

NIST Standard Enterprise Big Data Ecosystem

Wo Chang

Digital Data Advisor ISO/IEC JTC1/WG 9 Big Data, Convenor wchang@nist.gov

June 19, 2017



P(A|B) = P(B|A) P(A) P(B)

010011000010 01000111000 001011101010000111101010 1101000010 10 11111000001





Agenda

- Revisit Enterprise Computing
- What's the Computing Infrastructure Trend?
- What are the Big Data Architecture/Infrastructure Challenges?
- NIST Big Data Public Working Group Overview
- NIST Standard Enterprise Big Data Ecosystem
- NIST Big Data Interoperability Framework Version 1.0
- ISO/IEC JTC 1/WG 9 Big Data Standards Activities
- NIST Big Data Interoperability Framework Version 2.0
- NIST Big Data Reference Architecture (NBD-RA)
- Standard Big Data Analytics and Beyond





Revisit Enterprise Computing

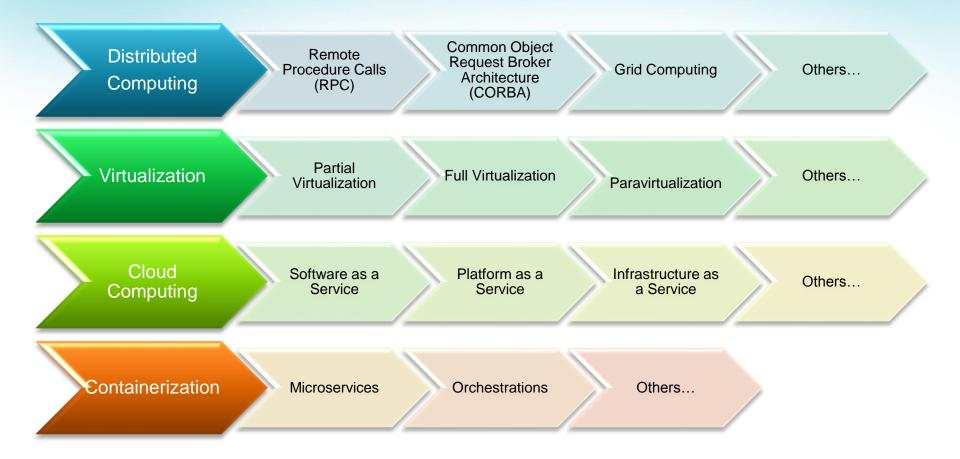
Enterprise computing is sometimes sold to business users as an entire platform that can be applied broadly across an organization and then further customized by users within each area. This means the *analytics, reporting, database management and other applications are standard across the system*, while the application packages being used and the data being accessed in each area will be different. In this sense, enterprise computing is a departure from finding single software solutions to specific business problems, such as inventory or accounting software. Instead, *enterprise computing is intended to offer integrated solutions to these problems*.

Source: https://www.techopedia.com/definition/27854/enterprise-computing





What's the Computing Infrastructure Trend?





Big Data Architecture/Infrastructure – Challenges (Computing Stack)

