**NIST BDWG Definitions Working Notes**

**M0024 Version 6 - 7/2/913**

**NIST Big Data Working Group: Definitions & Taxonomy Subgroup**

**Co-Chairs: Nancy Grady (SAIC), Natasha Balac(SDSC), Eugene Luster (R2AD)**

**Meetings: Mondays 11:00-13:00 EDT**

Guidelines:

Follow the Cloud Definitions document.

Fold taxonomy into reference architecture, again following cloud document

Sync with other subgroups for necessary and sufficient concepts

Restrict to what is different now that we have “big data”

Not trying to create taxonomy of the entire data lifecycle processes and all data types

Keep terms independent of a specific tool

Be mindful of terminology due to the context in different domains (e.g. legal)

“Big Data” and “Data Science” are currently composits of many terms.

Break down the concepts first, then define these two at the end

*Vertical Scalability* – increasing the capacity and capability of a single resource (faster CPU, bigger disk drive, more memory,…)

*Horizontal Scalability* – increasing the capacity and capability of a sytem by addition additional couple resources to the task at hand.

*Big Data* – representing the increasing scale of the dataset characteristics or the metrics on collection of the data into a system that forces a change in the architecture.

*Grid Computing* – extending a compute function horizontally across multiple resources

*Big Data Engineering* – the technologies for extending a system across multiple resources for data handling.

*Data Science* – extracting knowledge and benefit from a collection of data; a change in the traditional data lifecycle processes do a more exploratory or trending analysis

* Probabilistic or trending analysis; Correlation not causation;
* Data exploration - finding questions not finding answers
* Combining domain knowledge; analytics skills; programming expertise
* Data Characteristics for analysis – veracity, cleanliness, provenance, data types
* The process of sheparding data end-to-end through the data lifecycle to ensure the data is being handled properly and the desired results are correct.

Metrics

|  |
| --- |
| NOTE: not yet addressed guidance on how big to be called “Big” |

**Reference Architecture**

The Reference Architecture is discussing the use of a simple architecture, data comes in, you transform it using some infrastructure resources, and then you send the results out to be used by another transform.

We can reorder our concepts into these buckets, and the connectors between them. I’ve labeled them to help make it easier to follow.

1. Sources
2. Collection
3. Transformation
4. System topology for connecting transformation functions to infrastructure resources
5. Infrastructure
6. Export
7. Usage

**(1) Sources**

Sources have attributes that relate to the metadata for the origin of the data and the provider, characteristics (types) of the data, and the characteristics to describe the dataset as a whole.

Metadata Characteristics

*Meta-data* – data about data.

|  |
| --- |
| NOTE: There are subcategories of metadata, but there is no specific metadata related to “big data”. Data has always been described related to the origins, scope, etc of datasets. We will include only those concepts needed for completeness with new concepts. |

*Provenance* – the specific metadata that describes the origin (collection process) and subsequent transformations to the data…also can be considered the history of the data.

Data Element Characteristics

*Data type* is a description of the data file/record/etc.

Typically these are defined into:

*Structured* – like data in a database table, or semantically wrapped data

*semi-structured* – like XML with text as the elements inside of the markup

*unstructured* – like a document, that may in fact have formatting (like headings in Word) from which you can infer a way to break up a document into segments of text.

|  |
| --- |
| NOTE: A complete data taxonomy is beyond the scope of this work, and there are no new data types that have arisen for big data. |

For Analysts, data is broken into:

*Primary data* - Raw Data as originally collected

*Secondary data* – Data that has been organized into useful information

*Tertiary data* – Information that has been analyzed to produce knowledge/insight

*Value* – tertiary data that has been put into action to generate a benefit.

|  |
| --- |
| NOTE: Again these are useful distinctions for analysis, but don’t represent anything new in big data |

Data Formats:

*Text*

*Binary*

*Semantic*

|  |
| --- |
| NOTE: Again these are useful distinctions for analysis, but don’t represent anything new in big data |

Dataset as Rest

*Volume* – amount of data

*Variety* – different character/structure in different datasets

* data types (structured, unstructured, etc…)
* differing grids (like GIS data) on which the data resides (grid can be in metadata)
* differing time scales

scaling can force you into different technologies (DB lookup -> semantic)

|  |
| --- |
| NOTE: This concept is not new, but dynamic scalability is being addressed |

*Diversity* – numbers of datasets – data mashups across domains

*PII* – Personally identifiable information

*Mosaic Effect* - privacy from number of dataset – combining datasets that do not have PII and result in identification and loss of privacy

*Complexity* – inter-relatedness of data records (such as found in genome)

|  |
| --- |
| NOTE: This concept is not new, but does lead to engineering and data science complexity |

**(2) Collection - the process of bringing an external source into the transformation process**

Dataset in motion

*Velocity* – rate of flow of data

|  |
| --- |
| NOTE: Data streaming is the traditional term for high velocity data |

*Variability* – changing velocity

Variability (Context) - change structure, content, etc…

Data portability – data can be transmitted in a machine interpretable fashion?

