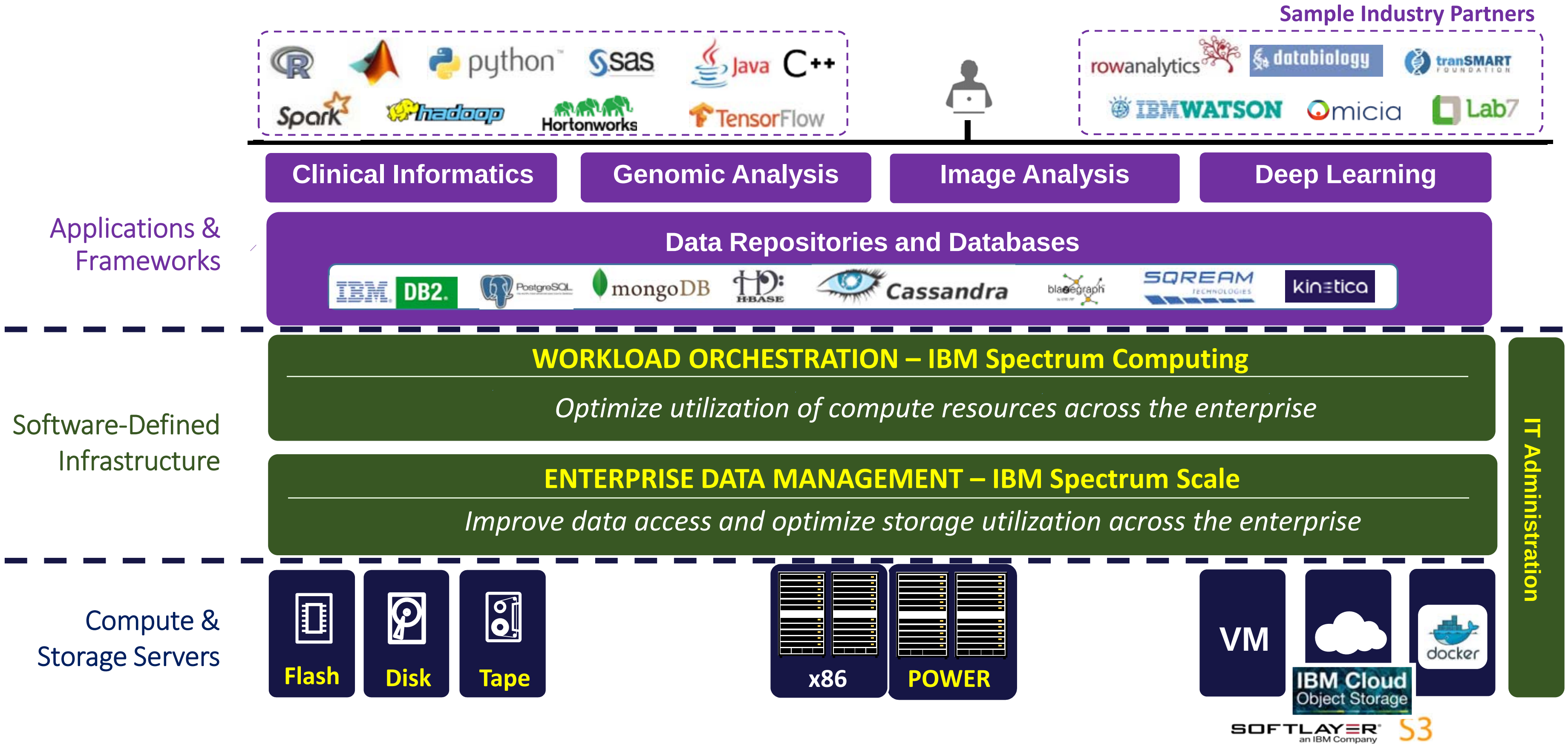
HPC 2.0 for Genomics

An Introduction to IBM HPDA Framework & Reference Architecture

Frank Lee, PhD
IBM Systems

IBM Systems Builds the Foundation for the Cognitive Era



Sample Industry Partners

Applications & Frameworks

Clinical Informatics

Genomic Analysis

Image Analysis

Deep Learning

Data Repositories and Databases

Software-Defined Infrastructure

WORKLOAD ORCHESTRATION – IBM Spectrum Computing

Optimize utilization of compute resources across the enterprise

ENTERPRISE DATA MANAGEMENT – IBM Spectrum Scale

Improve data access and optimize storage utilization across the enterprise

IT Administration

Compute & Storage Servers

Flash

Disk

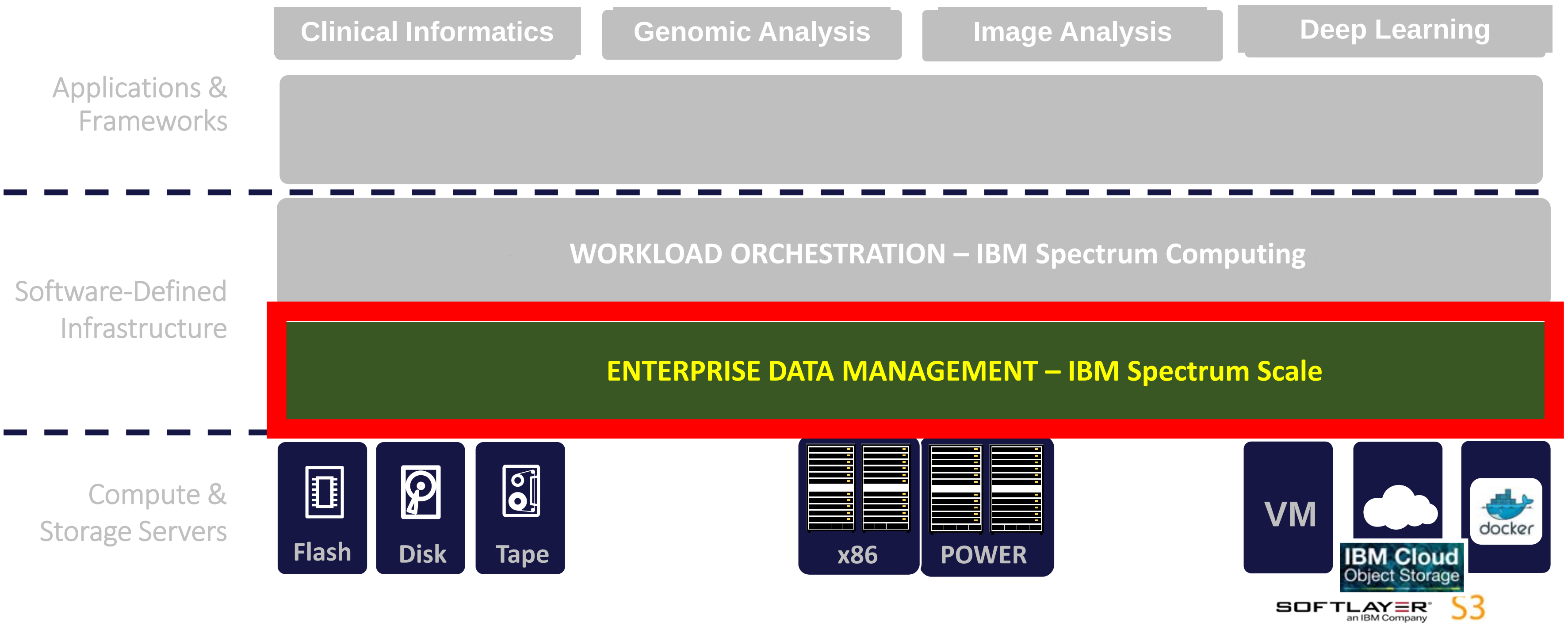
Tape

x86

POWER

VM

Foundation for Data



Clinical Informatics

Genomic Analysis

Image Analysis

Deep Learning

Applications & Frameworks

Software-Defined Infrastructure

WORKLOAD ORCHESTRATION – IBM Spectrum Computing

ENTERPRISE DATA MANAGEMENT – IBM Spectrum Scale

Compute & Storage Servers

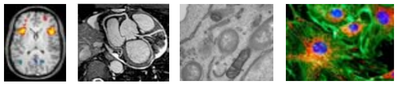
Challenge 1: High Speed for Big Data

Next Generation Sequencing (NGS)

- Raw Data: Up to 200 GB/file (compressed)
- Processed "Variant" Data: Up to 500 MB/file

