

Standard Enterprise Big Data Ecosystem

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

March 22, 2017





Agenda

- Why Enterprise Computing is Important?
- What's the Computing Infrastructure Trend?
- What are the Challenges in Big Data Architecture and Infrastructure?
- What's the Standard Big Data Enterprise Ecosystem?
- NIST Big Data Public Working Group Foundation Documents to JTC 1/WG 9
- ISO/IEC JTC 1 Study Group on Big Data Report (year 2014)
- ISO/IEC JTC 1/WG 9 Working Group on Big Data Activities (2015 now)
- How Can You Help?





Why Enterprise Computing is Important?

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 and Infrastructure – Challenges (Computing Stack)

Cross Cutting Capabilities				Machine Learning Mahout , MLlib , MLbase CompLearn (NA)		se	Data Analytics Libraries: Statistics, Bloinformatics Imagery Linear Algebra R, Bioconductor (NA) ImageJ (NA) Scalapack, PetSc (NA)							
nito		stributed Coordi	essage Protocols	High Level (Integrated) Systems for Data Processing										
ring				Hive (SQL on Hadoop) Heatalog (PL on Hadoop)		(Pro Lanj	Pig cedural guage)	Shark (SQL on Spark, NA)	(SQL o Ham	MRQL on Hadoop, ia, Spark)	Impal	a (NA) Idera	Swazall (Log Files Google NA	
		nati		Parallel horizontally scalable Data Processing										
T	•	9		Hadoop Sp (Map (Ite Reduce) N	native Strat MR) Itera	Twister ospher tive Mi	e (DAG)	Hama (BSP)	Storm	S4 Yahoo	Samza LinkedIn	Giraph ~Prege	n Pegasu on Hadoo (NA)	
mb	e		Thơ	Batch Stream Graph										
an.	c			ABDS Inter-process Communication HPC Inter-process Communication										
	u			Hadoop, Spark Communications MPI(NA)										
6	i			& Reductions Harp Collectives(NA) Pub/Sub Messaging Netty/NA)/ZeroMO(NA)/ActiveMO/OPid/Kafka										
ang	t	200		i any sum tite	r ant and successful international and and an antique and antique and an antique and antique and an									
ia, N	Y	Kee	ft, Pr	In memory o	In memory distributed databases/caches: GORA (general object from NoSQL), Memcached									
Bio	8	per, JG	oto	(NA), Redis(N	A) (key val	ue), Ha	azelcast (NA), Encact	1e (NA);	21.12				
5	123		uf (ORM Object	ORM Object Relational Mapping: Hibernate(NA), OpenJPA and JDBC Standard									
Ca ()	Priv	roup	NA)	Extraction Tools SQL		QL		SciDB	NoSQ	HoSQL: Column HBase Accumulo Cassandra Data on (Data on (DHT) HDFS) HDFS)			Solandr (Solr+	
VA)		5		(Entities) (C (Watson)	Tika MySQI (Content) (NA)		(SQL on HBase)	Arrays, R,Python	(Data d HDFS			ssandra (DHT)	Cassandra +Documen	
	а			NoSQL: Document NoSQL: Key Value (all NA)										
	e Y			MongoDB (NA)	CouchDB	Luce So	ine Ir	Berkeley DB	Azure	Dynar Amazo	no l on ~Dy	Riak /namo	Voldemort ~Dynamo	
				NoSQL: General Graph NoSQL: TripleStore RDF SparkQL File										
H				Neo4J	Yarcda	ta		Sesame	Alle	eroGraph	RYA RD	Eon	anagemer	
				Java Gnu (NA)	Commercial (NA)		Jena	(NA)	Commercial		Accumulo		IRODS(NA)	
				Data Transport BitTorrent, HTTP, FTP, SSH Globus Online (GridFTP)										
				ABDS Cluster Resource Management HPC Cluster Resource Management										
Non Apache				Mesos, Yarn, Helix, Llama(Cloudera)					Condor, Moab, Slurm, Torque(NA)					
ects				ABDS File Syst	tems	Use	r Level	_	HPC	File System	5 (NA)			
Jha, bur	/Fox	/ nuva		HDFS, Swift Obje	t, Ceph act Stores	F	USE(NA) DSIX Inter	face	Gh	uster, Lustr Distribute	e, GPFS, d, Parall	GFFS el, Federa	ated	
ree	lay	ers a	re	Interoper DevOps/0	ability Lay	er oymei	Whirr /	JClouds Puppe	et/Chef,	OCCI CDN /Boto/Clou	ll (NA) dMesh (N	JA)		
che	/Cor	nme	rcial	laaS Platforn	Manager	Op	en Source		Comm	ercial Cloue	ds		Bare	