Data availability – can be accessed externally (like open data initiative)

DaaS

APIs

? data services

**(3) Transformation**

Data Lifecycle

* Collection –> data
* Curation –> information
* Analysis –> knowledge
* Action -> value

*Transformation* – change in the state of a data element or dataset encompassing any one of three traditional categories

* *Curation* – turning the raw data into organized information; cleansing the data, removing outliers, converting to standard units or vocabulary
* *Pre-analytics* – preparing the data for analytics which changes the data into new forms to prepare for analytics; interpolation, regridding, …
* *Analytics* – turning information into knowledge, which can include synthesizing new data not just a re-packaging of the information; trending, predictive modeling, aggregation, summarization,…

**(4) ??? topology connecting transform resource(s) to infrastructure resource(s)**

Data Process Changes

* Data Warehouse -> Curation=ETL with storage after curation
* Volume -> storage before curation; storing raw data; ELT
* Velocity -> collection+curation+analytics (alerting) before storage
* Downsizing
* Splitting applications, moving applications

Combining Big and not-Big

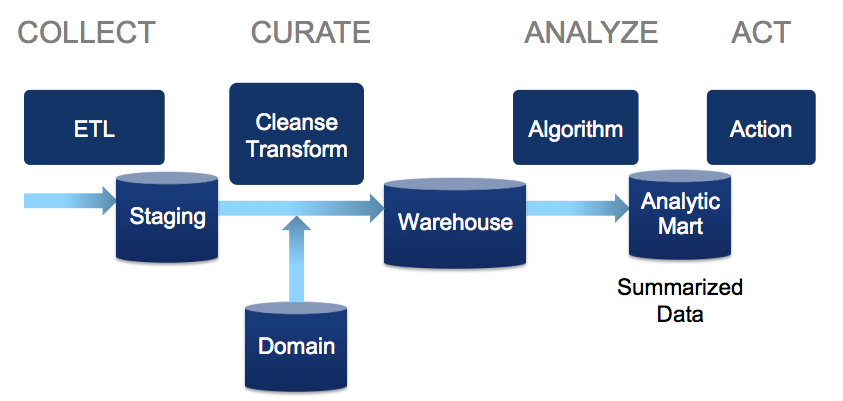
MapReduce

Replication

Changing Process Ordering based on dataset characteristics

Not sure how this one folds in, since it can represent the ordering of multiple single step (source, collection, transform, topology, infrastructure, export, usage) cycles

