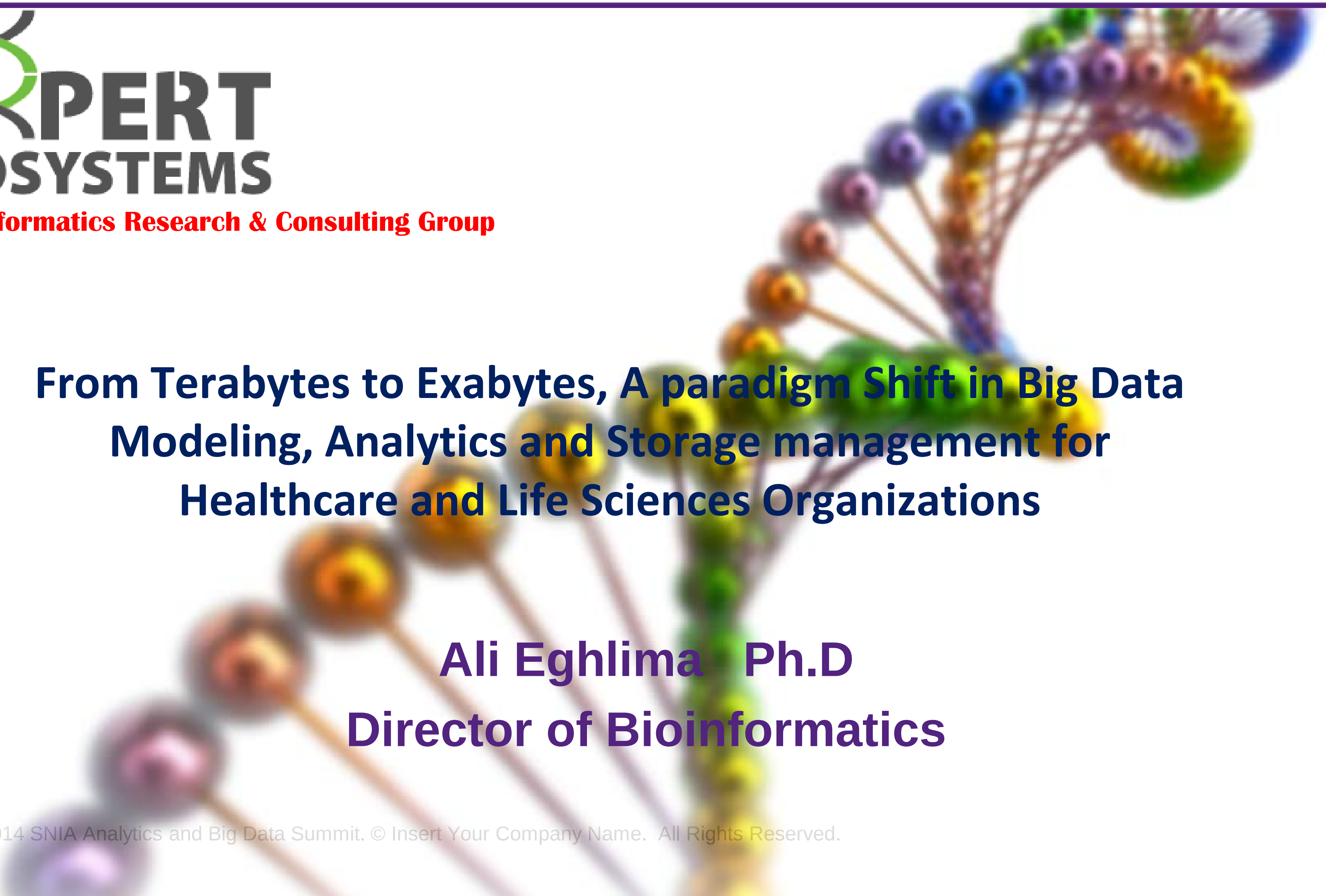
ANALYTICS BIG & DATA SUMMIT

EXPERT BIOSYSTEMS

A Bioinformatics Research & Consulting Group



# From Terabytes to Exabytes, A paradigm Shift in Big Data Modeling, Analytics and Storage management for Healthcare and Life Sciences Organizations

