centimeter-scale ceramic sample is known to depend on the distribution of micron-scale processing-induced defects in the material. Extensive materials characterization provides the size and orientation distribution of these defects (which are typically carbonaceous inclusions in boron carbide) in the material. This data is provided to both modelers and to those attempting to improve the processing route, and the fidelity of the models is determined by comparing the large datasets generated by simulations using those models with the large datasets obtained from dynamic mechanical experiments using Kolsky bar techniques.

Data sharing in MEDE

With the objective to provide an online platform that assists and accelerates iterative materials engineering across MEDE, the initial step is to outline current research group collaboration and understand the materials research cycle. A high-level description of the data flow between groups help us see the requirements for a data integration service. For

each research group, we collect general information on their procedures and the nature of the data they use and produce as well as the data volumes involved in their research.

An overview of the data dependency in the Ceramics group is illustrated in Fig. 1. They can be sorted into three approximate categories: modeling, processing, and experiments. These groups share a variety of data types including textual simulation results, image-based histories of experiments, processing histories, equation parameters, and recommendations of actors. While currently the heaviest scale of data sharing is imagery on the order of tens to hundreds of gigabytes (GB), several scientists have expressed plans to collaborate with new datasets on the order of terabytes (TB).

With the goal to alleviate materials researchers' data management burdens while introducing minimal changes to existing workflows, it is good practice to design the collaborative platform to ingest data as close to the source as possible and provide familiar platform-based data processing tools. For example, given a Kolsky Bar physical experiment, the data is produced by sensors monitoring the event. In terms of data reliability, provenance, backup recovery, and collaborator access, it is advantageous to immediately upload the raw sensor data to a storage server. A collaboration platform will track any changes and operations performed on the datasets; good data provenance enumerates the steps taken to reach results and is useful for peer review, the verification of data integrity, and is educational for students learning the art. However, researchers' current practices include performing laptop- or personal computer-based analyses and aggregations on the raw data before gaining reportable results. If similar analysis tools could be provided to the researchers via the online collaborative platform, not only would a burden be lifted from the individual researcher's laptop but the workflow would be available for team members to examine and contribute to. We surveyed MEDE scientists and found that

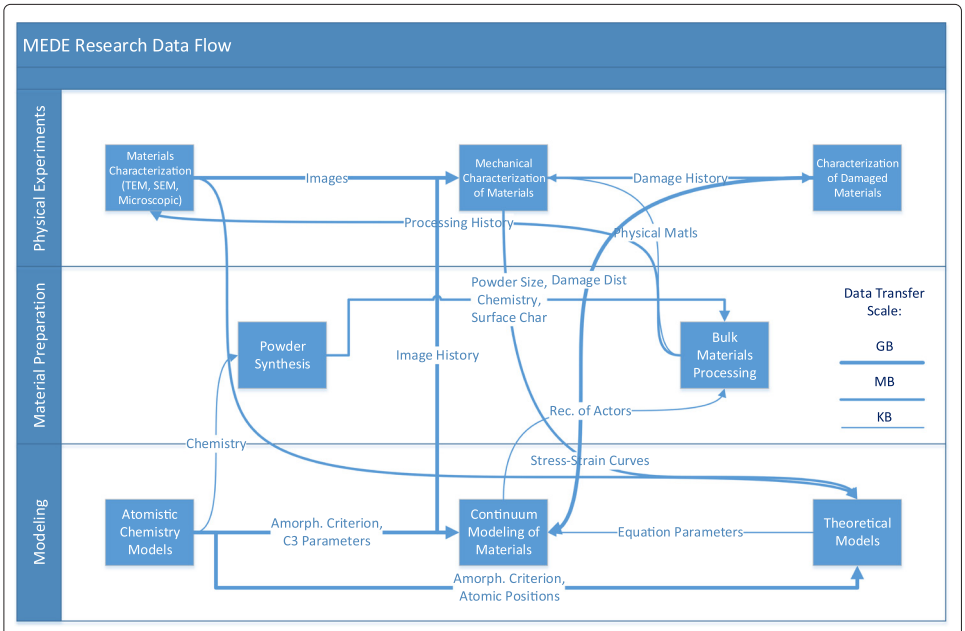

Fig. 1 MEDE research group data flow. Our first task was to form a high-level understanding of data sharing between MEDE research groups. With these existing practices in mind, we identified where SciServer could have an impact. Our initial case study focused on the interaction between Ceramics' Continuum Modeling and Physical Experiments

