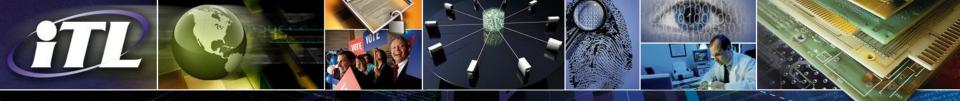


# Standard Big Data Architecture and Infrastructure

Wo Chang Digital Data Advisor Information Technology Laboratory (ITL) National Institute of Standards and Technology (NIST) wchang@nist.gov

May 20, 2016





# Agenda

- Brief Introduction background, projects, interests
- Big Data Architecture and Infrastructure challenges and opportunities
- Samples of Independent Big Data Activities
- Collaboration Focus Areas Join Big Data international standards development
- Other Topics: Make application domains (Big Data, IoT, CPS, Smart Cities, etc.) available



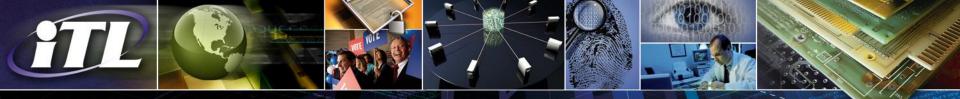


# **Brief Introduction: Wo Chang**

### Mr. Wo Chang, Digital Data Advisor, NIST Information Tech. Lab. (ITL)

- Responsibilities: Open data and big data interoperability. Mr. Chang currently the Convener of the ISO/IEC JTC 1/WG9 Working Group on Big Data, co-chairs the NIST Big Data Public Working Group, and chairs the ISO/IEC JTC/1 SC 29 WG11 (MPEG) Multimedia Preservation AHG.
- Prior to joining ITL Office, Mr. Chang was manager of the Digital Media Group in ITL and his duties included oversees several key projects including digital data archival and preservation, management of electronic health records, motion image quality, cloud computing, and multimedia standards. In the past, Chang was the Deputy Chair for the US INCITS L3.1, chaired several other key projects for MPEG, participated with the HL7 and ISO/IEC TC215 for health informatics, IETF for the protocols development, and was one of the original members of the W3C's SMIL and developed one of the SMIL reference software.





**Brief Introduction: Projects and Interests** 

 NIST Big Data Public Working Group, Co-Chair ISO/IEC JTC 1/WG 9 Working Group on Big Data, Convener

Develop a reference architecture that is vendor-neutral, technology- and infrastructure-agnostic to enable any stakeholders to perform analytics processing for their given data sources without worrying about the underlying computing environment.

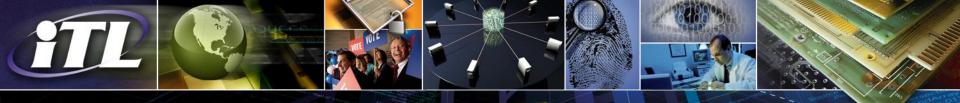
NIST Ubiquitous Data Interoperability, Lead Architect

Develop an interoperable data infrastructure that is scalable to enable automatic data mashups between heterogeneous datasets from various domains without worrying about the data source and structure.

# Research Interests

Scalable graph mining algorithms and visual analytics for massive audiovisual content, digital data mashup, cloud computing, content metadata description, multimedia synchronization, and Internet protocols.