Ali Eghlima Ph.D
Director of Bioinformatics

# Before we start: About me

- Ali Eghlima
- Expert BioSystems, EVP, and Director of Bioinformatics
- Data Scientist, Software/System/Solution Architect
- Five Years as Sr. Principal Engineer at Raytheon
    - Leading R&D Projects in Enterprise Architecture, Cyber Security, “Huge” Big Data Analytics, Real-Time Distributed Big Data Collection and Analysis
- 20-years Career as Senior Consulting Engineer at DEC, Compaq, and HP
- Primary Technical Expertise –
    - Big Data Analytics, Real-time Distributed Computing, High Availability, Cyber Security, Cluster and Cloud Technology, High Performance Computing, Numerical Analysis
- Pioneer and Advocate in Cluster & Cloud Computing
- Ph.D  from RPI, MS and Engineering degrees from MIT

# Agenda

- Characteristics of Healthcare & life sciences data
- Review, Data integrity/Privacy/Cyber Security concerns of major healthcare/research Centers
- Review current technology, and common systems architecture used for Big Data Analytics in Health Sciences vs other industries.
- Issues, challenges and potential solutions for real-time and archived data storage managements
- Present scalable open source computing platform to manage Exabyte class datasets
- Concluding Remarks

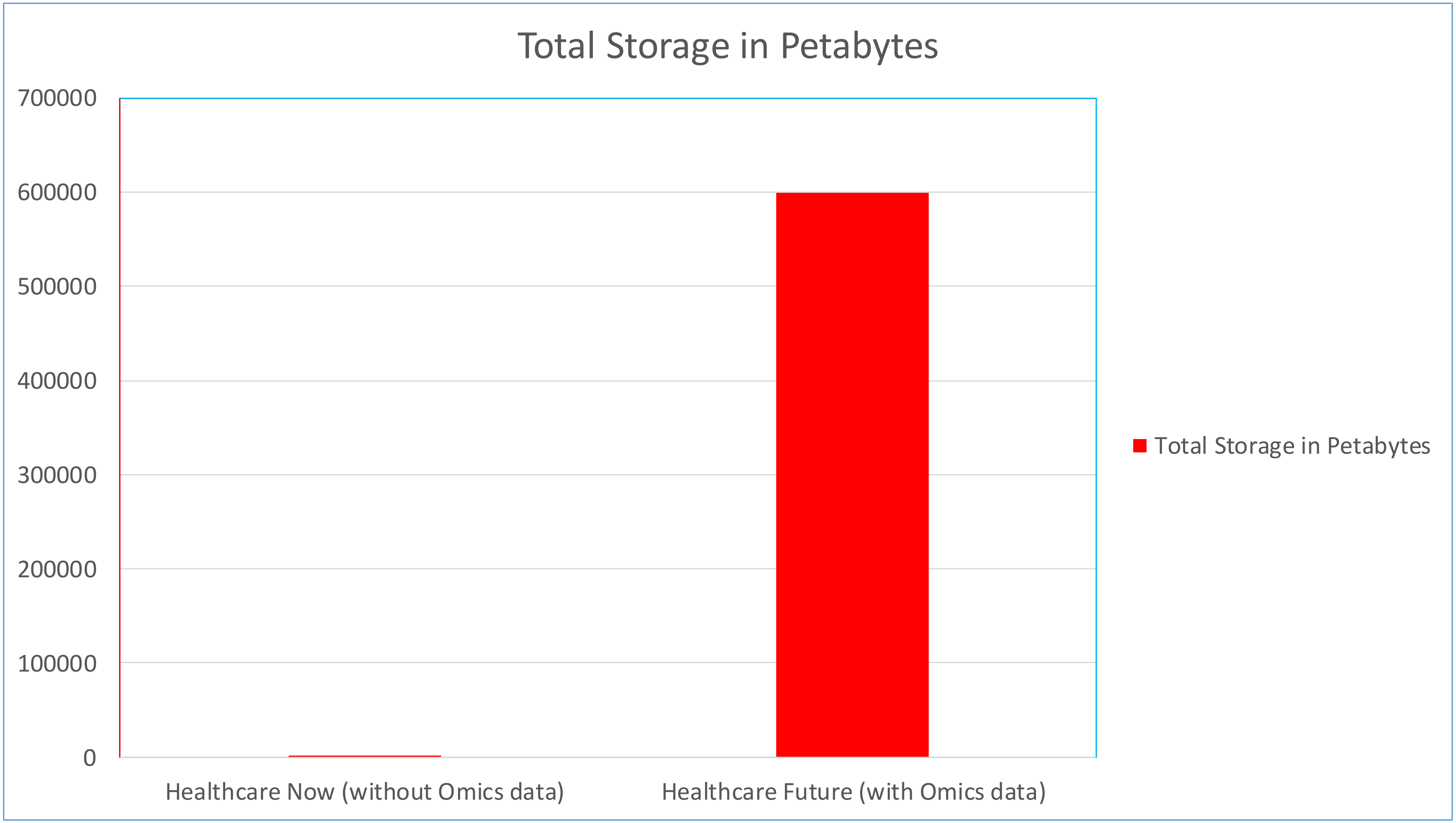
# Characteristics of Healthcare Data vs Other Data

