SNIA Global Education

Protecting Data in the “Big Data” World

David A. Chapa / CTO, Thought Leader

Author:
SNIA - Data Protection & Capacity Optimization (DPCO) Committee

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

About the SNIA DPCO Committee

- This tutorial has been developed, reviewed and approved by members of the Data Protection and Capacity Optimization (DPCO) Committee, which any SNIA member can join for free
- The mission of the DPCO is to foster the growth and success of the market for data protection and capacity optimization technologies
  - Online DPCO Knowledge Base: www.snia.org/dpco/knowledge
  - Online Product Selection Guide: http://sniadataprotectionguide.org
- 2016 goals include educating the vendor and user communities, market outreach, and advocacy and support of any technical work associated with data protection and capacity optimization

Check out these SNIA-DPCO Tutorials at www.snia.org/education/tutorials

- **Intro to Data Protection – Backup to Tape Disk, & Beyond**
- **Trends in Data Protection & Restoration Technologies**
- **Privacy vs Data Protection: The Impact of the EU Data Protection Legislation**
- **Advanced Data Reduction Concepts**
- **Intro to Encryption and Key Management: Why, What and Where?**

Abstract

The continued growth of data and these "Big Data" repositories need to be protected. In addition, new regulations are mandating longer data retention, and the job of protecting these ever-growing data repositories is becoming even more daunting. This presentation will outline the challenges and the methods that can be used for protecting "Big Data" repositories.

Topics will include:

- The unique challenges of managing and protecting "Big Data"
- The various technologies available for protecting "Big Data"
- The various data protection considerations for "Big Data“, for various environments, including Disaster Recovery/Replication, Capacity Optimization, Cloud, etc.

Terminology

Big Data [Storage System]

A characterization of datasets that are too large to be efficiently processed in their entirety by the most powerful standard computational platforms available.

Compression [General]

The process of encoding data to reduce its size. Lossy compression (i.e., compression using a technique in which a portion of the original information is lost) is acceptable for some forms of data (e.g., digital images) in some applications, but for most IT applications, lossless compression (i.e., compression using a technique that preserves the entire content of the original data, and from which the original data can be reconstructed exactly) is required.

Data Deduplication [Storage System]

The replacement of multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth.

Terminology (Cont.)

Data Protection [Data Management]

Assurance that data is not corrupted, is accessible for authorized purposes only, and is in compliance with applicable requirements.

Structured Data [Data Management]

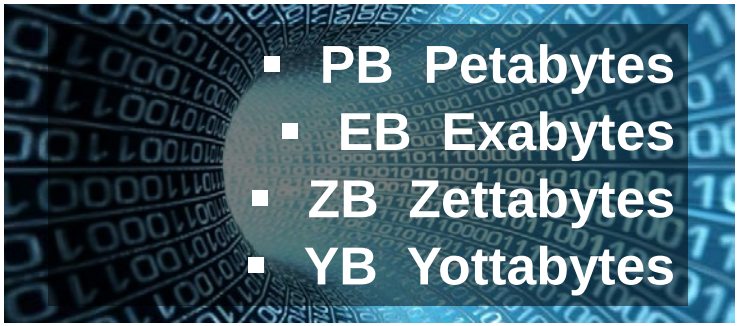
Data that is organized and formatted in a known and fixed way. The format and organization are customarily defined in a schema. The term "structured" data is usually taken to mean data generated and maintained by databases and business applications.

Unstructured Data [Data Management]

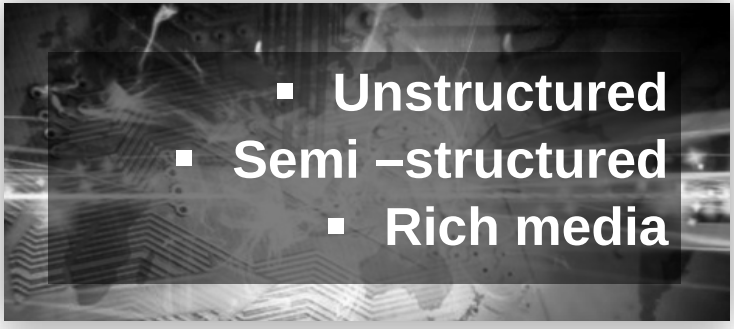
Data that cannot easily be described as structured data.