Orchestration & Workflow Oozie, ODE, Airavata and OODT (Tools) NA: Pegasus, Kepler, Swift, Taverna, Trident, ActiveBPEL, BioKepler, Galaxy																	
Cross Cutting Capabilities				Machine Learning Mahout , MLlib , MLbase CompLearn (NA)		Data Analytics Libraries: Statistics, Bioinformatics Imager R, Bioconductor (NA) Imagel (N				-							
lonite		İstrib	Message	High Level (Integrated) Systems for Data Processing													
Monitoring		Distributed Coordination	age Protocols	Hive (SQL on Hadoop)	Hcatal	og ces (Pro	Pig ocedural nguage)		Shark (SQL on park, NA)	(SQL	MRQL on Hadoo ma, Spark)	р,	mpala (i Cloude		Swazall (Log Files Google NA)		
A	s	ation		Hadoop	Iterative	NA:Twiste Stratosphe Iterative M	re (DA	ez AG)	Processir Hama (BSP)	Storr	Yahoo	Link		Giraph "Pregel	Pegasus on Hadoop (NA)		
Ambari	e			ABDS	Inter-prod	ess Com	- Batc					Stream — Graph er-process Communication					
	c u			Hadoop, S	-			cion			MPI(NA)	cess c	sconnunication				
	r			& Reductio				Colle	ctives(NA	0							
Gan	<u> </u>		=	Pub/Sub N	lessaging		Netty	(NA),	/ZeroMQ	(NA)/A	ctiveMQ/	QPid/I	Cafka				
Ganglia, Nagios, Inca (NA)	t V Keeping V Ke							NoSQL)), Memo	ached							
, si	& P	, JGroups	L L L	ORM Object Relational Mapping: Hibernate(NA), OpenJPA and JDBC Standard													
			A	Extractio		SQL			SciDB		QL: Colur				Solandra (Solr+		
VA)	r i v	S		UIMA (Entities) (Watson)	Tika (Content)	MySQL (NA)	Phoe (SQI HBa	on	(NA) Arrays, R,Python	HBa (Data HDI	on (D	umulo ata on DFS)	Cassa (DH	(T)	(Solr+ Cassandra) +Document		
	a NoSQI				NoSQL: Document NoSQL: Key Value (all NA)												
	Y			MongoDB (NA)	CouchD		ene olr	В	erkeley DB	Azur Tabl		namo azon	Rial ~Dyna		oldemort Dynamo		
				NoSQL: Ge	eneral Gr	aph	NoS	QL: T	ripleSto	re RD	DF Spar	kQL			File		
						Neo4J Java Gnu (NA)	Com	cdata mercial (NA)	Jen		Sesam (NA)	e All	egroGrap mmercia	h RY	A RDF o	on is	nagement
Data Transport BitTorrent, HTTP, FTP, SSH Globus Online (Gr							(GridFT	Р)									
				ABDS	ABDS Cluster Resource Management HPC Cluster Resource Management												
– No ojects	– Non Apache jects			Mesos, Yarn, Helix, Llama(Cloudera) Condor, Moab, Slurm, Torque(NA)													
./Iba	/lha/Eav/			ABDS File Systems User Level HPC File Systems (NA)													
mbur	/Jha/Fox/ nburugamuva rch 9 2014			HDFS, Swift, Ceph FUSE(NA) Gluster, Lustre, GPFS, GFFS Object Stores POSIX Interface Distributed, Parallel, Federated													
	Green layers are			Interoperability Layer Whirr / JClouds OCCI CDMI (NA) DevOps/Cloud Deployment Puppet/Chef/Boto/CloudMesh (NA)													
loud	ache/Commercial oud (light) to HPC arker) integration				laaS Platform Manager Open Source Commercial Clouds OpenStack, OpenNebula, Eucalyptus, CloudStack, vCloud, Amazon, Azure, Google							Bare Metal					
	lavers			Apac	Apache Big Data Stack (ABDS) with HPC Integration/Enhancement												

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





Big Data Architecture/Infrastructure – Challenges (Analytics Stack)

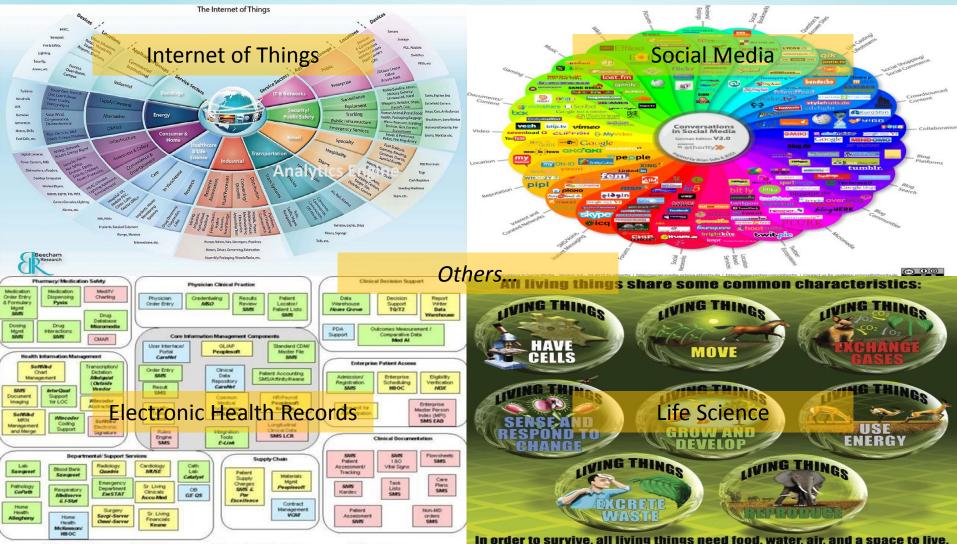


Source: http://1.bp.blogspot.com/-PKiTQa0mrn4/T_mGb6Al3yl/AAAAAAAAAAAAQ/TtH7xyjQ3FA/s640/analytics+tools+landscape.bmp