Biomedical Imaging

- Medical Imaging: MRI, CT, Ultrasound, ...
- Microscopy

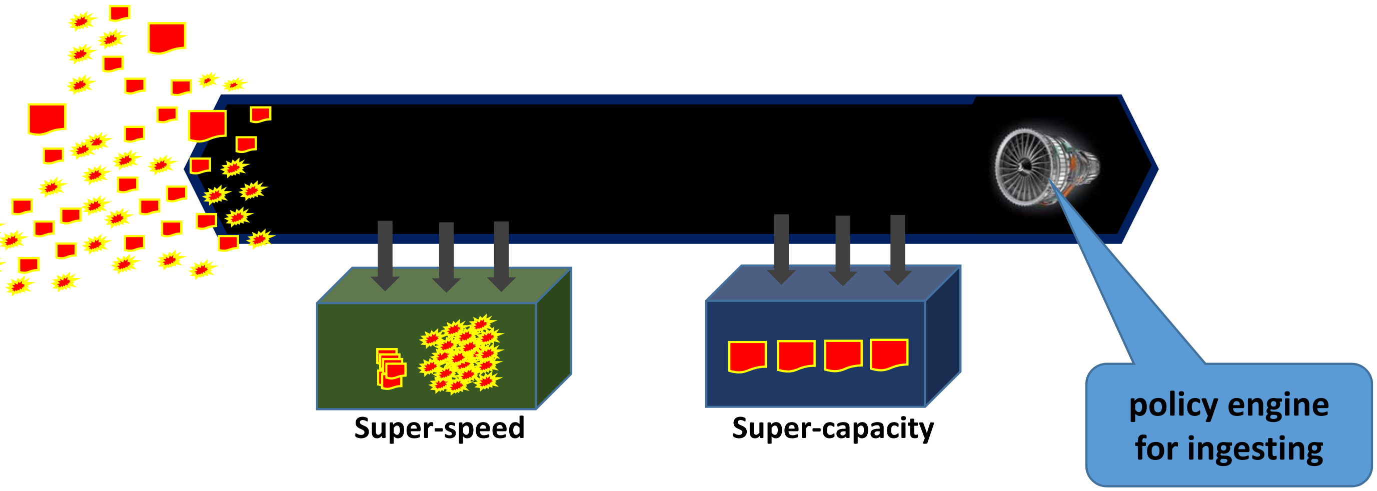


Time-Varying Sensors

- Medical Monitors
- Personal Sensors

Curated Scientific Literature

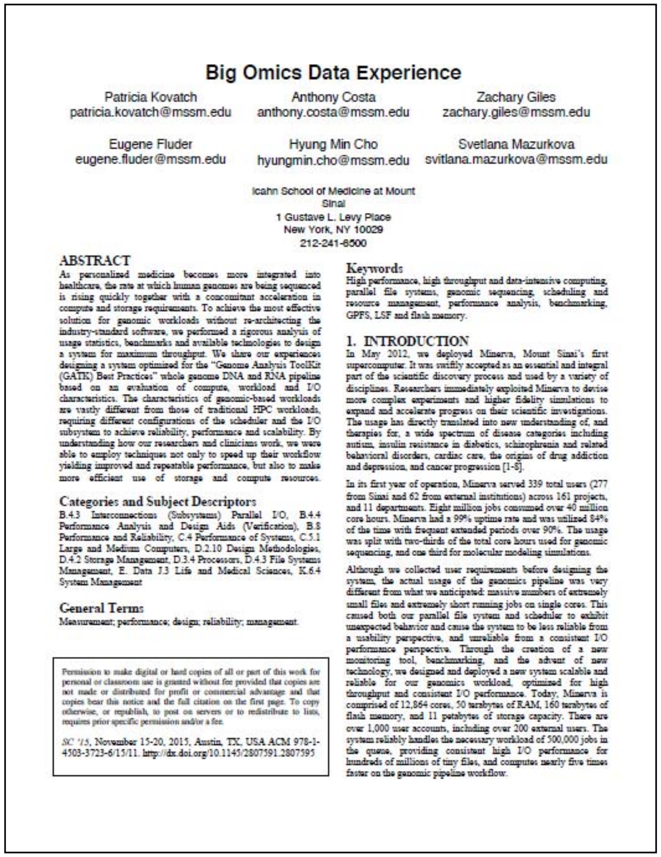
- Text files: CSV, TXT
- Online Web Crawls



Client Reference

Mount Sinai

Before	After
50 hours using 1 Node ~24cores, 1 QDR link, 256GB RAM	5 hours using 1 Node ~12cores, 1 FDR Link, 64GB RAM



Big Omics Data Experience

Patricia Kovatch
patricia.kovatch@mssm.edu

Anthony Costa
anthony.costa@mssm.edu

Zachary Giles
zachary.giles@mssm.edu

Eugene Fluder
eugene.fluder@mssm.edu

Hyung Min Cho
hyungmin.cho@mssm.edu

Svetlana Mazurkova
svitlana.mazurkova@mssm.edu

Icahn School of Medicine at Mount
Sinai
1 Gustave L. Levy Place
New York, NY 10029
212-241-6500

ABSTRACT

As personalized medicine becomes more integrated into healthcare, the rate at which human genomes are being sequenced is rising quickly together with a concomitant acceleration in compute and storage requirements. To achieve the most effective solution for genomic workloads without re-architecting the industry-standard software, we performed a rigorous analysis of usage statistics, benchmarks and available technologies to design a system for maximum throughput. We share our experiences designing a system optimized for the "Genome Analysis ToolKit (GATK) Best Practices" whole genome DNA and RNA pipeline based on an evaluation of compute, workload and I/O characteristics. The characteristics of genomic-based workloads are vastly different from those of traditional HPC workloads, requiring different configurations of the scheduler and the I/O subsystem to achieve reliability, performance and scalability. By understanding how our researchers and clinicians work, we were able to employ techniques not only to speed up their workflow yielding improved and repeatable performance, but also to make more efficient use of storage and compute resources.

Categories and Subject Descriptors

B.4.3 Interconnections (Subsystems) Parallel I/O, B.4.4 Performance Analysis and Design Aids (Verification), B.8 Performance and Reliability, C.4 Performance of Systems, C.5.1 Large and Medium Computers, D.2.10 Design Methodologies, D.4.2 Storage Management, D.3.4 Processors, D.4.3 File Systems Management, E. Data J.3 Life and Medical Sciences, K.6.4 System Management

General Terms

Measurement; performance; design; reliability; management.

SC '15, November 15-20, 2015, Austin, TX, USA ACM 978-1-4503-3723-6/15/11. http://dx.doi.org/10.1145/2807591.2807595

Keywords

High performance, high throughput and data-intensive computing, parallel file systems, genomic sequencing, scheduling and resource management, performance analysis, benchmarking, GPFS, LSF and flash memory.

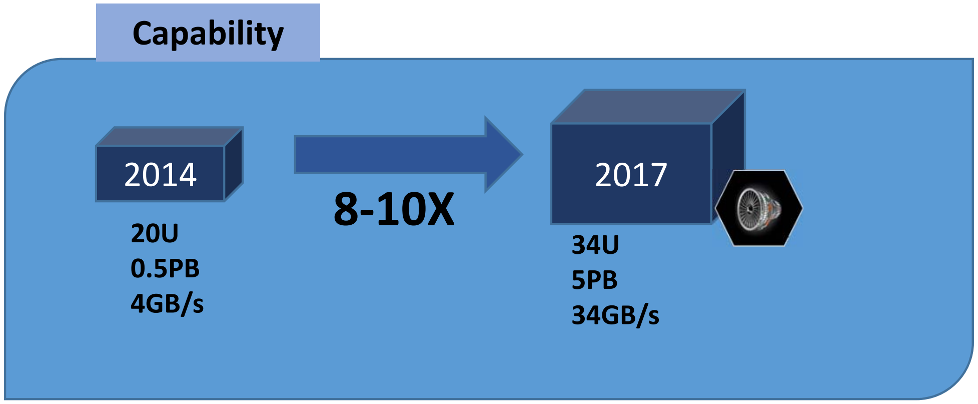
1. INTRODUCTION

In May 2012, we deployed Minerva, Mount Sinai's first supercomputer. It was swiftly accepted as an essential and integral part of the scientific discovery process and used by a variety of disciplines. Researchers immediately exploited Minerva to devise more complex experiments and higher fidelity simulations to expand and accelerate progress on their scientific investigations. The usage has directly translated into new understanding of, and therapies for, a wide spectrum of disease categories including autism, insulin resistance in diabetics, schizophrenia and related behavioral disorders, cardiac care, the origins of drug addiction and depression, and cancer progression [1-8].

In its first year of operation, Minerva served 339 total users (277 from Sinai and 62 from external institutions) across 161 projects, and 11 departments. Eight million jobs consumed over 40 million core hours. Minerva had a 99% uptime rate and was utilized 84% of the time with frequent extended periods over 90%. The usage was split with two-thirds of the total core hours used for genomic sequencing, and one third for molecular modeling simulations.

Although we collected user requirements before designing the system, the actual usage of the genomics pipeline was very different from what we anticipated: massive numbers of extremely small files and extremely short running jobs on single cores. This caused both our parallel file system and scheduler to exhibit unexpected behavior and cause the system to be less reliable from a usability perspective, and unreliable from a consistent I/O performance perspective. Through the creation of a new monitoring tool, benchmarking, and the advent of new technology, we designed and deployed a new system scalable and reliable for our genomics workload, optimized for high throughput and consistent I/O performance. Today, Minerva is comprised of 12,864 cores, 50 terabytes of RAM, 160 terabytes of flash memory, and 11 petabytes of storage capacity. There are over 1,000 user accounts, including over 200 external users. The system reliably handles the necessary workload of 500,000 jobs in the queue, providing consistent high I/O performance for hundreds of millions of tiny files, and computes nearly five times faster on the genomic pipeline workflow.

Data Machine for HPC 2.0



Capability

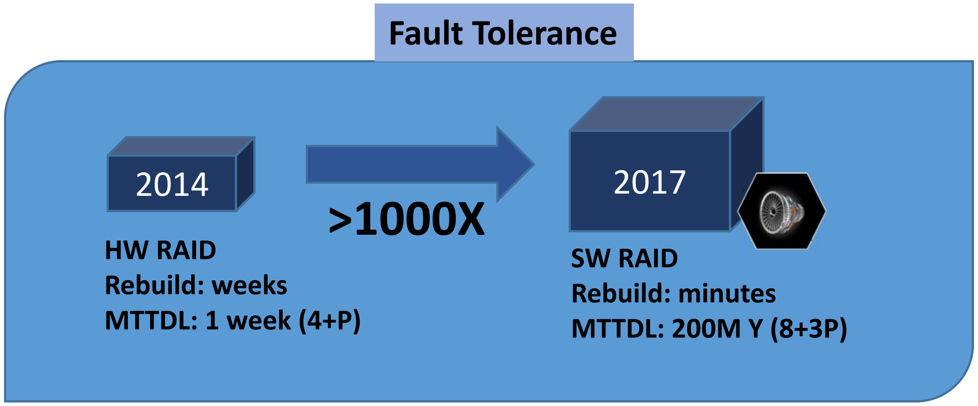
2014

20U
0.5PB
4GB/s

8-10X

2017

34U
5PB
34GB/s



Fault Tolerance

2014

HW RAID
Rebuild: weeks
MTTDL: 1 week (4+P)

