INFORMATION TECHNOLOGY LABORATORY

NIST Big Data Reference Architecture for Analytics and Beyond

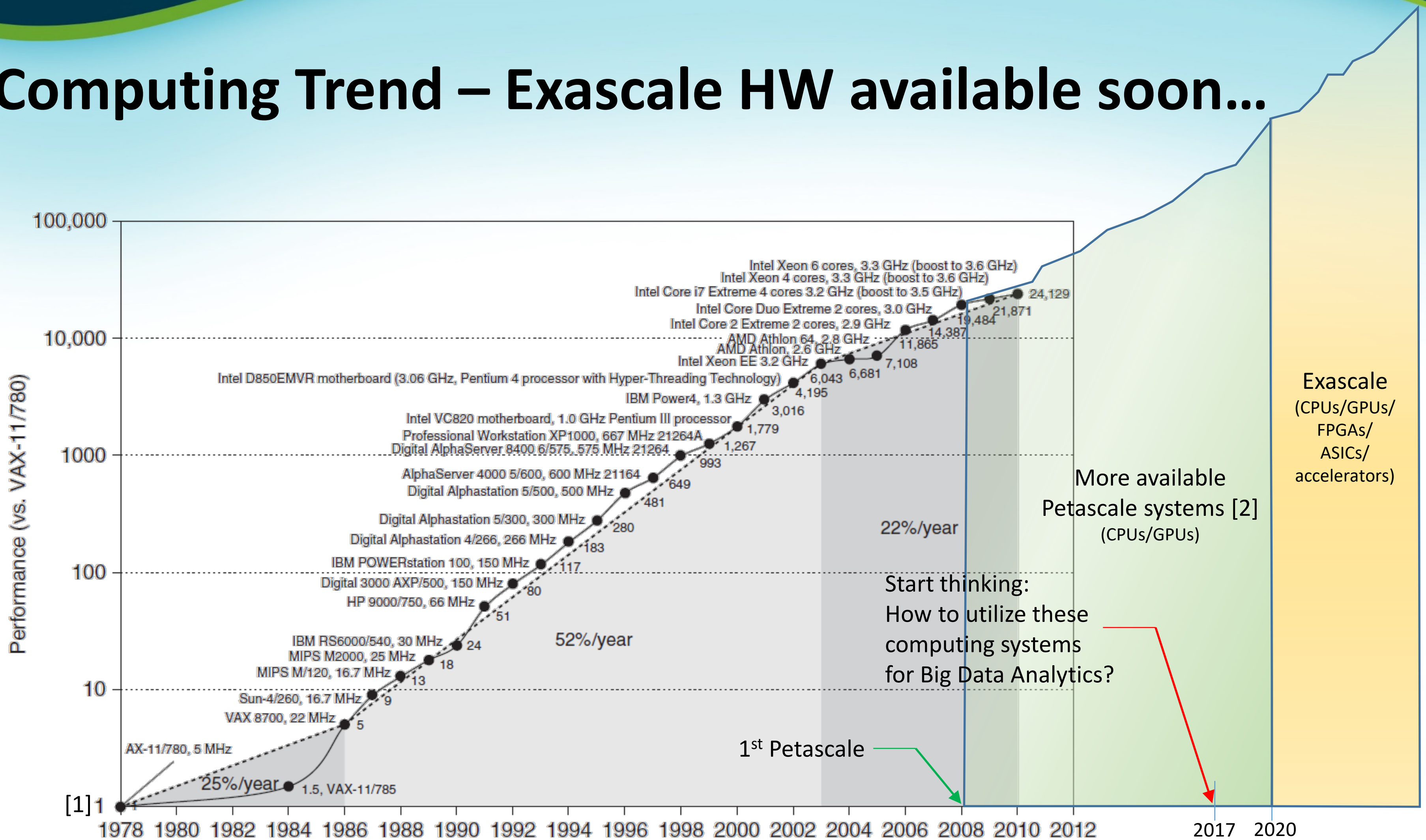
Wo Chang
Digital Data Advisor
wchang@nist.gov

