Hadoop and data integration with System z

Dr. Cameron Seay, Ph.D — North Carolina Agricultural and Technical State University

Mike Combs — Veristorm

March 2, 2015, Session 16423

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides **education**, professional **networking** and industry **influence**.

SHARE in Seattle 2015

The Big Picture for Big Data

The Big Picture for Big Data

"The Lack of Information" Problem

1 in 3 — Business leaders frequently make decisions based on information they don't trust, or don't have

1 in 2 — Business leaders say they don't have access to the information they require to do their jobs

83% — Of CIOs cited "Business Intelligence and Analytics" as part of their visionary plans to enhance competitiveness

60% — Of CIOs need to do a better job capturing and understanding information rapidly in order to make swift business decisions

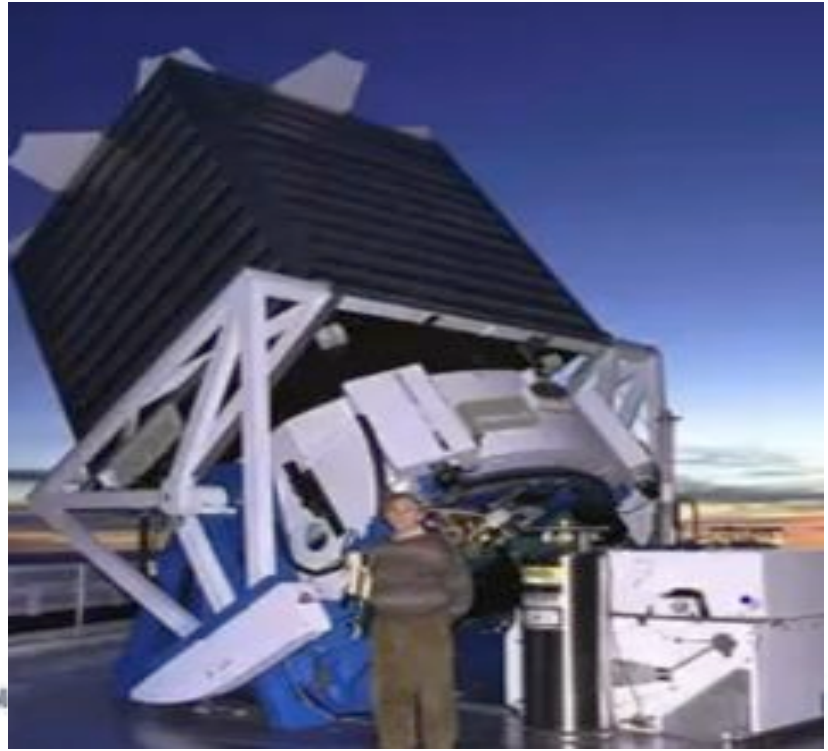
"The Surplus of Data" Problem

- "The 3 V's" of Big Data
  - Volume: More devices, higher resolution, more frequent collection, store everything
  - Variety: Incompatible Data Formats
  - Velocity: Analytics Fast Enough to be Interactive and Useful

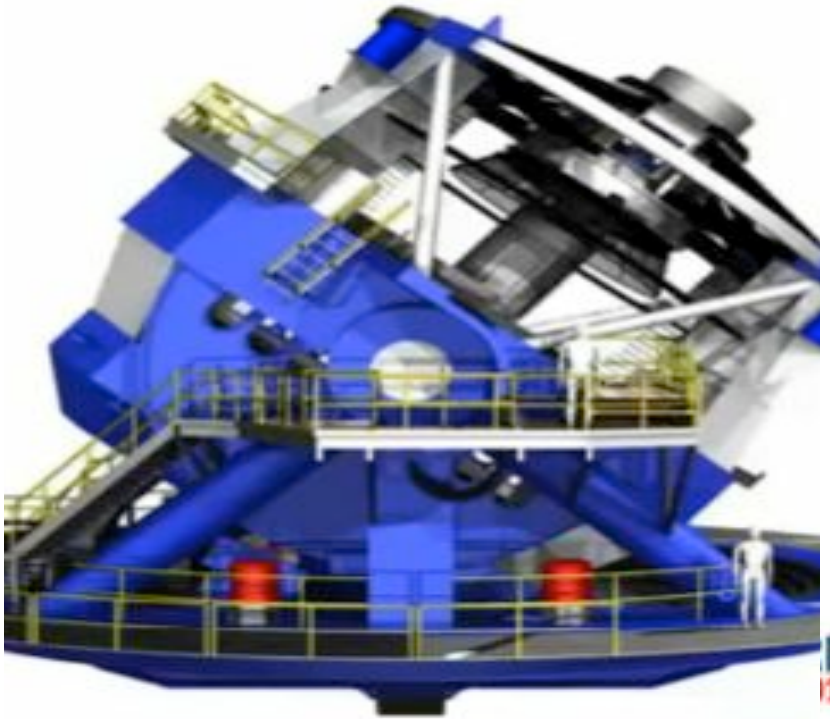
(Doug Laney, Gartner Research)

