

# BIG DATA USE CASE TEMPLATE 2

## *NIST Big Data Public Working Group*

This template 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 template 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 template can also be completed in the Google Form for Use Case 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>.

## TEMPLATE OUTLINE

1	OVERALL PROJECT DESCRIPTION .....	2
2	BIG DATA CHARACTERISTICS .....	4
3	BIG DATA SCIENCE .....	5
4	GENERAL SECURITY AND PRIVACY .....	7
5	CLASSIFY USE CASES WITH TAGS .....	9
6	OVERALL BIG DATA ISSUES .....	11
7	WORKFLOW PROCESSES .....	12
8	DETAILED SECURITY AND PRIVACY .....	16

### General Instructions:

Brief instructions are provided with each question requesting an answer in a text field. For the questions offering check boxes, please check any that apply to the use case. .

No fields are required to be filled in. Please fill in the fields that you are comfortable answering. The fields that are particularly important to the work of the NBD-PWG are marked with \* .

Please email the completed template to Wo Chang at [wchang@nist.gov](mailto:wchang@nist.gov).

**NOTE:** No proprietary or confidential information should be included.

**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.

Yelp Insights by Abraca Data

## 1.2 USE CASE DESCRIPTION \*

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

Yelp Inc. is an online portal which provides crowd-sourced reviews for businesses and extends other services like restaurant table reservation. During the COVID-19 pandemic situation, ordering food online was the preferred source for people due to shelter in place orders issued by the Government and their safety. Yelp being the most popular review portal provided an edge to make informed decisions for the people during the Pandemic. This can be an opportunity for business owners to expand their sales by keeping their Yelp profile competitive with increase in dependency of people on reviews and ratings.

Our project focuses on analyzing the current trends of correlation between income and restaurants business, impact of business attributes, customer 's reviews on business ratings and their price range

In this project we will be utilizing AWS-provided services and third-party software, Python, MySQL and Tableau

## 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
Vani Kancherlapalli		vani.kancherlapalli@sjsu.edu	Author	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Manisha Palliwal		manisha.paliwal@sjsu.edu	Author	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chidroop Sagar		chidroop.sagar@sjsu.edu	Author	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
			Author	<input type="checkbox"/>	<input type="checkbox"/>

## 1.4 DOMAIN ("VERTICAL") \*

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

Restaurant Business

## 1.5 APPLICATION \*

Summarize the use case applications.

The project findings will help investors to figure out the feasibility of a new venture and guide existing owners how to upscale their business margins.

This can be an opportunity for business owners to expand their sales by keeping their Yelp profile competitive with increase in dependency of people on reviews and ratings.

## 1.6 CURRENT DATA ANALYSIS APPROACH \*

Describe the analytics, software, hardware approach used today. This section can be qualitative with details given in Section 3.6.

We have used ETL tool AWS GLUE, python notebook and AWS Redshift, ' COPY ' command to cleanse, flatten and filter unwanted columns and create ER relationship between entities as part of the data wrangling process.

There are two main datasets as part of the project (Yelp and US Census Household Income) and one supporting dataset (Zip Code Information).

Yelp file -Business (JSON file) Data was uploaded in S3 bucket and AWS Glue ETL was used to cleanse, filter required rows and process data into AWS Redshift Cluster

## 1.7 FUTURE OF APPLICATION AND APPROACH \*

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

Integrate Yelp Dataset with US population for each County to find correlation with Businesses Integrate Yelp

Dataset with US area for each County to find correlation with Businesses

Create an interactive Web Application of Data Analytics

Perform ML Sentiment Analysis on Business Reviews Perform deeper analysis of Reviewer comments

Identify areas of improvements using advanced statistical tools that businesses can incorporate for market value

## 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 discussed in a separate part of this template.

All team members participated in identifying the data sources and data wrangling, data visualisations and the Web app

## 1.9 PROJECT GOALS OR OBJECTIVES

Please describe the objectives of the use case.

Yelp Inc. is an online portal which provides crowd-sourced reviews for businesses and extends other services like restaurant table reservation.

During the COVID-19 pandemic situation, ordering food online was the preferred source for people due to shelter in place orders issued by the Government and their safety.

Yelp being the most popular review portal provided an edge to make informed decisions for the people during the Pandemic.