The 3 Original "V"s of Big Data

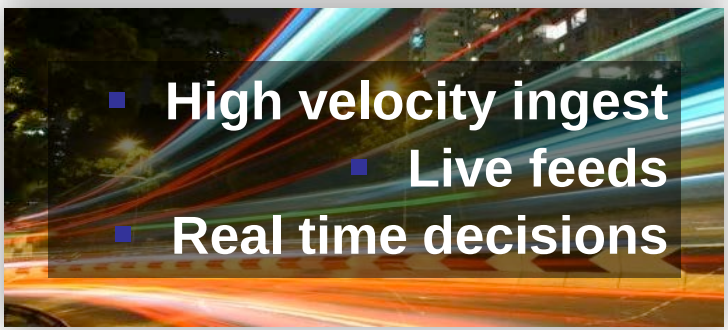
Volume



Variety

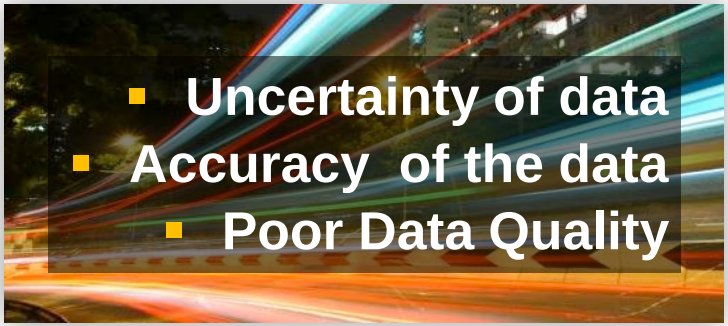


Velocity



Looking Ahead…

Veracity



Value

The Challenge of Big Data: Volume

SNIA Education

Volume

- PB Petabytes
- EB Exabytes
- ZB Zettabytes
- YB Yottabytes

$1PB = 10^{15}$ bytes
$1EB = 10^{18}$ bytes
$1ZB = 10^{21}$ bytes
$1YB = 10^{24}$ bytes

- Jet engine produces ~10TB of data every 30 minutes of flight time

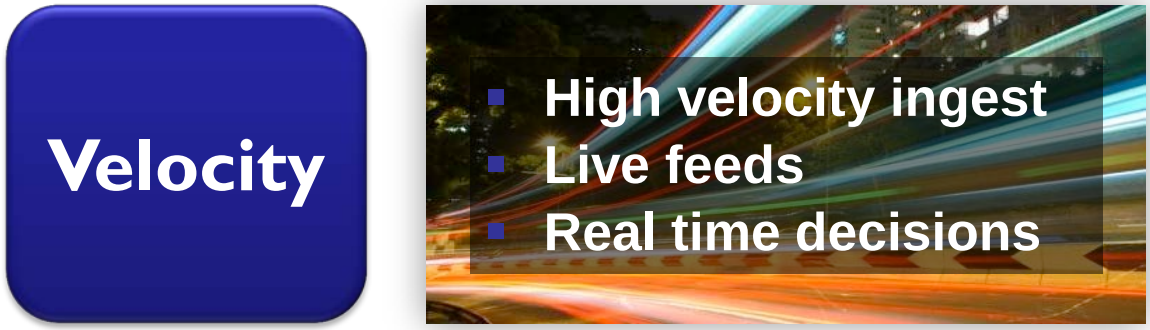
- Google processes ~20PB of data per day

- If one exabyte's worth of data were placed onto DVDs in slimline jewel cases, and then loaded into Boeing 747 aircraft, it would take 13,513 planes to transport this one exabyte

The Challenge of Big Data: Variety

- Unstructured data: office files, video files, audio files, etc.
- Semi-structured data: integrated text/media files, Web/XML files, etc.
  - Easy to generate but difficult to query, optimize, etc.
- Rich Media: Streaming media, Flash videos, etc.

The Challenge of Big Data: Velocity



- High velocity ingest:
  - Easy to generate but difficult to query, optimize, etc.
- Live feeds:
  - News, audio/video television programming, surveillance video, etc.
- Real-time decisions:
  - Retail: discounts, suggest other items to purchase, etc.
  - Security
  - Etc.

Veracity

- 1 in 3 business leaders do not trust the info that they use to make decisions
- Poor data quality costs the US $3.1 Trillion per year
- 27% unsure of how much of their data was accurate

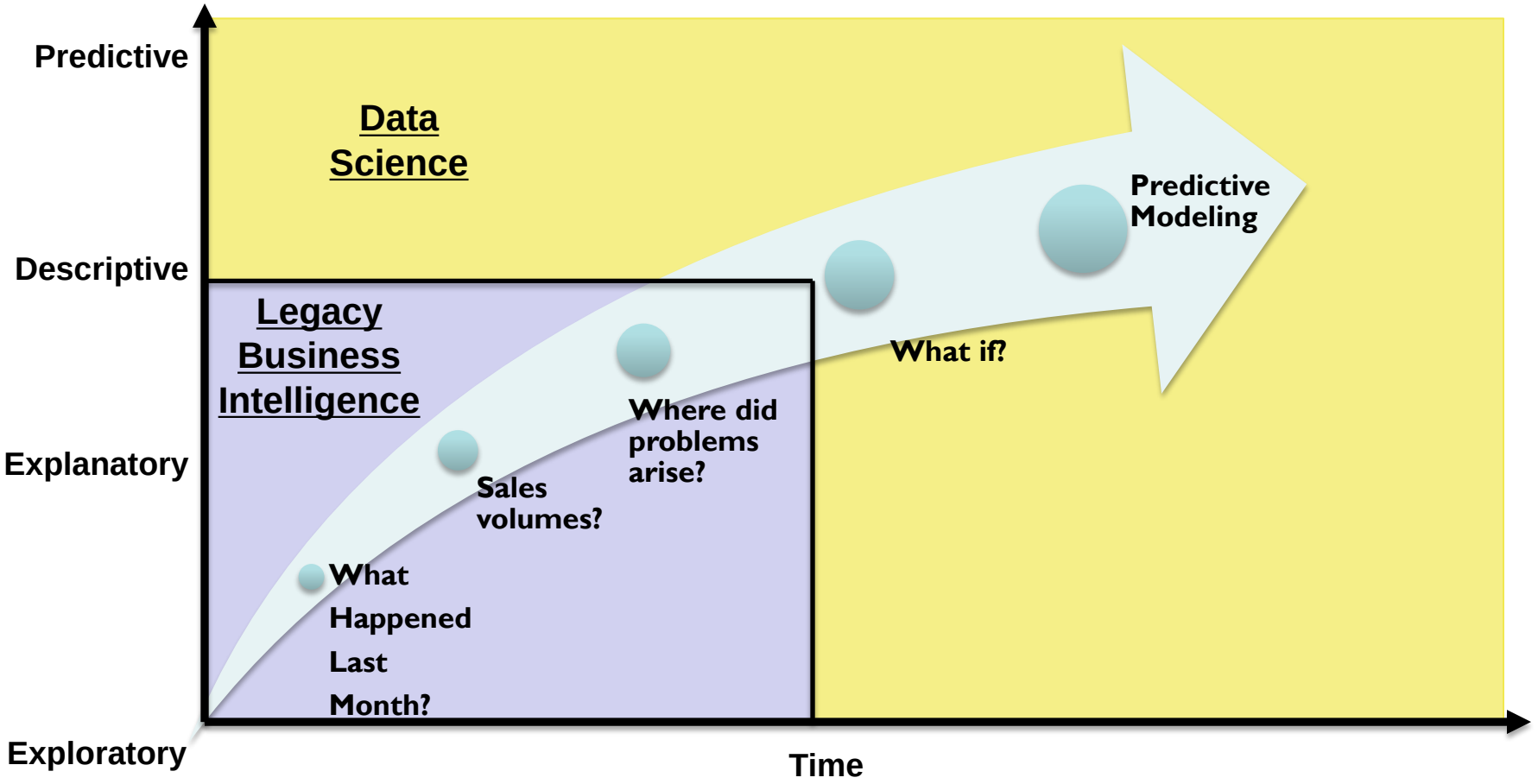
Source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Value: Industries Impacted by Big Data