- Almost permanent
- It is being owned by individual
- Data ownership after individual death is unknown ( offspring, siblings, other family members)

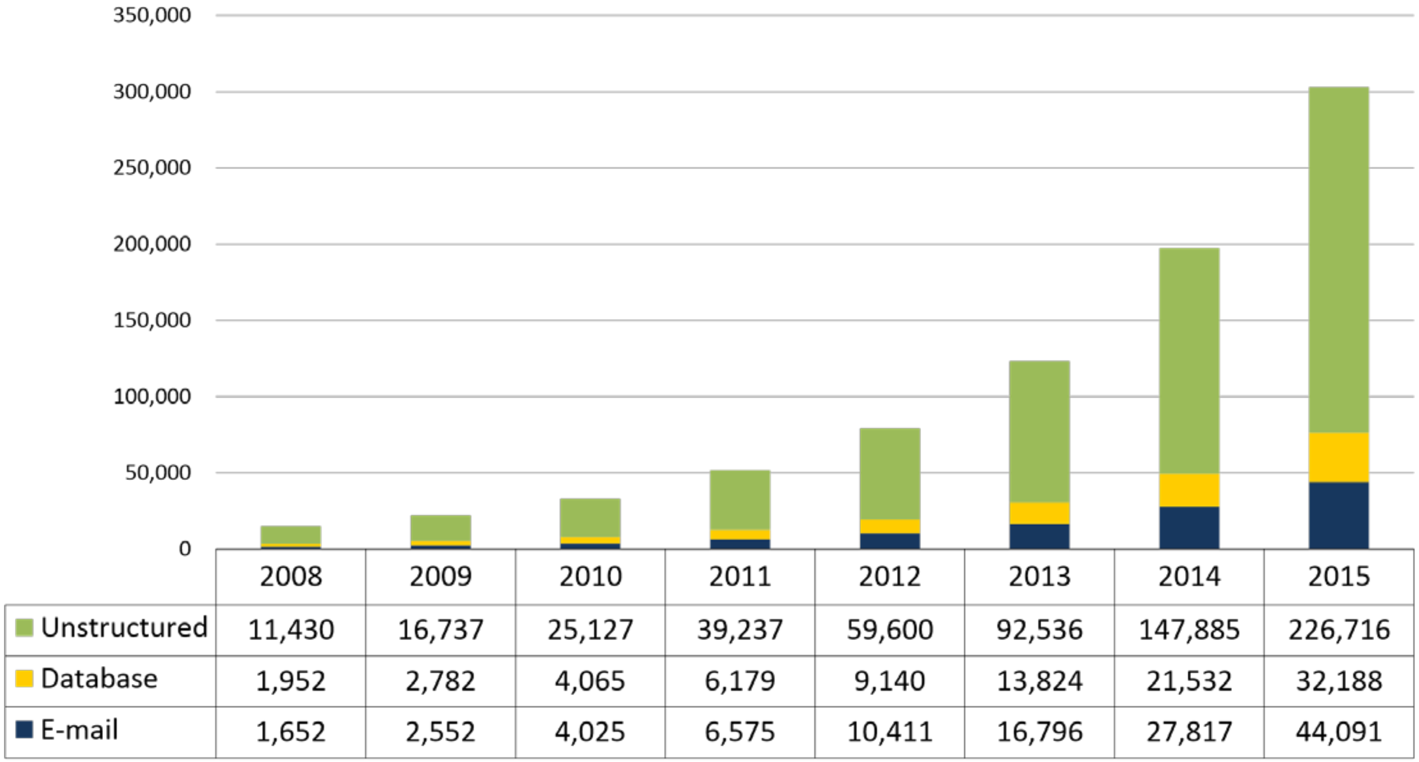
# Example: Storage/Dataset Size Health Sciences vs Other Industries