data manipulative scripting and plotting tools such as Excel, Matlab, and R are commonly used to process the raw data into results.

System design

A useful data platform should fulfill many requirements that match the needs and constraints of current practices. We envision a completely online, secure single-sign on system with three key connected components: file storage, scalable database services, and a templated scripting environment for plotting and analysis. At each component, interactive access to the data should be browser-based and shareable either privately, with immediate peers, or among research groups.

For the most basic collaboration, there must be an online file storage system where researchers upload and share their data with whomever they choose. This service will handle file sizes ranging from KBs to TBs, and act as the data gateway to the rest of the platform. Beyond simple file storage, the database component will serve as the core engine for processing and aggregating the raw data. The database interface will also provide a history of operations performed on the datasets, allowing students and fellow researchers to learn each others' methods and track dataset provenance.

Familiar tools used for collaborative materials engineering must also be made available. Commonly performed analysis codes previously executed on researcher's own computers, such as the derivation of a peak stress versus strain rate plot, should instead be performed on the servers where the data is stored. An online scripting environment, similar to what researchers already use, will provide the main interaction between researchers and their data. Frequently used codes will be encapsulated in pre-written template scripts as an example starting point for new users.

Materials researchers work with multitudes of file formats to store data. While text-based and comma-separated value (CSV) files are frequently used, many instruments that capture experiments produce proprietary data formats that are only officially supported by the manufacturing companies' software. Initially, the collaboration platform will be able to automatically parse and load commonly used text-based file formats into the database and scripting environments; however, the file storage service is designed to be quickly extensible with custom plugins for parsing other data formats. These plugins are snippets of codes that inform the database on how a specific file format's data is organized so that the database may appropriately read and load the data into tables. Knowledge of a file format data scheme may not always be readily available, especially with some proprietary data formats. With this worst-case scenario in which developing a parsing plugin is not possible, the researcher is encouraged to both upload the data file to the file storage service for backup and to then use any available external proprietary software to produce a CSV or textual format for database ingestion.

Providing these three services, all customized for use in materials engineering with minimal change to existing workflows, is no small feat. Yet, MEDE may draw from decades of expertise in big data management tools originally developed in astrophysics but now adapted for general scientific use. We have identified and adapted pieces of the SciServer [1] project to fit the requirements of a MEDE data collaboration platform. Materials scientists may use one or more of these components, whether that is to share files with collaborators or interactively develop plots and analyses of data in an online environment.

Building blocks of SciServer

Developed at Hopkins' Institute for Data Intensive Engineering and Science (IDIES) as part of an NSF Data Infrastructure Building Blocks (DIBBs) 5-year program, SciServer is a publishing platform geared for scientific datasets [2, 3]. SciServer aims to bring the analysis to the data; rather than downloading terabytes before being able to work with the data, SciServer provides tools for codes to be run server-side where the data is stored.

SciServer's development paradigm is to go from 'working to working,' that is, SciServer is being developed by adapting and enhancing existing, already working tools. Each of SciServer's components is functional from the start and are refined as the user experience demands [2]. Many of the existing tools grew from and were proven in the Sloan Digital Sky Survery (SDSS) project and have been generalized as part of the SciServer project. Our own efforts build upon the SciServer platform to create a solution tailored to MEDE research needs.

Below is a description of several SciServer web applications and our adaptations for the use of MEDE scientists. Each of these applications is available using a registered account after a single-sign on gateway.

