BIG DATA USE CASE SURVEY 2

NIST Big Data Public Working Group

This survey was designed by the NIST Big Data Public Working Group (NBD-PWG) to gather Big Data use cases. The use case information you provide in this survey will greatly help the NBD-PWG in the next phase of developing the NIST Big Data Interoperability Framework. We sincerely appreciate your effort and realize it is nontrivial.

The survey can also be completed in the Google Form for UC Template#2: http://bit.ly/1ff7iM9

More information about the NBD-PWG and the NIST Big Data Interoperability Framework can be found at http://bigdatawg.nist.gov.

SURVEY OUTLINE

Survey instructions are provided with each question

* Denotes fields that are important and must be filled in.

NOTE: No proprietary or confidential information should be included.

Please submit the completed form to Wo Chang at wchang@nist.gov.

Submit Form

1 Overall Project Description

1.1 Use Case Title *

Please limit to one line. A description field is provided below for a longer description.

Web-Enabled Landsat Data (WELD) Processing

1.2 Use Case Description *

Summarize all aspects of use case focusing on application issues (later questions will highlight technology).

The use case is specific to the part of the project where data is available on the HPC platform and processed through the science workflow. It is a 32 stage processing pipeline that includes two separate science products

1.3 Use Case Contacts *

Add names, phone number, and email of key people associated with this use case. Please designate who is authorized to edit this use case.

Name	Phone	Email	PI	Author	Edit rights?	Primary
Andrew Michaelis	650-604-5413	andrew.r.michaeli		Author	☑	☑
				Author	□	□
				Author	□	□
				Author	□	□

1.4 Domain ("Vertical") *

What application area applies? There is no fixed ontology. Examples: Health Care, Social Networking, Financial, Energy, etc.

Land use science: image processing

1.5 Application *

Summary of use case applications.

The product of this use case is a dataset of science products of use to the land surface science community that is made freely available by NASA. The dataset is produced through processing of images from the Landsat 4, 5

1.6 Current Data Analysis Approach *

Describe the analytics, software, hardware approach used today.

>> Compute System: Shared High Performance Computing (HPC) system at NASA Ames Research Center (Pleiades)

1.7 Future of Application and Approach *

Describe the analytics, software, hardware, and application future plans, with possible increase in data sizes/velocity.

Processing will be improved with newer and updated algorithms. This process may also be applied to future datasets and processing systems (Landsat 8 and Sentinel-2 satellites, for example).

1.8 Actors / Stakeholders

Please describe the players and their roles in the use case. Identify relevant stakeholder roles and responsibilities. Note: Security and privacy roles are surveyed in a separate part of this survey.

South Dakota State University – science, algorithm development, QA, data browse visualization and distribution framework; NASA Advanced Supercomputing Division at NASA Ames Research Center – data processing at scale; USGS – data access and data distributi

1.9 Project Goals or Objectives

Please describe the objectives of the use case.

The WELD products are developed specifically to provide consistent data that can be used to derive land cover as well as geophysical and biophysical products for assessment of surface

1.10 Use Case URL(s)

Include any URLs associated with the use case. Please separate with semicolon (;).

http://globalmonitoring.sdstate.edu/projects/weldglobal/gweld.html
http://globalweld.cr.usgs.gov/
https://nex.nasa.gov

1.11 Picture and Diagrams?

Please email any pictures or diagrams with this survey.

2 Big Data Characteristics

Big Data Characteristics describe the properties of the (raw) data including the four major 'V's' of Big Data described in NIST Big Data Interoperability Framework: Volume 1, Big Data Definition.

2.1 Data Source

Describe the origin of data, which could be from instruments, Internet of Things, Web, Surveys, Commercial activity, or from simulations. The source(s) can be distributed, centralized, local, or remote.

Satellite Earth observation data from Landsat 4, 5, and 7 missions. The data source is remote and centralized – distributed from USGS EROS Center.

2.2 Data Destination

If the data is transformed in the use case, describe where the final results end up. This has similar characteristics to data source.