Traditional Data Warehouse; ETL

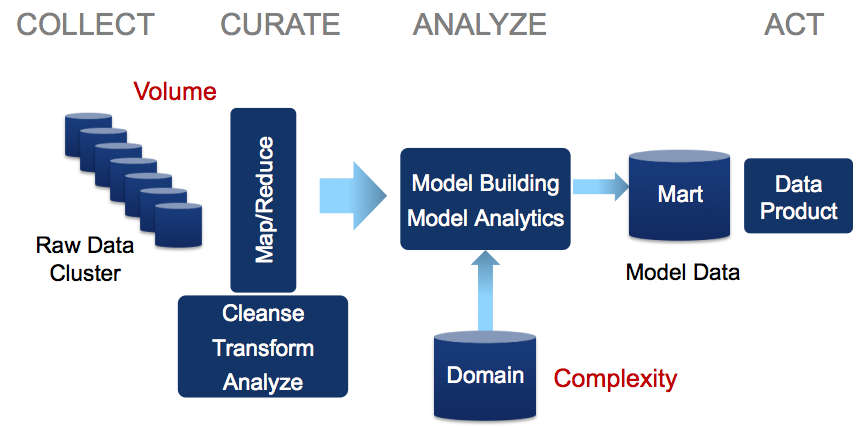


Volume

Store raw before transform

ELT – process driven

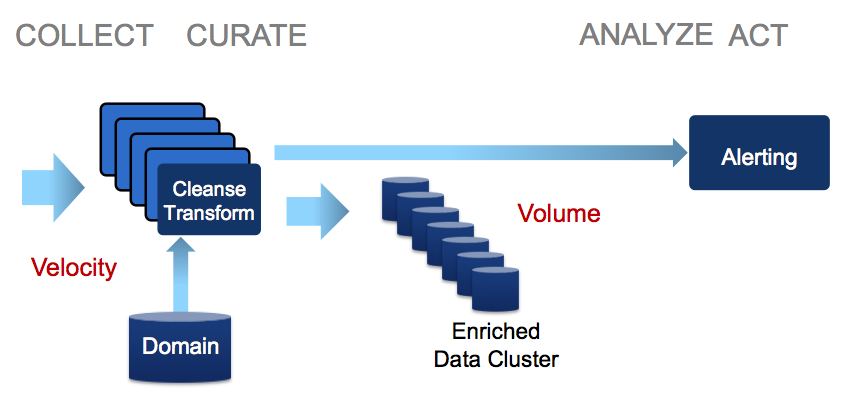
Schema on read



Velocity

Data streams

Persist after analyze

****

Variety – many datasets

Don’t ETL until runtime?

also where you do the filtering

look at these for ideas in this topic in Wikipedia – communications between stage

activeMQ

TIBCO

UIMA

<Put in data consistency up in persistence section as a capability not a techonolgy?>

overlay networks, command and control networks, peer-to-peer

Put in concepts of synchronizing data?

Old style was master-slave, now peer-to-peer

**(5) Infrastructure – not yet been addressed**

*Persistence* – structure for storing data

Flat files (text, binary)

HDFS

Messages

Markup

Relational database – settled on SQL

Content Management Systems – documents, messages, etc

- is this just another form of RDB

NoSQL (no SQL, new SQL, not only SQL)

* Big table
* Name-value pairs
* Graph – node/link
* Document

Tier-ed storage:

In-memory

Cache

SSD

hard disk drive

networked drives

archive -

Distributed:

Local -

Cluster – multiple local or distributed resources in a system

Distributed -

Federated -

network-based (horizontal scalability)

Fault tolerance

**(6) Export**

**(7) Usage**

**(8) Security**

cell, row, column, dataset, perimeter

Aggregation

Waves of Technology – <sync this with Roadmap>

Horizontal Scaling

Vertical Scaling

Indexing – row/column

**(9) Management**

**(10) Cloud Computing**

**(11) Networking**

**(7) Changing Analytics Paradigm – Data Science**

Statistics – rigorous causal analysis of carefully sampled data

Data Mining – approximate causal analysis of repurposed data carefully sampled

Data Science – probabilistic analysis/trending of large selection or even entire dataset

Data Science – correlation not necessarily causation

Data Science – determine the questions and not the answers

Data Science – getting an answer by solving a simpler problem

Data Science - Venn diagram – domain, analytics, programming – these can go to roles

Data Scientist -

? what curriculum is a core for saying you’re a “data scientist”

qualitative characterization

Do we need certification to distinguish this, or is it just implying you need to work collaboratively with more skill sets than before

Veracity – precision/accuracy/timeliness of the data

Provenance – a particular kind of metadata about the history (pedigree) of the dataset (how analyzed, etc) – <need to make this specific to big data>

Cleanliness/Quality – more data vs more

Obsolescence

Filtering

Data integration/matching – different primary, but secondary fields that can be correlated

Crawlers

Bots

Network Throtting

Filtering

Metrics (to understand when you need a “new” architecture)

Service Level Agreements

May require data reduction

Scalability

[review requirements and reference architecture subgroups]

**(10) Taxonomy**

Role

Responsibilty

Component

Subcomponent

**Data Process**

Collection -> raw data

Curation -> information

Analysis -> knowledge

Action -> value

**Architecture Layers**

Data Source

Transformation

Infrastructure

Usage/consumption

Role – an external entity outside the scope of the component of interest

Capabilities – the internal entity performing a function (person, software, infrastructure) – logical/functional

Cloud taxonomy was from an engagement model

Are we talking business roles here, or technical roles

Role – data provider

Role – data transformer

Role – data user

Infrastructure

Security

Business – governance

Express concepts of vertical and horizontal scalability

vertical is shared nothing? shared data?