- Financial
  - Number of Accounts – From 10000 to 300 Millions
  - Storage per Account - ~Gigs or less
  - Total Storage – From ~Tens of Terabytes to ~300 Petabytes
- Healthcare
  - Number of Patients – From 10000 to 300 Millions
  - Storage per Patient – From ~Gigabytes Today to ~ Many Terabytes in future
  - Total Storage – From ~ 20 Petabytes to ~600 Exabyte

# Example: Storage/Dataset Size Healthcare: Now vs. Future



# Total Archived Capacity



| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|
| 🟩 Unstructured | 11,430 | 16,737 | 25,127 | 39,237 | 59,600 | 92,536 | 147,885 | 226,716 |
| 🟨 Database | 1,952 | 2,782 | 4,065 | 6,179 | 9,140 | 13,824 | 21,532 | 32,188 |
| 🟦 E-mail | 1,652 | 2,552 | 4,025 | 6,575 | 10,411 | 16,796 | 27,817 | 44,091 |

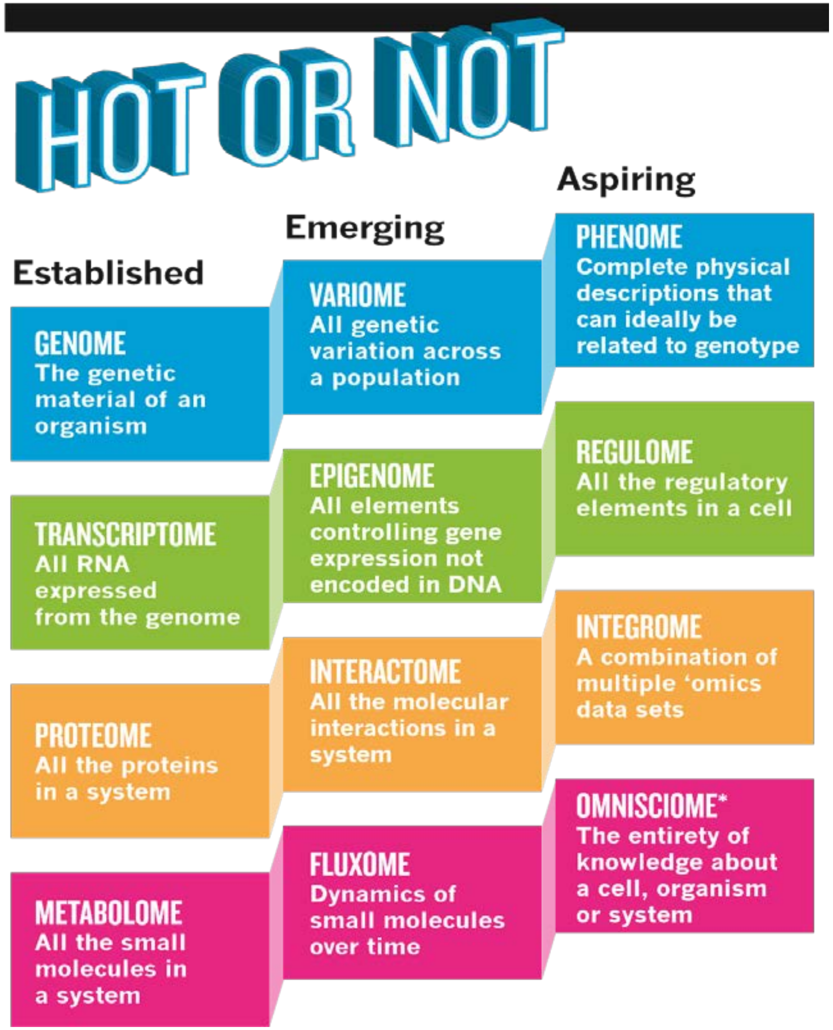
*Source: Enterprise Strategy Group, 2010.*

**Total Archived Capacity, by Content Type, Worldwide, 2008-2015 (Petabytes)**

# It is not just genomic data

**Nature – 494**
**February 2013**

**Big biology: The ’omes puzzle**
Where once there was the genome, now there are thousands of ’omes. *Nature* goes in search of the ones that matter.