The final data is distributed by USGS EROS Center – a remote centralized data system. It is also available on the NEX platform for further analysis and product development.

2.3 Volume

Size	30PB of processed data through the pipeline (1PB inputs, 10PB intermediate, 6PB outputs)
Units	Petabytes of data that flow through the processing pipeline
Time Period	Data was collected over a period of 27 years and is being processed over a period of 5 years
Proviso	The data represent the operational time period of 1984 to 2011 for the Landsat 4, 5, and 7 satellites

Size: Quantitative volume of data handled in the use case
Units: What is measured such as "Tweets per year", Total LHC data in petabytes, etc.?
Time Period: Time corresponding to specified size.
Proviso: The criterion (e.g. data gathered by a particular organization) used to get size with units in time period in three fields above

2.4 Velocity

Enter if real time or streaming data is important. Be quantitative: this number qualified by 3 fields below: units, time period, proviso. Refers to the rate of flow at which the data is created, stored, analyzed, and visualized. For example, big velocity means that a large quantity of data is being processed in a short amount of time.

Unit of measure	Terabytes processed per day during processing time periods: 150 TB/day
Time Period	24 hours
Proviso	Based on programmatic goals of processing several iterations of the final

Unit of Measure: Units of Velocity size given above. What is measured such as "New Tweets gathered per second", etc.?
Time Period: Time described and interval such as September 2015; items per minute
Proviso: The criterion (e.g., data gathered by a particular organization) used to get Velocity measure with units in time period in three fields above

2.5 Variety

Variety refers to data from multiple repositories, domains, or types. Please indicate if the data is from multiple datasets, mashups, etc.

This use case basically deals with a single dataset.

2.6 Variability

Variability refers to changes in rate and nature of data gathered by use case. It captures a broader range of changes than Velocity which is just change in size. Please describe the use case data variability.

Not clear what the difference is between variability and variety. This use case basically deals with a single dataset.

3 Big Data Science

3.1 Veracity and Data Quality

This covers the completeness and accuracy of the data with respect to semantic content as well as syntactical quality of data (e.g., presence of missing fields or incorrect values).

This data dealt with in this use case are a high quality, curated dataset.

3.2 Visualization

Describe the way the data is viewed by an analyst making decisions based on the data. Typically visualization is the final stage of a technical data analysis pipeline and follows the data analytics stage.

Visualization is not used in this use case per se, but visualization is important in QA processes conducted outside of the use case as well as in the ultimate use by scientists of the product datasets that result from this use case

3.3 Data Types

Refers to the style of data, such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.

structured image data

3.4 Metadata

Please comment on quality and richness of metadata.

Metadata adhere to accepted metadata standards widely used in the earth science imagery field.

3.5 Curation and Governance

Note that we have a separate section for security and privacy. Comment on process to ensure good data quality and who is responsible.

Data is governed by NASA data release policy; data is referred to by the DOI and the algorithms have been peer-reviewed. The data distribution center and the PI are responsible for science data support.

3.6 Data Analytics

In the context of these use cases, analytics refers broadly to tools and algorithms used in processing the data at any stage including the data to information or knowledge to wisdom stages, as well as the information to knowledge stage.

There are number of analytics processes through out the processing pipeline. The key analytics is identifying best available pixels for spatio-temporal composition and spatial aggregation processes as a part of the overall QA. The analytics algorithms are custom

4 Security and Privacy

Questions in this section are designed to gather a comprehensive image of security and privacy aspects (e.g., security, privacy, provenance, governance, curation, and system health) of the use case. Other sections contain aspects of curation, provenance and governance that are not strictly speaking only security and privacy considerations. The security and privacy questions are grouped as follows:

- Roles
- Personally Identifiable Information
- Covenants and Liability
- Ownership, Distribution, Publication
- Risk Mitigation
- Audit and Traceability
- Data Life Cycle
- Dependencies
- Framework provider S&P
- Application Provider S&P
- Information Assurance | System Health
- Permitted Use Cases

For the questions with checkboxes, please select the item(s) that apply to the use case.