Big Data Architecture/Infrastructure – Challenges (Applications Stack)



Birding Moderate | Birding Durchased - Not Inclemented Documents



Big Data Architecture/Infrastructure – Challenges (Integration)

BIG DATA LANDSCAPE 2017

	DIG DATA LANDSCAT L 2017	
INFRASTRUCTURE	ANALYTICS	APPLICATIONS - ENTERPRISE
HADOOP ON PREMISE Cloudera Hotonwork MCRR Pivotal IBM InfoSphere' bluedata jethro Columbia Scheme Cloud Platform Distriction Columbia Columbia Scheme Cloud Platform Distriction Columbia Scheme Cloud Platform Distriction Dist	DATA ANALYST PLATFORMS DATA SCIENCE PLATFORMS DATA S	SALES MARKETING-B2B MARKETING-B2C CUSTOMER SERVICE CUSTOMER SERVICE MARKETING-B2B MARKETING-B2B MARKETING-B2B Zet& Discommenden Discommenden Customer Service Zet& Discommenden Discommenden Customer Service Zet& Discommenden Discommenden Customer Service ACTIONO SALESC MARKETING-B2B Zet& Discommenden Discommenden Customer Service ACTIONO SALESC MARKETING-B2B Zet& Discommenden Discommenden Customer Service ACTIONO SALESC MARKETING-B2B Zet& Discommenden Customer Service ACTIONO SALESC MARKETING-B2B Zet& Discommenden MEDALLIA ACTIONO SALESC MARKETING-B2B MARKETING-B2C Zet& Discommenden MEDALLIA ACTIONO SALESC MEDALLIA MEDALLIA MEDALLIA Salesc MEDALLIA Customer Service MEDALLIA MEDALLIA MEDALLIA Salesc MEDALLIA MEDALLIA Salesc MEDALLIA MEDALLIA Salesc MEDALLIA MEDALLIA Salesc MEDALLIA MEDALI
NOSOL DATABASES Coogle Cloud Planform Creating Coogle Cloud Planform Creating Cloud Planform	BIPLATFORMS WISUALIZATION WISUALIZ	HUMAN CAPITAL HUMAN
DATA TRANSFORMATION talend @ Pentobe alteryx @ Tersacta @ Streamses @ Streamses	MACHINE LEARNING Machine LEARNING Machine Learning Machine Structure Machine Structur	APPLICATIONS - INDUSTRY ADVERTISING EDUCATION CITIEGO ITIEGO CITIEGO CITIEGO CITIEGO CITIEGO CITIEGO
STORAGE CLUSTER SERVICES APP DEV CROWDSOURCING HARDWARE Ø condict anistry Maccoder Anistry © condict Anistry © condict Support March Support CROWDSOURCING HARDWARE Ø condict Anistry © condict Anistry © condict Anistry © condict Condict Anistry March Support Condict Anistry Google TPU ARM (Condict Anistry) March Support Ø condict Anistry © condict Anistry @ condict Anistry @ condict Anistry March Support March Support Ø condict Anistry @ condict Anistry @ condict Anistry @ condict Anistry March Support Ø condict Anistry @ condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ condict Anistry Ø condict Anistry @ condict Anistry @ conistry	SOCIALANALYTICS SOCIAL	TATRADO Dela XIO sumguinguinguinguinguinguinguinguinguinguin
CROSS-INFRASTRUCTURE/		
Score Store	STREAMING Lalend Control Sport Spo	Milling Control Solution Solution Mathematical M
	DATA SOURCES & APIS	DATA RESOURCES
HEALTH JAWBONE'	AIR / SPACE / SEA	CATION INTELLIGENCE COTHER Segment Place[] Segment States Segment States
Last updated 4/5/2017 © Matt Turck (@r	nattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)	mattturck.com/bigdata2017 FIRSTMARK

Source: http://mattturck.com/wp-content/uploads/2017/04/Big-Data-Landscape-2017-Matt-Turck-FirstMark.png





NIST Big Data Public Working Group Overview

Goal

Work with industry, academia and government to create a consensus-based extensible NIST Big Data Interoperability Framework (NBDIF) which is a vendorneutral, technology- and infrastructure-independent ecosystem. It can enable Big Data stakeholders (e.g. data scientists, researchers, etc.) to utilize the best available analytics tools to process and derive knowledge through the use of standard interfaces between swappable architectural components.