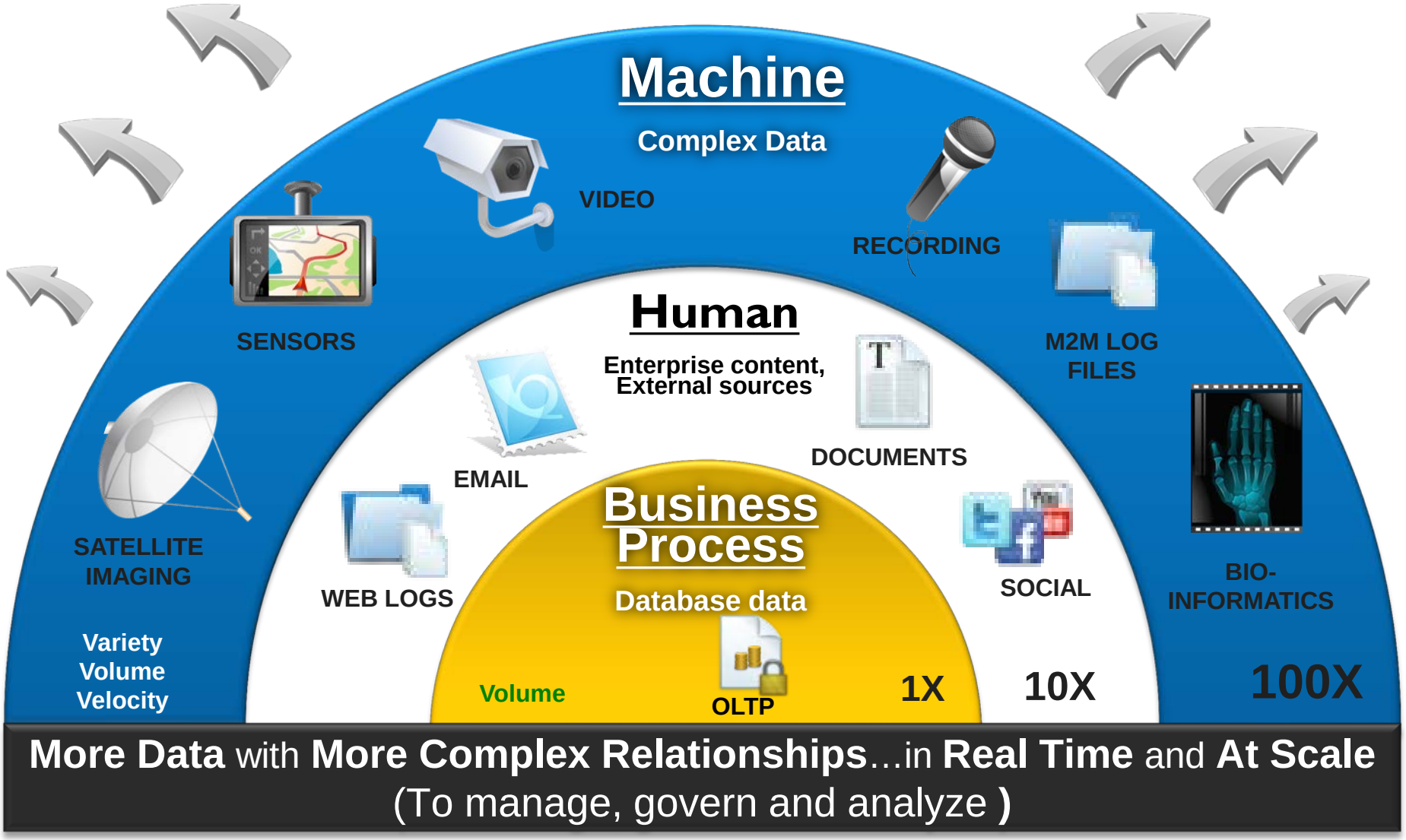
SNIA Education

Value

Thought Leading Analytics

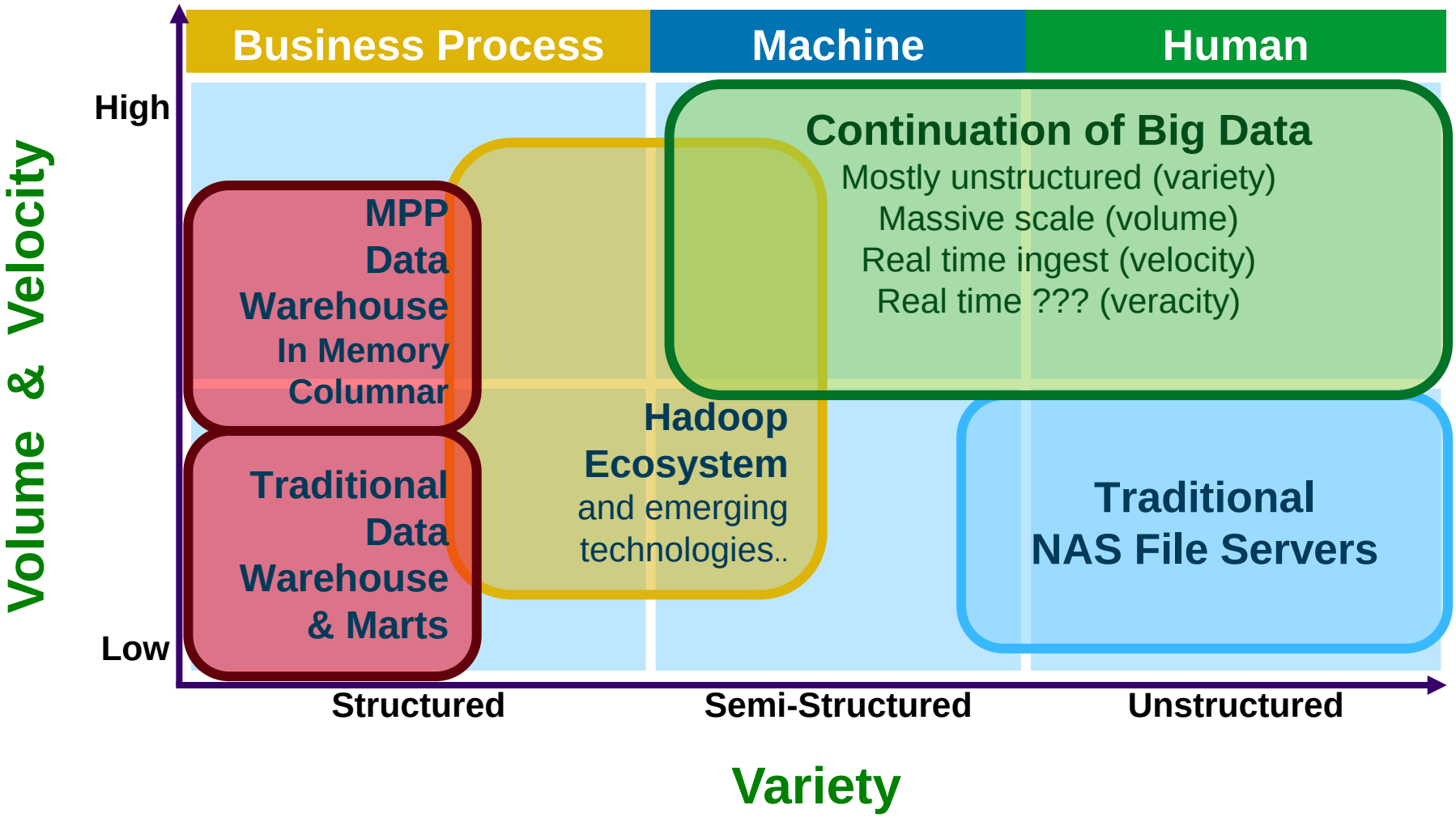


Sources of Big Data Growth

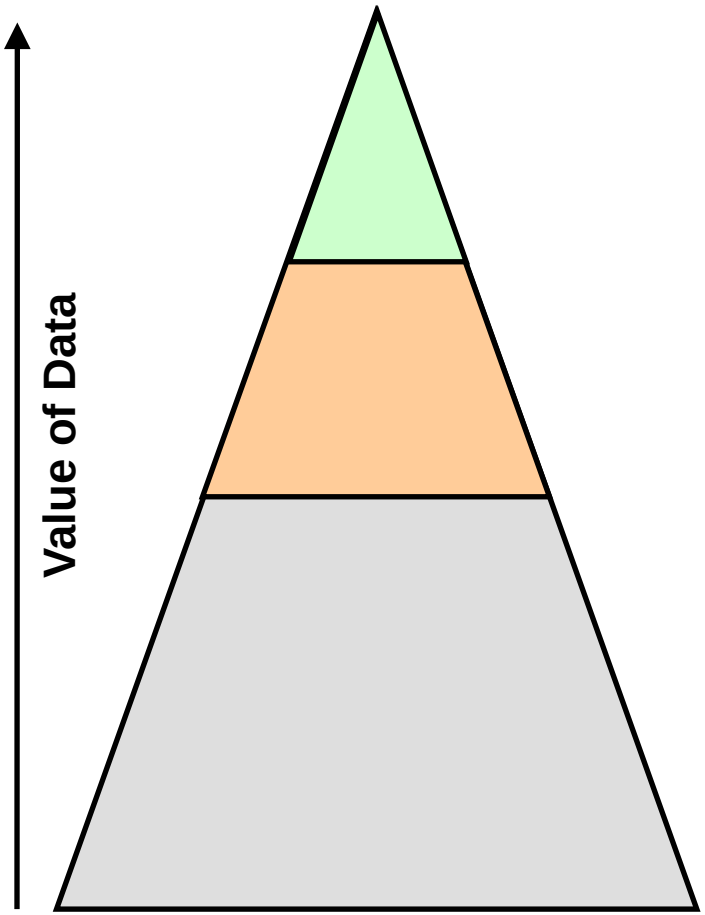


More Data with **More Complex