4.1 Roles

Roles may be associated with multiple functions within a big data ecosystem.

4.1.1 Identifying Role

Identify the role (e.g., Investigator, Lead Analyst, Lead Scientists, Project Leader, Manager of Product Development, VP Engineering) associated with identifying the use case need, requirements, and deployment.

PI; Project sponsor (NASA EOSDIS program)

4.1.2 Investigator Affiliations

This can be time-dependent and can include past affiliations in some domains.

Andrew Michaelis, NASA, NEX Processing Pipeline Development and Operations
David Roy, South Dakota State University, Project PI

4.1.3 Sponsors

Include disclosure requirements mandated by sponsors, funders, etc.

NASA EOSDIS project

4.1.4 Declarations of Potential Conflicts of Interest

None

4.1.5 Institutional S/P duties

List and describe roles assigned by the institution, such as via an IRB.

None

4.1.6 Curation

List and describe roles associated with data quality and curation, independent of any specific Big Data component. Example: Role responsible for identifying US government data as FOUO or Controlled Unclassified Information, etc.

Joint responsibility of NASA, USGS, and Principal Investigator

4.1.7 Classified Data, Code or Protocols

- □ Intellectual property protections
- □ Military classifications, e.g., FOUO, or Controlled Classified
- ☑ Not applicable
- □ Other:

4.1.8 Multiple Investigators | Project Leads *

- □ Only one investigator | project lead | developer
- □ Multiple team members, but in the same organization
- ☑ Multiple leads across legal organizational boundaries
- □ Multinational investigators | project leads
- □ Other:

4.1.9 Least Privilege Role-based Access

Least privilege requires that a user receives no more permissions than necessary to perform the user's duties.

- □ Yes, roles are segregated and least privilege is enforced
- □ We do have least privilege and role separation but the admin role(s) may be too all-inclusion
- □ Handled at application provider level
- □ Handled at framework provider level
- □ There is no need for this feature in our application
- □ Could be applicable in production or future versions of our work
- ☑ Other: Not used

4.1.10 Role-based Access to Data *

Please describe the level at which access to data is limited in your system.

- ☑ Dataset
- □ Data record / row
- □ Data element / field
- □ Handled at application provider level
- □ Handled at framework provider level
- □ Other:

4.2 Personally Identifiable Information (PII)

4.2.1 Does the System Maintain PII? *

□ Yes, PII is part of this Big Data system
☑ No, and none can be inferred from 3rd party sources
□ No, but it is possible that individuals could be identified via third party databases
□ Other:

4.2.2 Describe the PII, if applicable

Describe how PII is collected, anonymized, etc. Also list disclosures to human subjects, interviewees, or web visitors.

4.2.3 Additional Formal or Informal Protections for PII

4.2.4 Algorithmic / Statistical Segmentation of Human Populations

□ Yes, doing segmentation, possible discrimination issues if abused. Please also answer the next question.
□ Yes, doing segmentation, but no foreseeable discrimination issues.
☑ Does not apply to this use case at all (e.g., no human subject data)
□ Other:

4.2.5 Protections afforded statistical / deep learning discrimination

Identify what measures are in place to address this concern regarding human populations, if it applies. Refer to the previous question.

Not applicable to this use case

4.3 Covenants, Liability, Etc.

4.3.1 Identify any Additional Security, Compliance, Regulatory Requirements *

Refer to 45 CFR 46: http://1.usa.gov/1bg6JQ2

□ FTC regulations apply
□ HHS 45 CFR 46
□ HIPAA
□ EU General Data Protection (Reference: http://bit.ly/1Ta8S1C)
□ COPPA
□ Other Transborder issues
□ Fair Credit Reporting Act (Reference: http://bit.ly/1Ta8XSN)
□ Family Educational Rights and Protection (FERPA)
☑ None apply
□ Other:

4.3.2 Customer Privacy Promises

Select all that apply. E.g., RadioShack promise that is subject of this DOJ ruling: http://bit.ly/1f0MW9t