Big Data: Volume

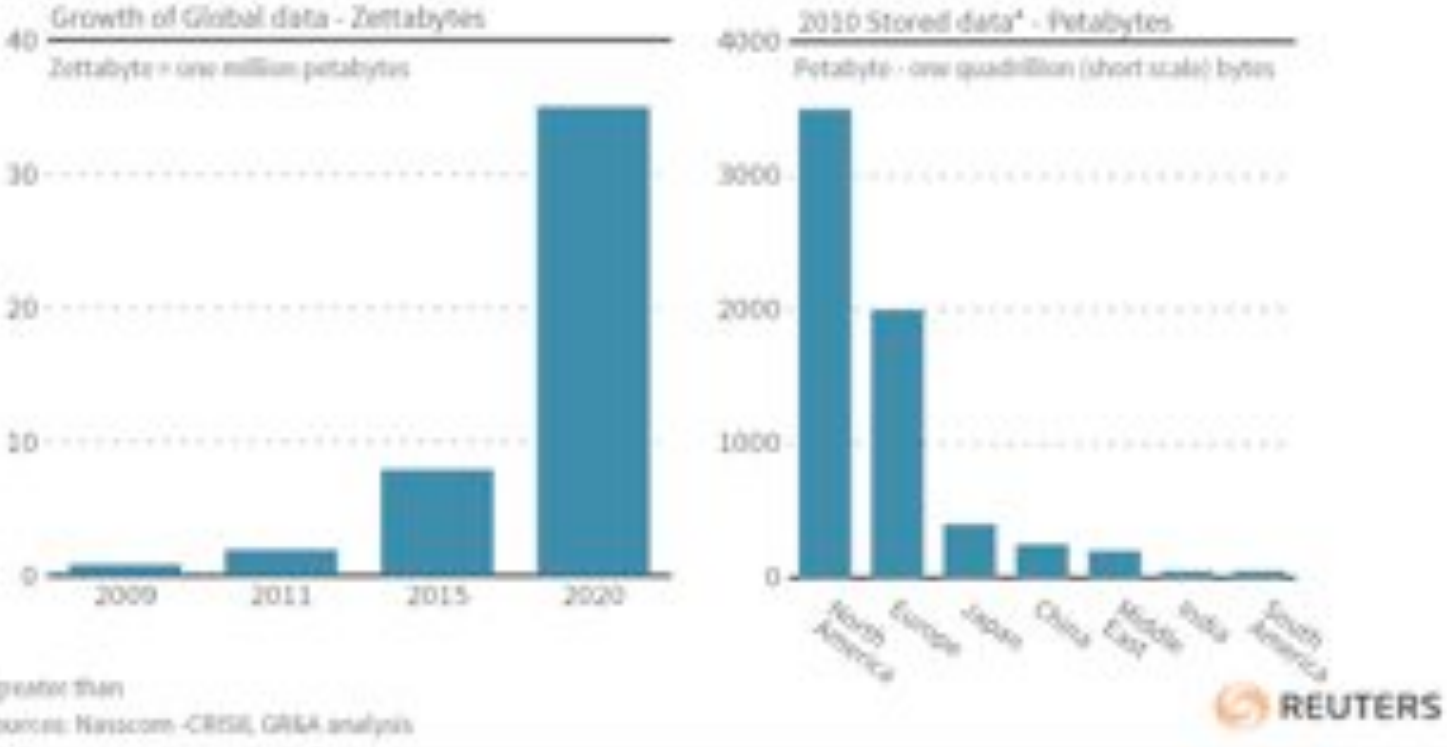
SDDS telescope, 80 TB in 7 years



LSST telescope, 80 TB in 2 days



Big Market, Big Growth



Big Data: Variety

20% is “Structured”

- Tabular Databases like credit card transactions and Excel spreadsheets
- Web forms

80% is “Unstructured”

- Pictures: Photos, X-rays, ultrasound scans
- Sound: Music (genre etc.), speech
- Videos: computer vision, cell growing cultures, storm movement
- Text: Emails, doctor’s notes
- Microsoft Office: Word, PowerPoint, PDF

Big Data: Velocity

- To be relevant, data analytics must timely
- Results can lead to new questions; solutions should be interactive
- Information should be searchable

Increasing needs for Detailed Analytics

- Baselining & Experimenting
  - Parkland Hospital analyzed records to find and extend best practices
- Segmentation
  - Dannon uses predictive analytics to adapt to changing tastes in yogurt
- Data Sharing
  - US Gov Fraud Prevention shared data across departments
- Decision-making
  - Lake George ecosystem project uses sensor data to protect $1B in tourism
- New Business Models
  - Social media, location-based services, mobile apps

Big Data Industry Value

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

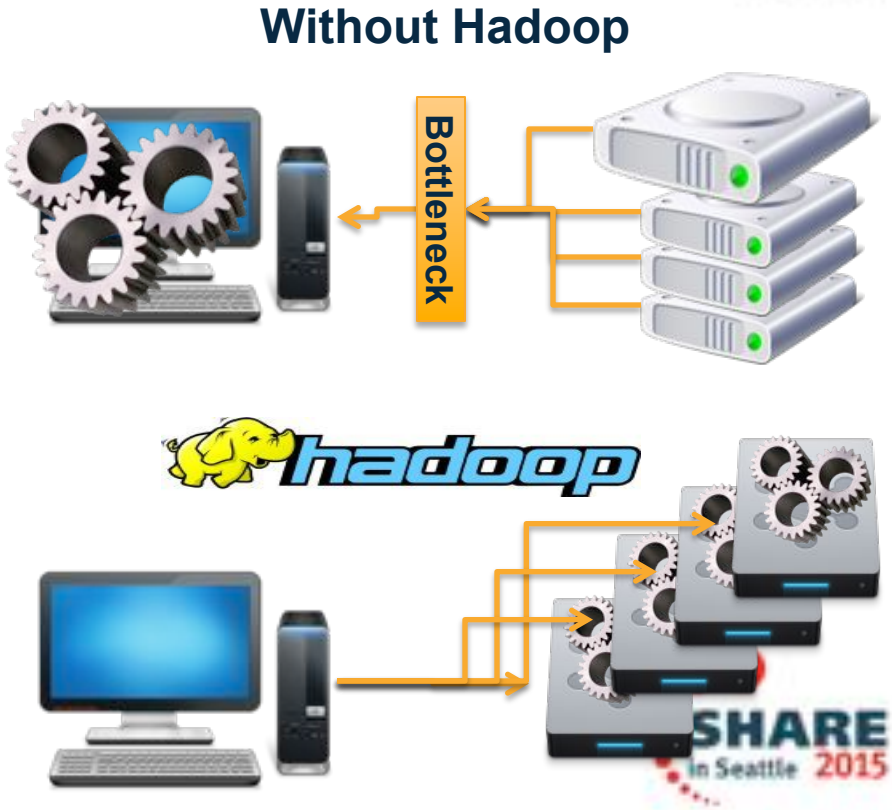
Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