This can be an opportunity for business owners to expand their sales by keeping their Yelp profile competitive with increase in dependency of people on reviews and ratings.

Our project focuses on analyzing the current trends of correlation between income and restaurants business, impact of business attributes, customer ' s reviews on business ratings and their price range.

### 1.10 USE CASE URL(S)

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

1) [https://data.census.gov/cedsci/table?](https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true)

[q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true](https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSST1Y2019.S1901&hidePreview=true)

### 1.11 PICTURES AND DIAGRAMS?

Please email any pictures or diagrams with this template.

## 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.

1) <https://data.census.gov/cedsci/table?q=Income%20%28Households,%20Families,%20Individuals%29&g=0400000US01%240500000,02%240500000,04%240500000&tid=ACSSST1Y2019.S1901&hidePreview=true>

2) <https://www.yelp.com/dataset>

3) <https://www.unitedstateszipcodes.org/>

### 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.

Yelp file -Business (JSON file) Data was uploaded in S3 bucket and AWS Glue ETL was used to cleanse, filter required rows and process data into AWS Redshift Cluster

Yelp file -Business Ambience Attributes (nested string format) was fetched from Yelp business file and converted into JSON format using python 3 and after converting to csv using external tool uploaded in S3 in csv format.

Further data was copied from S3 to AWS Redshift cluster

US Census Household Income (csv file) was uploaded in S3 bucket and Glue ETL was used to cleanse, filter required rows and process data into AWS Redshift Cluster

### 2.3 VOLUME

Size	~ 3.3 GB total
Units	The yelp Dataset is from 2004 – 2020 The Household income is for 2019
Time Period	Static Data
Proviso	Yelp, US Household income, Zip Code Dataset

**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	NA as we implemented static data
Time Period	NA as we implemented static data
Proviso	NA as we implemented static data

**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.

We have extracted the data from 3 different sources and in different formats. The data was in JSON, csv and Text Format

## 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.

NA we implemented static data

# 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).

We have used ETL tool AWS GLUE, python notebook and AWS Redshift, 'COPY' command to cleanse, flatten and filter unwanted columns and create ER relationship between entities as part of the data wrangling process.

There are two main datasets as part of the project (Yelp and US Census Household Income) and one supporting dataset (Zip Code Information).

Yelp file -Business (JSON file) Data was uploaded in S3 bucket and AWS Glue ETL was used to cleanse, filter

### 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.

We have implemented Tableau as our third party software to show our Visualizations.

We connected Tableau to our Redshift Cluster and have conducted analysis.

We implemented Tableau Story method to provide visualization insights.

### 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.

- 1)The Yelp Business file had an ambience attribute field which was defined as a string instead of JSON.
- 2)The leading zeros in the zipcode dataset was being truncated hence was converted into a pipedelimited file using Python Notebook and later loaded into private S3 bucket.
- 3)We have the CSV file for the household data.
- 4)One of the Fields in the JSON was stored as a string, we had to write a Python program to flatten it.

### 3.4 METADATA

Please comment on quality and richness of metadata.

The Yelp Review and Business Dataset was in JSON format ( Semi - Structured) which was flattened into Tabular format

We have CSV and Text data. One of the Fields in the JSON was stored as a string, we had to write a Python program to flatten it.

### 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.

NA

### 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. This section should be reasonably precise so quantitative comparisons with other use cases can be made. Section 1.6 is qualitative discussion of this feature.

Tableau was our main Software where we have implemented for data Visualization . We used Tableau Story to provide insights.

We also hosted a website which consists our data visualizations.

## 4 GENERAL SECURITY AND PRIVACY

The following questions are intended to cover general security and privacy topics. Security and privacy topics are explored in more detail in Section 8. For the questions with checkboxes, please select the item(s) that apply to the use case. Depending on the answers below, the questions in Section 8 may not apply to your use case.

### 4.1 CLASSIFIED DATA, CODE OR PROTOCOLS

- ☐ Intellectual property protections
- ☐ Military classifications, e.g., FOUO, or Controlled Classified
- ☐ Not applicable
- ☒ Creative commons/ open source
- ☐ Other:

### 4.2 DOES THE SYSTEM MAINTAIN PERSONALLY IDENTIFIABLE INFORMATION (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:

### 4.3 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:

### 4.4 IS THERE AN EXPLICIT DATA GOVERNANCE PLAN OR FRAMEWORK FOR THE EFFORT?

Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.