□ Yes, we're making privacy promises to customers or subjects
□ We are using a notice-and-consent model
☑ Not applicable
□ Other:

4.4 OWNERSHIP, IDENTITY AND DISTRIBUTION

4.4.1 Publication rights

Open publisher; traditional publisher; white paper; working paper

□ Open publication
□ Proprietary
□ Traditional publisher rights (e.g., Springer, Elsevier, IEEE)
□ "Big Science" tools in use
☑ Other: Datasets produced are available to the public with a requirement for

4.4.2 Chain of Trust

Identify any chain-of-trust mechanisms in place (e.g., ONC Data Provenance Initiative.) Potentially very domain-dependent; see the ONC event grid for instance. Reference: http://bit.ly/1f0PGDL

None

4.4.3 Delegated Rights

Example of one approach: “Delegation Logic: A Logic-based Approach to Distributed Authorization”, Li, N., Grosof, B.N., Feigenbaum, J.(2003) https://www.cs.purdue.edu/homes/ninghui/papers/thesis.pdf

None

4.4.4 Software License Restrictions

Identify proprietary software used in the use case Big Data system which could restrict use, reproducibility, results, or distribution.

None

4.4.5 Results Repository

Identify any public or private / federated consortia maintaining a shared repository.

The datasets produced from this dataset are distributed to the public from repositories at the USGS EROS Center and the NASA EOSDIS program.

4.4.6 Restrictions on Discovery

Describe restrictions or protocols imposed on discoverable end points.

None

4.4.7 Privacy Notices

Indicate any privacy notices required / associated with data collected for redistribution to others

□ Privacy notices apply
☑ Privacy notices do not apply
□ Other:

4.4.8 Key Management

□ A key management scheme is part of our system
□ We are using public key infrastructure.
□ We do not use key management, but it could have been useful
☑ No readily identifiable use for key management
□ Other:

4.4.9 Describe Key Management Practices

4.4.10 Is an identity framework used?

□ A framework is in place. (See next question.)
□ Not currently using a framework.
☑ There is no perceived need for an identity framework.
□ Other:

4.4.11 CAC / ECA Cards or Other Enterprise-wide Framework

□ Using an externally maintained enterprise-wide identity framework
□ Could be used, but none are available
☑ Not applicable

4.4.12 Describe the Identity Framework.

4.4.13 How is intellectual property protected?

□ Login screens advising of IP issues
□ Employee or team training
□ Official guidelines limiting access or distribution
□ Required to track all access to, distribution of digital assets
□ Does not apply to this effort (e.g., public effort)
☑ Other: Believe there are standards for citation of datasets that apply to use of the

4.5 Risk Mitigation

*4.5.1 Are measures in place to deter re-identification? **

□ Yes, in place
□ Not in place, but such measures do apply
☑ Not applicable
□ Other:

4.5.2 Please describe any re-identification deterrents in place

4.5.3 Are data segmentation practices being used?

Data segmentation for privacy has been suggested as one strategy to enhance privacy protections. Reference: http://bit.ly/1P3h12Y

- ☐ Yes, being used
- ☐ Not in use, but does apply
- ☑ Not applicable
- ☐ Other:

4.5.4 Is there an explicit governance plan or framework for the effort?

- ☐ Explicit governance plan
- ☐ No governance plan, but could use one
- ☐ I don't think governance contributes anything to this project
- ☑ Other: Resulting datasets are governed by the data access policies of the USGS

4.5.5 Privacy-Preserving Practices

Identify any privacy-preserving measures that are in place.

None

4.5.6 Do you foresee any potential risks from public or private open data projects?

Transparency and data sharing initiatives can release into public use datasets that can be used to undermine privacy (and, indirectly, security.)

- ☐ Risks are known.
- ☐ Currently no known risks, but it is conceivable.
- ☑ Not sure
- ☐ Unlikely that this will ever be an issue (e.g., no PII, human-agent related data or subsystems.)
- ☐ Other:

4.6 Provenance (Ownership)

Provenance viewed from a security or privacy perspective. The primary meaning for some domains is digital reproducibility, but it could apply in simulation scenarios as well.