December 8, 2017

Agenda

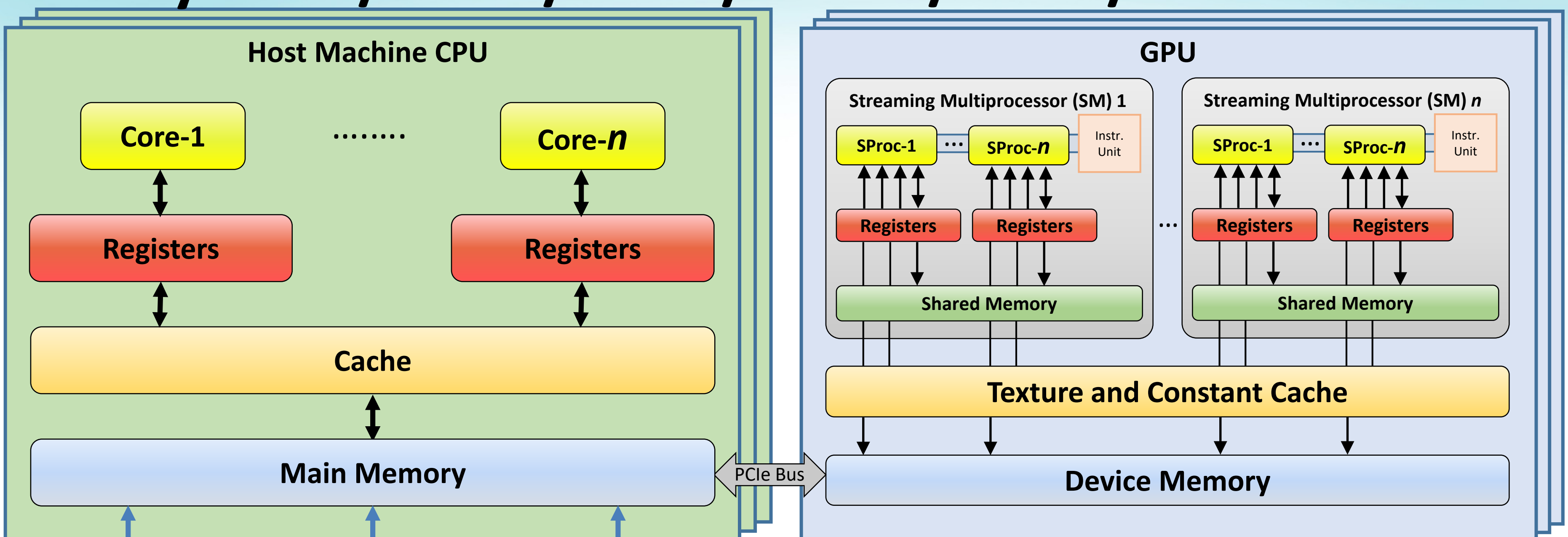
- Computing Trend – Exascale HW available soon…
- Computing Trend – Current HPC and Big Data Stacks
- Exascale Big Data Analytics Opportunities and Challenges
- NIST Big Data Public Working Group (NBD-PWG)
  - Goals and Deliverables
  - Big Data Architecture Challenges: Computing Stack
  - NIST Big Data Reference Architecture
  - NIST Big Data Standards Roadmap
- Goals for Big Data Analytics and Beyond
- Enable Convergence of Data + Compute

Computing Trend – Exascale HW available soon…

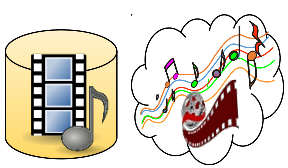


[1] Computer Architecture A quantitative Approach (5th edition) by John L. Hennessy and David A. Patterson
[2] Top500 Supercomputing Listing: https://www.top500.org/lists/2016/11/
FPGAs=Field Programmable Gate Arrays
ASICs=Application Specific Integrated Circuits

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



Data Sources | Resource Mgmt | Feature Tools



Audiovisual Content

High-/Low-level Job Scheduler

Feature extraction Libraries

Goal: maximize giving computing resources with proper load balancing between tasks (CPUs vs GPUs, etc.) via intelligence scheduler.

Targets: Legacy HPC apps + BD analytics

Applications: simulation, modeling, streaming, social media, IoT, etc.

Computing Trend – Current HPC and Big Data Stacks

High-Performance Computing (HPC)		Big Data
Fortran /Parallel C/MP/MPI Applications *High Performance*	Programming Model	Java, Python, etc. Applications *Scalable and Simple to Use*
SLURM/Torque *Resource Management and Scheduler*	Resource Management	YARN *Resource Management and Scheduler* *(More resilient of hardware failures)*
Lustre Parallel Distributed File Systems *Remote Storage*	File System	HDFS, SPARK *Local Storage*
Compute & Memory Focused *High Performance Components* *(Sever, Storage, Switch Fabric, Infrastructure)*	Hardware	Scalable Computing and Storage Focused *Standard Server Components* *(Sever, Storage, Switch Fabric, Infrastructure)*

Exascale Big Data Analytics Challenges and Opportunities

Differences	HPC	Big Data
Interconnect Hardware	RDMA (Remote Direct Memory Access) via Infiniband and OmniPath	Conventional hardware for horizontal scaling
Programming Language	C, C++, etc. required recompile with different OS	JVM, Python, etc. for portability between OS
Computing	Large computation loads	Large and complex datasets (order of terabytes/exabytes)
Filesystems	Mostly NSF	Distributed file systems between cluster nodes (e.g., HDFS)
Storage	Not much, mainly on computing	Very high demand
Fault Tolerance	Needs to enforce to handle system failures and soft errors	Built-in
Execution Control Flow	MPI directly execute on target machines; much better control	Spark uses descriptive API managed by Spark driver and submit job to cluster nodes for execution
Scalability	Mostly vertical	Mostly horizontal
Others…	…	…

Common Goals	HPC	Big Data
Optimize code for performance, energy, and reliability	YES	YES
Reduce data in motion with dynamic tasks scheduler	YES	YES
Others…	…	…

Questions:

1. How best to combine the two stacks (HPC inside Big Data, Big Data inside HPC, or hybrid)?
2. What best standards interface to support them?
3. *Others…*

NIST Big Data Public Working Group (NBD-PWG)

Goal:

