Data management challenges in today's Healthcare and Life Sciences ecosystems

Jose L. Alvarez
Principal Engineer, WW Director Life Sciences
jose.alvarez@seagate.com
Evolution of Data Sets in Healthcare – RIS Example (1995!)

- Specialized Medicine data silos
- Multiple modalities yield separate DICOM images repositories
- Minimal Data distribution
- HL7 transfer for Demographics from HIS System
- No single data repository - VNA
Data Aggregation from multiple clinical disciplines

- Heavy HL7 interfaces to pass data between systems.
- Separate metadata sources that eventually yield an aggregate data set.
- Beginnings of Clinical portals with some decision support.
There are integration engines but data management issues are still present...

- Data integrity across all sources
- Chain of trust
- Security
- Data encryption on transit and at rest
- Data durability and longevity
The New paradigm of the connected individual

- Distributed data sources
- Always on mode – instant gratification
- Challenge for data storage and curating
- Effective Biometrics for meaningful patient history
- Chain of trust
- Security
- Long term data preservation
Why Genomics data sets are important and disruptive

$4.27b by 2017!
Seagate Cloud Systems and Solutions Portfolio

**HPC**
- Engineered to optimize capacity and performance
- 40% fewer racks require
- 1TB/s + file system performance

**Scale-Out Systems**
- Engineered solutions for object storage
- Validated architectures for open source and software-defined storage
- Private cloud appliances for backup and recovery
- Modular, scalable components for DIY customers

**OEM**
- 2+ million enclosures
- 17+ Exabytes shipped
- Drive Variety (HDD, SAS, SATA, SSD, hybrid)
- Enclosures, controllers
- Customer-driven partnership
- Services: Logistics, fulfillment, warranty, design, supply chain

**Cloud Services**
- Backup as a service
- Disaster recovery as a Service
- Archive as a service
- Endpoint backup
- Managed services
Our Technology Investment Roadmap

- Lowest Cost/ TB Storage
- Increasing Total Performance
- Flash Controllers
- Lowering System Power and Cooling Costs
- Lowering System Volume
- Self-Healing Components and Systems
- Lowest Total System Lifecycle Cost (Deployment, Operation, Disposal)
Seagate® brings an open approach to Intelligent Information Infrastructure™ accelerating bioinformatics pipelines and helping manage next-generation workloads—with scale, performance, security and cost aligned to today's challenges.

Our multi-tiered data storage solutions enable high-throughput, scalable geo-distributed storage, while meeting the complex compliance and data management challenges of high performance computing in bioinformatics.
Science is not completed in totally separate stages anymore.
But is rather a flowing cycle.
Work and reorganization is done between physical media.
Data on physical media is analyzed, reduced, and manipulated.
Intelligent Information Infrastructure - Powered by Seagate

Delivering with PARTNERS’ components through complete systems and into the cloud
Building a Secure High Performance Data Fabric

Intelligent Information Infrastructure

Accelerating Customer Workflows
Acquire, Analyze, Store, Distribute and Archive.
### The Big Data Transition in Genomics

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective:</strong></td>
<td>Research</td>
<td>Clinical</td>
</tr>
<tr>
<td><strong>Time to sequence</strong></td>
<td>1 week</td>
<td>10 human genomes per day</td>
</tr>
<tr>
<td><strong>Cost to sequence</strong></td>
<td>$100K</td>
<td>$1000</td>
</tr>
<tr>
<td><strong>Applications:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSIX</td>
<td>Casava</td>
<td>Casava, GATK (Broad),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SOAPdenovo (BGI)</td>
</tr>
<tr>
<td>Analytics</td>
<td>None</td>
<td>SAP Hana – Drug discovery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hadoop, Others.</td>
</tr>
<tr>
<td>Cloud</td>
<td>None</td>
<td>Instrument to Illumina</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BaseSpace, Amazon, Google</td>
</tr>
<tr>
<td><strong>Leaders – Instruments</strong></td>
<td>Life Tech, illumina, 454</td>
<td>illumina, Life Tech, PacBio</td>
</tr>
<tr>
<td><strong>Leaders – Compute</strong></td>
<td>Dell, HP and IBM</td>
<td>BGI (1PF), Google (600K cores), Amazon</td>
</tr>
<tr>
<td><strong>Leaders – Storage</strong></td>
<td>CIFS, NFS, GPFS, Lustre</td>
<td>NFS, CIFS, GPFS, Lustre, Object Stores</td>
</tr>
</tbody>
</table>
Life Science Solutions : At Scale!

Until the arrival of petascale supercomputers, no one could piece together the entire HIV capsid – an assemblage of more than 1,300 identical proteins – in atomic-level detail. The simulations that added the missing pieces to the puzzle were conducted during testing of Blue Waters, a new supercomputer at the National Center for Supercomputing Applications at the University of Illinois.”

Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics
Nature 497, 643–646 (30 May 2013) doi:10.1038/nature12162
Received 02 November 2012 Accepted 05 April 2013 Published online 29 May 2013

Exponentially less cost, space, cooling and power – less is more!
The converged Infrastructure Solution

- HPC Data CS9000 (Scale out Parallel file system)
- Clinical
- HPC Compute and Analytics farm
- HSM policy engine (e.g. iRODS, TSM, DMF)
- Object Storage (Scality for Unstructured data CIFS, NFS and Archive)
- Object Servers (S3)
- Sequencers and other Lab Instruments
- Cloud Backup/DR and Long term Archive
The converged Infrastructure Solution

- HPC Data CS9000 (Scale out Parallel file system)
- Clinical
- HPC Compute and Analytics farm
- HSM policy engine (e.g. iRODS, TSM, DMF)
- Object Storage (Scality for Unstructured data CIFS, NFS and Archive)
- Object Servers (S3)
- Cloud Backup/DR and Long term Archive
- Sequencers and other Lab Instruments
Integration across Parallel File System and Grid Data services

**iRODS**

- **iRODS Resource**
  - **ClusterStor File System**
  - **Parallel Application**
  - **Scratch and Medium term Storage**

- **iRODS Resource**
  - **Seagate StrataStor**
  - **Virtual Machine Environment**
  - **VM’s**

- **iRODS Resource**
  - **Cleversafe, Scality, Swiftstack**
  - **Grid Services**
  - **Object Store**
Secure Access to Data on the Clinical Side

1) Clinician request for sequence reads on patient X
2) Patient id lookup to obtain subject id
3) Subject id lookup in iRODS
4) Data sets packaged in zip file and retrieved
5) Data unzipped and displayed within secure workspace
NGS Reference Implementation

- **Initialization**: iRODS will apply sample IDs and results (or links to results) of automated processing.
- **Sequencing**: iRODS will kick off each process in the pipeline, or launch a workflow engine for more complex tasks.
- **Formatting and Cleaning**: iRODS will automatically compile reports upon schedule or request.
- **Quality Control**: iRODS will stage files for processing, evaluation on a secure workspace, and archiving.
- **Standard Analytical Processing**: iRODS will search on metadata.
- **Querying**: iRODS will manage complex, dynamic user permissions across multiple workgroups.
- **Interpretation**: Additional Action (ex. Treatment)
- **Consultation**: Archive/Replication
ClusterStor HSM Partners – iRODS, Cray TAS and SGI DMF

- ClusterStor v2.0 or greater w/Lustre 2.5 is HSM Ready
  - Requires HSM application and support from a partner
- HSM partner options
  - SGI DMF
  - Cray TAS
  - iRODS
The converged Infrastructure Solution

HPC Data CS9000 (Scale out Parallel file system)

Clinical

HPC Compute and Analytics farm

HSM policy engine (e.g. iRODS, TSM, DMF)

Object Storage (Scality for Unstructured data CIFS, NFS and Archive)

Object Servers (S3)

Cloud Backup/DR and Long term Archive

Sequencers and other Lab Instruments

Seagate Confidential
New predictive data Analytics architectures
Hadoop Workflow Accelerator on ClusterStor

Overview

• Hadoop compute uses the ClusterStor storage system for primary input and output operations.

• MapReduce temporary files are saved using minimal local storage or can be configured to save to ClusterStor eliminating all local storage.

• The Hadoop Workflow Accelerator eliminates the reliance on direct attached storage, allowing for truly diskless compute nodes and independent scaling of compute and storage.

• Hadoop Workflow Accelerator optimizes Hadoop performance for you workloads and applications while improving TCO.
The converged Infrastructure Solution

- HPC Data CS9000 (Scale out Parallel file system)
- Clinical
- HPC Compute and Analytics farm
- HSM policy engine (e.g. iRODS, TSM, DMF)
- Object Storage (Scality for Unstructured data CIFS, NFS and Archive)
- Object Servers (S3)
- Cloud Backup/DR and Long term Archive
- Sequencers and other Lab Instruments
Software Defined CSS Validated Architectures

Scality is Validated & Supported By Seagate For Key Customer Use Cases

**Validated Architecture**
- Use Cases
- Tailored Hardware Configurations
- Shipped with SW installed
- Seagate options for Billing, etc
- Tested solutions
- Benchmarks
- End to end support
- Managed Services option
Software Defined CSS Validated Architecture Family

SAN & NAS
- Nexenta

Terabytes
- Basho

500 TBs
- Cleversafe

Object Storage
- OpenStack Swift

Petabytes
- Ceph
- Scality
Single Pane of Glass for Infrastructure and data Management

Fully Integrated Solution Visibility and Management
Low level diagnostics, embedded monitoring, proactive alerts

Easy to Manage

Real Time Monitoring
Thank you!