- ☐ Explicit data governance plan
- ☐ No data governance plan, but could use one
- ☒ Data governance does not appear to be necessary
- ☐ Other:

**4.5 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 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:

**4.7 UNDER WHAT CONDITIONS DO YOU GIVE PEOPLE ACCESS TO YOUR DATA?**

Database and S3 buckets are limited to developers use only while visualizations are accessible to public

**4.8 UNDER WHAT CONDITIONS DO YOU GIVE PEOPLE ACCESS TO YOUR SOFTWARE?**

Our visualizations are accessible for public viewing as Tableau stories. We have hosted a website which is accessible to anyone.



## 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

We did not face any access or sharing or management or storage issues while working on this project using AWS.

## 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).

Tableau Desktop and Public

## 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.

Python and AWS GLUE.

**5.8 PLEASE ESTIMATE RATIO I/O BYTES/FLOPS**

Specify in text box with units.

~ 43 milliseconds for 2.5 MB

**5.9 DESCRIBE MEMORY SIZE OR ACCESS ISSUES**

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

Approximately of 3.5 GB of memory size was used in AWS S3

The Bucket was made accessible to required tools

## 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.

The Yelp Business file had an ambience attribute field which was defined as a string instead of JSON. As a result this attribute was parsed using Python Notebook and created into separate columnar values.

The leading zeros in the zipcode dataset was being truncated hence was converted into a pipedelimited file using Python Notebook and later loaded into private S3 bucket.

**6.2 USER INTERFACE AND MOBILE ACCESS ISSUES**

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

Website and Tableau Visualisation is accessible to everyone over multiple platforms

### 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?

Our project focuses on analyzing current patterns of Restaurant Business on basis of Yelp Dataset and supporting US Census Household income dataset.

We brought together the holistic view of below parameters which will enable existing business owners and new investors in Restaurant ventures to make informed decisions

Most popular cuisine

## 7 WORKFLOW PROCESSES

Please answer this question if the use case contains multiple steps where Big Data characteristics, recorded in this template, 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.

Downloaded dataset from Yelp, US Household Income, Zipcode  
 Load into AWS S3 bucket  
 AWS GLUE ETL to cleanse, filter, flatten data; Python Notebook for parsing nested JSON string  
 Load into AWS Redshift; data normalization for visualization  
 Tableau for visualization  
 Website to host application

### 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	Download datasets and load to AWS S3 private bucket (Source Files)
Data Source(s)	Yelp Dataset, unitedstateszipcode and Householdincome data which are mentioned in 2.1 (DATA source)
Nature of Data	Static Data
Software Used	AWS
Data Analytics	Required information was extracted from AWS Glue
Infrastructure	Local Machine - 16GB
Percentage of Use Case Effort	~ 10%
Other Comments	

**7.2.2 Workflow Details for Stage 2**

Stage 2 Name	Extract - Transform - Load
Data Source(s)	AWS S3
Nature of Data	Static data in JSON, CSV and text files
Software Used	AWS Glue, Python Notebook (Nested String Format)
Data Analytics	
Infrastructure	Local Machine - 16GB
Percentage of Use Case Effort	~ 40%
Other Comments	

**7.2.3 Workflow Details for Stage 3**

Stage 3 Name	Destination files / Tables
Data Source(s)	AWS Glue and S3 Bucket
Nature of Data	Cleaned data
Software Used	AWS RedShift, AWS RedShift COPY command
Data Analytics	
Infrastructure	Local Machine - 16GB
Percentage of Use Case Effort	~ 10%
Other Comments	

**7.2.4 Workflow Details for Stage 4**

Stage 4 Name	Data Analysis and Visualization
Data Source(s)	AWS RedShift Table
Nature of Data	Static
Software Used	Tableau
Data Analytics	
Infrastructure	Local Machine - 16GB
Percentage of Use Case Effort	~ 30%
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	Create and deploy website
Data Source(s)	Visualization
Nature of Data	Static
Software Used	HTML, CSS and JQuery
Data Analytics	
Infrastructure	Local Machine - 16GB
Percentage of Use Case Effort	~10%
Other Comments	

## 8 DETAILED 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 answers will be very beneficial to the NBD-PWG in understanding your use case. However, if you are unable to answer the questions in this section, the NBD-PWG would still be interested in the information gathered in the rest of the template. 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

### 8.1 ROLES

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

#### 8.1.1 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.

All team members participated in identifying the data sources and data wrangling, data visualisations and the Web app

#### 8.1.2 Investigator Affiliations

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

NA



**8.1.3 Sponsors**

Include disclosure requirements mandated by sponsors, funders, etc.