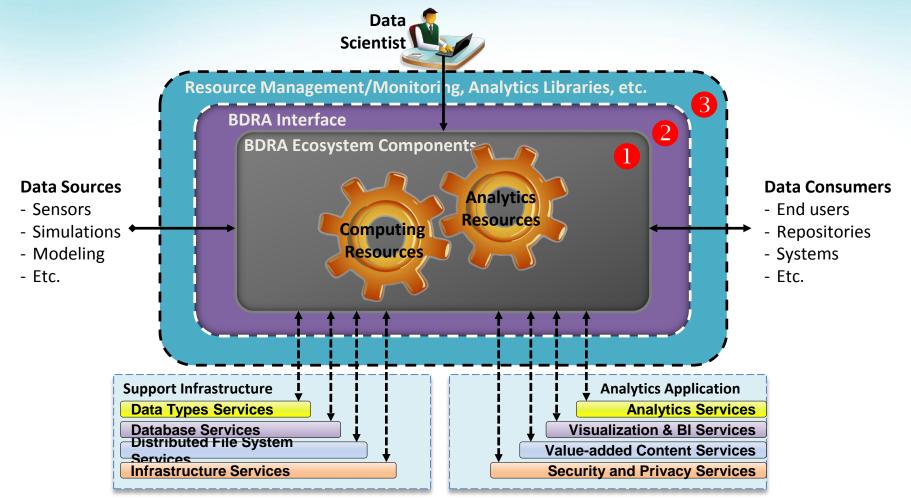
Approach – Three Stages (refers to three versions)

- Identify the high-level NIST Big Data Reference Architecture (NBDRA) key components, which are technology-, infrastructure-, and vendor-agnostic [Done]
- 2. Define general interfaces between the NBDRA key components [Ongoing]
- 3. Validate the NBDRA by building Big Data general applications through the general interfaces.





NIST Standard Enterprise Big Data Ecosystem







NIST Big Data Interoperability Framework Version 1 Deliverable: Stage-1 – High-level Reference Architecture https://bigdatawg.nist.gov/V1_output_docs.php (Sep. 16, 2015)

NIST SP1500-1: Definitions	NIST SP1500-2: Taxonomies	NIST SP1500-3: Use Cases & Requirements			
NIST SP1500-4: Security & Privacy	NIST SP1500-5: Architecture Survey – White Paper	NIST SP1500-6: Reference Architecture			
	NIST SP1500-7: Standards Roadmap				





ISO/IEC JTC 1/WG 9 Big Data Standards Activities

ISO/IEC JTC 1/WG 9 Working Group on Big Data (Jan. 2015 – now)

- 180+ from 26 NBs: Australia, Austria, Brazil, Canada, China, Finland, France, Germany, India, Ireland, Israel, Japan, Korea, Luxembourg, Mexico, Netherlands, Norway, Russian Federation, Saudi Arabia, Singapore, Slovenia, South Africa, Spain, Sweden, UK, US
- Current Projects
 - ISO/IEC 20546 Information technology Big data Definition and vocabulary (Committee Draft #3 as June 2017)
 - ISO/IEC 20547 Information Technology Big data Reference architecture (5 Parts as June 2017)

Part 1: (TR) Framework and Application Process (2nd WD) Part 2: (TR) Use Cases and Derived Requirements (Recommended for Publication) Part 3: (IS) Reference Architecture (5th WD) Part 4: (IS) Security and Privacy Fabric (2nd ED, under SC 27/WG 4) Part 5: (TR) Standards Roadmap (Recommendation for Publication)

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, IIC, OGC, BDVA (coming up)





ISO/IEC JTC 1/WG 9 Big Data Standards Activities

1st International Workshop on Big Data Standards, March 7, 2016, Dublin, Ireland 16 Speakers, One Panel Discussion (4 panelists), 80+ Registered, 50+ attended



