

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant: _____ Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	3 2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	1	Expert level of the user (e.g., experienced, intermediate or new/casual)	3+1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	1	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	1	Data privacy concerns	2
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	1
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	2
View detailed metadata explaining the data fields next to the data itself	2	Availability of data for a variety of systems (e.g., accelerators)	3+1
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	2
Partial anonymization of the data (sensitive fields in the dataset like user name and application name <u>will be removed</u> but the institution and machine names will be available)	3	A large consumer base for the data in the research community	2
Data uploaded should be <u>easy to cite</u> and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
input data
output data
executables

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	✓
Campus clusters	✓
XSEDE systems	
Open Science Grid (OSG)	
BlueWaters	
Commercial cloud services	
Other (please specify)	Shahae II, Server

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	only on my server
Researcher in academia	Y
Computational end-user of HPC systems	N
System vendor	N
Other (please specify)	—

Name of the survey participant: [REDACTED]

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	3	Expert level of the user (e.g., experienced, intermediate or new/casual)	3
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	2	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	2
Cost of deploying measurement tools	2	Data privacy concerns	2
Cost of storing, maintaining and updating such data	2	Other (please write in)	
Cost of documenting failure events	2		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	2	Visualize the data from search	1
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	2	Other desired features (please write in)	

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data (the data sets will be non-identifiable to the actual source/person who contributed)	1	Prominent public recognition of the PIs and institutions contributing data	1
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	1	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	1	A large consumer base for the data in the commercial community	3

Other issues? (please write in)

The real question might be to obfuscate the data so that general trends remain, but individual data points have some amount of noise.

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
Inputs cannot be shared (Programs cannot be)

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	Y
Commercial cloud services	Y
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	Y
Researcher in academia	Y
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	N

Name of the survey participant: _____ Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	2
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	3	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	3
Cost of storing, maintaining and updating such data	2	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	3
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	2
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	2	Prominent public recognition of the PIs and institutions contributing data	3
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	2	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
Data that comes from resiliency research.

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	Y
Open Science Grid (OSG)	N
BlueWaters	Y
Commercial cloud services	N
Other (please specify)	I

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	Y
Computational end-user of HPC systems	N
System vendor	N
Other (please specify)	I

Name of the survey participant: [REDACTED]

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	3	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	2
Cost of deploying measurement tools	3	Data privacy concerns	3
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	2	Visualize the data from search	1
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	1
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	3
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)		A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	3
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
Sadly, we cannot share anything :(

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	
Campus clusters	
XSEDE systems	
Open Science Grid (OSG)	
BlueWaters	
Commercial cloud services	
Other (please specify)	DOE supercomputers

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	
Researcher in academia	Y - government, not academia
Computational end-user of HPC systems	
System vendor	
Other (please specify)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant [REDACTED]

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	3
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in) <i>"Parsed logs" - need to capture meaning - not coded messages</i>	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools		Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	
Cost of deploying measurement tools		Data privacy concerns	
Cost of storing, maintaining and updating such data		Other (please write in)	
Cost of documenting failure events			

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	2	Visualize the data from search	2
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	2
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)		Prominent public recognition of the PIs and institutions contributing data	
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)		A large consumer base for the data in the research community	
Data uploaded should be easy to cite and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	N
XSEDE systems	Y
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	N
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	
Researcher in academia	X
Computational end-user of HPC systems	
System vendor	
Other (please specify)	

Name of the survey participant: _____ Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	3	Expert level of the user (e.g., experienced, intermediate or new/casual)	1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	1
Cost of storing, maintaining and updating such data	1	Other (please write in)	
Cost of documenting failure events	1		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	1	Visualize the data from search	1
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	2	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	1	Prominent public recognition of the PIs and institutions contributing data	2
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	2	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	2	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.
 <<To be answered by Participants>>

Data that cannot be shared
Only security or privacy related information.
personal

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	Y
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	Y
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	

Name of the survey participant: [REDACTED] Date: Nov. 18, 2015**Part I: Utilization of data**

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	3	Expert level of the user (e.g., experienced, intermediate or new/casual)	3
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	2	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	1	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	1
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	3
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this?
 (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	2	Prominent public recognition of the PIs and institutions contributing data	3
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	2
Data uploaded should be easy to cite and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	Y
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	Y
Researcher in academia	Y
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	

Name of the survey participant: _____ Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	2
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in) Power information	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	2
Cost of storing, maintaining and updating such data	2	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	2	Visualize the data from search	2
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	2	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	2
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	2
Data uploaded should be easy to cite and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	Y
Open Science Grid (OSG)	Y
BlueWaters	Y
Commercial cloud services	N
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	Y
Computational end-user of HPC systems	N
System vendor	N
Other (please specify)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant: [REDACTED]

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	3	Expert level of the user (e.g., experienced, intermediate or new/casual)	1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in) <i>agree w/ question from audience that data correlation from other system components will be critical (storage, network, etc)</i>	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	3	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	2
Cost of deploying measurement tools	2	Data privacy concerns	3
Cost of storing, maintaining and updating such data	2	Other (please write in) <i>in a lab environment, getting buy in from users + sys admins before deploying collection tools = releasing data sets</i>	
Cost of documenting failure events	2		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	3
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	N
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	N
Other (please specify)	lab systems LLNL ORNL LANL ANL SDSC NERSC

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	N
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	lab researcher

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant: [REDACTED]

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	2
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in) Virtualized Resources	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	3
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	3
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	1
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	2	A large consumer base for the data in the commercial community	3
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
Datasets that are part of emerging architecture designs.