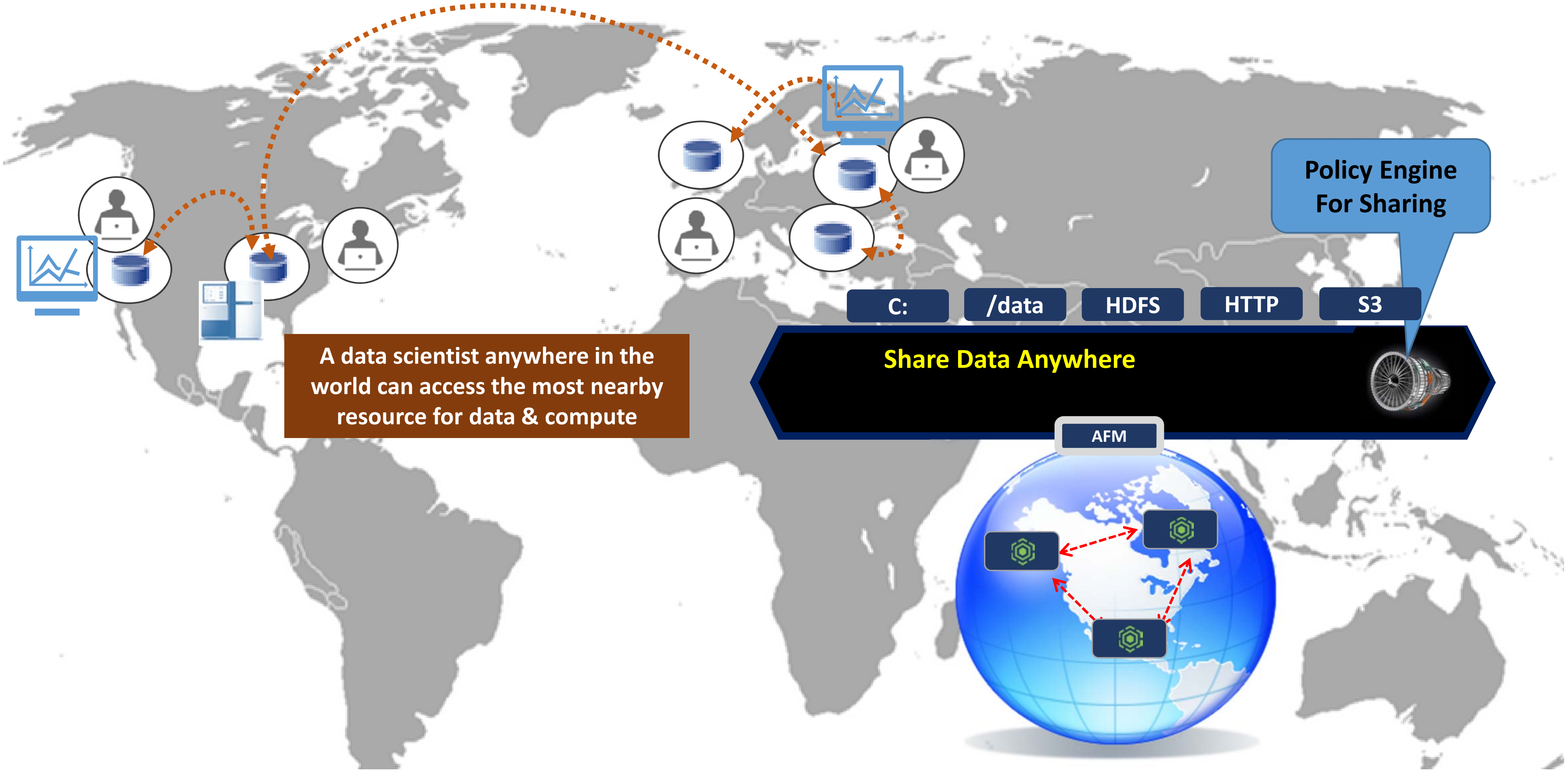
>1000X

2017

SW RAID
Rebuild: minutes
MTTDL: 200M Y (8+3P)

MTTDL: for 50,000 disk

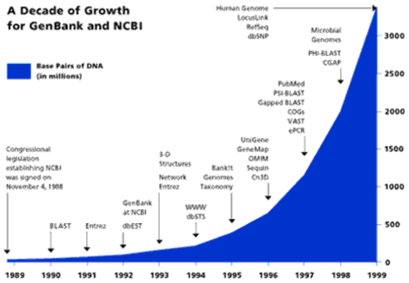
Challenge 2: Data Sharing for Global Collaboration



A data scientist anywhere in the world can access the most nearby resource for data & compute

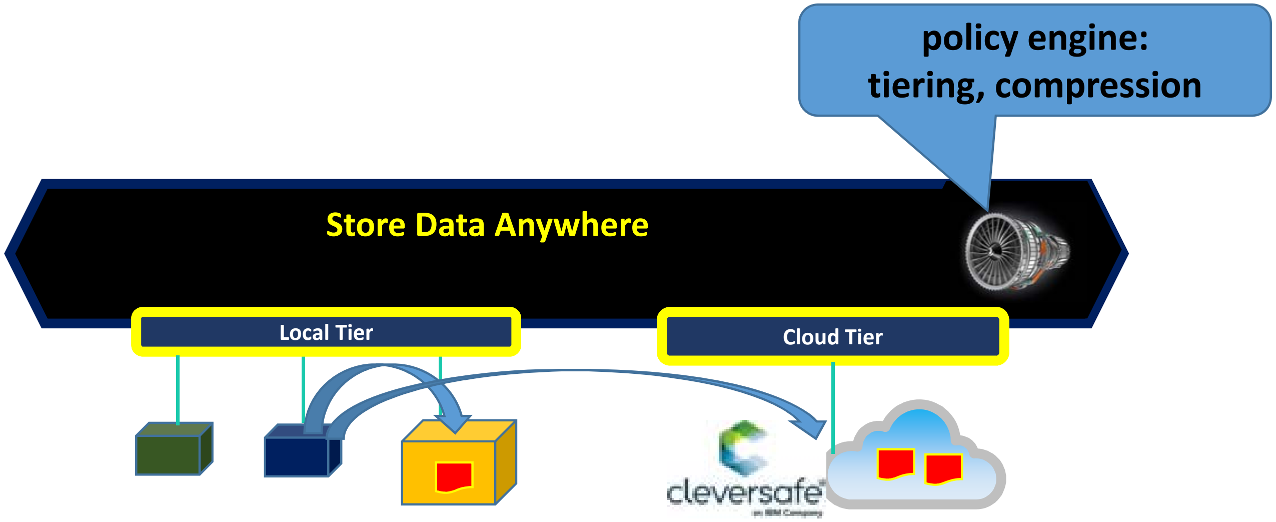
Policy Engine For Sharing

Challenge 3: Cost Control



Genomic data doubling
6-12 months

But budgets are not!

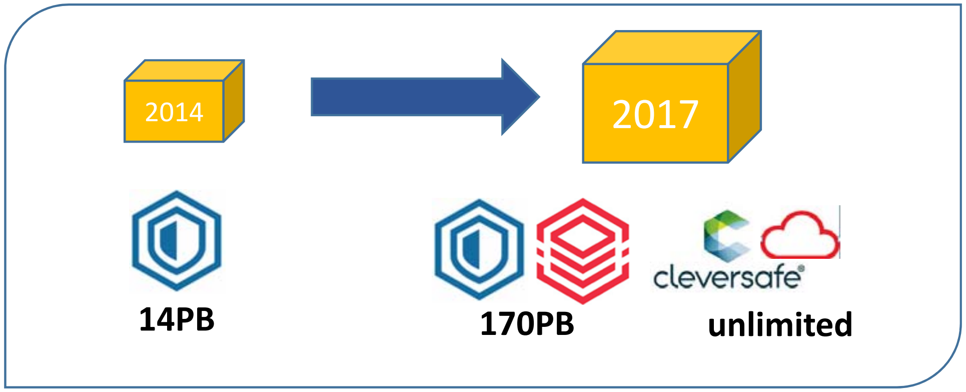


policy engine:
tiering, compression

Store Data Anywhere

Local Tier

Cloud Tier



14PB

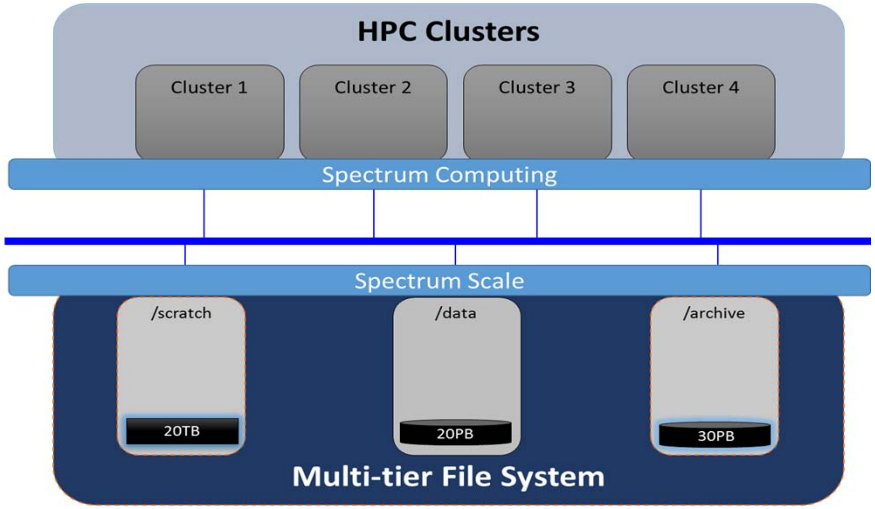
170PB

unlimited

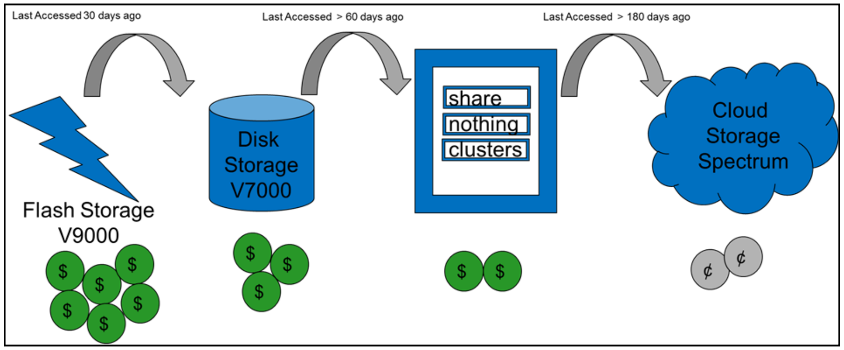
Client Reference: Scale - Cost

IBM

THE UNIVERSITY OF TEXAS
MD Anderson
~~Cancer~~ Center



City of Hope™



Performance	Cost
10X Better performance on the same hardware	90% Reduction of storage cost