Leo Clancy

IDA Head of Technology Welcome, Speakers, Delegates and Visitors



Ian Cowan NSAI Ireland *ICT and Big Data Standardization in Ireland*

Jean Stride

BSI UK

Wo Chang ISO IEC JTC1 WG9 Convenor *WG9 Big Data Standard*





Maurice Buckley

CEO NSAI & Vice-President Technical CENELEC Supporting the Digital Single Market through Standardization: A European and Irish perspective on Industry 4.0 and Big Data

Colette Maloney HoU Smart Cities and Sustainability DG CONNECT EU Smart Cities and Big Data



Dave Lewis Spokes Director - ADAPT Ireland *Standards for the Multilingual Web*



Big Data: A Market Strategy







ISO/IEC JTC 1/WG 9 Big Data Standards Activities

1st International Workshop on Big Data Standards, March 7, 2016, Dublin, Ireland 16 Speakers, One Panel Discussion (4 panelists), 80+ Registered, 50+ attended



Phil Archer W3C UK *Big Data Europe, an infrastructure for all*

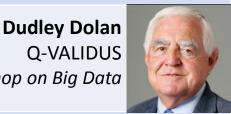


Rob Brennan EU ALIGNED ALIGNED - Bringing together Software and Data Engineering for Data-intensive Systems

Wael William Diab FutureWei Big Data Ecosystem



Q-VALIDUS Towards a CEN Workshop on Big Data





Moez Draief Huawei France *Parsimonious network monitoring*



Ismael Caballero UCLM Spain Data Quality for Big Data

Souleiman Hasan Insight Ireland Big Data Technical Priorities



John Strassner Futurewei Engineering Value from Big Data







NIST Big Data Interoperability Framework Version 2 Deliverable: Stage-2 – Reference Architecture Interface https://bigdatawg.nist.gov/V2_output_docs.php (Nov. 2017)

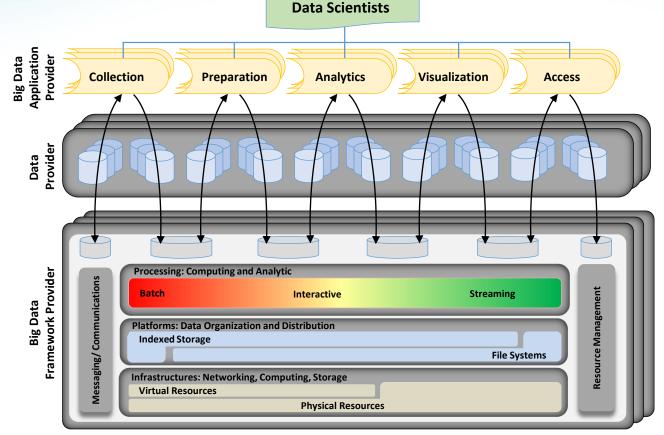
NIST SP1500-1: Definitions	NIST SP1500-2: Taxonomies	NIST SP1500-3: Use Cases & Requirements			
NIST SP1500-4: Security & Privacy	NIST SP1500-5: Architecture Survey – White Paper	NIST SP1500-6: Reference Architecture			
NIST SP1500-7: Standards Roadmap	NIST SP1500-8: Reference Architecture Interface	NIST SP1500-9: Adoption & Modernization			





NIST Big Data Reference Architecture (NBD-RA)

Enable data scientists, engineers, researchers, etc. to increase productive and enhance quality in data science through standard modularized Big Data Analytics tools.







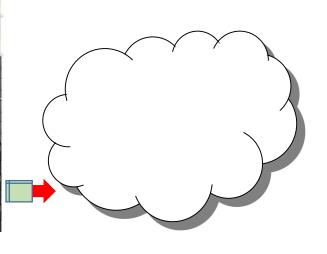
Standard Big Data Analytics and Beyond

Enable Big Data analytics tools for *interoperability, portability, reusability, and extensibility.* Practical Aspect: Analytics tools can be *reusable, deployable, and operational* (max. use of resources) [similar approach as to HTML, PCI bus, etc.]

Standard Big Data Reference Architecture Interface Specification – *interoperable across various platforms*









Desktop/Server

Data Center Many CPUs/Cores/GPUs/ FPGAs/ASICs/accelerators

.....

Cloud