SciDrive

SciDrive is a free scientific data publishing platform and implements the Dropbox API for safe storing and sharing of files [3]. SciDrive is not just another version of Dropbox; while Dropbox public use is limited to sharing relatively small files, SciDrive is currently hosted on a 100 TB cluster at Johns Hopkins University [3] and focuses on scalable storing of large scientific datasets. Granted, with increased usage, a 100 TB cluster may become insufficient. However, an analysis of cloud storage prices shows that in the long run it is more cost efficient to build one's own cluster with commodity hardware and open-source solutions such as SciDrive [3, 4].

In a collaboration platform for MEDE, SciDrive will fulfill the scientist's file storage needs and serve as the data ingest point. Each scientist with an account will have a private scratch space as well as common group-shared storage and can publish results to non-users by giving an access hyperlink to the desired file. Figure 2 shows the drag and drop upload interface.

SciDrive provides interfaces for tools that work with the contents of the stored files, and features automated metadata extraction (Fig. 3). Default content type extraction is based on characteristic symbol sequences in the first few lines of a file [3]. Alternatively, SciDrive can extract specific content based on the file extension type. Users with file format knowledge may write plugins that allow SciDrive to perform file-type specific data extraction routines. This adaptation point is immediately useful as a staging platform for further analysis of MEDE data; within SciDrive, users may designate folders to be a linked to CasJobs, the database component of SciServer. For common file types used in MEDE, custom plugins will provide the file format specifications needed for SciDrive to appropriately parse the file and load the extracted data into a CasJobs database table. These plugins are how SciDrive is extensible in handling new or proprietary data file formats. While SciDrive is able to store any type of file, a file-type-specific plugin enables automatic parsing and loading into the database for analysis. In the absence of a file format-specific plugin, such as in the case where a file format is proprietary and prohibitively difficult