Finance and Insurance
- **~1.5 to 2.5 percent annual productivity growth**
- **$828 billion industry**

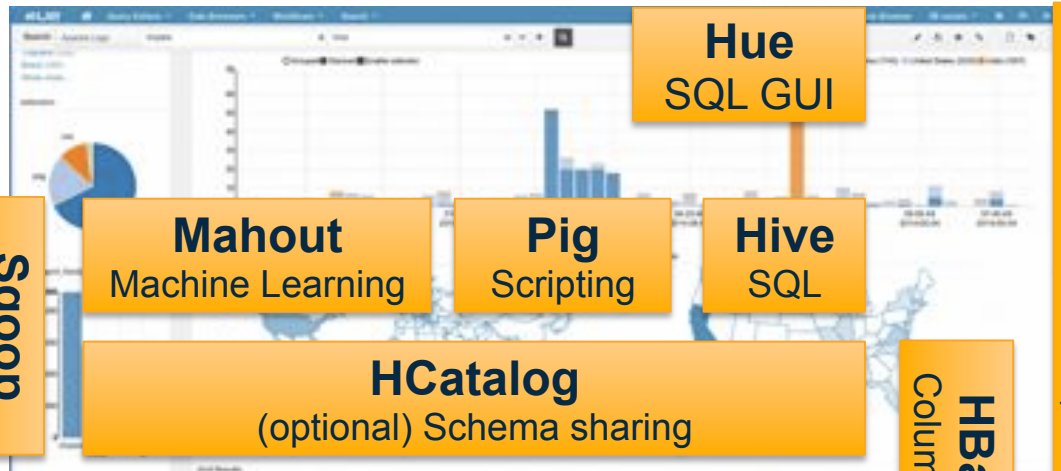
SOURCE: McKinsey Global Institute analysis

What is Hadoop and Why is it a Game Changer?

- Hadoop solves the problem of moving big data
  - Eliminates **interface** traffic jams
  - Eliminates **network** traffic jams
  - New way to move Data
- Hadoop automatically divides the work
  - Hadoop software divides the job across many computers, making them more productive



