# **NIST Big Data Taxonomies**

**Note**: this outline is an update to M0353, a conversion of the mindmap **[link]**[**:**](http://www.mindmeister.com/322462463)[**https://www.mindmeister.com/322462463**](https://www.mindmeister.com/322462463)hierarchies [created in November of 2014]. Some adjustments to the mindmap have occurred since; this doc will be updated shortly to reflect the current state of the mindmap. Also, some additional notes not in the mindmap, are annotated here.

Characteristics? Processes?

## Big Data RA Taxonomy - Level 1: Roles - Level 2: Activities - Level 3: Components - Level 4: Sub Components

### Data Provider, actors: - Enterprises - Public Agencies - Researchers & Scientists - Search Engines - Web, FTP, etc Applications - Network Operators - End Users *the RA taxon starts with system orchestrator (move?)*

#### (Sources) Data Capture from Sources

##### Online

###### Web Browsers

###### Sensors

###### Deep Packet Inspection Devices (e.g. Bridge, Router, Border Controller)

##### Offline

###### Public Records

###### Internal Records

#### Data Persistence

##### hosting

###### Internal hosting

###### External hosting

###### Cloud hosting

##### Data Retention Policy

#### (Prep) Data Scrubbing

##### Removing Personally Identifiable Information

##### Randomization (for implicit PII)

#### Data Annotation / Metadata creation

##### ontology

##### within a data file format

#### Access Rights Management

##### Specified by Data Sources

##### Specified by Data producer

##### Personal Identifier (PII) access rights

#### Access Policy Contracts

##### Policy

###### Primary Rights

###### Secondary Rights

##### Agreements

#### Data Distribution APIs

##### File Transfer Protocol

###### FTP

###### Streaming

##### compression techniques

###### single compressed file

###### Split compressed file

need to describe this one, to functionally describe the new BZIP2

##### Authentication

##### Authorization

#### Capabilities Hosting

##### Provide query access without transferring the data

##### Allow analytic tools to be sent to operate on the data sets

#### Data Availability Publication

##### Web description

##### Services Catalog

##### data dictionaries

##### Advertising

### System Orchestrator, actors: - Business Leadership - Consultants - Data Scientists - Information Architects - Software Architects - Security Architects - Privacy Architects - Network Architects