Relationships**…in **Real Time** and **At Scale**
(To manage, govern and analyze)

Big Data: In Context

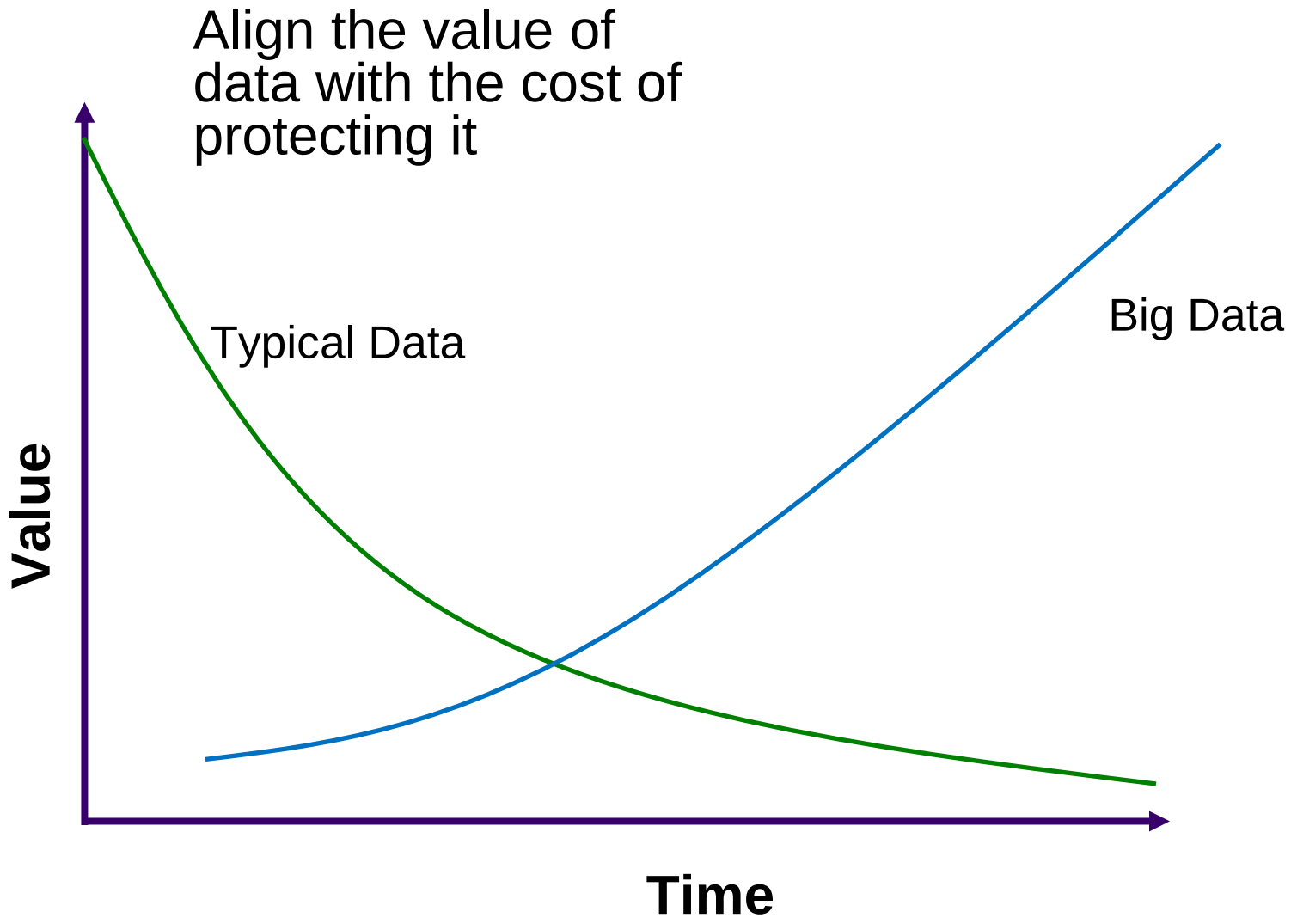


All Data is not Equal



- **Mission Critical Data**
  - Every minute down impacts revenue
  - Complete protection is a high priority
- **Important Data**
  - Needs to be recovered quickly to avoid impact
  - Justifiable protection is required
- **Non-Critical Data**
  - 24+ hour recovery time is acceptable
  - Cost-effective protection needed