Hadoop Projects & Ecosystem



Hue
SQL GUI

Mahout
Machine Learning

Pig
Scripting

Hive
SQL

Sqoop
DB import

HCatalog
(optional) Schema sharing

HBase
Column DB

Flume
Log import

MapReduce/Tez on YARN
Distributed processing framework

HDFS
Distributed, redundant file system

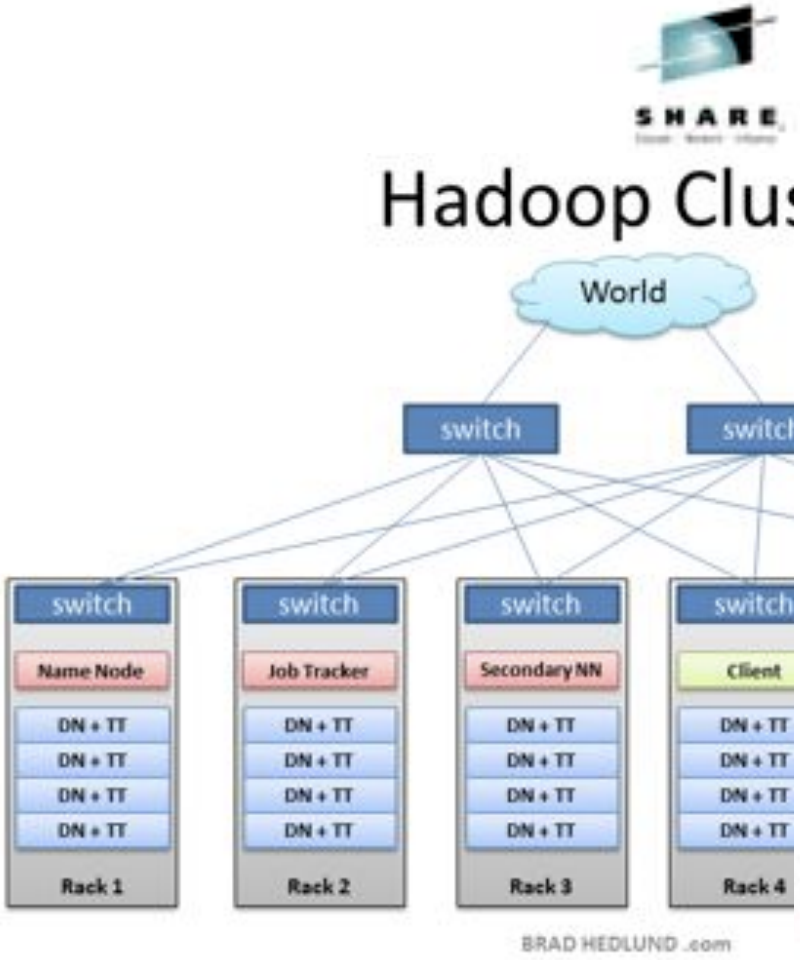
Oozie
Orchestration, workflow

ZooKeeper
Configuration, coordination

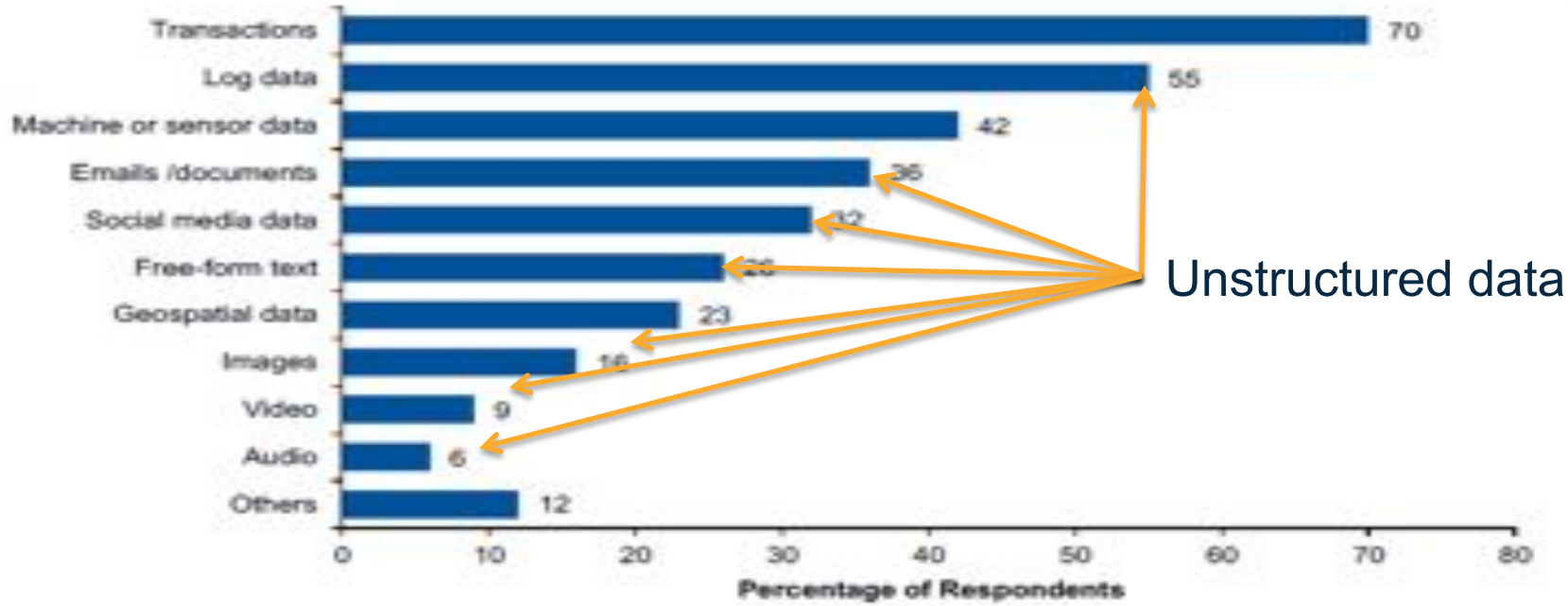
Ambari
Deploy, configure, monitor

Typical Hadoop Cluster

- NameNode
  - Files, metadata in RAM, logs changes
- Secondary NameNode
  - Merges changes. Not a backup!
- JobTracker
  - Assigns nodes to jobs, handles failures
- Per DataNode
  - DataNode— Files and backup; slave to NameNode
  - TaskTracker— Manages tasks on slave; slave to JobTracker



Today's Big Data Initiatives: Transactions, Logs, Machine Data



N =465 (multiple responses allowed)

Source: Gartner (September 2013)

IT leads Big Data usage.. but departments are catching up...

"What groups or departments are currently using Big Data/planning to use Big Data?"

Operations 54% (e.g. supply-demand)	Marketing 47% (e.g. campaigns)	IT Analytics 47% (e.g. network secure)	Product Dev. 22% (e.g. social feedback)
Finance 46% (e.g. risk exposure)	Sales 37% (e.g. cross/upsell)	Research 30% (e.g. simulation)	Logistic & Distr. 18% (e.g. route opt.)
Customer Service 26% (e.g. segmentation)	Manufacturing 18% (e.g. process opt)	GRC 15% (e.g. auditing)	Human Resources 11% (e.g. head hunting)
Procurement 15% (e.g. best buy)	Supply Chain 15% (e.g. sourcing)	Other 7%	Don't know 3%