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

5 Subgroups (July 2013 – now):

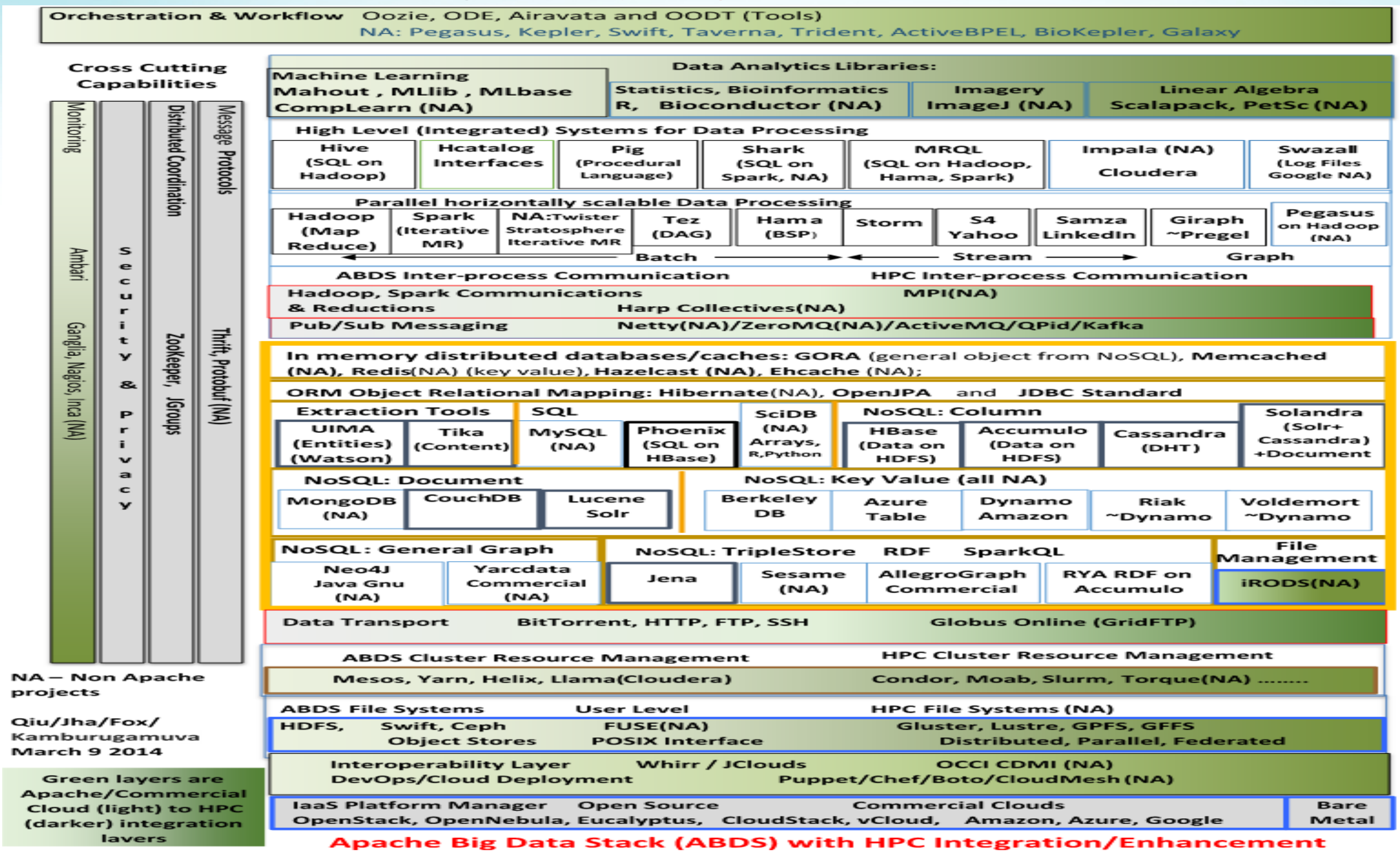
1. Definitions & Taxonomies
2. UC & Requirements
3. Security & Privacy
4. Reference Architecture
5. Standards Roadmap

Deliverables:

1. Big Data Definitions
2. Big Data Taxonomies
3. Big Data Requirements & Use Cases
4. Big Data Security & Privacy
5. Big Data Architectures White Paper Survey
6. Big Data Reference Architecture
7. Big Data Standards Roadmap
8. Big Data Reference Architecture Interfaces (new)
9. Big Data Adoption and Modernization (new)