Client References Cont'd: Scale - Cost

THE
GENOME
INSTITUTE
at Washington University

NEW YORK
GENOME CENTER®

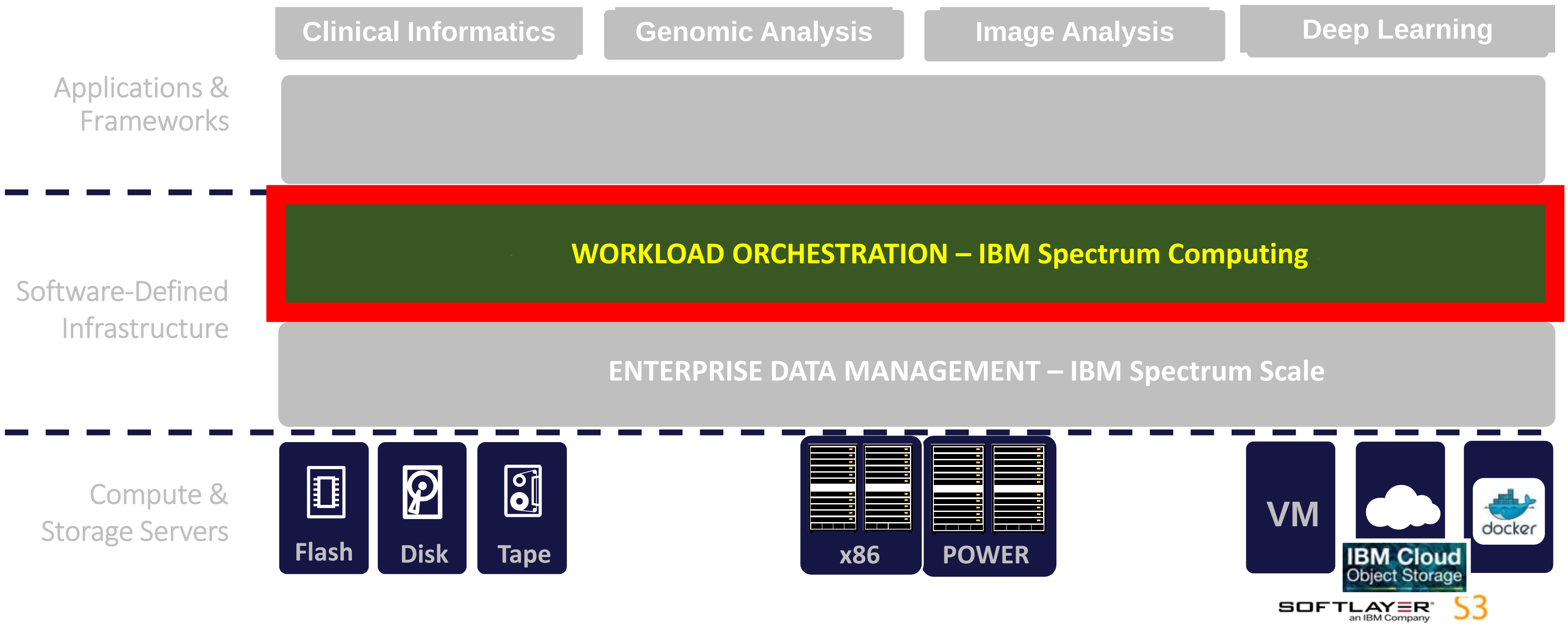
University
of Basel

السدرة
Sidra
مركز السدرة للطب والبحوث
Sidra Medical and Research Center

Research Computing
UNIVERSITY OF COLORADO BOULDER

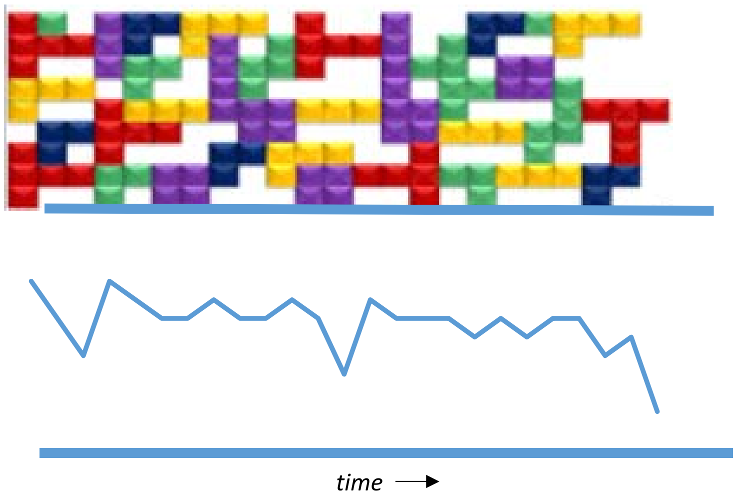
W
UNIVERSITY of
WASHINGTON

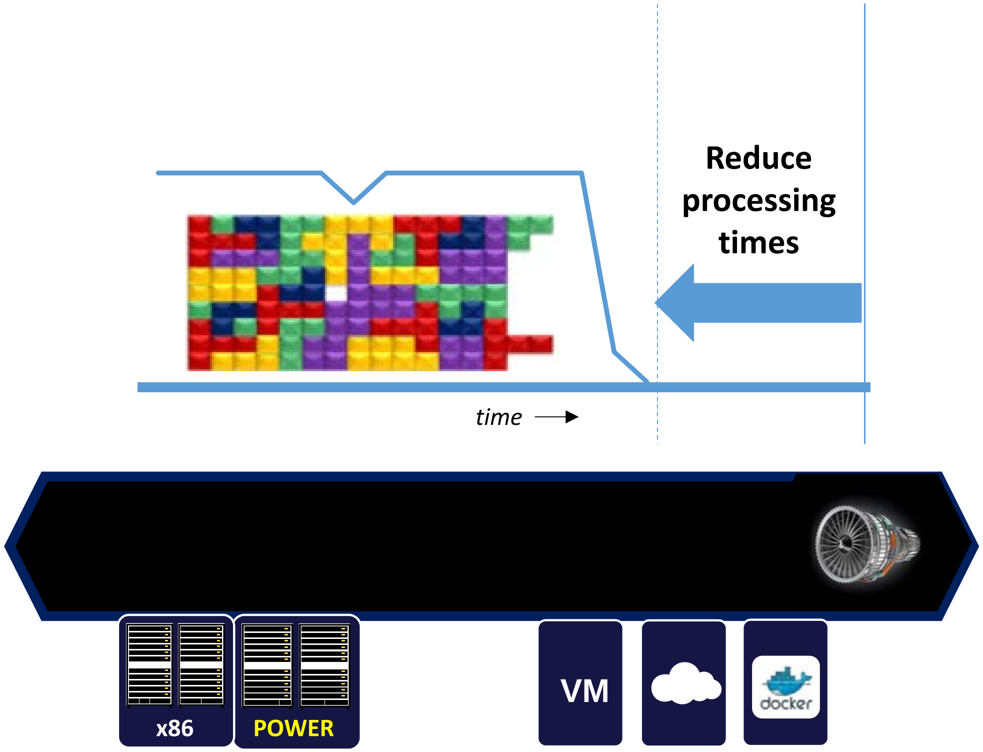
Foundation for Workload


Clinical Informatics
Genomic Analysis
Image Analysis
Deep Learning
Applications & Frameworks
Software-Defined Infrastructure
WORKLOAD ORCHESTRATION – IBM Spectrum Computing
ENTERPRISE DATA MANAGEMENT – IBM Spectrum Scale
Compute & Storage Servers
Flash
Disk
Tape
x86
POWER
VM
docker
IBM Cloud Object Storage
SOFTLAYER an IBM Company
S3

Challenge 4: Workflow Optimization

Example #1: Resource Utilization for Workflow





Client References: Workflow Optimization

wellcome trust sanger institute

Data-aware scheduling with API

THE GENOME INSTITUTE at Washington University

IO-aware scheduling with real-time data

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

IO-aware scheduling with some math

$$\frac{\partial f}{\partial t} = \lim_{h \to 0} \frac{f(t+h,\vec{x}) - f(t,\vec{x})}{h}$$