Source: Forrsights BI/Big Data Survey

Base: 176 big data users and planners

Transaction Data = Mainframe Computers

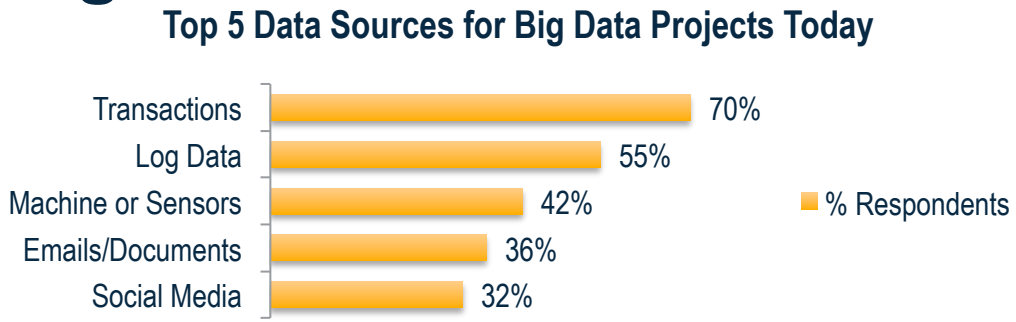
- Mainframes run the global core operations of
  - 92 of top 100 banks
  - 21 of top 25 insurance
  - 23 of top 25 retailers
- Process 60% of **all** transactions
- Mainframe computers are the place for essential enterprise data
  - Highly reliable
  - Highly secure

- IBM's Academic Initiative
  - 1000 higher education institutions
  - In 67 nations
  - Impacting 59,000 students
- However, mainframe data uses proprietary databases which must be translated to talk to formats familiar in "Big Data"

The ROI Behind the Hype

The most relevant insights come from enriching **your** primary enterprise data

Integration Problem For Data Scientists



"Survey Analysis - Big Data Adoption in 2013 Shows Substance Behind the Hype", Gartner, 2013, Link

"By most accounts **80%** *of the dev effort in a big data project goes into data integration*

…and only **20%** *goes towards data analysis."*

"Extract, Transform, and Load Big Data With Apache Hadoop", Intel, 2013, Link

- Enterprise data analysts require near-realtime mainframe data
- Mainframe users wish to off-load batch processes to Hadoop for cost savings

Obstacles to Include Mainframe Data

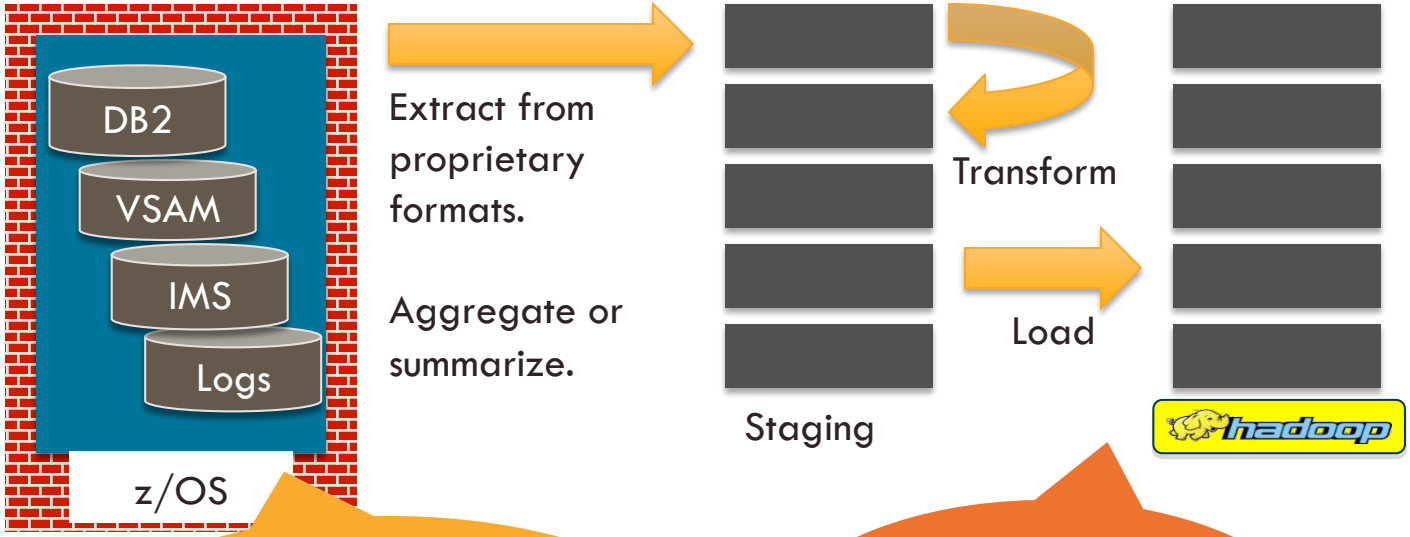
1/ Data Governance as the data moves off z/OS operational systems

2/ Data Ingestion from z/OS into Hadoop (on or off platform) is a bottleneck (MIPS & ETL cost, Security around data access and transfer, Process Agility)