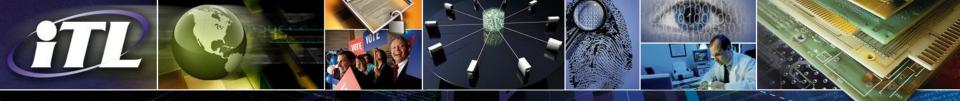


# **Big Data Architecture and Infrastructure – Challenges (Computing Stack)**

Orcl	hestr	atior	1 & W	orkflow Oozie, Ol NA: Pega						tiveBPEL, B			,	
С	oss ( apat	ilitie	es	Machine Learning Mahout , MLlib , MLbase		Data Analytics Librar Statistics, Bioinformatics R. Bioconductor (NA)			atics	Imagery		Linear Algebra alapack, PetSc (NA)		
Aonit		Distri	Mess	CompLearn (NA)			_	•		inages (iv)	~/ 30	агараск,	reise (IVA)	
Monitoring		Distributed Coordination	Message Protocols	High Level (Inte Hive (SQL on Hadoop)	Pig Shark (SQL on Spark, NA)		MRQL Impala (N (SQL on Hadoop, Hama, Spark) Clouder				Swazall (Log Files Google NA)			
Ar	s			Parallel ho Hadoop Spark (Map (Iteration Reduce) MR)		er Tez	z G)	Processin Hama (BSP)	g Storm		Samza .inkedIn		on Hadoop	
Ambari	e			ABDS Inter-process Communication HPC Inter-process Communication										
	c u r			Hadoop, Spark Communications MPI(NA) & Reductions Harp Collectives(NA)										
Ganglia,	t	8	룿	Pub/Sub Messaging Netty(NA)/ZeroMQ(NA)/ActiveMQ/QPid/Kafka										
lia, Nagio	у &	ZooKeeper,	Thrift, Protobuf (NA)	In memory distributed databases/caches: GORA (general object from NoSQL), Memcached (NA), Redis(NA) (key value), Hazelcast (NA), Ehcache (NA);										
os, In	P r i	JGroups	buf (	ORM Object Relational Mapping: Hibernate(NA), OpenJPA and JDBC Standard										
Nagios, Inca (NA)			NA)	Extraction Too UIMA Tik (Entities) (Cont	a MySQL	Phoe (SQL		SciDB (NA) Arrays,	NoSC HBas (Data			ssandra	Solandra (Solr+ Cassandra)	
	·			(Watson)		HBas		R,Python	HDF			(DHT)	+Document	
	a c			NoSQL: Docum	NoSQL: Document NoSQL: Key Value (all NA)									
	y			MongoDB Cour (NA)		ene olr	В	erkeley DB	Azur Table			Riak ynamo	Voldemort ~Dynamo	
				NoSQL: General	NoSQL: TripleStore RI				DF SparkQL			File		
				Neo4J Java Gnu ( (NA)	Yarcdata Commercial (NA)		Jena					DF on nulo	lanagement iRODS(NA)	
				Data Transport BitTorrent, HTTP, FTP, SSH Globus Online (GridFTP)										
				ABDS Cluster Resource Management HPC Cluster Resource Management										
A – No oject:		bach	e	Mesos, Yarn, Helix, Llama(Cloudera) Condor, Moab, Slurm, Torque(NA)										
iu/Jha	/Fox	/		ABDS File Systems User Level HPC File Systems (NA)										
mbu	rugar	nuva	•	HDFS, Swift, Ceph FUSE(NA) Gluster, Lustre, GPFS, GFFS Object Stores POSIX Interface Distributed, Parallel, Federated										
arch 9 Gree			are	Interoperability Layer Whirr / JClouds OCCI CDMI (NA) DevOps/Cloud Deployment Puppet/Chef/Boto/CloudMesh (NA)										
Cloud	(ligh	t) to	нрс	laaS Platform Manager Open Source Commercial Clouds Bare OpenStack, OpenNebula, Eucalyptus, CloudStack, vCloud, Amazon, Azure, Google Metal										
darke	r) int lave		tion	Apache Big Data Stack (ABDS) with HPC Integration/Enhancement										

Apache Big Data Stack (ABDS) with HPC Integration/Enhancement





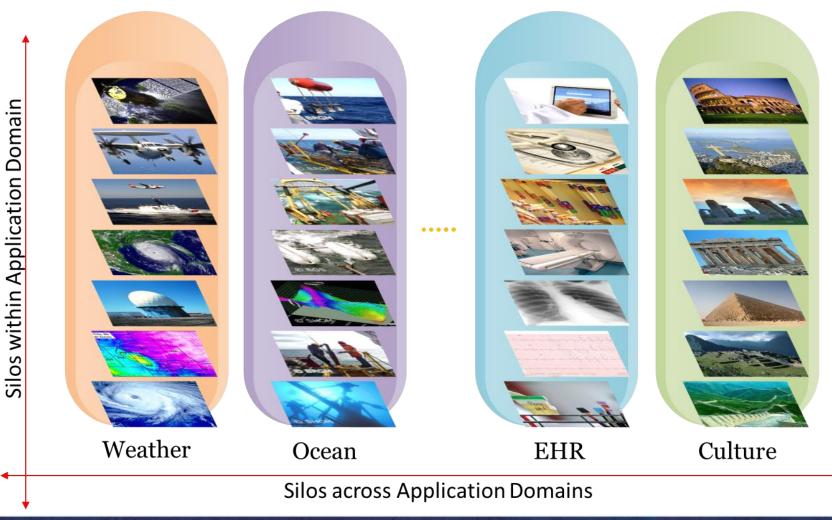
**Big Data Architecture and Infrastructure – Challenges (Analytics Stack)** 