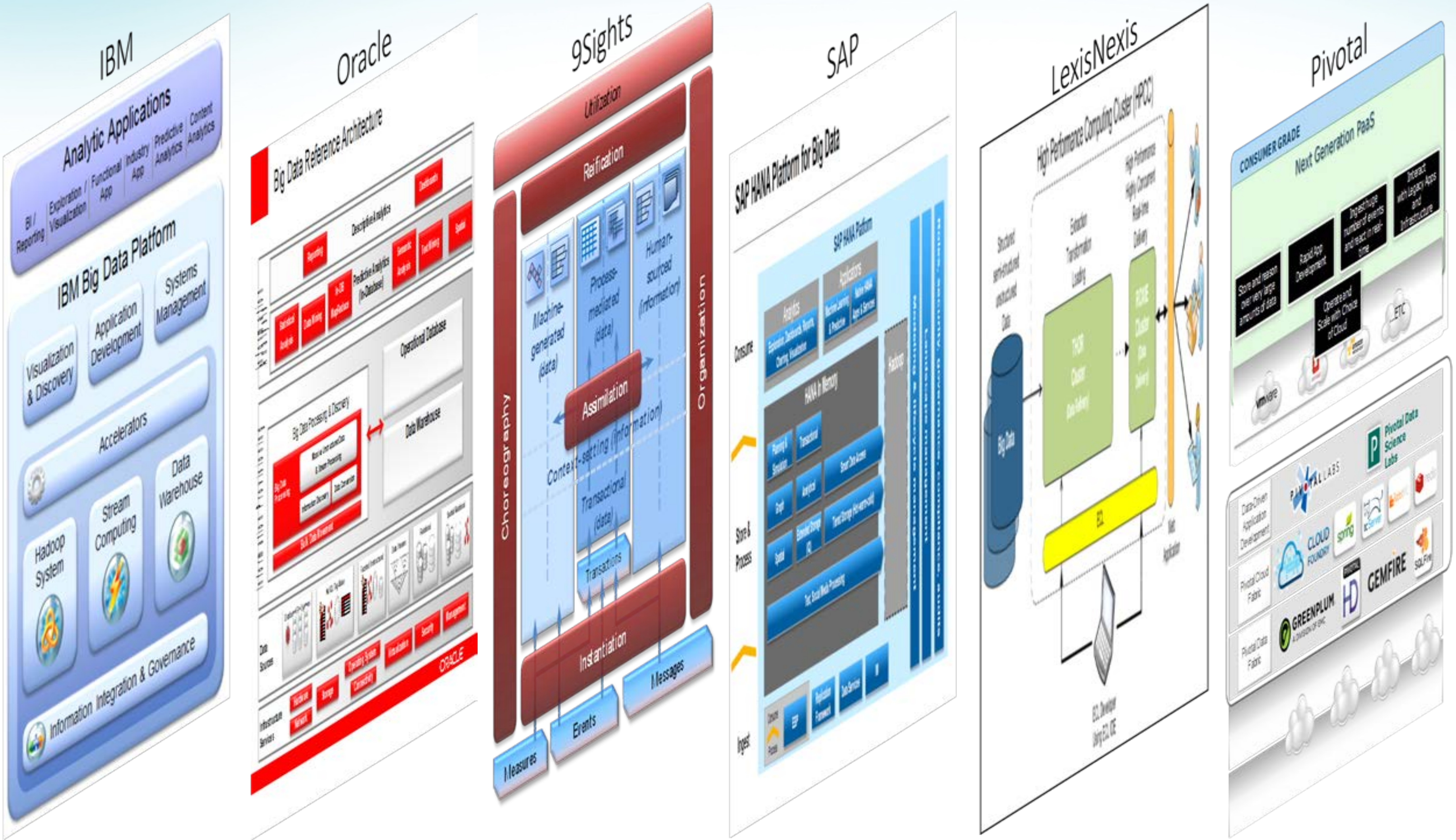
NIST Big Data Public Working Group (NBD-PWG)

Big Data Architecture Challenges: Computing Stack



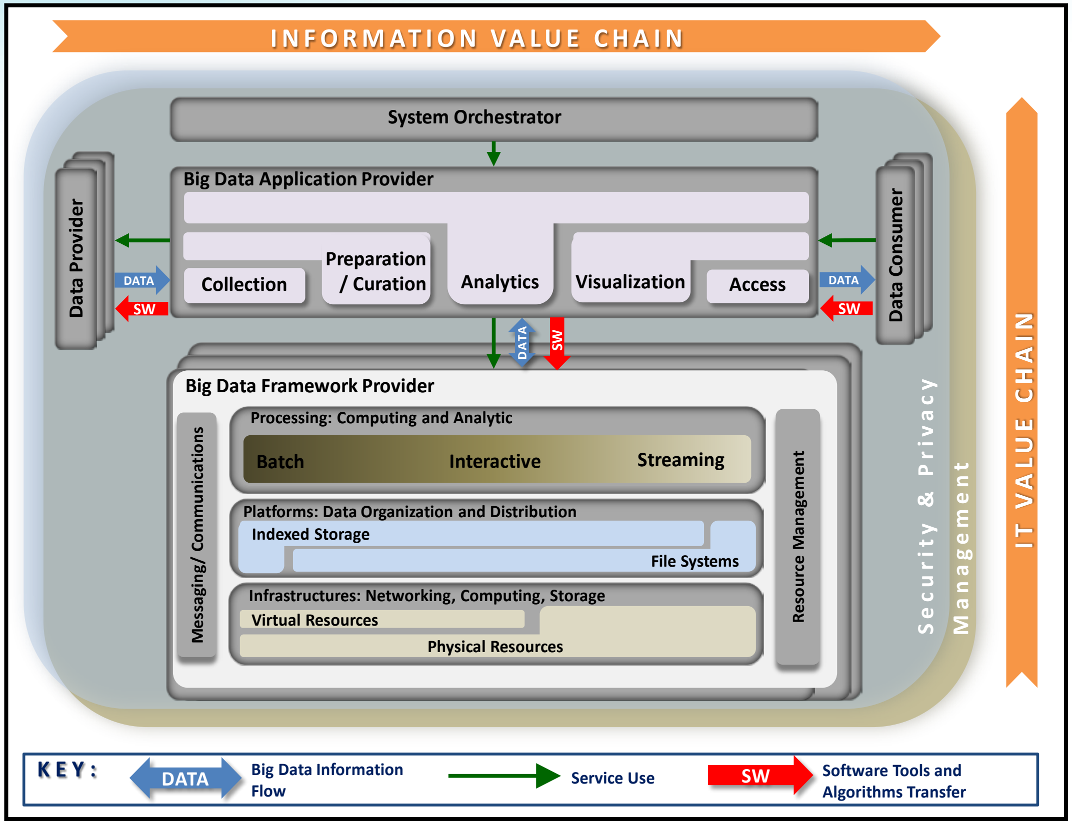
NIST Big Data Public Working Group (NBD-PWG)

Vendors Big Data architectures



NIST Big Data Public Working Group (NBD-PWG)

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



- 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

INFORMATION TECHNOLOGY LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

NIST Big Data Public Working Group (NBD-PWG)