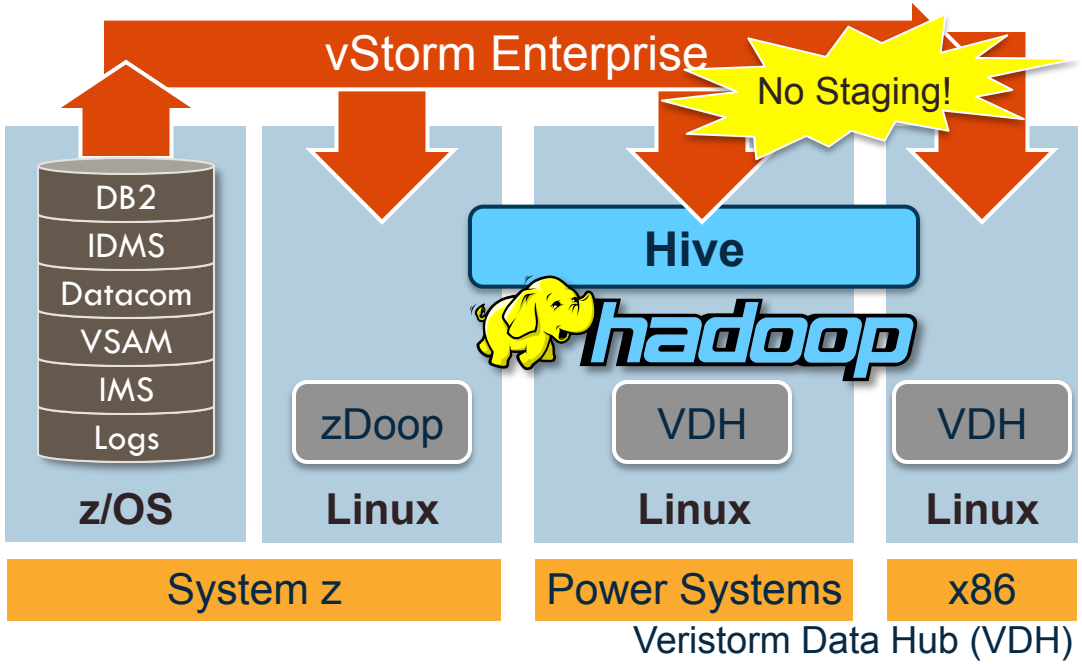
Lead to key requirements:

- Existing security policies must be applied to data access and transfer.
- There needs to be high speed / optimized connectors between traditional z/OS LPARs and the Hadoop clusters
- Ability to serve data transparently into Hadoop clusters on mainframe AND on distributed platform

Data Ingestion Challenges

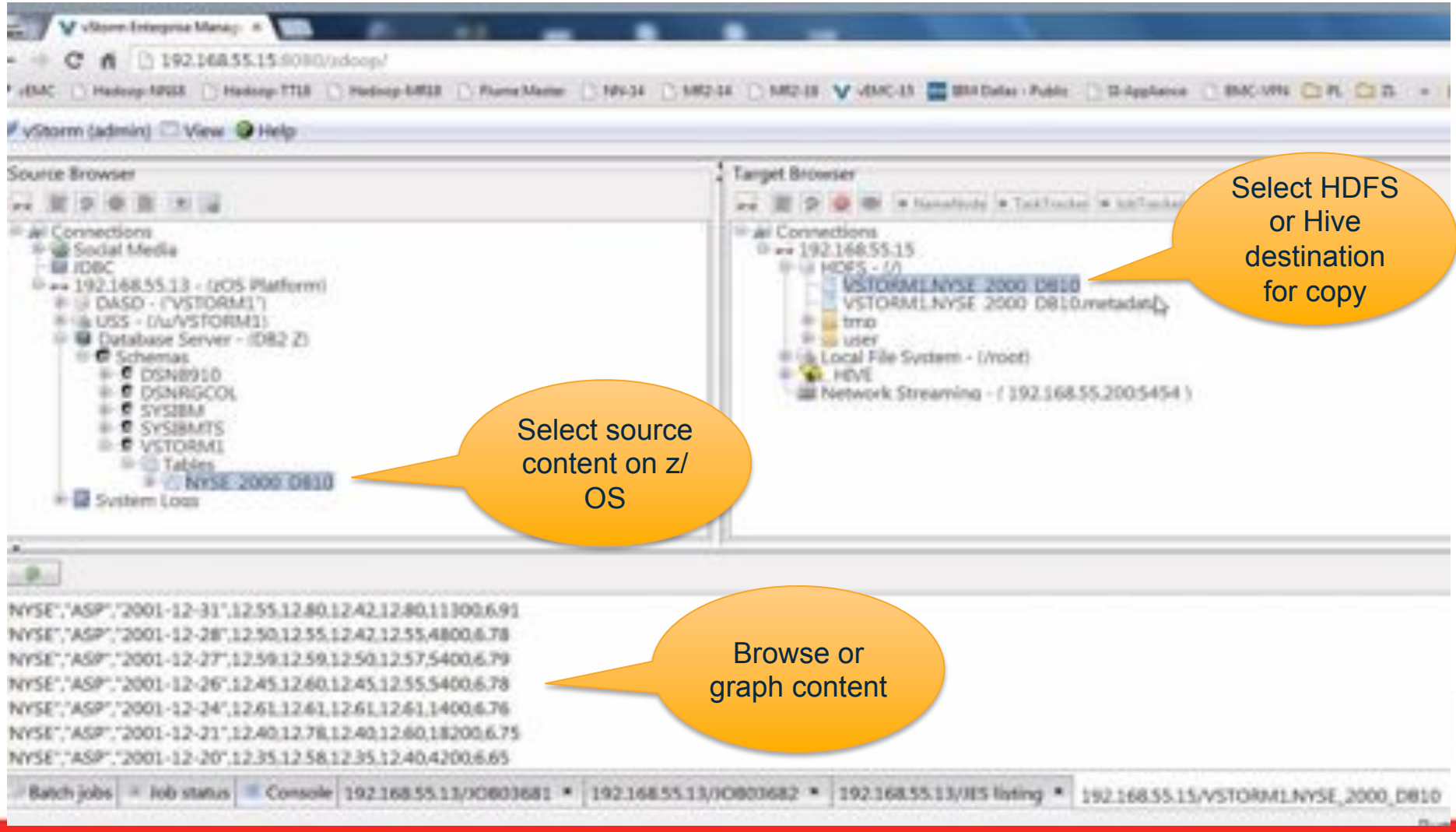


vStorm Enterprise – Mainframe data to Mainstream



Veristorm Data Hub (VDH)