HOT OR NOT

Established

GENOME
The genetic material of an organism

TRANSCRIPTOME
All RNA expressed from the genome

PROTEOME
All the proteins in a system

METABOLOME
All the small molecules in a system

Emerging

VARIOME
All genetic variation across a population

EPIGENOME
All elements controlling gene expression not encoded in DNA

INTERACTOME
All the molecular interactions in a system

FLUXOME
Dynamics of small molecules over time

Aspiring

PHENOME
Complete physical descriptions that can ideally be related to genotype

REGULOME
All the regulatory elements in a cell

INTEGROME
A combination of multiple ‘omics data sets

OMNISCIOME*
The entirety of knowledge about a cell, organism or system

# Data integrity/Privacy/Cyber Security concerns of major healthcare/research Centers

# Theft of Healthcare Identity Data Consequences

- Medical services, devices and prescription drugs
- Physician information to create fake prescriptions and then resell the medicine online.
- File false claims to insurance companies and government agencies

# Theft of Healthcare Identity Data Value

- Credit Card info $1
- Personal Identification Information (PII) for $10-$12
- Patient Records for $50

Source:
1 - Medical Identity Fraud Alliance, “The Growing Threat of Medical Identity Fraud: A Call To Action,” July 2013, accessed at http://medidfraud.org/wp-content/uploads/2013/07/MIFA-Growing-Threat-07232013.pdf.
2 - David Carr, “Healthcare Data Breaches to Surge in 2014,” InformationWeek Healthcare, Dec. 26, 2013, accessed at http://www.informationweek.com/healthcare/policy-and-regulation/healthcaredata-breaches-to-surge-in-2014/d/d-id/1113259.

# Theft of Healthcare Identity Data is Growing

- 2010 – 1.42 Million
- 2011 – 1.49 Million
- 2012 – 1.85 Million

Source:
**Ponemon Institute**, “Fourth Annual Benchmark Study on Patient Privacy and Data Security,” March 2014, accessed at http://lpa.idexpertscorp.com/acton/attachment/6200/f-012c/1/-/-/-/-/ID%20Experts%204th%20Annual%20Patient%20Privacy%20%26%20Data%20Security%20Report%20FINAL%20%281%29.pdf

# Healthcare Data Security Threat

## (reported by healthcare provider)

- ❒ **Employee negligence**
- ❒ **Unsecured mobile devices**
- ❒ **Security gaps with business associates**
- ❒ **Evolving criminal threats**
- ❒ **New vulnerabilities under the Affordable Care Act**

Survey participants had strong reservations about the security of Health Information Exchanges (HIEs): **A third** said they don’t plan to participate in HIEs because they are not confident enough in the security and privacy of patient data shared on the exchanges

http://www2.idexpertscorp.com/ponemon-report-on-patient-privacy-data-security-incidents/

# Technology, and Common Systems Architecture used for Big Data Analytics in Health Sciences vs other industries

# Cloud Computing ?

- Private
- Public
- Community

# Private cloud ?

## From Webopedia

- *Private cloud* is the phrase used to describe a cloud computing platform that is implemented within the corporate firewall, under the control of the IT department.
- A private cloud is designed to offer the same features and benefits of public cloud systems, but removes a number of objections to the cloud computing model including control over enterprise and customer data, worries about security, and issues connected to regulatory compliance.

# Cloud or Public cloud

- **Network Cloud**

  In telecommunications, a cloud refers to a public or semi-public space on transmission lines (such as T1 or T3) that exists between the end points of a transmission

- **Cloud Computing**

  Cloud computing is a type of computing that relies on *sharing computing resources* rather than having local servers

  - **Consumer -** **S**oftware **as a S**ervice (**SaaS**)
  - **Developers and Architects** – **P**latform **as a** Service (**PaaS**)
  - **IT Pros and system administrators** - **I**nfrastructure **as a S**ervice (**IaaS**)