Deliverable: Stage 1 & 2 – Reference Architecture + Interface

https://bigdatawg.nist.gov/V2_output_docs.php **(drafts as June 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	New NIST SP1500-8: Reference Architecture Interface	New NIST SP1500-9: Adoption & Modernization

Documents Transferred to JTC 1

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 International Standard (DIS) as July 2017)

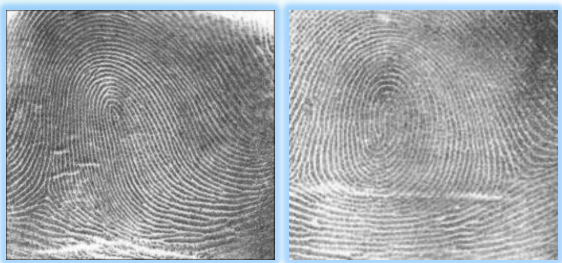
  - **ISO/IEC 20547 Information Technology – Big data Reference architecture (5 Parts as June 2017)**
    Part 1: (TR) Framework and Application Process (2nd WD)
    Part 2: (TR) Use Cases and Derived Requirements (under Publication)
    Part 3: (IS) Reference Architecture (CD as August 2017)
    Part 4: (IS) Security and Privacy Fabric (2nd ED, under SC 27/WG 4)
    Part 5: (TR) Standards Roadmap (under 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

NIST Big Data Public Working Group (NBD-PWG)

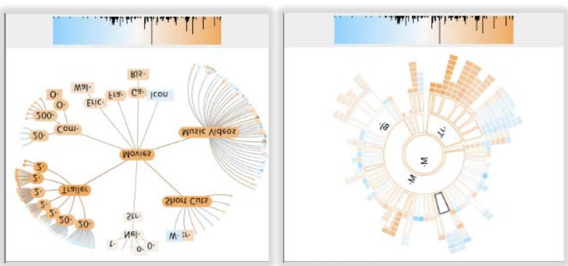
V2 NIST Big Data Development Strategies

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

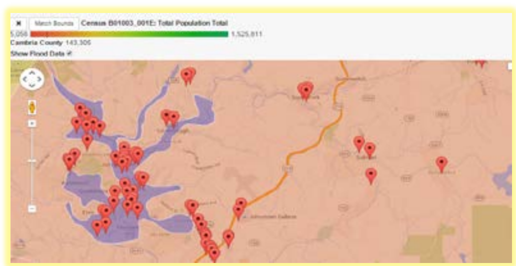


Fingerprints Matching

Human and Face Detection from Video



Twitter Feeds



Spatial Big Data/GIS

Healthcare Payment Fraud

- Data warehousing
- Global Cities
- Earth Science
- Life Science
- IoT
- *Others…*

NIST Big Data Public Working Group (NBD-PWG)

V2 NIST Big Data Development Strategies

- Use Cases Implementation
  - Identify small-scale implementable use cases with datasets and analytics algorithms which are available to public
  - Apply DevOps environment to implement selected use cases based on the NBD-RA components by using any given commercial/public Big Data technologies and tools (objective is to observe interactions and dataflow between NBD-RA components)
  - Under Development
    - Drug Discovery (HPC + Big Data)
    - Numeric Weather Prediction (HPC + Big Data)
    - Healthcare Fraud Detection (Big Data)
- Use Cases White Paper
  - Document step-by-step how a given use case be implemented
  - Make available implementation codes to public
  - Publish under NIST Special Publication