github/stjude

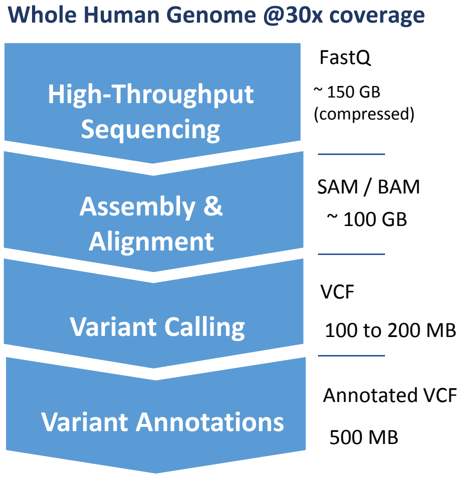
Challenge 5 Workflow Automation

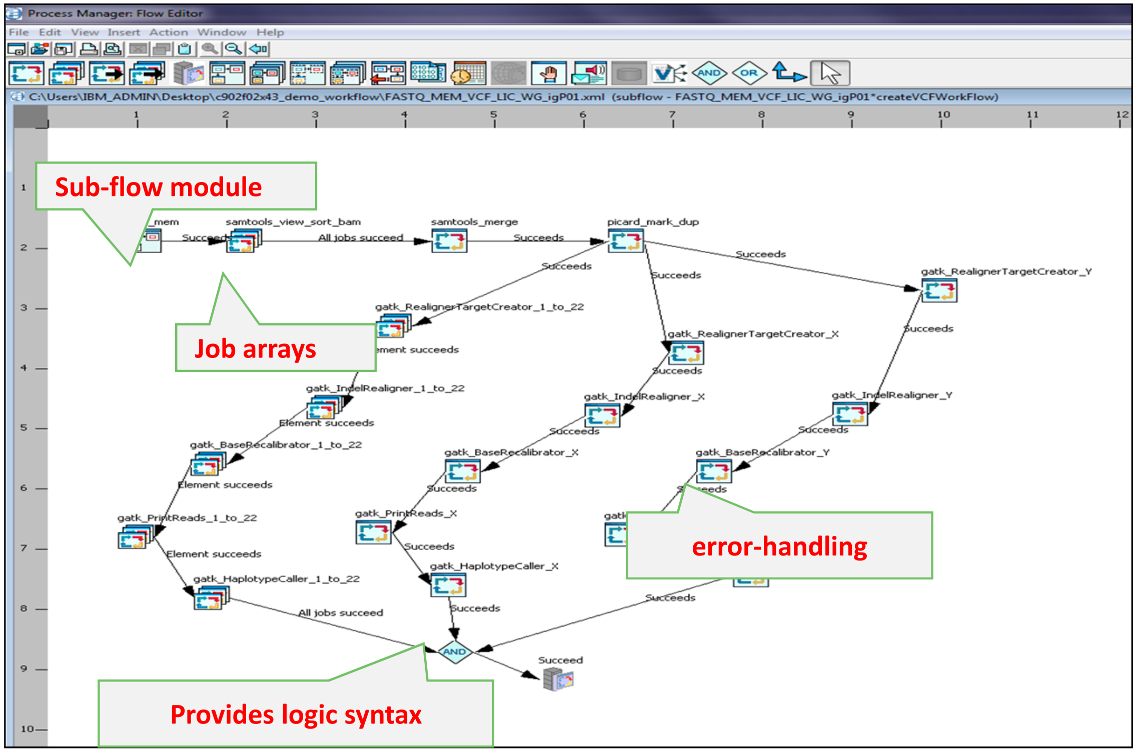
Example #1: Genomic Analysis Pipelines

Processing time per genome

1 to 100 hours*

on 1 compute node





C:\Users\IBM_ADMIN\Desktop\c902f02x43_demo_workflow\FASTQ_MEM_VCF_LIC_WG_igP01.xml (subflow - FASTQ_MEM_VCF_LIC_WG_igP01*createVCFWorkFlow)

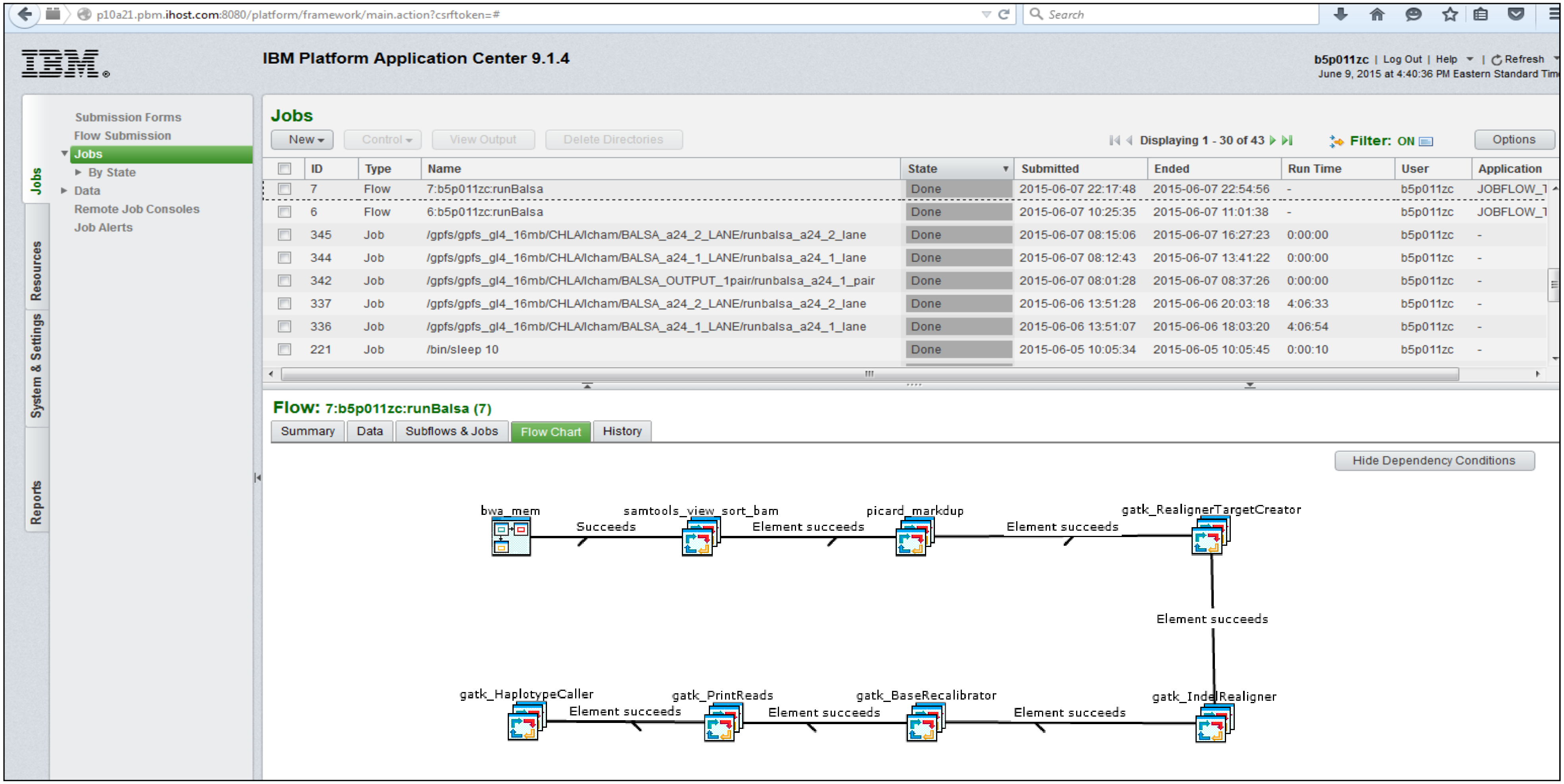
Sub-flow module

Job arrays

error-handling

Provides logic syntax

Client Reference: Workflow Automation

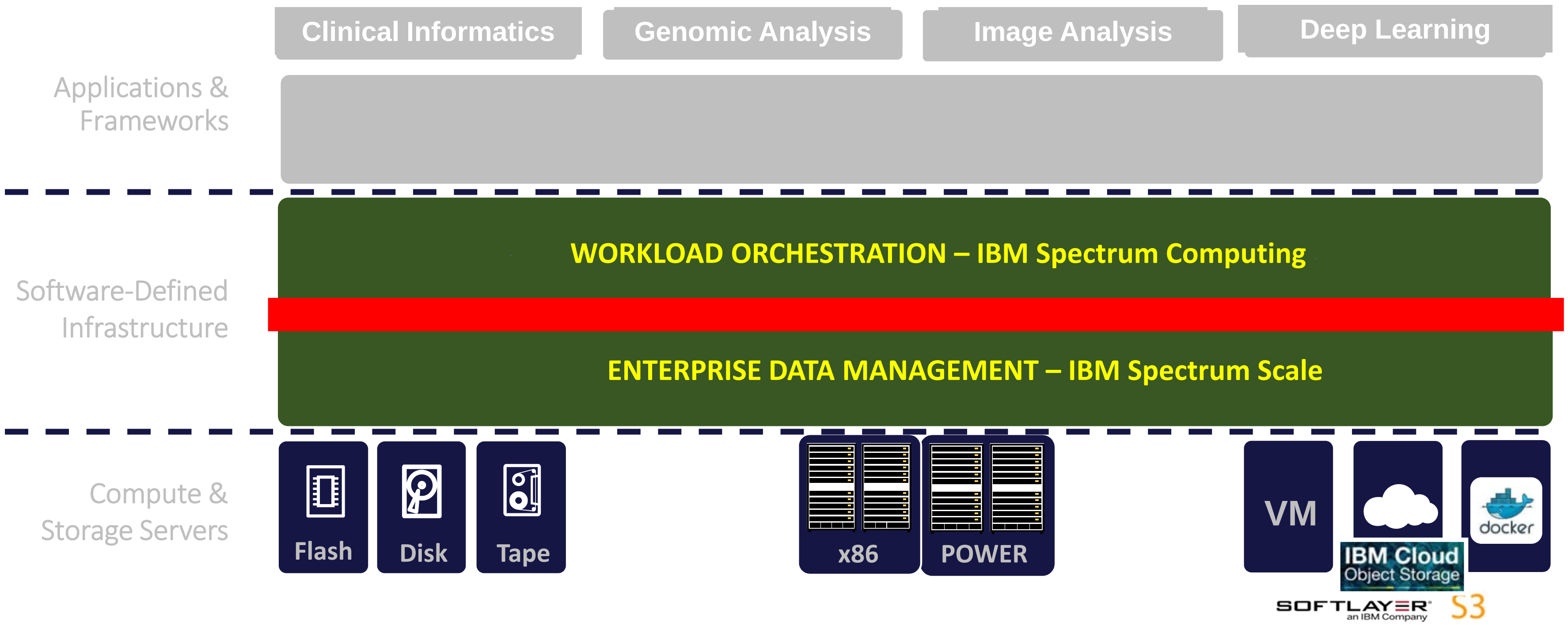


Jobs

☐	ID	Type	Name	State	Submitted	Ended	Run Time	User	Application
☐	7	Flow	7:b5p011zc:runBalsa	Done	2015-06-07 22:17:48	2015-06-07 22:54:56	-	b5p011zc	JOBFLOW_1
☐	6	Flow	6:b5p011zc:runBalsa	Done	2015-06-07 10:25:35	2015-06-07 11:01:38	-	b5p011zc	JOBFLOW_1
☐	345	Job	/gpfs/gpfs_gl4_16mb/CHLA/lcham/BALSA_a24_2_LANE/runbalsa_a24_2_lane	Done	2015-06-07 08:15:06	2015-06-07 16:27:23	0:00:00	b5p011zc	-
☐	344	Job	/gpfs/gpfs_gl4_16mb/CHLA/lcham/BALSA_a24_1_LANE/runbalsa_a24_1_lane	Done	2015-06-07 08:12:43	2015-06-07 13:41:22	0:00:00	b5p011zc	-
☐	342	Job	/gpfs/gpfs_gl4_16mb/CHLA/lcham/BALSA_OUTPUT_1pair/runbalsa_a24_1_pair	Done	2015-06-07 08:01:28	2015-06-07 08:37:26	0:00:00	b5p011zc	-
☐	337	Job	/gpfs/gpfs_gl4_16mb/CHLA/lcham/BALSA_a24_2_LANE/runbalsa_a24_2_lane	Done	2015-06-06 13:51:28	2015-06-06 20:03:18	4:06:33	b5p011zc	-
☐	336	Job	/gpfs/gpfs_gl4_16mb/CHLA/lcham/BALSA_a24_1_LANE/runbalsa_a24_1_lane	Done	2015-06-06 13:51:07	2015-06-06 18:03:20	4:06:54	b5p011zc	-
☐	221	Job	/bin/sleep 10	Done	2015-06-05 10:05:34	2015-06-05 10:05:45	0:00:10	b5p011zc	-