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	
Open Science Grid (OSG)	
BlueWaters	N
Commercial cloud services	Y
Other (please specify)	Commercial clusters

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	
Researcher in academia	
Computational end-user of HPC systems	
System vendor	
Other (please specify)	

Data Analytics
& Machine Learning @ Intel
Path finding

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant _____ Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	3	Expert level of the user (e.g., experienced, intermediate or new/casual)	1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	1	Data privacy concerns	2
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	2	Visualize the data from search	3
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	2
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey
FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	2	Prominent public recognition of the PIs and institutions contributing data	1
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	1	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared
user names
App names (probably)

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	N
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	Y
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant: _____

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	3
Hardware performance counter measurements	3	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	2
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	1
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	3	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	3	Data privacy concerns	3
Cost of storing, maintaining and updating such data	2	Other (please write in)	
Cost of documenting failure events	2		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	2
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	2	Availability of data for jobs representing applications from diverse domains	3
View detailed metadata explaining the data fields next to the data itself	3	Availability of data for a variety of systems (e.g., accelerators)	3
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	2
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	3	A large consumer base for the data in the commercial community	2
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared	<i>user information</i>

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	Y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	Y
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	N
Researcher in academia	Y
Computational end-user of HPC systems	Y
System vendor	N
Other (please specify)	

SC15 BoF Participant Survey

FRESCO: An Open Data Repository for Dependability Research and Practice

Name of the survey participant: _____

Date: Nov. 18, 2015

Part I: Utilization of data

1a. Indicate the usefulness of the following types of data in an open systems and workload data repository (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Job-level activity and performance data (libraries, executables and environment user accessed, performance measurement of IB, CPU, memory, filesystem during job runtime)	3	Syslog messages	2
Hardware performance counter measurements	2	Type of application executed (eg. Genomics, Weather Forecast, Structural analysis, Image processing, etc)	3
Measurements from system monitoring tools like Nagios or Ganglia	2	Expert level of the user (e.g., experienced, intermediate or new/casual)	2
Accounting logs for job submission (e.g., how long did a job run, did it terminate successfully or not)	3	Other (please write in)	

1b. What are the challenges in collecting such datasets from a cluster? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Challenges	Answer	Challenges	Answer
Degradation of job performance by the use of measurement tools	2	Difficulty in determining what to collect and store, unless a researchers approaches with specific requests	3
Cost of deploying measurement tools	2	Data privacy concerns	2
Cost of storing, maintaining and updating such data	3	Other (please write in)	
Cost of documenting failure events	3		

1c. What would be useful usability features for the data repository? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important)

Useful features	Answer	Useful features	Answer
Run analysis scripts on the server without downloading the data	3	Visualize the data from search	2
Selection and download data in small manageable chunks of a few 100 MBs (e.g. over a short period)	3	Availability of data for jobs representing applications from diverse domains	2
View detailed metadata explaining the data fields next to the data itself	2	Availability of data for a variety of systems (e.g., accelerators)	2
Tools for filtering, extracting and classify error data from various sources	3	Other desired features (please write in)	

Part II: Data sharing

2a. What issues are important to you when you consider sharing data through a repository like this? (enter a number 1-3, with 3=Important, 2=Neutral, 1=Not important):

Type of data	Answer	Type of data	Answer
Complete anonymization of the data(the data sets will be non-identifiable to the actual source/person who contributed)	3	Prominent public recognition of the PIs and institutions contributing data	2
Partial anonymization of the data (sensitive fields in the dataset like user name and application name will be removed but the institution and machine names will be available)	3	A large consumer base for the data in the research community	3
Data uploaded should be easy to cite and the contributor credited for the dataset	2	A large consumer base for the data in the commercial community	3
Other issues? (please write in)			

2b. What type of data can you not share at all?

Ex: Application names, Library names, any framework used like mpi, hadoop, etc.

<<To be answered by Participants>>

Data that cannot be shared

Part III: Your role in the computational environment

3a. What kind of computational infrastructure do you have access to? (check all that applies)

Computational infrastructure	Answer (Y or N)
Desktop, lab servers	Y
Campus clusters	y
XSEDE systems	N
Open Science Grid (OSG)	N
BlueWaters	N
Commercial cloud services	y
Other (please specify)	

3b. What is your role in it? (Ex: Cluster administrator, Researcher from academia, Practical user of the system, etc.)

Computational infrastructure	Answer (Y or N)
System (cluster) administrator	y
Researcher in academia	y
Computational end-user of HPC systems	y
System vendor	N
Other (please specify)	