#### Business Ownership Requirements and Monitoring

##### Business Goals

##### Targeted Business Action

##### Data Provider contracts and SLAs

##### Data Consumer contracts and SLAs

##### Capabilities Provider Negotiation

##### Make/Buy Cost Analysis

#### Governance Requirements and Monitoring

##### Policy compliance requirements and monitoring

##### Change management process definition

#### Data Science Requirements and Monitoring

##### Data Source Selection

###### Data Description

###### Data Location

###### File Types

###### File Attributes

###### data provenance evaluation

##### Data Collection Requirements and Monitoring

##### Data Curation Requirements and Monitoring

##### Data Analysis Requirements and Monitoring

##### Data Visualization Requirements and Monitoring

##### Application Type [almost in order?]

###### ex. Streaming

###### ex. Aggregation

###### ex. Integration

###### ex. Transfer

###### ex. Search

###### ex. Statistics

###### ex. RT Analytics

###### ex. Batch Analytics

###### ex. Interactive Annotation

###### ex. Others

#### System Architecture Requirements and Monitoring

##### Data Process requirements

##### Software component determination

##### Hardware component determination

##### Logical Data Modeling and Partitioning

##### Data export requirements

##### Scaling Requirements

###### Terabytes

###### Petabytes

###### Exabytes

###### Zettabytes

###### Etc.

### Big Data Application Provider, actors: - Application Specialists - Platform Specialists - Consultants

#### Data Collection Processes

##### transport

###### secure transport

###### protocol

###### Network

##### Data

###### message

###### binary blob

###### text blob

##### Metadata

###### Structural

SQL catalog tables

W3C XML Schema

W3C XPath and Xquery

Avro schema

###### Catalog

Dublin Core

Provenance

who created/modified/when

Security

Who can Access

Audit

Who did Access

Domain

namespace

Media

EXIF

###### Semantic

ISO 5218 Representation of Sexes

ISO 3166 Names of Countries and Subdivisions

Taxonomies

Ontologies

RDFS/OWL

###### Correlation

relational Model referential constraints

RDF Schema

Links

FOAF

###### Integration

field equivalence

###### Application/Domain Specific

temporal information

valid time

transaction time

geospatial information

spheroid

grid

retail

price histories

supplier information

analytic tools for data

##### Data Parsing

#### (Prep) Data Preparation Processes

##### Data Conversion

##### (Cleaning) Data Cleansing

##### Data Integration

##### Calculated field creation

##### (Extract) Feature Extraction

###### time series

###### audio

Fourier Transform

###### image

biometric markers

fingerprint points

object identification

edge detection

###### video

object in motion

###### text

keywords

n-grams

###### GIS

###### graph

Connection

density

reachability

connectedness

Distance

Geodesic distance

Flow

##### Data summarization

##### Data partition implementation

##### Data Storage preparation

##### Data Virtualization Layer

##### Schema-on-read

##### query

#### (Analysis) Data Analytics Processes

##### Metadata Matching Processes

###### Lexical

Exact Match

Synonym match

Pattern Match

###### Semantic

semantic similarity

###### Statistical

distinct value analysis

data distribution analysis

##### Data In-Motion Characteristics *sub of analysis*

###### Input/Output data rate

##### Data Analysis Complexity *sub of analysis*

###### Computational Complexity

complexity of the processing task

###### Machine Learning Technique

###### Data Extent Complexity

size of the data that has to be accessed

###### Data Location Complexity

in memory or across multiple locations

##### Analytics Latency Requirement *total disconnect here*

###### Real-Time Analytics - Streaming

Low Latency ~ tens of milliseconds

###### Near Real-Time

Medium to high latency ~ few seconds to minutes

###### Batch

High Latency ~ seconds to hours

Supervised Learning

Linear Algebra

Sparse

Dense

Machine Learning implementation

Correlation

Classification

Data Fusion

Data Mining

Artificial Intelligence

Pattern Recognition

Predictive Modeling

Regression

Clustering

Spatial Analysis

Audio Analysis

Time Series Analysis

Visual Analysis

ex. Images

ex. Video

Natural Language Processing

Parts of Speech

statistical

Others

##### Visualization *issue of having all vis12n nested under one term / location.*

###### Exploratory Data Visualization

Discover - Data Browsing

Multidimension

ex. 2D/3D

Multiresolution

Interaction

Animation

Simulation

Statistical Graphics

Surface Rendering

Volume Rendering

Heat Maps

Outliers

Boundary Conditions

###### Explicatory Data Visualization

Confirm - Examining Analytics Output

Charting

Graphing

Near real time presentation

View analytics-extracted features

###### Explanatory Data Visualization

Conclude - Telling the Story

Reports

Infographic

Business Intelligence

Customer summarization presentation

#### Human-in-the-loop rapid analytics process

Isn't this really Workflow support?  I like that better than human in the loop as some of it automated for a user.

##### discovery

##### hypothesis

##### hypothesis testing

##### alternate methods

#### (Access) Access

##### Data export API Processes

###### API

###### Protocol

###### Query Language

###### Others

##### Data Charging Mechanisms

##### Consumer Analytics Hosting

##### Analytics-as-a-service hosting

##### Note: This one should mirror everything in the Data Provider, since the Data Consumer views this system as their data provider

### Big Data Framework Provider, actors: - In-house Clusters - Data Centers - Cloud Providers

#### Infrastructures

##### Cluster Implementation

###### Operating System

To me Operating System is an attribute of the cluster implementation - not of the framework itself. Ideally, the framework shields the external actors from knowing the underlying OS and cluster implementation. This is certainly true for Hadoop.

Linux

Machintosh

Windows

Etc.

###### Physical Resource Implementation

Server

Cluster

multi-site cluster

###### Virtual Resource Implementation

###### Logical Distribution

peer-to-peer

Master-Slave

Multiple masters with slave nodes

###### System Administration

##### Storage Implementation

The Roadmap document breaks this out into more granularity as: Local Disks, /Filesystems, HW/SW, RAID, SAN, NAS, Distributed File systems, Distributed Object Stores *which are grandchildren, where are the parents*

###### Local Disks

###### HW/SW RAID

###### SAN

###### NAS

###### Distributed Filesystems

###### Distributed Object Stores

###### Storage Administration

Block and Object stores are generally more logical implementations.