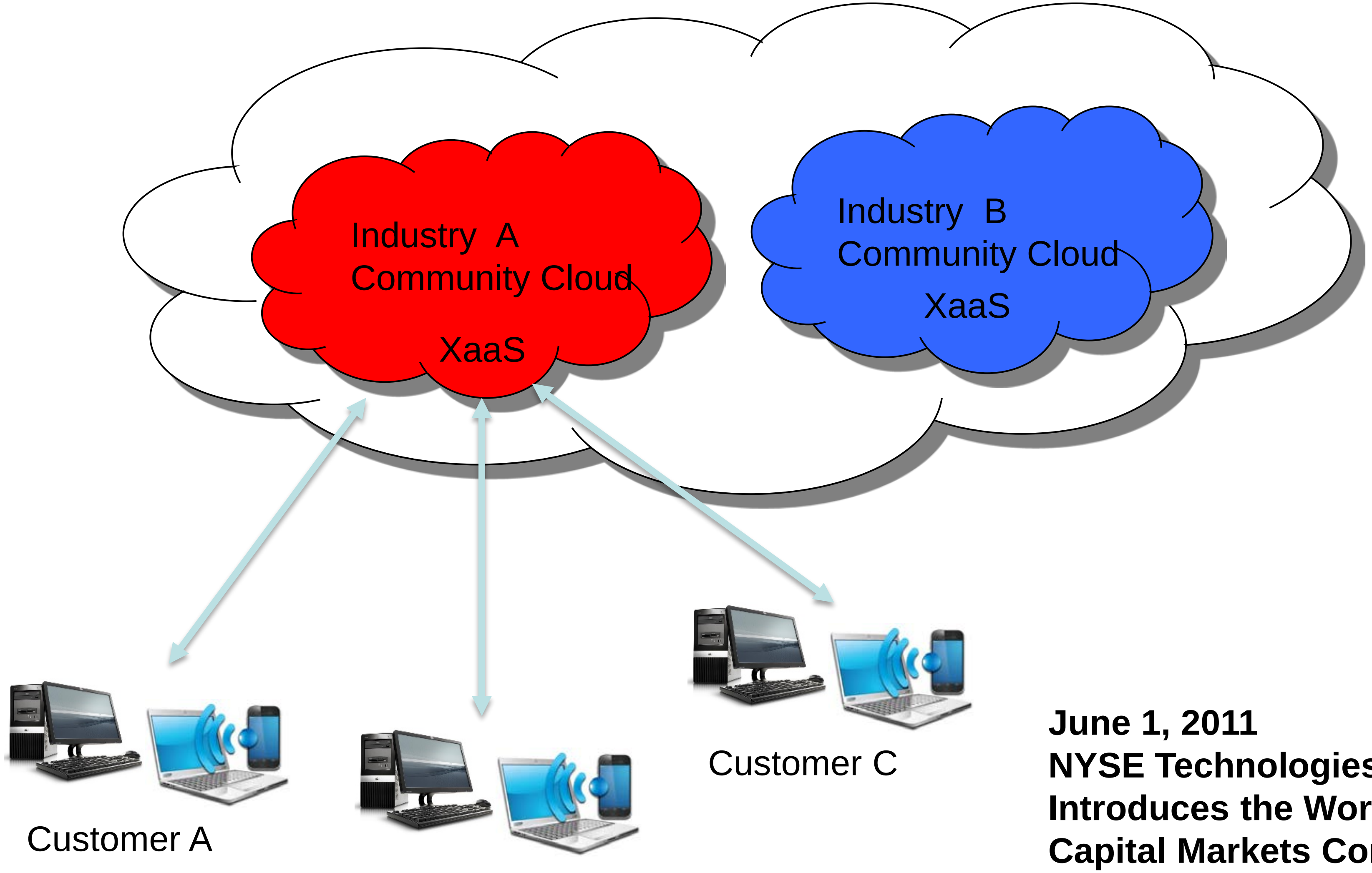
# Community Cloud ?