Seek Implementers, Collaborators, and Early Adopters

Use Case Implementation: Long DNA Sequence Alignment

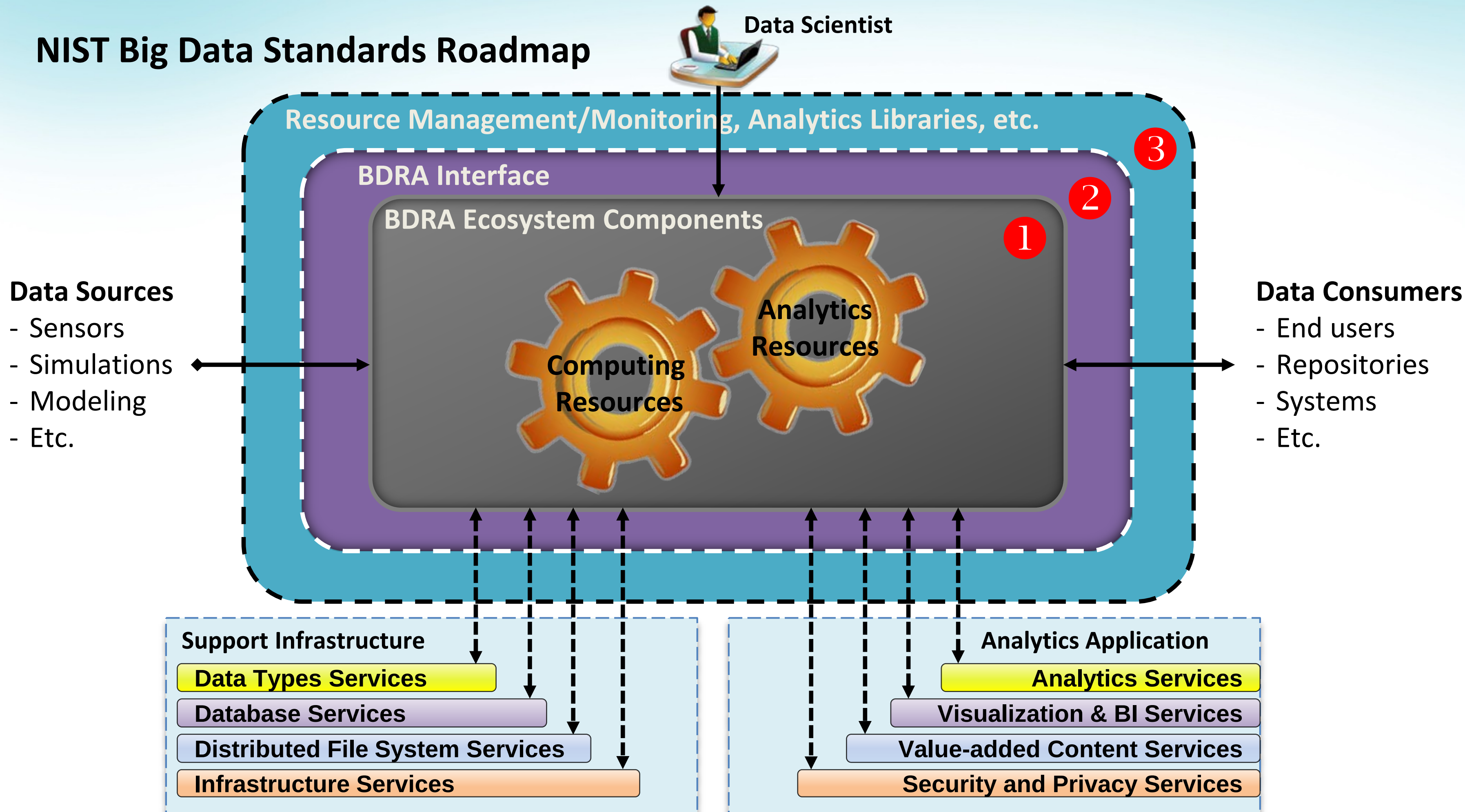
Sequence Pair sizes	1 KNC (SW1) [Using 237 threads] †		2 KNC (SW1) [Using 237* threads] †		4 KNC (SW1) [Using 237* threads] †		1 KNL (SW1+) [Using 41* Threads]		Nvidia K80 (SW2)		Nvidia TitanX (SW2)		Speedup (compare from 1KNC)				
	Time (s)	GCUPS	Time (s)	GCUPS	Time (s)	GCUPS	Time (s)	GCUPS	Time (s)	GCUPS	Time (s)	GCUPS	2 KNC	4 KNC	1 KNL	K80	TitanX
D44M vs. D46M 4.4×10^6 X 4.6×10^6	700 11.66m	29.2	396 6.60m	51.7	200 3.33m	100.7	1,258 29.9m	16.2	225 3.75m	91.0	120 2.00m	170.6	1.8	3.4	0.55	3.1	5.8
D23M vs. D33M 23×10^6 X 33×10^6	25,166 6.99h	30.0	14,105 3.91h	53.5	6,855 1.90h	110.1	26,397 7.33h	28.5	8,190 2.27h	92.1	4,930 1.36h	154.0	1.8	3.7	0.95	3.0	5.1
D23M vs. D42M 23×10^6 X 42×10^6	32,209 8.94h	30.0	17,958 4.98h	53.9	8,746 2.42h	110.6	33,978 9.43h	28.4	10,553 2.93h	91.6	6,342 1.76h	152.5	1.8	3.7	0.94	3.0	5.1
D23M vs. D50M 23×10^6 X 50×10^6	38,452 10.67h	30.0	21,291 5.91h	54.1	10,385 2.88h	111.0	40,519 11.25h	28.4	12,646 3.51h	91.1	7,754 2.15h	152.5	1.8	3.7	0.94	3.0	4.9
D33M vs. D42M 33×10^6 X 42×10^6	45,868 12.74h	30.1	25,553 7.09h	54.0	12,381 3.43h	111.4	45,617 12.67h	30.2	15,352 4.26h	89.9	9,043 2.51h	152.4	1.8	3.7	1.01	2.9	4.9
D33M vs. D50M 33×10^6 X 50×10^6	54,582 15.16h	30.1	30,402 8.44h	54.0	15,499 4.30h	106.0	54,340 15.09h	30.2	18,175 5.04h	90.3	10,751 2.98h	152.7	1.8	3.5	1.00	3.0	5.0
D42M vs. D50M 42×10^6 X 50×10^6	70,053 19.45h	30.0	38,875 10.79h	54.1	18,902 5.25h	111.4	67,564 18.76h	31.15	22,990 6.38h	91.5	13,615 3.78h	154.5	1.8	3.7	1.03	3.0	5.1

SW1=*SWAPHI-LS* for KNC, SW1+=modified SW1 for KNL, SW2=*SW#* using CUDA/GPU, *=maximum threads used before performance decrease, "h"=hour, "m"=minute
GCUPS=billion cell updates per second. †- http://ieeexplore.ieee.org/document/6968772/?arnumber=6968772&tag=1

Credit to: Michelle Luo, Yuechen Chen, Eddie Banuelos-Casillas, Cory Wang (all George Washington University intern students at NIST)