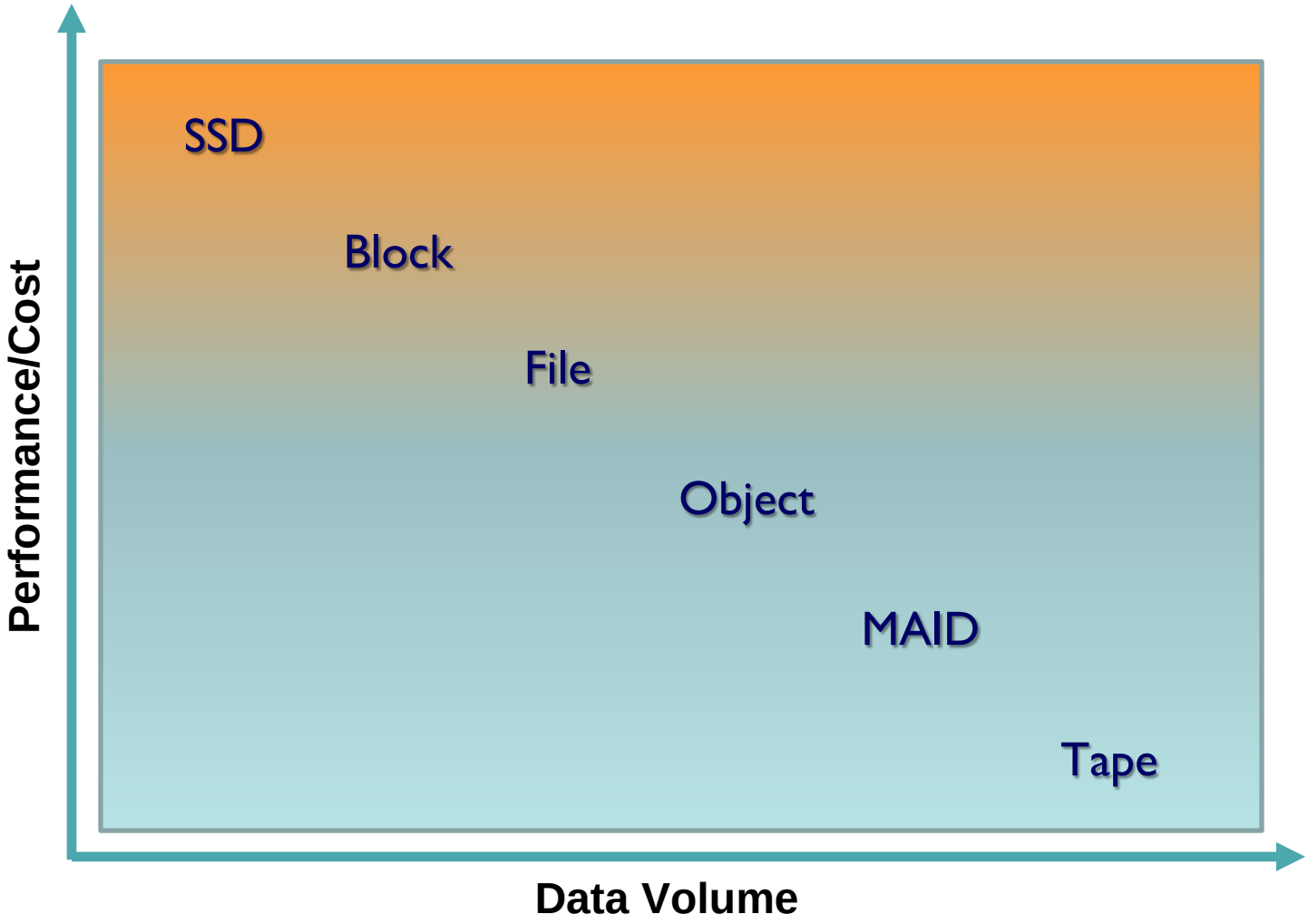
Data often "Loses" Value over Time



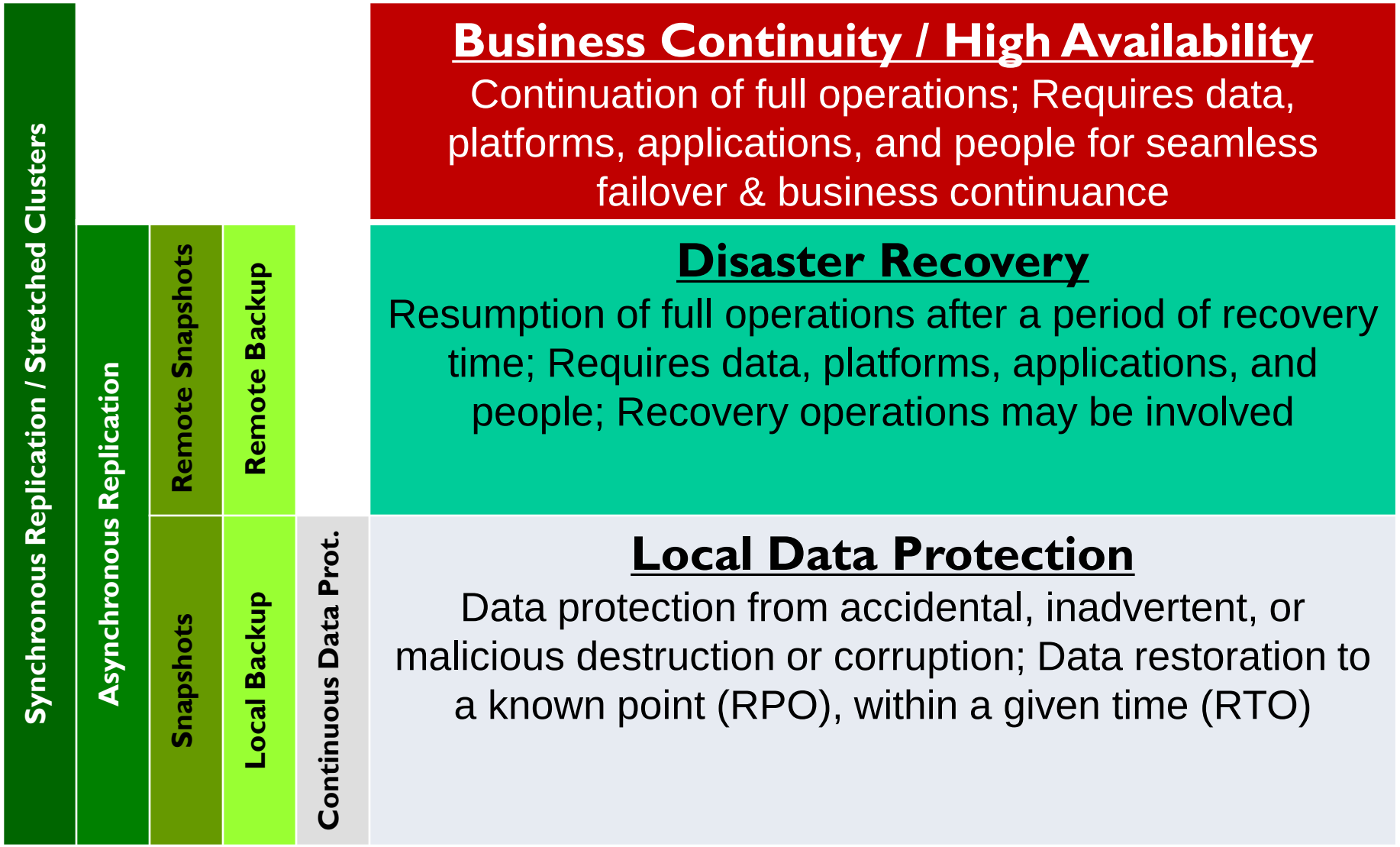
Active vs Passive Archive

- Active archive is data that needs to be accessible
  - Data is accessible to users, e.g.:
    - Healthcare records
    - Land deeds
    - Business intelligence data
- Passive archive ("deep archive") is offline or offsite
  - Data does not need to be "readily" accessible
  - User usually knows that the data is offline

Tiering Examples



What is Data Protection?



Data Protection in the Typical Data Center

- Backup to tape still common
  - Multi-stream to tape to decrease backup windows
  - Compression technology native on the tape drives
- Backup to disk more prevalent for "quick" restores
  - Deduplication is also very common (for space savings, not speed)
  - Satisfy more stringent RPO / RTO requirements
  - Backup to VTL – way to leverage existing tape operations with disk media
  - Snapshots are common for file servers and VMs
- Remote replication of disk-based backups, replacing off-site tape
- Application-aware backup methodologies (e.g., database, email)

Challenges of Protecting Big Data

- Big Data is too "big" for traditional backup schemes
  - Takes too much space to backup
  - Not enough time to backup all the data
  - Consumes too many resources
- Is it Necessary to Backup all data?
  - Temporary (intermediate) data
  - Non-mission critical application data
- Regulatory requirements (e.g., privacy)
- SLAs / SLOs, where data needs to be:
  - Restored within a specified time frame, or Business Continuity Plan (BCP)
  - Different data types require different RTO/RPO ("data importance mapping")

Methodologies of Protecting "Big Data"

- Protection against multiple failures / corruptions
  - RAID-6
  - Geo-distributed Parity
  - New forms of erasure coding
  - Snapshots
- Geographically dispersed copies
  - Active/active
  - Active/passive
- Replication
  - Active/active
  - Active/passive
  - Active/standby
- Any combination of the above

Deduplication and Compression

SNIA Education

Dedupe and compression are similar

- Both are dependant on data patterns
  - Results can vary from little/no optimization to high percentage
- Both consume system resources
- Both can optimize required storage capacity or bandwidth utilization