Flow: 7:b5p011zc:runBalsa (7)

Foundation for Metadata & Provenance



Applications & Frameworks

Clinical Informatics

Genomic Analysis

Image Analysis

Deep Learning

Software-Defined Infrastructure

WORKLOAD ORCHESTRATION – IBM Spectrum Computing

ENTERPRISE DATA MANAGEMENT – IBM Spectrum Scale

Compute & Storage Servers

Flash

Disk

Tape

x86

POWER

VM

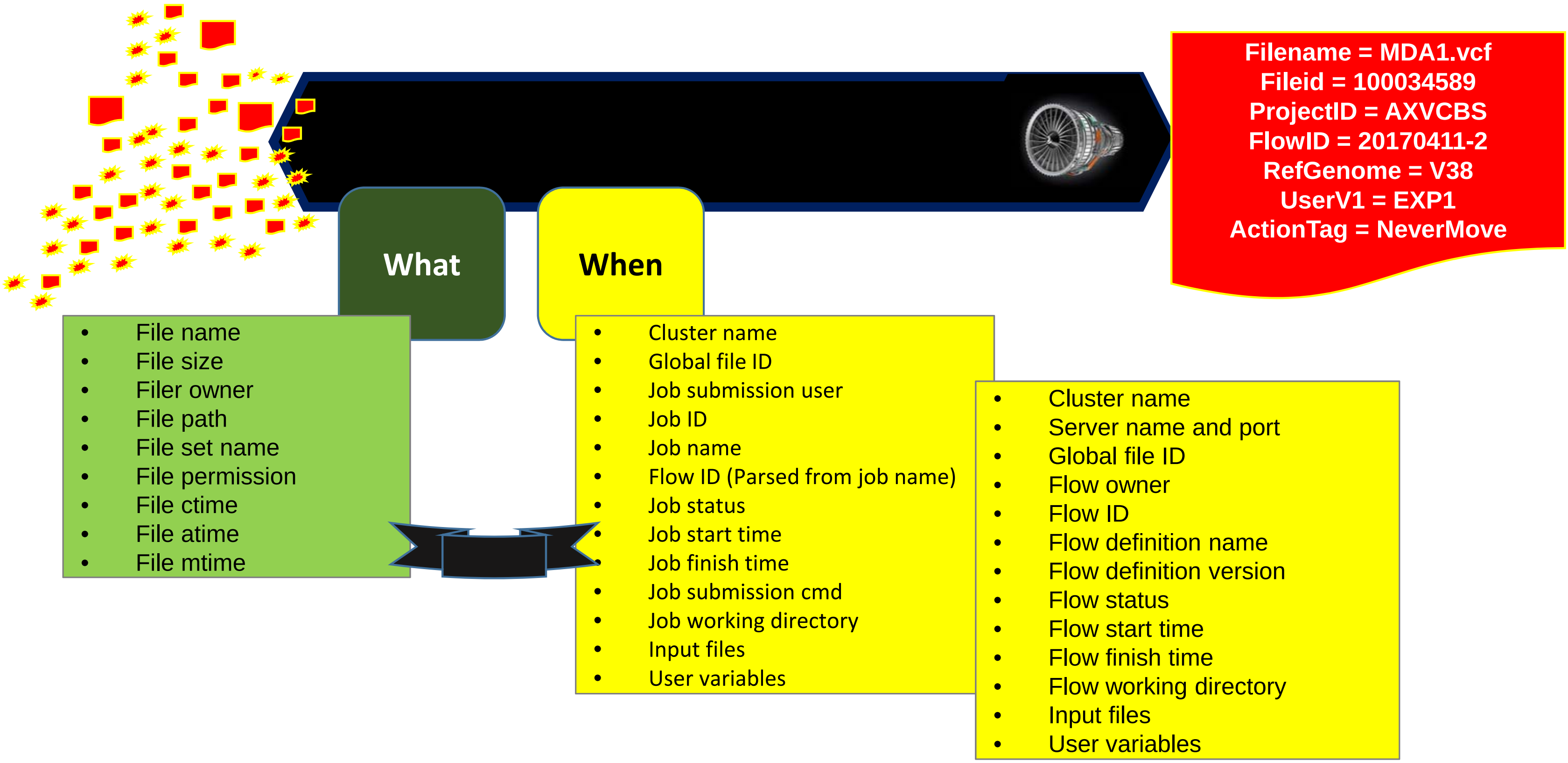
IBM Cloud Object Storage

docker

SOFTLAYER an IBM Company

S3

Challenge 6: Provenance



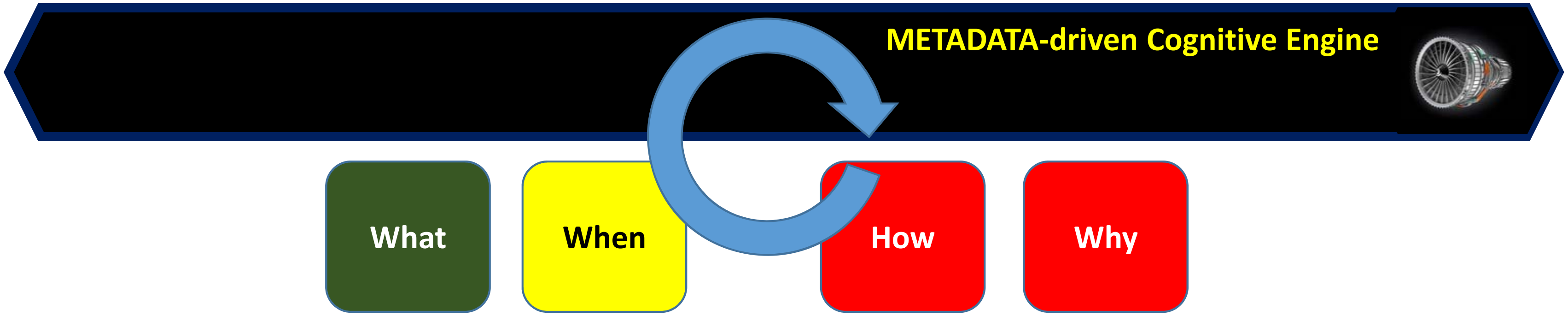
Filename = MDA1.vcf
Fileid = 100034589
ProjectID = AXVCBS
FlowID = 20170411-2
RefGenome = V38
UserV1 = EXP1
ActionTag = NeverMove

- File name
- File size
- Filer owner
- File path
- File set name
- File permission
- File ctime
- File atime
- File mtime

- Cluster name
- Global file ID
- Job submission user
- Job ID
- Job name
- Flow ID (Parsed from job name)
- Job status
- Job start time
- Job finish time
- Job submission cmd
- Job working directory
- Input files
- User variables

- Cluster name
- Server name and port
- Global file ID
- Flow owner
- Flow ID
- Flow definition name
- Flow definition version
- Flow status
- Flow start time
- Flow finish time
- Flow working directory
- Input files
- User variables