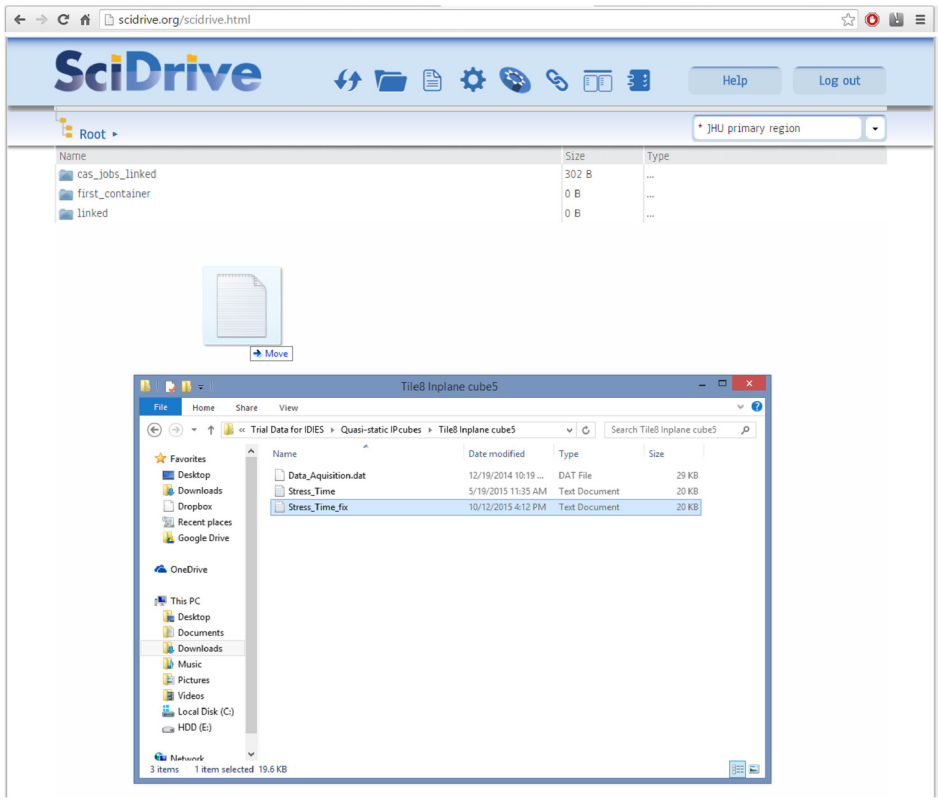

Fig. 2 SciDrive user interface. SciDrive serves as the file storage and data ingest point in the MEDE collaboration platform. This figure shows how users may drag and drop the files they wish to upload to the cloud file system. SciDrive folders may also be linked to CasJobs, so that the data in the file will be automatically extracted and inserted into a CasJobs database for further analysis

to reverse engineer, users may upload both the original data file and a parsable textual version generated with the relevent external proprietary software.

CasJobs

Catalog Archive Server Jobs (CasJobs) is the core of the SciServer infrastructure, providing web-based query management and database services [5]. CasJobs excels when working with large datasets too cumbersome for personal computer-based processing; filtering and aggregations are most efficient when handled by a database designed for such operations. Several features make CasJobs stand out as a collaboration platform. Each user is given their own database context called “MyDB” where they can prototype queries and store personal tables. Groups are allocated contexts as well, where all users belonging to the group can manage and work with the same data. To ensure that each user gets their fair time, queries are executed according to a built-in scheduling subsystem.

CasJobs is the component largely responsible for data provenance tracking in the system. For every dataset in the CasJobs database, the CasJobs interface displays the source file in SciDrive that the table was generated from along with a history of operations performed on the table; with this information, a reviewer may track any changes made to the data from within the collaboration platform and examine the analysis performed.

While materials scientists experienced with SQL have the option to use the CasJobs interface to encode their analysis into queries, SQL knowledge is not necessary in order

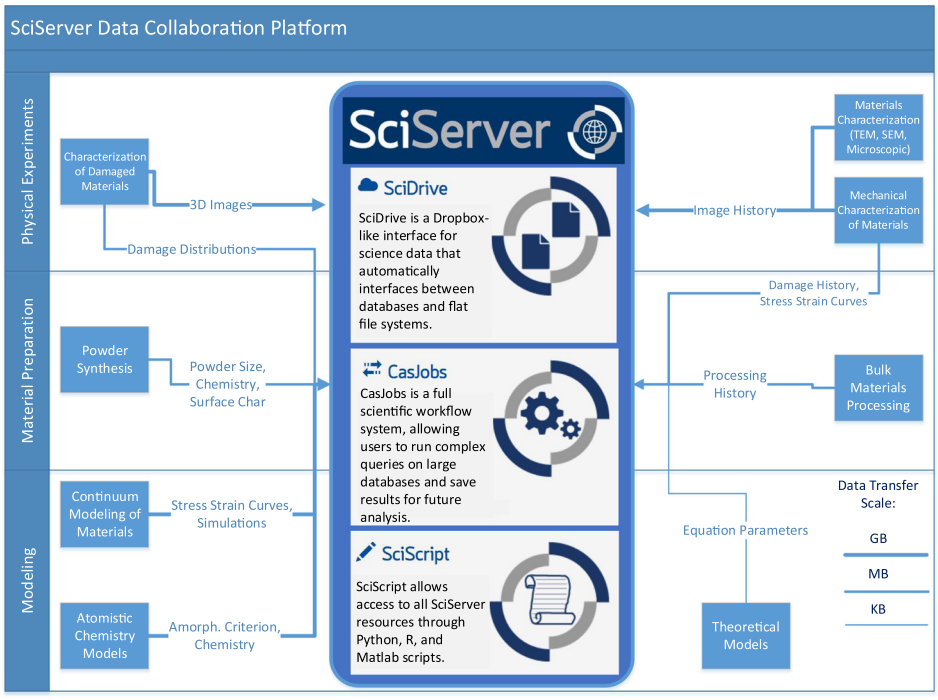

Fig. 3 Application of SciServer components. Based on the size and type of data being shared between MEDE groups, we identified which components of SciServer would be useful as a collaborative data-sharing platform. While most image data would be directly stored and shared via SciDrive, numerical and textual data can be uploaded to the CasJobs Database via the SciDrive drag and drop interface. Data stored in CasJobs is readily accessible to SciScript, enabling scientists to apply their own analysis online via SQL queries, Python scripts, and Matlab plots