Horizontal is “independent” resources

**Roles/Capabilities/Person or agent**

Data reSource Provider – Role? Characteristics

Data Transporter - Connection or messaging from external role to internal role for capability

-> data in motion characteristics (Velocity/Variability,…)

Data element

Dataset characteristics (volume/variety/diversity…)

Metadata

Data Transforms - capability

* Curation & pre-analytics
* Analytics

Data infrastructure – capability

Persistence format

Persistence location

Data user - Role

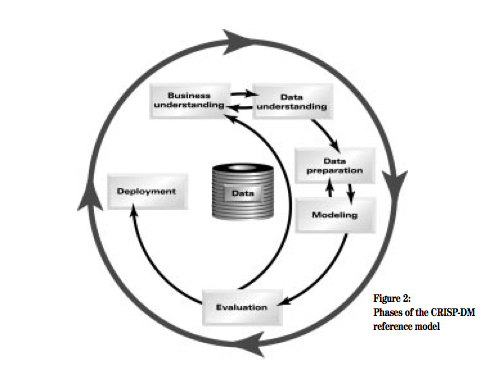
Appendix A The following is for reference purposes only.

We are not trying to follow any specific process, this one is given as an example when we want to determine what are the **new processes** from “big data”

There are many definitions of a data lifecycle. For the taxonomy we’ll need to see what processes are new, and what of the current processes we need to include.

We can look **for example** at the CRISP-DM set of data processes (will upload pdf to NBDWG site) to see if there are any changes due to “big data” from this lifecycle set of processes. CRISP-DM was a consortium led by NCR and SPSS and OHRA in 1999/2000, which created a description of the processes in the data lifecycle (not the tools or techniques).

We can look through these to help guide our taxonomy, or alternatively see what cannot be accommodated in this set of processes now that we have big data.



Outline of Data processes (not necessarily big data, just general data processes)

Notice that every step may determine that you step back to a “prior” process.

Business Understanding

Objectives goals

Data Mining goals

Plan

Data Understanding

Collect initial data

Describe data

Explore data

Verify Data Quality

Data Preparation

Select data

Clean data

Construct data

Integrate data

Format data

Modeling

Select modeling technique

Generate test design

Evaluation

Evaluate results

Review process

Determine next steps

Deployment

Plan deployment

Plan monitoring and maintenance

Produce final report

Review project

Appendix B

**Open questions:**

How do we define metrics to indicate when it’s “big data”?

How do we define processes and metrics to guide procurement?

What of the cloud infrastructure/terms/etc do we need to modify for big data?

How much do we need to consider data types?

(see definition category 1)

Correspondingly do we need to consider the objectives of the data analysis?

Is there something in the scalability of the internet of things we need to consider?

(see category 6)

What security concepts are needed for “big data”?

Say something like data element/row/column security?

What concepts do we need to include from the open data initiative?

What concepts do we need from data repositories, e.g. data.gov

Is there value in other collaborative tools like social for asynchronous discussion?

**Appendix C Initial Approach - Break concepts into categories**

Data elements

* Concepts that are needed later,
* such as raw data -> information -> knowledge -> wisdom
* (metadata – not clear what’s different)
* Complexity – dependent relationships across records

Dataset at rest

* Characteristics: Volume, Variety (~~many datasets~~; data types~~; timescales~~)
  + Diversity (many datasets/domains)
* Persistence (flatfiles, RDB, NoSQL incl Big Table, Name-Value, Graph, Document)
* Tier-ed storage (in-memory, cache, SSD, hard disk, network,…\_
* Distributed: local, multiple local resources, network-based (horizontal scalability)
* Fault tolerance

Dataset in motion

* Velocity (flow rate), Variability (changing flow rate; structure; temporal refresh)
* Accessibility like Data-as-a-Service

Data Processes

* Collection –> data
* Curation –> information
* Analysis –> knowledge
* Action -> value

Data Process Changes

* Data Warehouse -> Curation=ETL with storage after curation
* Volume -> storage before curation; storing raw data; ELT
* Velocity -> collection+curation+analytics (alerting) before storage
* Downsizing
* Splitting applications, moving applications

Data Science – multiple terms

* Probabilistic or trending analysis; Correlation not causation; finding questions
* Combining domain knowledge; analytics skills; programming expertise
* Data Characteristics for analysis – veracity, cleanliness, provenance, data types

Metrics

Not worked on these yet – how big to be “Big”

Combining Big and notbig

Implicit PII

Taxonomy

waiting to follow Reference Architecture, Security

Line up hardware/software/network concepts with Reference Architecture

Line up roles with use cases – try to follow Cloud Taxonomy