Beyond Provenance



METADATA-driven Cognitive Engine

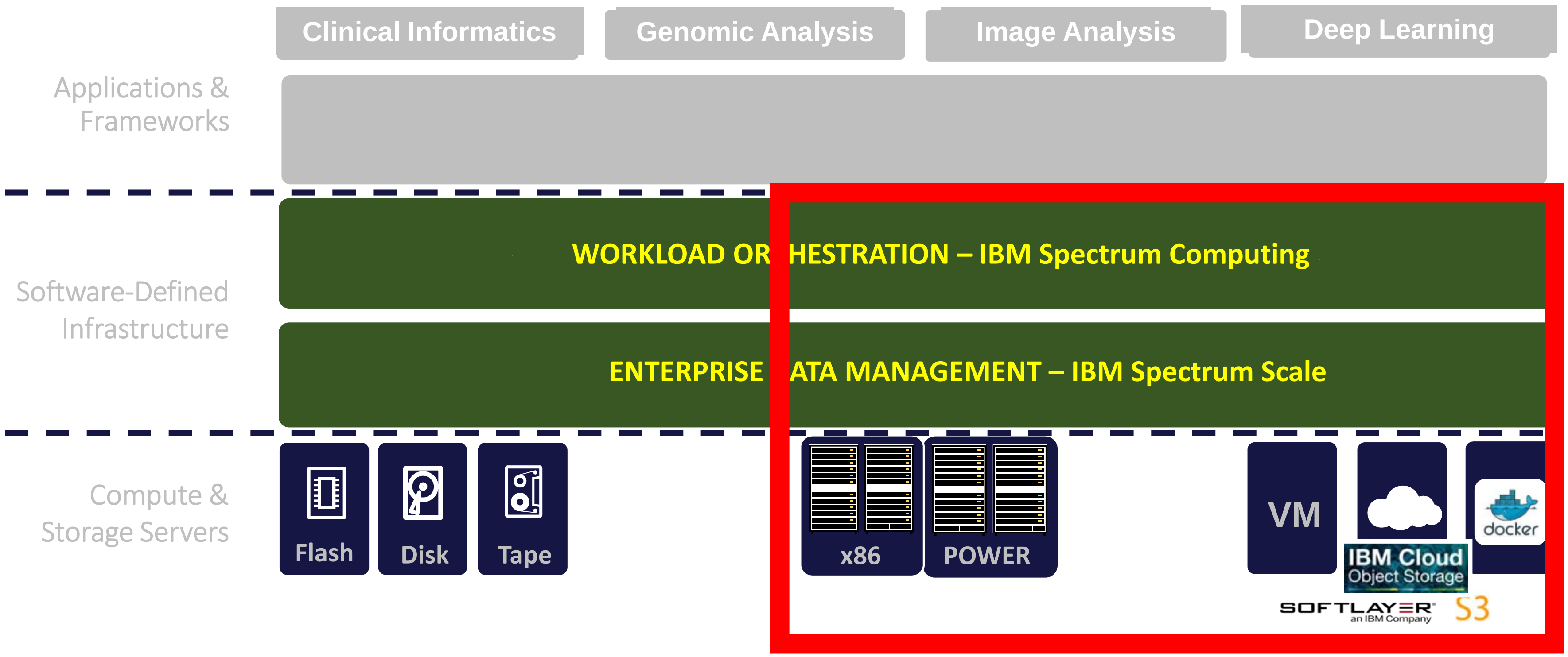
What

When

How

Why

Foundation for Cloud



Clinical Informatics

Genomic Analysis

Image Analysis

Deep Learning

Applications & Frameworks

Software-Defined Infrastructure

WORKLOAD OR HESTRATION – IBM Spectrum Computing

ENTERPRISE ATA MANAGEMENT – IBM Spectrum Scale

Compute & Storage Servers

Flash

Disk

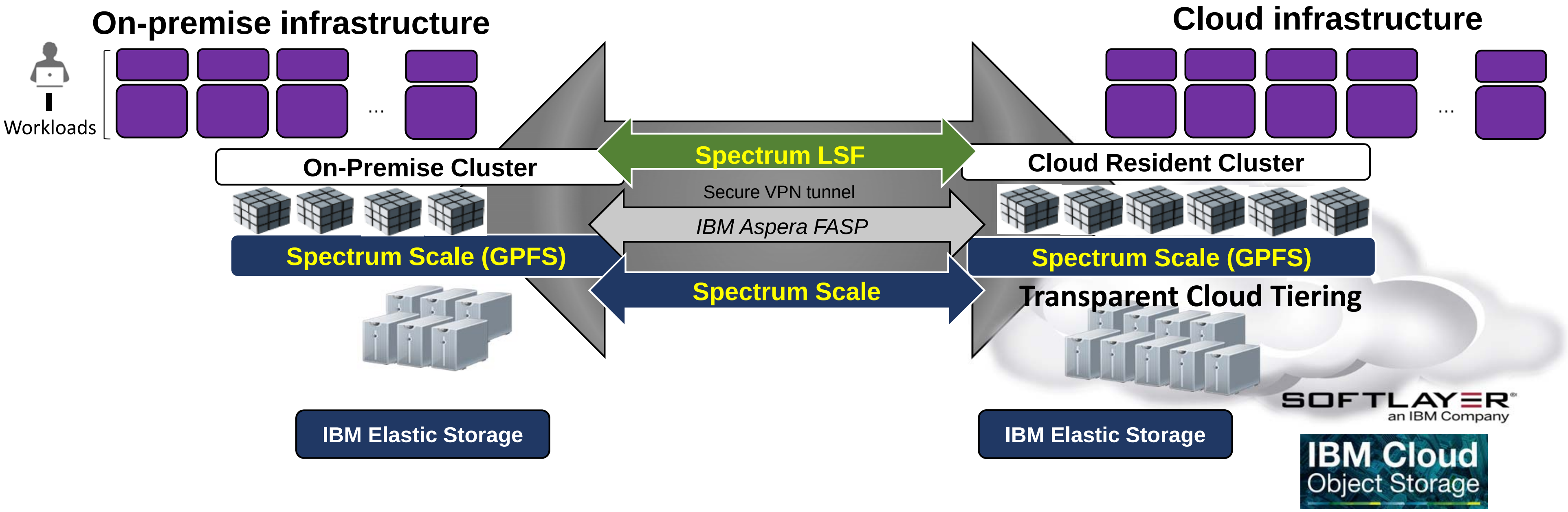
Tape

x86

POWER

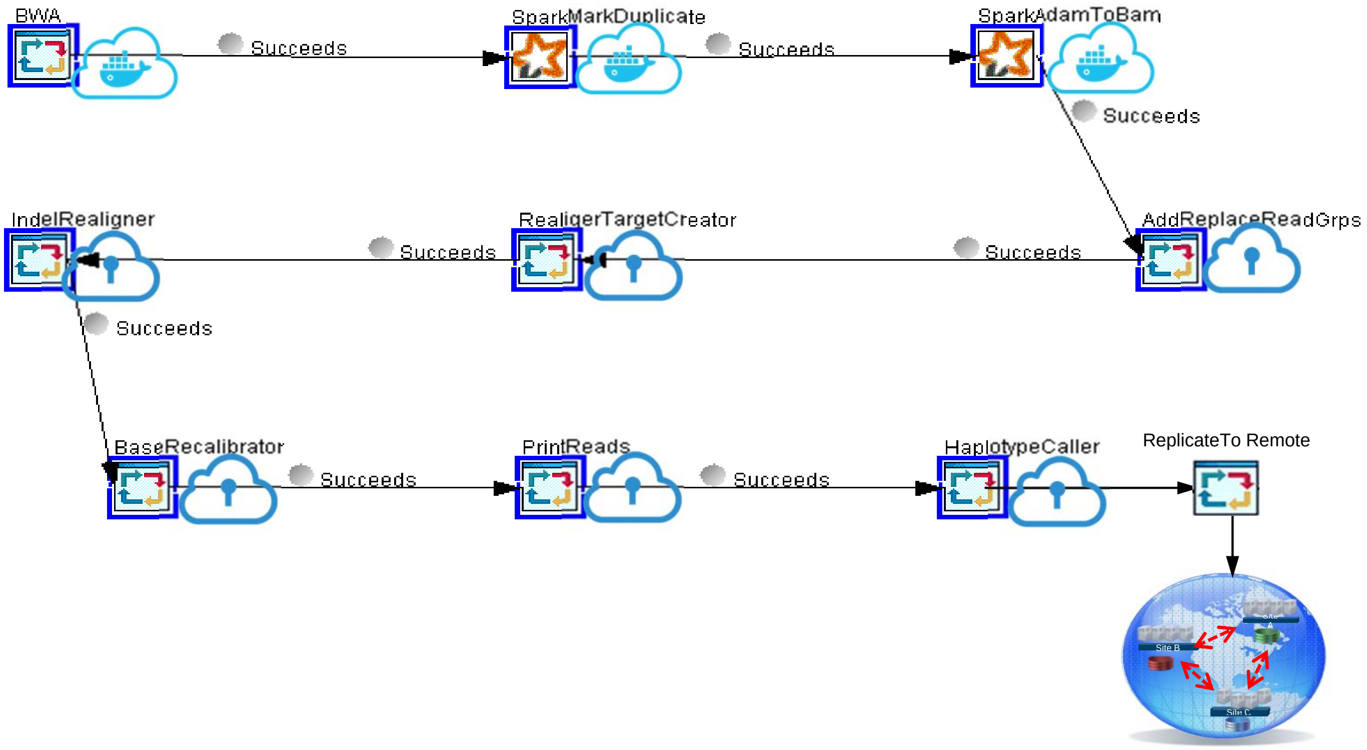
VM

A Hybrid Cloud Architecture

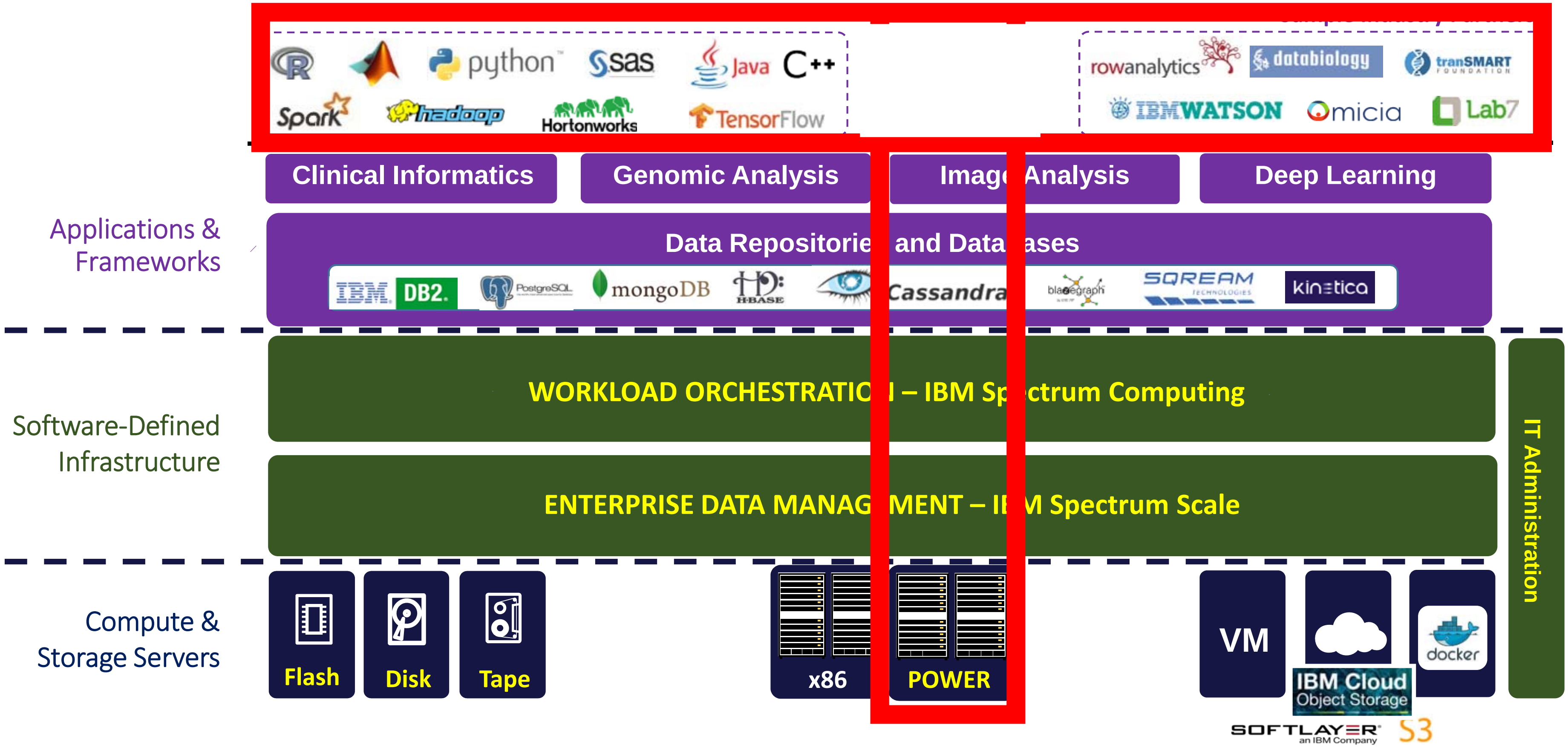


†AFM = Active File Management

Workflow in Multi-cloud



Application-level Optimization



Applications & Frameworks

Clinical Informatics

Genomic Analysis

Imag Analysis

Deep Learning

Data Repositorie and Data ases

Software-Defined Infrastructure

WORKLOAD ORCHESTRATIO – IBM Sp ctrum Computing

ENTERPRISE DATA MANAG MENT – IB M Spectrum Scale

IT Administration

Compute & Storage Servers

Flash

Disk

Tape

x86

POWER

VM

Application-level Optimization

IBM POWER continues to develop technologies that accelerate compute for the next generation of analytics, including the latest deep learning & machine learning algorithms

OpenPOWER™

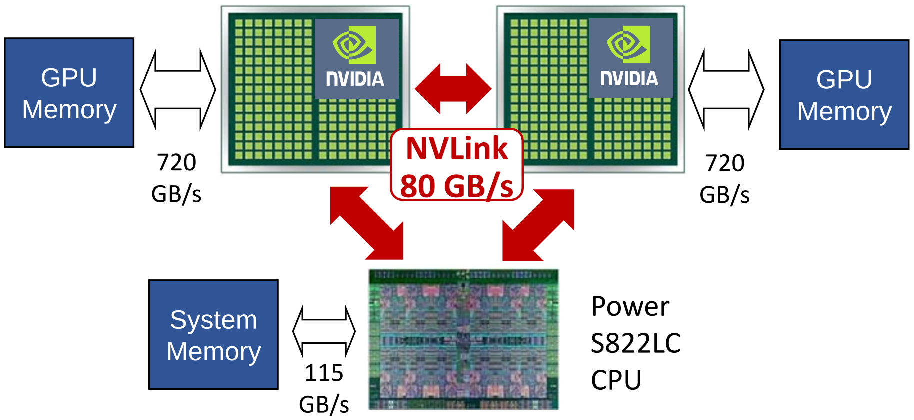
IBM POWER S822LC

Basic Advantages over Intel Haswell

Feature	Intel Haswell	IBM POWER8
SMT / Core	2 Threads	8 Threads
L1d Cache / Core	32 KB	64 KB
L2 Cache / Core	256 KB	512 KB
L3 Cache / Processor	16 to 45 MB	80 to 96 MB
L4 Cache / System	-	64 MB to 2 GB
Maximum Sustained Memory Bandwidth	53 GB/s	224 GB/s

Sample POWER-based Technology: NVLink

Nvidia GPU with NVLink



- Up to 4 integrated Nvidia P100 “Pascal” GPUs
- Delivering > 2.5X the bandwidth to GPUs
- Reduces common CPU-GPU bottlenecks