to use this platform. CasJobs can automatically extract and store file data from specified SciDrive folders and serve as a staging platform for the more commonly known scripting tools available in SciScript.

SciScript

The Python tools available in SciScript are similar in usage to the current scripting and plotting practices of MEDE researchers and will be more familiar than the SQL interface of CasJobs. The SciScript workflow is a well-known practice; a researcher imports their data into the scripting environment, manipulates the data in some way, and finally plots the results. An example of the SciScript interface is shown in Fig. 4.

SciScript allows users to save, store, and run their scripts online while immediately displaying script results. The ability to store past scripts in online notebooks is an important requirement for data provenance; students to the art or reviewers may examine scripts used to analyze or alter data and re-execute in order to verify the results. From a SciDrive source file upload to a SciScipt plot, a user may trace the path of the data through the collaboration platform.

Case description

We began by testing SciServer tools on a subset of research groups; as a starting point, we created a template workflow encompassing the data interaction between the continuum modeling group and physical experiments. The workflow template is a persistent,

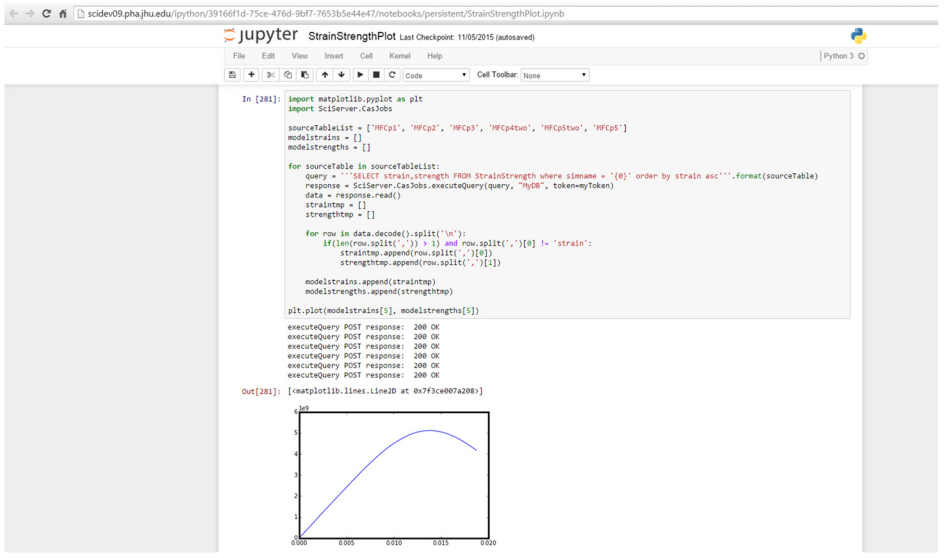

Fig. 4 SciScript interface. The SciScript online Python notebook environment provides easy scripting access to any data stored in CasJobs and SciDrive. In this short example, the python script queries, formats, and plots data from CasJobs

example use case of the SciServer chain from data ingestion to plot analysis, containing the necessary code and usage directions within each tool's interface. The examined comparison point between the model and experiment data is a material's peak stress with respect to strain rate. In this case study, SciServer ingests each group's datasets, performs aggregation in CasJobs, and produces comparative plots in SciScript.