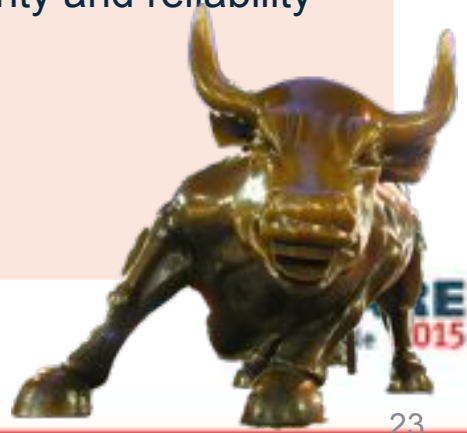
- IBM BigInsights
- Cloudera
- Hortonworks
- MapR


Source Browser
Target Browser
Select HDFS or Hive destination for copy
Select source content on z/OS
Browse or graph content

Financial Services Use Case

Problem	Solution	Benefits

• High cost of searchable archive on mainframe
  • $350K+ storage costs (for 40TB)
  • MIPS charges for operation
• $1.6M+ development costs due to many file types, including VSAM
• 2000+ man-days effort and project delay
	• Move data to Hadoop for analytics and archive
	• Shift from z/OS to IBM Linux (processors) on z to reduce MIPS
	• Use IBM SSD storage
	• Use IBM private cloud softlayer
	• Tap talent pool of Hadoop ecosystem
		• Reduction in storage costs
		• Dev costs almost eliminated
		• Quick benefits and ROI
		• New analytics options for unstructured data
		• Retains data on System z for security and reliability



Health Care Use Case

Problem	Solution	Benefits
• Relapses in cardiac patients • "One size fits all" treatment • $1M+ Medicare re-admission penalties • Sensitive patient data on Mainframe • No efficient way to offload and integrate	• Identify risk factors by analyzing patient data* • Create algorithms to predict outcomes	• 31% reduction in re-admissions • $1.2M savings in penalties • No manual intervention • No increase in staffing • 1100% ROI on $100K

* Mainframe database requires special skills to access without Veristorm

Public Sector Use Case

Problem	Solution	Benefits
• Mismanaged assets led to neighborhood, publicity, law & order issues • Post-2008 austerity measures reduced budget • Asset data was Mainframe based – no efficient offload and integration mechanism	• “Crowd source problem” reporting – cell phone photo, social media, GPS data • Integrate social media reports with asset / governance data on Mainframe, achieving regulation conformity	• Software cost of $400K compares to $2M consulting engagement • Better maintained neighborhoods yield $5.6M in higher taxes first year

* System z IMS database requires special skills to access without vStorm

Retail Use Case

Problem	Solution	Benefits
• Streams of user data not correlated • e.g. store purchases, website usage patterns, credit card usage, historical customer data • Historical customer data Mainframe based – no efficient, secure integration	• Secure integration of historical customer data, card usage, store purchases, website logs • Customer scores based on the various data streams • High scoring customers offered coupons, special deals on website	• 19% increase in online sales during slowdown • Over 50% conversion rate of website browsing customers • Elimination of data silos – analytics cover all data with no more reliance on multiple reports / formats

Big Data at NCAT

North Carolina Agricultural and Technical State University

- NC A&T State University
- Located in Greensboro, NC, enrollment approx. 10,500.
- One of the 100+ Historically Black Colleges and Universities
- Established in 1891 as a Land Grant College
- Still produces more African American engineers than any school in the world
- I am in the Computer Systems Technology Dept. in the School of Technology

Enterprise Systems Program, School of Technology at NC A&T:

- Mission: To support education, research, and business development in the System z space
- NCAT System z Environment:
  - Since 2010
  - Z9, 18 GPs, no IFLs, 128GB storage, 4 TB DASD (online)
  - 44 TB DS8300 (offline)
  - 2 LPARs (using 1)
  - z/VM is the base OS, all other OSes are guests of z/VM
  - Plan in the works with our business partners to acquire a BC12
  - Using GPs as IFLs (special no-MIPS deal with IBM)
  - Allocate GPs to the LPAR
  - VM 5.4
  - SUSE 11, Debian, RHEL
  - DB2, LAMP, SPSS for System z, Cognos, zDoop and more

System z as a Private Cloud

- Students & faculty need to rapidly deploy, clone, and turn down servers
  - Helps manage the student (user) learning process
- First university to adopt CSL Wave
- Adding rapid deployment of Hadoop clusters
- Early adopter of vStorm Enterprise
- On-demand scaling by simply adding IFLs
- No additional power or space
- Use existing skills, processes, security (RACF/LDAP), management tools

Research Support

- Several researchers at A&T have a focus on analytics
- Areas of Focus
  - Sentiment Analysis (opinion analysis)
  - Health Informatics (fraud detection, Medicare/Medicaid)
  - Predictive Analytics (student outcomes, product viability, etc.)
- Faculty have expressed interest in Hadoop
  - Need to manage larger data sets
  - Collecting unstructured/non-relational data
  - Want to pool data without pre-determined query in mind
  - Interactive/discovery and query