Also it [*what “it” is referred to?]* should either be under cluster implementation or data services.

##### Networking Implementation

###### Physical Network

Protocol

Fault Tolerance

Security/Encryption

###### Virtual Network

Protocol

Fault Tolerance

Security/Encryption

###### Network Administration

#### Platforms

##### Logical Storage

###### File Systems

file type

Delimited

Fixed Length

Binary

extent

local

distributed

###### Simple Tuple

Relational

Relational Database

when have known structured data with predictable production queries

NewSQL (ACID compliant in-memory)

Scalable RDB, queries efficiently

NoSQL (Non-Relational) tables

Row-oriented

Column-oriented

###### Complex Tuple

Field queries

NoSQL Document Store

Plain Text

JSON

BSON

Markup

HTML

XML

no Field queries

No SQL Key-Value

On Disk

In-memory

Queues

Can someone give me an example of a streaming logical storage organization. By definition to me streaming is data in motion not at rest. I can store streamed data in any number of logical models. If that is the case we should add FIFO and LIFO buffers here.

FIFO

LIFO

###### Graph

Property Graph

Non-Native (global index)

Native (index-free-adjacency)

Hyper-graph

Triple-Stores

##### Data Query Implementation

This is just way too limiting in nature and very focused on SQL style interfaces. What about query languages like the Lucene query syntax, or the MONGODB JSON query language.

This also ignores the HUGE variety of APIs used for bigtable and other DB interfaces (Hibernate's object store API for example).

###### SQL

Single Node

JDBC

SQL/CLI (ODBC)

Distributed Node

Indexed File System

Distributed File System

like new SQL on Hadoop tools

Distributed and Parallel

Standard

Parallel

Cluster-based

Distributed

Local Federated

###### API/UDF

###### SparQL

Xquery

###### Custom

##### Data Semantics

###### Data Registry

###### Semantic Representation

###### Ontology

##### Data Surety

###### Available, Consistent, Isolated, Durable

###### Basically Available, Soft state, Eventually Consistent

Session affinity

Logical Consistency

Eventual Consistency

Replication Durability

CAP tradeoffs

Write Quorum

###### Business Transactions

Compare and Set (CAS)

Version Stamp

Compare and Set

Counter

Random GUID

Content Hash

Time Stamp

Vector Stamp

Version Clock

Version Vector

##### Database Administration

#### Processing Frameworks

##### Schema Information (metadata)

How is this different than information in a Data Registry, or in a semantic catalog tied to the ontology.  I would combine the Data Domains and this area into one at this level.

###### Queryable

###### Schema on demand

###### Pre-knowledge

##### Query Frameworks

###### Relational

SQL

###### Arrays

Array Query Language

###### Distributed

batch

HiveQL

interactive

DrQL and Drill

real-time

StreamSQL

In-Memory database queries

F! with Distributed Transactional SQL

##### Application Framework

###### automation services

###### test services

###### analytics hosting services

###### data quality service

###### workflow services

##### Batch Processing Frameworks

In 2004, Phillip Colella working on DARPA’s High Productivity Computing Systems (HPCS) program developed a list of algorithms for simulation in the physical sciences that became known as the “Seven Dwarfs” (Colella, 2004).  More recently David Patterson and Katherine Yelick of the University of California – Berkley modified and extended this list to 13 shown in the table below based on the definition where “A dwarf is an algorithmic method that computes a pattern of computation and communication” (Patterson & Yelick)