The experiment sensor data was provided in tab-delimited text files, which CasJobs parses and inserts into a database table by default. Simulation results given by the modeling group are in a Uintah Data Archive (.uda) file format produced by the Uintah software suite [6]. A SciDrive plugin for parsing the UDA compressed files online is currently being tested; however, in the meantime, the simulation files are converted into CasJobs-injestable text files using the external VisIt software suite [7].

Once uploaded to a CasJobs-linked folder on SciDrive, the data is aggregated in the database with SQL queries stored and executed by a SciScript Python notebook. After interviewing modelers, we encoded their common simulation output processing practices into SQL and saved the queries as templates for further use. Each simulated particle's data is provided as timestamped 3×3 matrices representing force and stress. Using our SQL Server extensions which we have installed onto CasJobs, we calculate the strain tensor for each volume element i from the force tensors F_i as

$$\boldsymbol{\varepsilon_i} = \frac{1}{2}\left[\boldsymbol{F}_i^T\boldsymbol{F}_i - \boldsymbol{I}\right] \tag{1}$$

We solve the eigenproblem of these 3×3 matrices to obtain the eigenvalues $\{\varepsilon_i^1, \varepsilon_i^2, \varepsilon_i^3\}$ that can be used to calculate the scalar quantity

$$\bar{\varepsilon} = \frac{1}{N}\sum_i^N \left\{\frac{1}{2}\left[\left(\varepsilon_i^1 - \varepsilon_i^2\right)^2 + \left(\varepsilon_i^2 - \varepsilon_i^3\right)^2 + \left(\varepsilon_i^3 - \varepsilon_i^1\right)^2\right]\right\}^{1/2} \tag{2}$$

The calculation is done for all timesteps in the simulation to map out the changes over time $\bar{\varepsilon} = \bar{\varepsilon}(t)$. The stress is calculated similarly in the database, and the aggregation over the eigenvalues provides the empirical determination of the $\bar{\sigma}(t)$ function.

Next, we plot model-derived strain versus stress with scripts written on SciScript. Figure 5 shows the SciScript-generated plot of five different simulations, each executed with a specified strain rate. From the same scripting environment, the experimental data is queried and compared to the model data. Figure 6 is a comparison of peak stress versus strain rate from both of the modeling and physical realms. While these figures display nothing new in terms of materials science, their point is to illustrate SciServer's ability to serve as a collaborative platform containing the necessary tools for materials data processing. With the workflow templates created in this case study, researchers may upload their own similarly formatted datasets and simply change the template's source table name to aggregate their data and produce more meaningful comparative plots all in a matter of seconds.

Discussion and evaluation

SciServer hopes to enable faster collaboration between MEDE research groups and advance the materials-by-design approach. As a common file storage, database, and scripting platform for various MEDE researchers, SciServer can bring together data from various sources in order to draw new comparisons. While there are many fundamental benefits to a server-based data workbench, the downside is that researchers will still have to change their existing work habits in order to adapt to the new tools. Much of our effort has gone to observing and learning MEDE research practices so that we may augment SciServer to be as easy as possible to transition to. We hope to make SciServer attractive by streamlining and automating the researcher's data management; features such as pre-written workflow templates should shorten the time and effort between running