# **Big Data Architecture and Infrastructure – Challenges (Data Stack)**

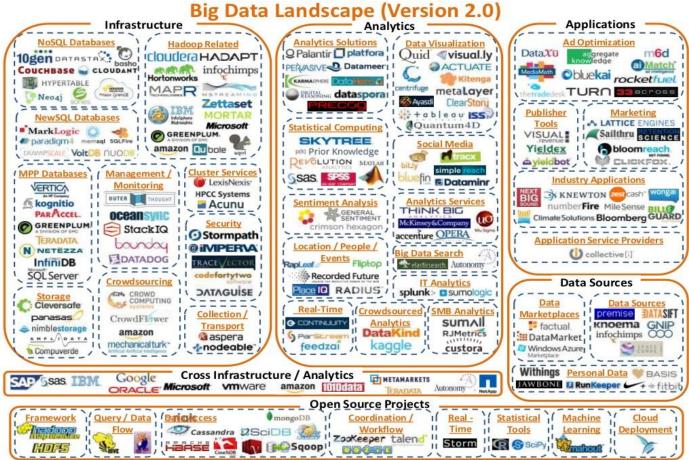




7



# **Big Data Architecture and Infrastructure – Challenges (Integration)**



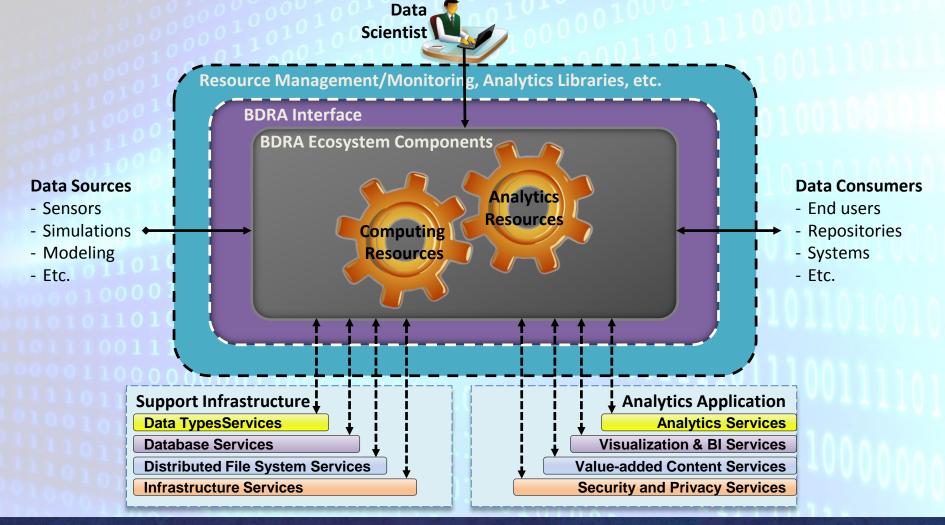
© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

Source: http://www.ongridventures.com/wp-content/uploads/2012/10/Big-Data-Landscape.jpg





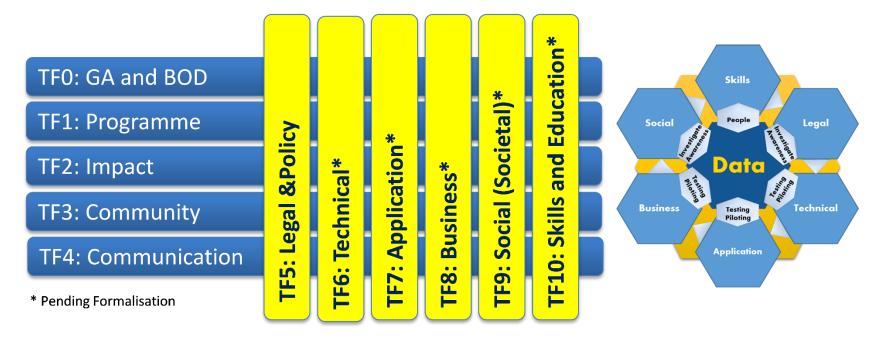








## EU Big Data Value Association (BDVA)



TF6: Technical: Identifying and refining the technical challenges of the programme – eg Data Management (Created Sub-Group 6 to deal with Big Data standardization)