Open Source Genomics Applications– Optimized with POWER8

- ALLPATHS-LG
- BarraCUDA
- bamtools
- Bedtools
- Bfast
- Bioconductor
- BioPerl
- BioPython
- BLAST (NCBI)
- Bowtie
- Bowtie2
- BreakDancer
- BWA
- Chimerascan
- Conda
- ClustalW
- Cufflinks
- DELLY2
- EMBOSS

- FASTA
- FastQC
- FASTX-Toolkit
- FreeBayes
- GenomicConsensus
- GenomeFisher
- GraphViz
- HMMER
- HTSeq
- Htslib
- IGV
- InterProScan
- ISAAC3
- iRODS
- Mothur
- MrBayes
- MrBayes5d
- MUSCLE
- Numpy

- Pandas
- PHYLIP
- PICARD
- Pindel
- PLINK
- PRADA
- Pysam
- Python
- R
- RNAStar/STAR
- RSEM
- SAMTools
- Sailfish
- Scalpel
- SHRiMP
- SIFT
- Snpeff
- SOAP3-DP
- SOAPaligner

- SOAPdenovo
- SoapFuse
- SQLite
- Sratoolkit
- STAR-fusion
- Tabix
- Tablet
- Tassel
- T-Coffee
- TMAP
- TopHat
- TranSMART
- Trinity
- Variant_tools
- Varscan
- Velvet/Oases

- bamkit
- bedops
- cutadapt
- diamond
- kraken
- lumpy
- parallel
- PLINK2
- primer3
- QIIME
- R cowplot
- R tidyverse

- Salmon
- Samblaster
- Scikit-bio
- Seqtk
- Spades
- Trimmonmatic
- Vcftools

Bio Builds http://biobuilds.org/

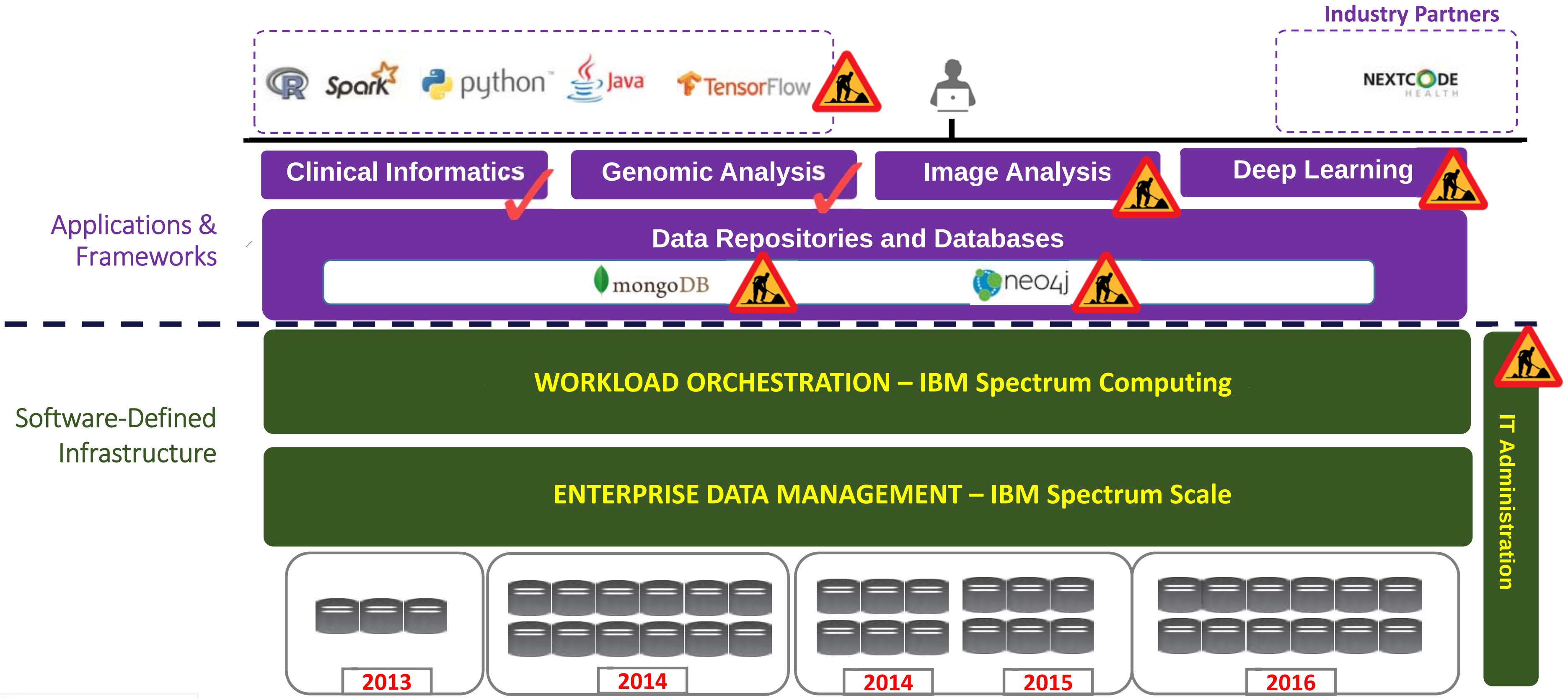
•**Turn-key:** Pre-built binaries and complete build scripts

•**Optimized:** POWER8 binaries

•**Long Term Support:** Community sponsorship and support contracts ensure ongoing support for tools

Qatar Genome Project

IBM



Industry Partners

NEXTCODE HEALTH

R Spark python Java TensorFlow

Clinical Informatics ✓

Genomic Analysis ✓

Image Analysis

Deep Learning

Applications & Frameworks

Data Repositories and Databases

mongoDB neo4j

Software-Defined Infrastructure

WORKLOAD ORCHESTRATION – IBM Spectrum Computing

ENTERPRISE DATA MANAGEMENT – IBM Spectrum Scale

IT Administration

2013

2014

2014

2015

2016

قطر جينوم
QATAR GENOME