4.6.1 Describe your metadata management practices

- ☐ Yes, we have a metadata management system.
- ☐ There is no need for a metadata management system in this use case
- ☐ It is applicable but we do not currently have one.
- ☑ Other: There is no metadata management system within this use case, but the

4.6.2 If a metadata management system is present, what measures are in place to verify and protect its integrity?

4.6.3 Describe provenance as related to instrumentation, sensors or other devices.

□ We have potential machine-to-machine traffic provenance concerns.
□ Endpoint sensors or instruments have signatures periodically updated
□ Using hardware or software methods, we detect and remediate outlier signatures
□ Endpoint signature detection and upstream flow are built into system processing
□ We rely on third party vendors to manage endpoint integrity
□ We use a sampling method to verify endpoint integrity
□ Not a concern at this time
□ Other:

4.7 Data Life Cycle

4.7.1 Describe Archive Processes

□ Our application has no separate "archive" process
□ We offload data using certain criteria to removable media which are taken offline
□ we use a multi-stage, tiered archive process
□ We allow for "forgetting" of individual PII on request
□ Have ability to track individual data elements across all stages of processing, including archive
□ Additional protections, such as separate encryption, are applied to archival data
□ Archived data is saved for potential later use by applications or analytics yet to be built
□ Does not apply to our application
☑ Other: Resultant datasets are not archived per se, but the repositories do have a

4.7.2 Describe Point in Time and Other Dependency Issues

□ Some data is valid only within a point in time,
□ Some data is only valid with other, related data is available or applicable, such as the existence of a building, the presence of a weather event, or the active use of a vehicle
□ There are specific events in the application that render certain data obsolete or unusable
□ Point and Time and related dependencies do not apply
☑ Other: Data are relevant and valid independent of when accessd/used, but all data

4.7.3 Compliance with Secure Data Disposal Requirements

Per NCSL: "at least 29 states have enacted laws that require entities to destroy, dispose. . ."
http://www.ncsl.org/research/telecommunications-and-information-technology/privacy-and-security.aspx

□ We are required to destroy or otherwise dispose of data
☑ Does not apply to us
□ Not sure
□ Other:

4.8 Audit and Traceability

Big Data use case: SEC Rule 613 initiative

4.8.1 Current audit needs *

- □ We have third party registrar or other audits, such as for ISO 9001
- □ We have internal enterprise audit requirements
- □ Audit is only for system health or other management requirements
- ☑ No audit, not needed or does not apply
- □ Other:

4.8.2 Auditing versus Monitoring

- □ We rely on third party or O.S. tools to audit, e.g., Windows or Linux auditing
- □ There are built-in tools for monitoring or logging that are only used for system or application health monitoring
- □ Monitoring services include logging of role-based access to assets such as PII or other resources
- □ The same individual(s) in the enterprise are responsible for auditing as for monitoring
- □ This aspect of our application is still in flux
- ☑ Does not apply to our setting
- □ Other:

4.8.3 System Health Tools

- □ We rely on system-wide tools for health monitoring
- □ We built application health tools specifically to address integrity, performance monitoring and related concerns
- □ There is no need in our setting
- ☑ Other: Systems employed in the use case are operated and maintained by the

4.8.4 What events are currently audited? *

- □ All data access must be audited
- □ Only selected / protected data must be audited
- □ Maintenance on user roles must be audited (new users, disabled user, updated roles or permissions)
- □ Purge and archive events
- □ Domain-dependent events (e.g., adding a new sensor)
- □ REST or SOAP events
- □ Changes in system configuration
- □ Organizational changes
- □ External project ownership / management changes
- □ Requirements are externally set, e.g., by PCI compliance
- □ Domain-specific events (patient death in a drug trial)
- ☑ Other: None