###### MapReduce

map only

classic mapreduce

iterative mapreduce

loosely synchronous

###### Spectral methods

###### N-Body methods

###### Grids

Structured

Unstructured

###### Combinational Logic

###### Graph Traversal

###### Backtrack and Branch-and-Bound

###### Dynamic Programming

###### Graphical Models

###### Finite State Machines

##### Stream Processing Frameworks

###### Complex Event Processing

##### Resource Management Frameworks

##### Processing Framework Administration

### Data Consumer, actors: - End Users - Researchers - Applications - Systems

#### Search & Retrieve

#### Download

#### Analyze locally

#### Reporting

#### Visualization

#### Data to use for their own processes

### Big Data Security and Privacy: - Corporate Security Officer - Security Specialist -

#### Security and Privacy Policy Requirements

##### International

##### National

##### Enterprise

#### Security and Privacy Monitoring

##### Security Protection Requirements and Monitoring

###### Security against Denial of Service

Cryptographic Protocols proactively resistant to DoS

###### Big Data for Security

Analytics for Security Intelligence

Data-driven Abuse Detection

Event Detection

Forensics

###### Security negotiation with actors

###### Security Compliance Analysis

##### Data Provenance Monitoring

###### End-point Input Validation

###### Communication Integrity

###### Authenticated Computations

Trusted Platforms

Cryptographically Enforced

###### Granular Audits

###### Control of Valuable Assets

Lifecycle Management

Retention, Disposition, Hold

Digital Rights Management

##### Data Privacy Requirements and Compliance Monitoring

###### Communication Privacy

###### Confidentiality

Access Policies

Systems Enforced

Cryptographically Enforced

###### Computing on Encrypted Data

Searching and Reporting

Outsourcing of Computation

###### Secure Data Aggregation

###### Key Management

###### Implicit PII determination

### System Management, actors: - In-house Staff - Data Center Management - Cloud Providers

I really see this as a component of the orchestration aspect.

#### Provisioning

#### Configuration

#### Package Management

#### Software Management

#### Backup Management

#### Capability Management

#### Resources Management

#### Data Management

##### Big Data Lifecycle Management

###### (needs helps from NARA)

##### Administration

#### Performance Management

#### Security & Policy Management

## System Data Taxonomy - Level 1: Object - Level 2: Attributes - Level 3: Characteristics - Level 4: Sub Characteristics

### Element

#### Data Format see ISO 195:2012(E)

#### value

#### Metadata

##### Veracity

###### temporal validity

##### units

##### See ISO 11179

### Record (collection of Elements)

#### format

##### Structured

###### e.g. comma separated list of elements

##### Semi-structured

###### e.g. document with sections

##### Unstructured

###### e.g. free text

#### Complexity

##### relationship between elements

#### Volume

##### e.g. DNA sequence

#### Metadata

### Datasets (Collection of Records)

#### Volume

##### collection over time

##### redundancy of data record (de-normalization)

##### lifecycle position

###### e.g. large raw data

###### big due to intermediate analysis

###### big derivative (big at the end) from data enrichment

#### Data Structure

##### format

#### Variety

##### distribution across domains

##### Distribution of types

#### Dataset Metadata

##### Provenance

##### Operational

##### Stage

###### Raw

###### Information

###### Knowledge

###### Actionability

###### Archiving

### Multiple Datasets (Collection of Datasets)

#### Variety

### Domain

#### Retail Transactions

#### Financial

#### Genome

#### etc...

#### linkage to what is going on in these domains

### Data Type

#### Sensor

#### Logs

#### Visual Media

#### Social Media

#### Web Pages

### Quality

#### Cleanliness

#### Missing

#### etc...

#### point to taxonomy of dirty data

#### ISO 8000

### Vocabulary Standards

#### bioontology.org

#### LinkedData.org

## Data Persistence Taxonomy - Level 1: State - Level 2: Attributes - Level 3: Characteristics - Level 4: Sub Characteristics

### (In Motion) In Motion

#### Variability

#### Velocity

##### rate of acquisition of new data records

###### e.g. click stream log

##### rate of update for existing data records

###### e.g. state changes for monitored objects

### (At Rest) At Rest

#### Physically at Rest

##### External

###### e.g. get dynamically from data provider

##### Local filesystem

###### e.g. HDFS/Hadoop

##### Distribution (on nodes)

###### CAP

Consistency

Availability

Partition Tolerance

###### update

append

update

#### Logically At Rest

##### Flat Files

##### markup

###### Semantic

###### XML

##### RDB

##### Table

##### Key-Value

##### Document

##### Graphical

###### Semantic/Ontological