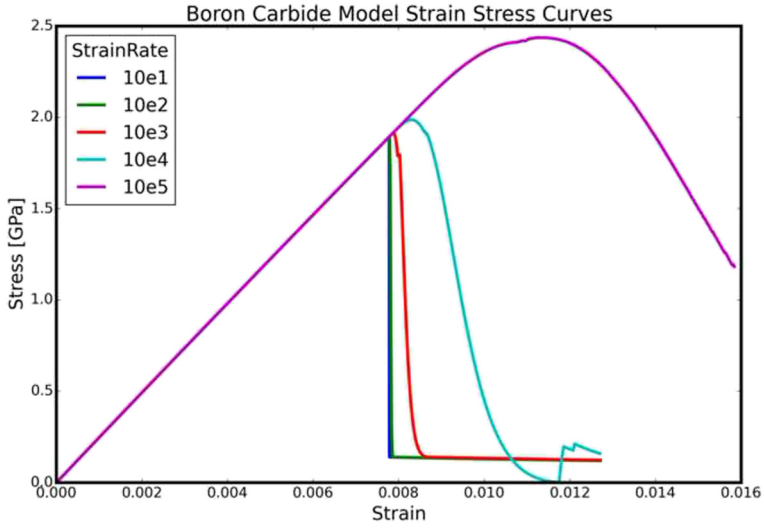

Fig. 5 Boron carbide model strain-stress curves. This figure was created online in SciScript. The researcher uploaded several simulation datasets generated from Uintah [6] to CasJobs. Pre-written aggregation and analysis queries are run on the data. Next, via SciScript, a researcher uses interactive python matplotlib scripts produce this common Strain-Stress curve

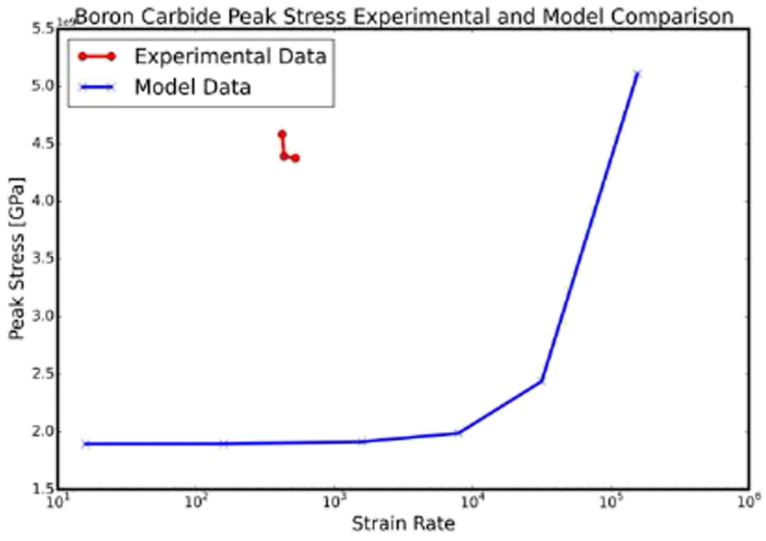

Fig. 6 Boron carbide peak stress experimental and model comparison. Collaborative analysis is readily available with the SciServer platform. A SciScript python notebook may query several CasJobs databases from different research groups. This example plot compares peak stress versus strain rate data from physical experiments and Uintah simulation models. Analysis of the sensor data, including deriving the strain rate of each experiment, was performed in short python scripts online in SciScript

a simulation or experiment and obtaining analysis plots, while developing plugins to parse specific file formats will automate database file ingestion. The SciServer approach excels at enabling large-scale data science and online collaboration; however, users will be limited to the tools available within the platform.

Although we have a case study of one collaboration point between two groups, there are still many use cases to examine and consider for SciServer. Striving to provide template scripts, queries, and tools general enough to encourage researchers to code their own analysis within SciServer is a difficult task requiring broad knowledge of how MEDE performs materials-by-design. We can further adapt SciServer to MEDE only through more researcher usage and support experience.

Conclusions

Data integration is undoubtedly a key aspect of materials research going forward; we systematically plan for a novel platform that requires little change in the everyday routine of scientists but provides a number of benefits from backups, data sharing, and connected analysis tools. We have carefully reviewed the current practices and designed a new platform that can accommodate existing studies while enabling new layers of abstractions for data analysis and dashboard-style reporting in the future.

Our approach is to build on existing general components of the SciServer [1] and create a materials research data platform where scientists can not only perform their own studies but also share the data and their tools with each other.