Education

- 4 undergraduate courses: intro, intermediate, advanced mainframe operations and z/VM
- 2 graduate courses (mainframe operations, z/VM)
- Proposed graduate certificate of enterprise systems (under review)
- High school outreach programs in enterprise systems
- 1 semester zVM class (CPCMS, Install Linux as a guest, getting Linux running, using VM as a deployment tool for Linux)
- VM will be increasingly important; key to preparing students for careers in Big Data

Student Example

- Over 70 students placed in enterprise systems positions
- Heavy focus on IBM's *Master the Mainframe* contest
- Two students participants in IBM's 50th Anniversary of the Mainframe
  - Dontrell Harris, a keynote speaker, capacity planning specialist at Met Life
  - Jenna Shae Banks, a judge for the first *Master the Mainframe* world championship
  - Placements at IBM, Met Life, USAA, BB&T, Fidelity, Wells Fargo, Bank of America, First Citizen's Bank, John Deere, State Farm and others

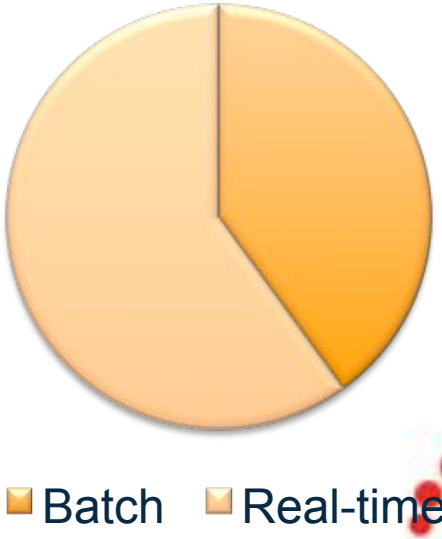
Big Data Initiative

- Challenge & opportunity
  - Saw the potential for zVM application for Hadoop; most people focused on x86
  - Hot topic for research; important to students
  - Provide easy & controlled access to mainframe data
  - Enable the developer community to take advantage of the enterprise primary data in a model they understand
  - Familiar environment: Linux, Java, SQL & the hot technology: Hadoop
- Getting buy-in for z
  - Most don't know z at all
  - Dean, Chair, Chancellor, Provost: Needed to be sold; not IT people
  - Simplify!

Unlock New Insight and Reduce Cost

- Do More
  - Analyze large amount of information in minutes
  - Offload batch processes to IFLs to meet batch window requirement
- Reduce Cost
  - Take advantage of IFLs price point to derive more insight
- Application extensibility achieved through newly available skillset

System z Workloads



White Papers and Articles

- "The Elephant on z",
  IBM & Veristorm, 2014
  - www.veristorm.com/go/elephantonz
- "Bringing Hadoop to the Mainframe",
  Paul Miller, Gigaom, 2014
  - www.veristorm.com/go/gigaom-2014
- "Inside zDoop, a New Hadoop Distro for IBM's Mainframe",
  Alex Woodie, Datanami, 2014
  - www.veristorm.com/go/datanami-2014-04
- "vStorm Enterprise Allows Unstructured Data to be Analyzed on the Mainframe",
  Paul DiMarzio, 2014
  - www.veristorm.com/go/ibmsystems-2014-06
- "vStorm Enterprise vs. Legacy ETL Solutions for Big Data Integration",
  Anil Varkhedi, Veristorm, 2014
  - www.veristorm.com/go/vsetl
- "Is Sqoop appropriate for all mainframe data?",
  Anil Varkhedi, Veristorm, 2014
  - www.veristorm.com/go/vssqoop
- IBM Infosphere BigInsights System z Connector for Hadoop
  - www.ibm.com/software/os/systemz/biginsightsz

  (Includes data sheet, demo video, Red Guide)
- Solution Guide: "Simplifying Mainframe Data Access"
  - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/tips1235.html

Videos

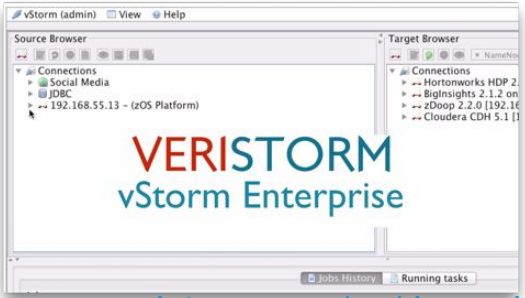
Introduction

www.veristorm.com/go/introvid11m

Webinar

www.veristorm.com/go/webinar-2014q3

Demo



www.veristorm.com/go/demovid2m

Any Questions?

- Mike Combs
  mcombs@veristorm.com
- Cameron Seay, Ph.D.
  cwseay@ncat.edu

- https://share.confex.com/share/124/webprogrameval/Session16423.html

Abstract

- Hadoop and data integration with System z
- http://www.veristorm.com/content/share-pittsburg-presentation
- Big Data technologies like Hadoop are transforming analytics and processing, but what is the role of System z? We'll examine System z advantages as a platform for Hadoop and as a rich source of enterprise data for processing in Hadoop both on and off the platform. How can System z and Hadoop respond quickly to the organizational needs to make data-driven decisions in near real-time, when the questions aren't well-known in advance?
- Dr. Cameron Seay, Ph.D., Assistant Professor, Computer Systems Technology, of North Carolina Agricultural and Technical State University will **share his experience with Hadoop on z applied to analytics and research projects**.
- Mike Combs, VP of Marketing, Veristorm, will **discuss how rapid, no-transformation access to mainframe data from Hadoop can enable new solutions, including lightweight performance management and capacity planning**.