- Centralized
- Distributed

# Centralized Community Cloud ?

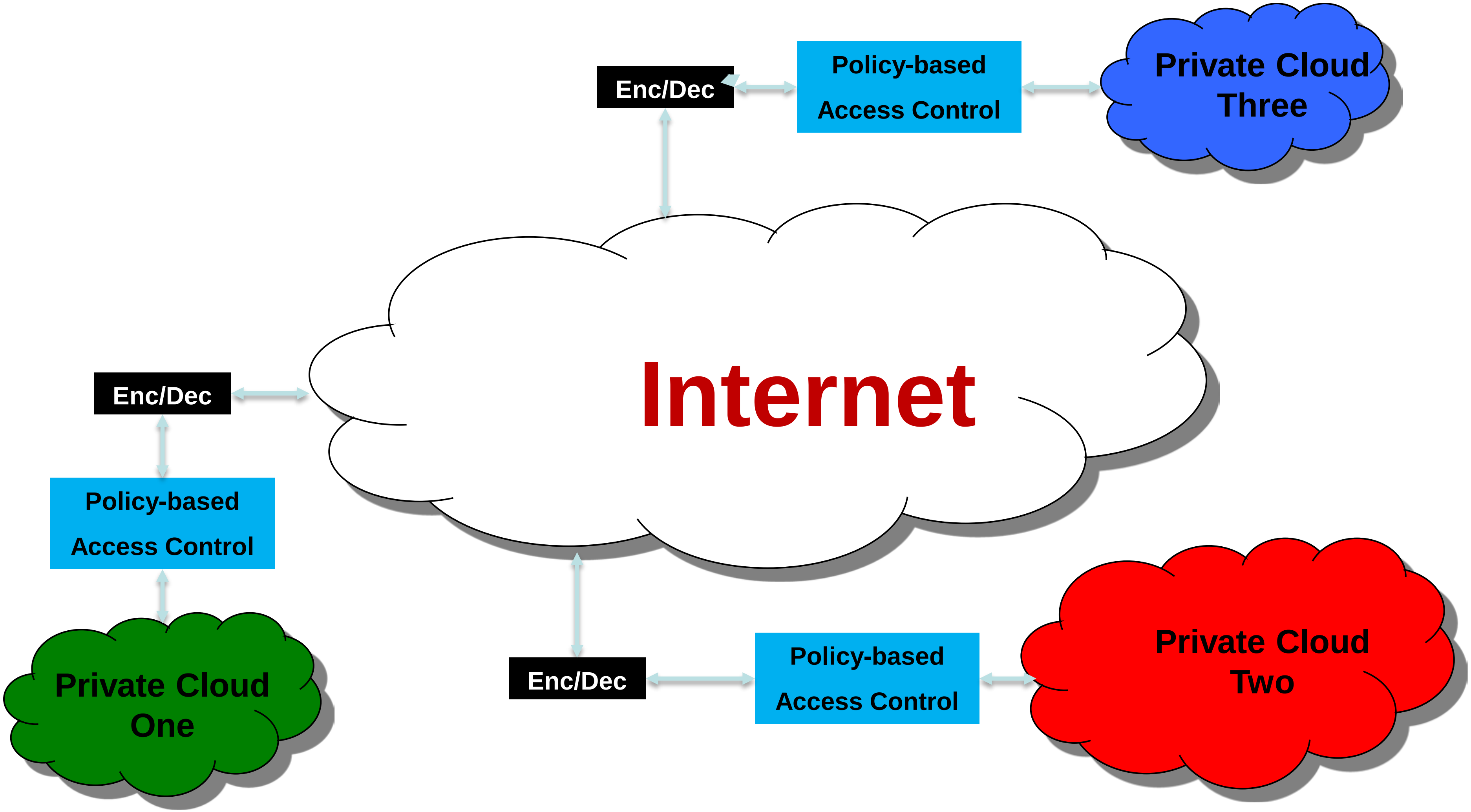
- Multi-Tenant Infrastructure
- Shared Among Several Organizations with Common Computing Concerns/Requirements
- Higher Level of Security, Privacy, and Performance (Compare to Public Cloud)
- Pay-as-you-go Billing Structure
- Cost, less than Private more than Public

# Examples: Centralized Community Cloud



**June 1, 2011**
**NYSE Technologies Introduces the World's First Capital Markets Community Platform**

# “Secure/Trusted” Distributed Community Cloud



# Issues, challenges and potential solutions for real-time and archived data storage managements

# Archiving, Tape Technology

## Ultirum LTO:

- ❒ Capacity per Tape – 6.25 Terabytes
- ❒ Cost (tape) – 1.3 cent per GB
- ❒ 250 million LTO tapes have been shipped
- ❒ Total, shipped Capacity ~ 100 Exabyte's