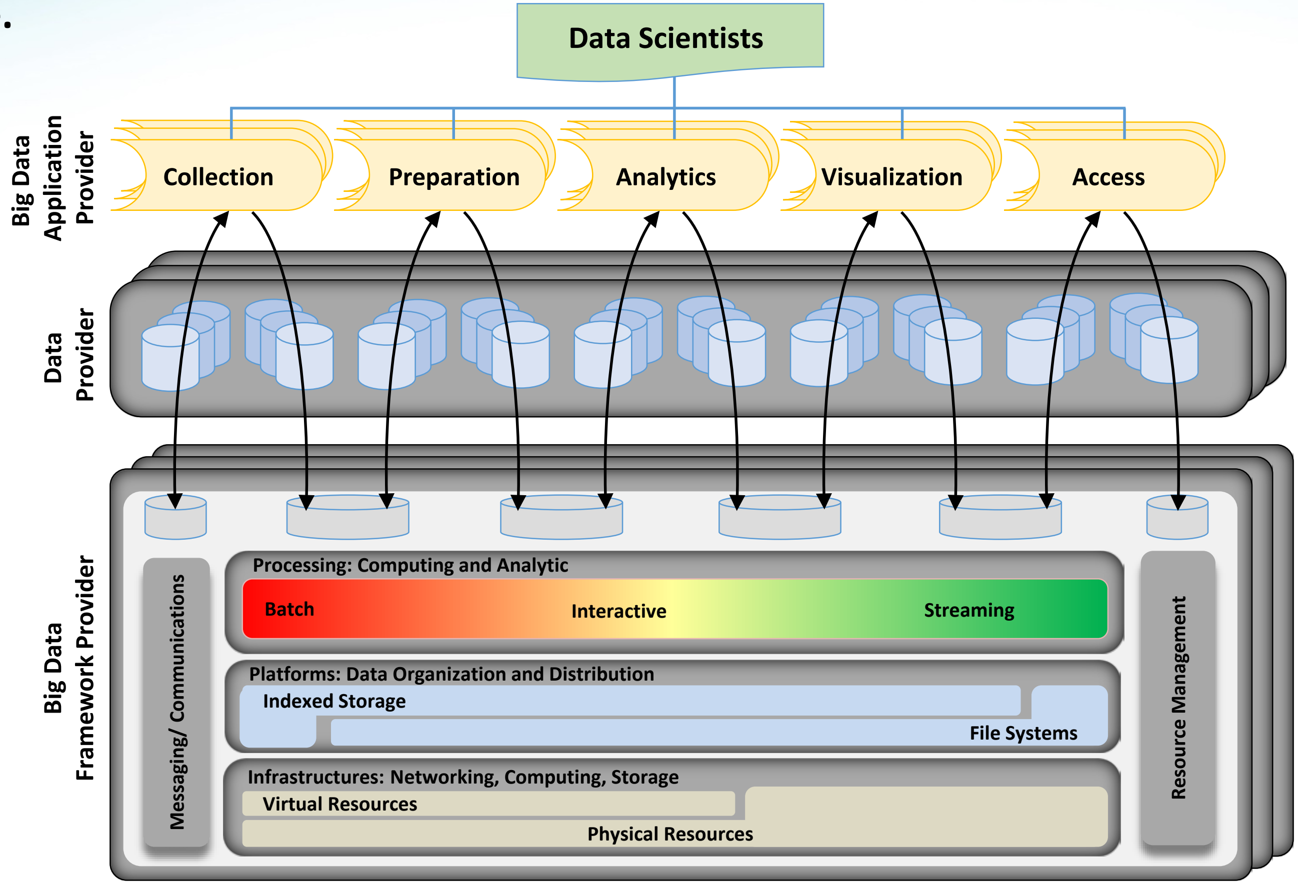
NIST Big Data Public Working Group (NBD-PWG)

NIST Big Data Standards Roadmap



Goals for Big Data Analytics and Beyond

Enable data scientists, engineers, researchers, etc. to increase productive and enhance quality in data science through modularized Big Data Analytics tools based from NIST Big Data Reference Architecture.