4.9 APPLICATION PROVIDER SECURITY

4.9.1 Describe Application Provider Security *

One example of application layer security is the SAP ERP application

☐ There is a security mechanism implemented at the application level
☐ The app provider level is aware of PII or privacy data elements
☐ The app provider implements audit and logging
☐ The app provider security relies on framework-level security for its operation
☑ Does not apply to our application
☐ Other:

4.10 FRAMEWORK PROVIDER SECURITY

One example is Microsoft Active Directory as applied across LANs to Azure, or LDAP mapped to Hadoop. Reference: http://bit.ly/1f0VDR3

4.10.1 Describe the framework provider security *

☐ Security is implemented at the framework level
☐ Roles can be defined at the framework level
☐ The framework level is aware of PII or related sensitive data
☑ Does not apply in our setting
☐ Is provided by the Big Data tool
☐ Other:

4.11 SYSTEM HEALTH

Also included in this grouping: Availability, Resilience, Information Assurance

4.11.1 Measures to Ensure Availability *

☐ Deterrents to man-in-the-middle attacks
☐ Deterrents to denial of service attacks
☐ Replication, redundancy or other resilience measures
☐ Deterrents to data corruption, drops or other critical big data components
☑ Other: System resources are provided by the NASA Advanced Supercomputing

4.12 PERMITTED USE CASES

Beyond the scope of S&P considerations presented thus far, please identify particular domain-specific limitations

4.12.1 Describe Domain-specific Limitations on Use

None

4.12.2 Paywall

☐ A paywall is in use at some stage in the workflow
☑ Not applicable

5 CLASSIFY USE CASES WITH TAGS

The questions below will generate tags that can be used to classify submitted use cases. See http://dsc.soic.indiana.edu/publications/OgrePaperv11.pdf (Towards an Understanding of Facets and Exemplars of Big Data Applications) for an example of how tags were used in the initial 51 use cases. Check any number of items from each of the questions.

5.1 DATA: APPLICATION STYLE AND DATA SHARING AND ACQUISITION

□ Uses Geographical Information Systems?
□ Use case involves Internet of Things?
□ Data comes from HPC or other simulations?
□ Data Fusion important?
□ Data is Real time Streaming?
☑ Data is Batched Streaming (e.g. collected remotely and uploaded every so often)?
□ Important Data is in a Permanent Repository (Not streamed)?
□ Transient Data important?
☑ Permanent Data Important?
☑ Data shared between different applications/users?
□ Data largely dedicated to only this use case?

5.2 DATA: MANAGEMENT AND STORAGE

☑ Application data system based on Files?
□ Application data system based on Objects?
□ Uses HDFS style File System?
☑ Uses Wide area File System like Lustre?
□ Uses HPC parallel file system like GPFS?
□ Uses SQL?
□ Uses NoSQL?
□ Uses NewSQL?
□ Uses Graph Database?

5.3 DATA: DESCRIBE OTHER DATA ACQUISITION/ ACCESS/ SHARING/ MANAGEMENT/ STORAGE ISSUES

5.4 ANALYTICS: DATA FORMAT AND NATURE OF ALGORITHM USED IN ANALYTICS

☑ Data regular?
□ Data dynamic?
□ Algorithm O(N^2) ?
□ Basic statistics (regression, moments) used?
□ Search/Query/Index of application data Important?
☑ Classification of data Important?
□ Recommender Engine Used?
□ Clustering algorithms used?
□ Alignment algorithms used?
□ (Deep) Learning algorithms used?
□ Graph Analytics Used?

5.5 ANALYTICS: DESCRIBE OTHER DATA ANALYTICS USED

Examples include learning styles (supervised) or libraries (Mahout).