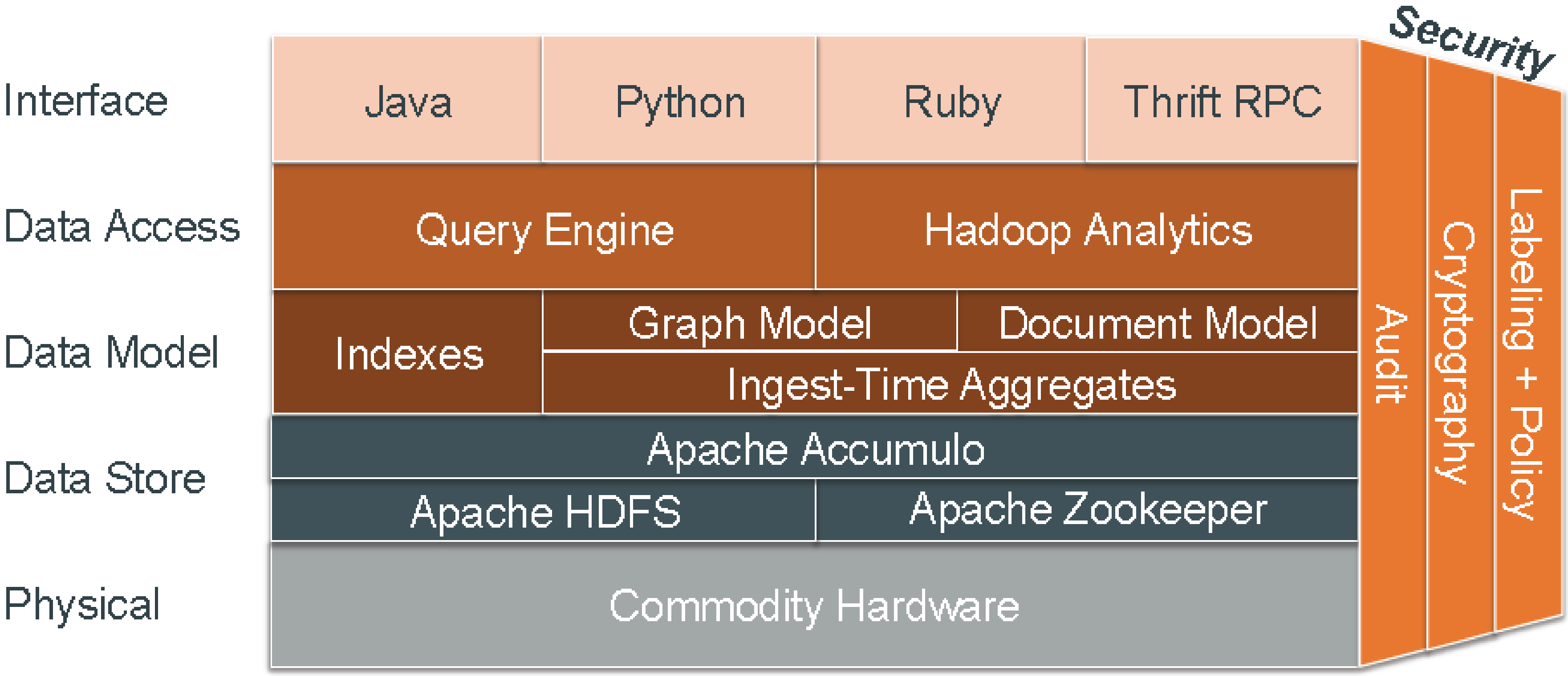
## Sony's new magnetic tape technology:

- ❑ Capacity - 185 TB per cartridges
- ❑ Announced at the INTERMAG Europe 2014

# Scalable Open Source Computing Platform to manage Exabyte class datasets

- Linux
- Hadoop
- MapReduce
- R
- Accumulo

# Technology Stack



Source: SQRRL Enterprise 2014

# Summary

- **Adding Omics (**Genomic...) **Data to the Patient EHR**

**Storage requirements, and associated computing power and network infrastructure performance will increase by at least three order of magnitude, just to keep up with today computing systems performance**

**Total Patient EHR, Data Storage ~ Zettabyte**

EXPERT BIOSYSTEMS
A Bioinformatics Research & Consulting Group

# Concluding Remarks