We created an end-to-end analysis use-case for the Ceramics group to compare experiments and simulations. The workflow is simple: Researchers drag-n-drop (or synchronize) raw data files to our storage facility (SciDrive), whose content are checked for known formats by our plugins. When enabled, these background processes can asynchronously

load the relevant data from the files into a relational database management system, which is accessible via SQL and web (CasJobs). Our custom extensions make filtering, aggregation, and analysis straightforward within the database so researchers can perform the data-intensive procedures on large remote computers. The results then can be transferred to the user or analyzed by our integrated scripting engine based on Python and Jupyter.

At every step along the way, multiple users may work with the same datasets, track provenance, and save or share written queries and scripts. Yet, there is still much to be done to further fit SciServer to MEDE's needs. While our first step was to provide researchers with familiar tools in an online environment, our next effort will be to collect more datasets from more groups in order to leverage CasJobs as a large-scale scientific database. Given more data, SciServer will allow collaborative large-scale analysis previously not possible.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NSC drafted the manuscript, developed the case study and the subsequent templated workflow pipeline. TB designed the adaptation of SciServer to MEDE and developed materials analysis database functions. ND contributed simulated results and assisted in design of the case study. KTR contributed MEDE organization information and details regarding common materials scientist practices. All authors read and approved the final manuscript.

Acknowledgements
Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-12-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Author details
[1]Department of Computer Science, Johns Hopkins University, 3400 N Charles St, 21218 Baltimore, MD, USA. [2]Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N Charles St, 21218 Baltimore, MD, USA. [3]Hopkins Extreme Materials Institute, Johns Hopkins University, 3400 N Charles St, Malone Hall 140, Baltimore, MD, USA. [4]Department of Mechanical Engineering, Johns Hopkins University, 3400 N Charles St, Latrobe Hall 223, Baltimore, MD, USA.

Received: 11 January 2016 Accepted: 23 February 2016
Published online: 29 April 2016

References

1. SciServer: Collaborative Data Driven Science. www.sciserver.org/. Accessed 5 Jan 2016.
2. SciServer: Big Data Infrastructure for Science. https://nsf.gov/discoveries/disc_summ.jsp?cntn_id=133526&org=NSF. Accessed 5 Jan 2016.
3. Mishin D, Medvedev D, Plante R, Graham M, Szalay S (2013) Data sharing and publication using the scidrive service. In: Manset N, Forshay P (eds). Astronomical Data Analysis Software and Systems XXIII, Waikoloa Beach Marriott, Hawaii, USA September 29–October 3, 2013, vol. 485. Taylor & Francis Group, Abingdon, England. http://www.aspbooks.org/a/volumes/table_of_contents/?book_id=553, http://www.aspbooks.org/publishing_with_asp/
4. Naldi M, Mastroeni L (2013) Cloud storage pricing: a comparison of current practices. In: Proceedings of the 2013 International Workshop on Hot Topics in Cloud Services. HotTopiCS '13. ACM, New York, NY, USA. p 27. doi:10.1145/2462307.2462315. http://doi.acm.org/10.1145/2462307.2462315
5. Li N, Thakar AR (2008) CasJobs and MyDB: a batch query workbench. Comput Sci Eng. 10(1):18–29
6. Berzins M, Luitjens J, Meng Q, Harman T, Wight CA, Peterson JR (2010) Uintah: A scalable framework for hazard analysis. In: Proceedings of the 2010 TeraGrid Conference. TG '10. ACM, New York, NY, USA. pp 3–138. doi:10.1145/1838574.1838577. http://doi.acm.org/10.1145/1838574.1838577
7. Childs H, Brugger E, Whitlock B, Meredith J, Ahern S, Pugmire D, Biagas K, Miller M, Harrison C, Weber GH, Krishnan H, Fogal T, Sanderson A, Garth C, Bethel EW, Camp D, Rübel O, Durant M, Favre JM, Navrátil P (2012) VisIt: an end-user tool for visualizing and analyzing very large data. In: High Performance Visualization–Enabling Extreme-Scale Scientific Insight. Taylor & Francis Group, Abingdon, England. pp 357–372