### **US National Science Foundation (NSF) Big Data Hubs**

#### Northeast

- Education
- Cities / Regions
- Discovery Science
- Energy
- Finance
- Health

#### South

- Health Disparities
- Coastal Hazards
- Industrial
- Materials and Manufacturing
- Habitat Planning

#### Midwest

- Food-Water-Energy
- Health Sciences, Life Sciences, Bioinformatics, Genomics
- Smart Cities and Communities
- Digital Agriculture (precision farming, sustainability, ...)
- Advanced Manufacturing
- Network Science
- Transportation
- Business Analytics
- •Ring 1: Tools and Services
- Ring 2: Data Science

#### West

- •Big Data technology
- Managing natural resources and hazards
- Precision medicine
- Metro data science
- Data-enabled scientific discovery and learning





## NIST Big Data Public Working Group (June 2013 – now)

**Goal:** Develop a consensus-based reference architecture that is vendorneutral, technology and infrastructure agnostic to enable any stakeholders to perform analytics processing for their given data sources without worrying about the underlying computing environment.

### Activities – 5 Subgroups

- 1. Definitions & Taxonomies
- 2. Use Cases & Requirements
- 3. Security & Privacy
- 4. Reference Architecture
- 5. Standards Roadmap

V1 (high-level NBD-RA components and descriptions) Big Data Interoperability Framework, Released September 16, 2015

### http://bigdatawg.nist.gov







### ISO/IEC JTC 1/WG 9 Working Group on Big Data



### Summary Key Findings:

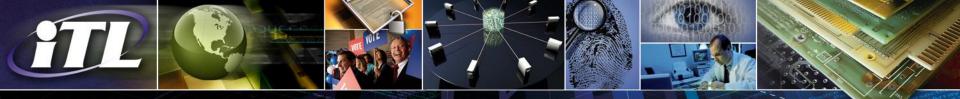
- Big Data systems are difficult to construct tools and applications that integrate data from multiple Big Data sources. The systems should be designed with security in mind from the ground up rather than have it emerge as an afterthought.
- Identified 16 potential standardization gaps to enable Big Data Systems interoperability.

### **Organizational Drivers to Provide:**

- 1. Insight: enable discovery of deeper, fresher insights from all enterprise data resources
- 2. Productivity: improve efficiency, effectiveness, and decision-making
- 3. Speed: facilitate more timely, agile response to business opportunities, threats, and challenges
- 4. Breadth: provide a single view of diverse data resources throughout the business chain
- 5. Control: support tighter security, protection, and governance of data throughout its lifecycle
- 6. Scalability: improve the scale, efficiency, performance, and costeffectiveness of data/analytics platforms

\* JTC 1 Big Data Report: http://www.iso.org/iso/big\_data\_report-jtc1.pdf





# **Collaboration Focus Areas**

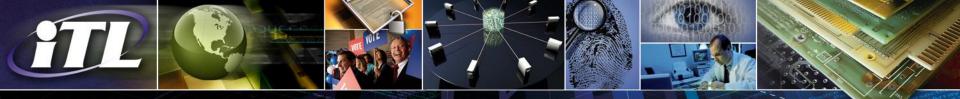
### Join Big Data international standards development

- 140+ from 22 NBs: Australia, Austria, Brazil, Canada, China, Finland, France, Germany, India, Ireland, Italy, Japan, Korea, Luxembourg, Netherlands, Norway, Russian Federation, Spain, Singapore, Sweden, UK, US
- Current Projects
  - ISO/IEC 20546 Information technology Big data Overview and vocabulary
  - ISO/IEC 20547 Information Technology Big data Reference architecture (5 Parts)

Part 1: (TR) Framework and Application Process Part 2: (TR) Use Cases and Derived Requirements Part 3: (IS) Reference Architecture Part 4: (IS) Security and Privacy Fabric (under SC 27/WG 4) Part 5: (TR) Standards Roadmap

 ISO/IEC Liaisons: SC 6/WG 7, SC 27, SC 29, SC 32, SC 36, SC 38, SC 39, ISO/TC 69, ISO/TC 204, ITU-T SG13





# **Other Topics**

### Make application domains (Big Data, IoT, CPS, Smart Cities, etc.) available

- Share public accessible use cases
- Share public accessible non-PII datasets
- Share public accessible *analytics tools*