None

5.6 PROGRAMMING MODEL

- □ Pleasingly parallel Structure? Parallel execution over independent data. Called Many Task or high throughput computing. MapReduce with only Map and no Reduce of this type
- □ Use case NOT Pleasingly Parallel -- Parallelism involves linkage between tasks. MapReduce (with Map and Reduce) of this type
- □ Uses Classic MapReduce? such as Hadoop
- □ Uses Apache Spark or similar Iterative MapReduce?
- □ Uses Graph processing as in Apache Giraph?
- □ Uses MPI (HPC Communication) and/or Bulk Synchronous Processing BSP?
- □ Dataflow Programming Model used?
- □ Workflow or Orchestration software used?
- □ Python or Scripting front ends used? Maybe used for orchestration
- □ Shared memory architectures important?
- □ Event-based Programming Model used?
- □ Agent-based Programming Model used?
- □ Use case I/O dominated? I/O time > or >> Compute time
- □ Use case involves little I/O? Compute >> I/O

5.7 OTHER PROGRAMMING MODEL TAGS

Provide other programming style tags not included in the list above.

5.8 PLEASE ESTIMATE RATIO I/O BYTES/FLOPS

Specify in text box with units.

Do not have the data to develop this ratio

5.9 DESCRIBE MEMORY SIZE OR ACCESS ISSUES

Specify in text box with any quantitative detail on memory access/compute/I/O ratios

None

6 Overall Big Data Issues

6.1 Other Big Data Issues

Please list other important aspects that the use case highlights. This question provides a chance to address questions which should have been asked.

6.2 User Interface and Mobile Access Issues

Describe issues in accessing or generating Big Data from clients, including Smart Phones and tablets.

No mobile access is applicable to this use case.

6.3 List Key Features and Related Use Cases

Put use case in context of related use cases. What features generalize and what are idiosyncratic to this use case?

6.4 Project Future

How are application and approach (e.g., hardware, software, analytics) expected to change in future?

Processing will be improved with newer and updated algorithms. This process may also be applied to future datasets and processing systems (Landsat 8 and Sentinel-2 satellites, for example)

7 Workflow Processes

Please answer this question if the use case contains multiple steps where Big Data characteristics, recorded in this survey, vary across steps. If possible flesh out workflow in the separate set of questions. Only use this section if your use case has multiple stages where Big Data issues differ significantly between stages.

7.1 Please comment on workflow processes

Please record any overall comments on the use case workflow.

The processing for this use case is a 32 stage pipeline. The WELD-Overview diagram presents a five stage high level workflow. Workflow details are not available at this time, but may be provided in the future if time allows. A top level workflow diagram is being

7.2 Workflow details for each stage *

Description of table fields below:

Data Source(s): The origin of data, which could be from instruments, Internet of Things, Web, Surveys, Commercial activity, or from simulations. The source(s) can be distributed, centralized, local, or remote. Often data source at one stage is destination of previous stage with raw data driving first stage.

Nature of Data: What items are in the data?

Software Used: List software packages used

Data Analytics: List algorithms and analytics libraries/packages used

Infrastructure: Compute, Network and Storage used. Note sizes infrastructure -- especially if "big".

Percentage of Use Case Effort: Explain units. Could be clock time elapsed or fraction of compute cycles

Other Comments: Include comments here on items like veracity and variety present in upper level but omitted in summary.

7.2.1 Workflow Details for Stage 1

Stage 1 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.2 Workflow Details for Stage 2

Stage 2 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.3 Workflow Details for Stage 3

Stage 3 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.4 Workflow Details for Stage 4

Stage 4 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

7.2.5 Workflow Details for Stages 5 and any further stages

If you have more than five stages, please put stages 5 and higher here.

Stage 5 Name	
Data Source(s)	
Nature of Data	
Software Used	
Data Analytics	
Infrastructure	
Percentage of Use Case Effort	
Other Comments	

Description of NIST Public Working Group on Big Data

NIST is leading the development of a Big Data Technology Roadmap. This roadmap will define and prioritize requirements for interoperability, portability, reusability, and extendibility for big data analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data. To help develop the ideas in the Big Data Technology Roadmap, NIST is creating the Public Working Group for Big Data.

Scope: The focus of the NBD-PWG is to form a community of interest from industry, academia, and government, with the goal of developing a consensus definitions, taxonomies, secure reference architectures, and technology roadmap. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable Big Data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.

For more, refer to the web site at http://bigdatawg.nist.gov/home.php