NA

**8.1.4 Declarations of Potential Conflicts of Interest**

NA

**8.1.5 Institutional S/P duties**

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

NA

**8.1.6 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.

NA

**8.1.7 Classified Data, Code or Protocols (Read only, question answered in Section 4.1)**

- ☐ Intellectual property protections
- ☐ Military classifications, e.g., FOUO, or Controlled Classified
- ☐ Not applicable
- ☒ Creative commons/ open source
- ☐ Other:

**8.1.8 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:

**8.1.9 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:

**8.1.10 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:

**8.2 PERSONALLY IDENTIFIABLE INFORMATION (PII)****8.2.1 Does the System Maintain PII? \* (Read only, question answered in Section 4.2)**

- ☐ 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:

**8.2.2 Describe the PII, if applicable**

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

NA

**8.2.3 Additional Formal or Informal Protections for PII**

NA

**8.2.4 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:

**8.2.5 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.

NA

**8.3 COVENANTS, LIABILITY, ETC.****8.3.1 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:

**8.3.2 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:

**8.4 OWNERSHIP, IDENTITY AND DISTRIBUTION****8.4.1 Publication rights (Read only, question answered in Section 4.3)**

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:

**8.4.2 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>

NA

**8.4.3 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>

NA

**8.4.4 Software License Restrictions**

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

NA

**8.4.5 Results Repository**

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

<https://github.com/vanikancherlapalli/Abraca-Data-228>

**8.4.6 Restrictions on Discovery**

Describe restrictions or protocols imposed on discoverable end points.

No restrictions

**8.4.7 Privacy Notices**

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

- ☐ Privacy notices apply  
☒ Privacy notices do not apply  
☐ Other:

**8.4.8 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:

**8.4.9 Describe and Key Management Practices**

NA

**8.4.10 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:

**8.4.11 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

**8.4.12 Describe the Identity Framework.**

NA

**8.4.13 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:

**8.5 RISK MITIGATION**

---

**8.5.1 Are measures in place to deter re-identification? \***

- ☐ Yes, in place
- ☐ Not in place, but such measures do apply
- ☒ Not applicable
- ☐ Other:

**8.5.2 Please describe any re-identification deterrents in place**

NA

**8.5.3 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:

**8.5.4 Is there an explicit data governance plan or framework for the effort?**

Data governance refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. (Read only, question answered in Section 4.4)

- ☐ Explicit data governance plan  
☐ No data governance plan, but could use one  
☒ Data governance does not appear to be necessary  
☐ Other:

**8.5.5 Privacy-Preserving Practices**

Identify any privacy-preserving measures that are in place.

NA

**8.5.6 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.) (Read only, question answered in Section 4.5)

- ☐ 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:

**8.6 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.

**8.6.1 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:

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

NA

**8.6.3 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: 
**8.7 DATA LIFE CYCLE****8.7.1 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: 
**8.7.2 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: 
**8.7.3 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:



**8.8 AUDIT AND TRACEABILITY**

Big Data use case: SEC Rule 613 initiative

**8.8.1 Current audit needs \* (Read only, question answered in Section 4.6)**

- ☐ 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:

**8.8.2 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:

**8.8.3 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:

**8.8.4 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:

## 8.9 APPLICATION PROVIDER SECURITY

---

### 8.9.1 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:

## 8.10 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>

### 8.10.1 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:

## 8.11 SYSTEM HEALTH

---

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

### 8.11.1 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:

## 8.12 PERMITTED USE CASES

---

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

### 8.12.1 Describe Domain-specific Limitations on Use

NA

### 8.12.2 Paywall

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

## **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 created 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>