V2 Big Data Architecture and Infrastructure – Challenges (Applications Stack)





Big Data Architecture and Infrastructure – Challenges (Integration) Big Data Landscape 2016



Source: http://www.199it.com/wp-content/uploads/2016/02/matt_turck_big_data_landscape_full.pngheading





NIST



Focus: 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 with 5 Subgroups:

- 1. Definitions & Taxonomies
- 2. Use Cases & Requirements

- 4. Reference Architecture
- 5. Standards Roadmap

3. Security & Privacy

Development:

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





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

http://bigdatawg.nist.gov







Vendors Big Data architectures





V2 focuses on interface between NBD-RA components through use cases by

	System Orchestrator			
Provider	Big Data Application Provider Preparation		ousumer	
	Big Data Framework Provider Processing: Computing and Analytic Batch Interactive Streaming Platforms: Data Organization and Distribution		& Privacy ment	
	Indexed Storage File Systems Infrastructures: Networking, Computing, Storage Virtual Resources Physical Resources	Resource Mar	Security Manager	

NIS

- Analyze activities diagrams
- Analyze functional diagrams
- Apply DevOps/Containers on small scale implementations

Goals:

- Aggregate low-level interactions into high-level general interfaces
- Produce set of white papers to demo how NBD-RA can be used



V2 NIST Big Data Reference Architecture Interface Interaction and workflow





V2 NIST Big Data Reference Architecture Interface Goals – Portability and Scalability (With support of High Performance Computing + Big Data Stacks)



Laptop



Desktop/Server

Data Center Many CPUs/Cores/GPUs

Cloud





V2 NIST Big Data Development Strategies

Selection of use cases: (a) available of datasets and (b) available of analytics codes



Fingerprints Matching



Twitter Feeds

- Data warehousing
- Global Cities







Human and Face Detection from Video



Spatial Big Data/GIS

- Earth Science
- Life Science



Healthcare Payment Fraud

- loT
- Others...





ISO/IEC Big Data Standardization



16



ISO/IEC JTC 1 Study Group on Big Data Report (Year 2014)



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: <u>http://www.iso.org/iso/big_data_report-jtc1.pdf</u>



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

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 Overview and vocabulary (Committee Draft #2 as Dec. 2016)
 - ISO/IEC 20547 Information Technology Big data Reference architecture (5 Parts)

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

 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





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

ISO/IEC JTC 1/SC 29/WG 11 (MPEG) on Big Media

Create AHG between SC 29/WG11 and WG9 to



VIDEO Identify MPEG tools relevant for Big Media





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

ISO/TC69 – Applications of Statistical Methods

- Apply standard statistical methodologies (CRISP, SEMMA, etc.)
- Create AHG between TC69, WG9, and NIST Big Data PWG to:

Explore new Big Data statistical methods

Identify use cases (healthcare fraud, live twitter feeds, etc.)

Implement use cases using best practice Big Data computing ecosystem

Document findings

Standardize new Big Data statistical methodologies





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

ISO/TC204 – Intelligent Transportation

- Apply standard statistical methodologies (CRISP, SEMMA, etc.)
- Create AHG between TC204 and WG9 to



Review SDOs in the Big Data area particularly architecture models, semantic definitions, metadata issues and APIs



Identify "Big Data topics" needed for transport data exchange and external data sources; gather and / or generate use cases related to big data topics for ITS



Examine TC204 work that support the Big Data areas and identify the gaps to fit into the foundation / architecture currently under development by SDOs (e.g., ISO/IEC, IEEE, SAE)





Recommend future work items (if any) to be developed by TC204



Examine security, privacy, ownership, and usage issues related to Big Data ITS applications



Recommend liaisons with SDOs for which collaboration is needed





How Can You Help?

- Join and contribute to NIST and/or ISO/IEC JTC 1/WG 9.
- Review and comment NIST and/or ISO/IEC JTC 1/WG 9 documents
- Provide Big Data use cases and requirements
- Make available datasets and analytic tools to public so NIST and WG 9 can use
- Others...