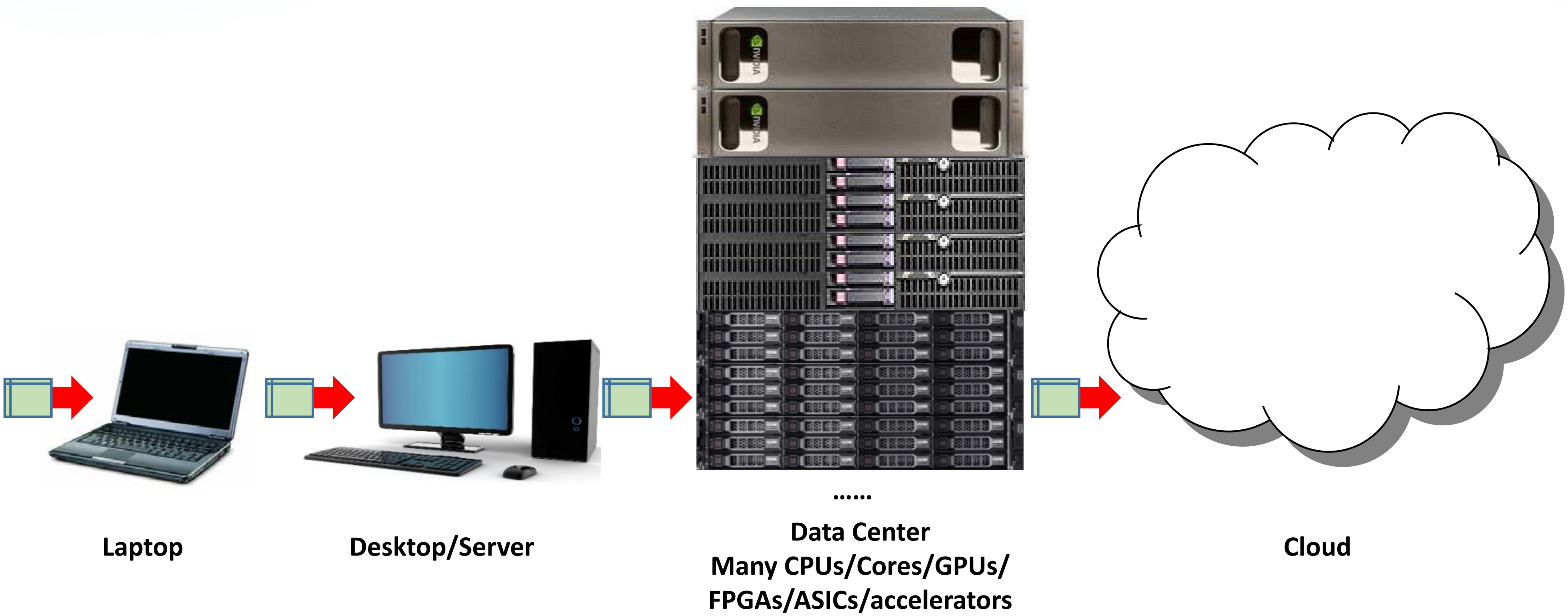


NIST Big Data Reference Architecture for Analytics and Beyond, Wo Chang, NIST/ITL, December 8, 2017

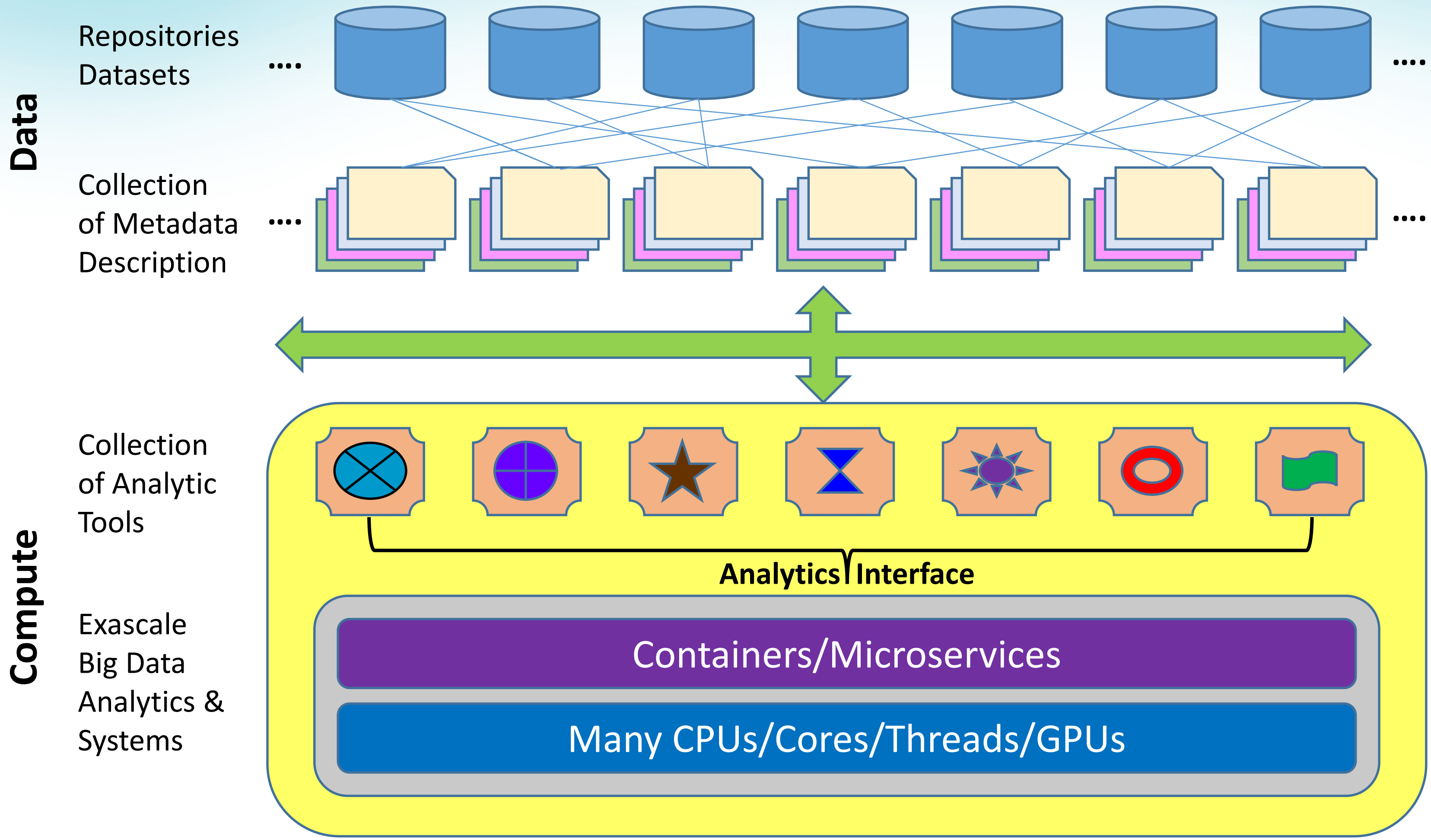
Goals for Big Data Analytics and Beyond

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) for HPC and Big Data (AI, deep learning, machine learning, etc.) computing environment.



Enable Convergence of Data + Compute



Questions ?