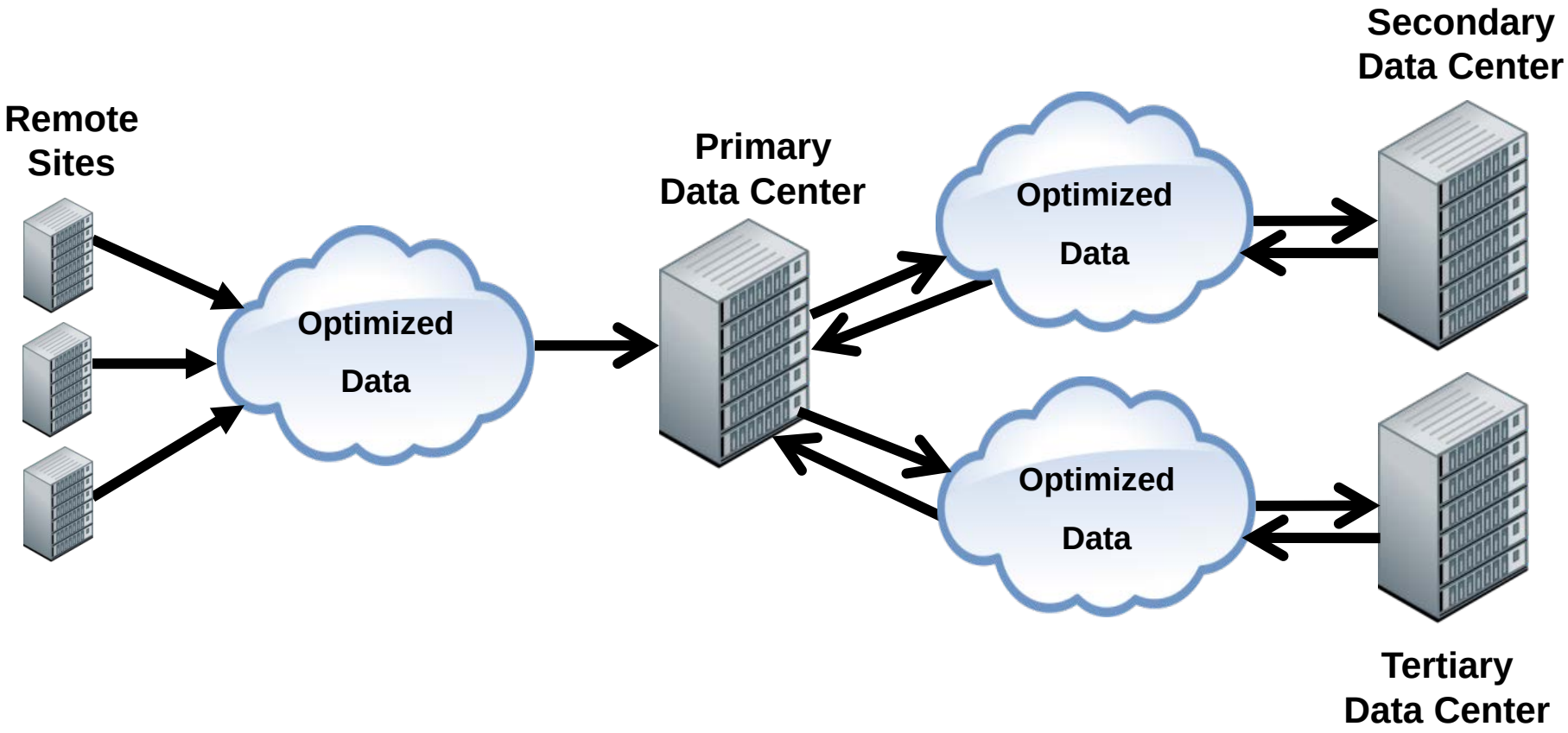
Dedupe and compression are different

- Dedupe and compression can be complementary
- But some knowledge about the data pattern is helpful
- Some data is best optimized via dedupe
- Some data is best optimized via compression
- Some data can be optimized via dedupe **and** compression

Sequence of optimization when encryption is used

- Dedupe is first, encryption is last

Considerations for Protecting Big Data: Remote / Dispersed Replication



- Can be one-way, bi-directional, multi-hop, cascade or n-way
- Utilizing data reduction techniques is “mandatory” for big data

Considerations for Protecting Big Data: Remote / Dispersed Replication (Cont.)

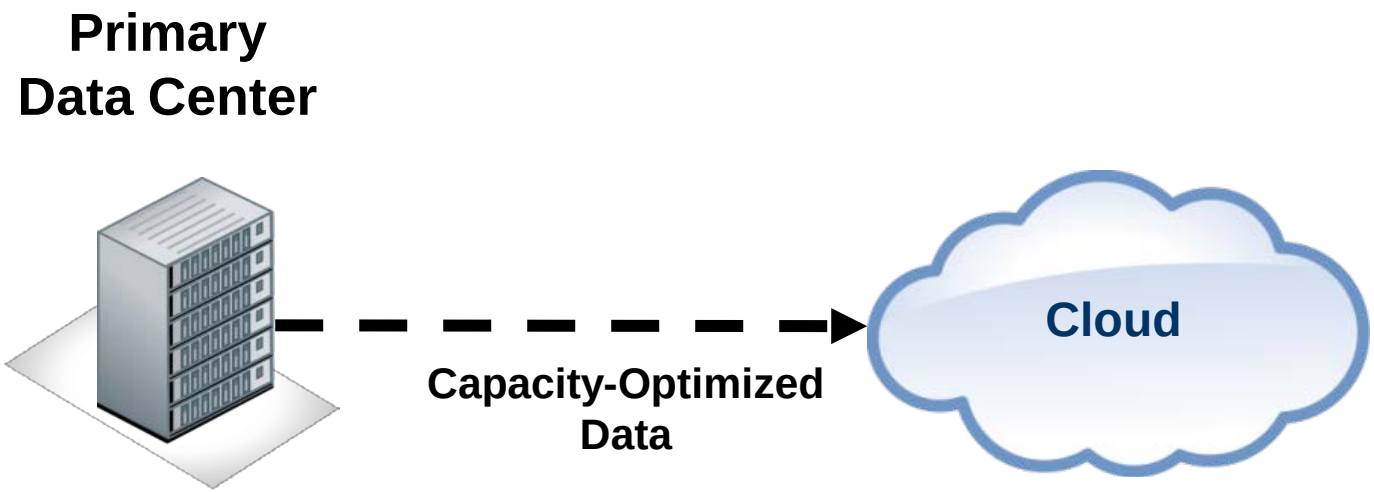
- Focus on your Service Level Agreements (SLAs) first
  - Needs to meet window for *Replication*
  - Needs to meet SLA for System *Recovery or Data Restore*
- Is DR site planned as failover site?
  - If so, need to consider handling of data reduction re-hydration
- Is it Necessary to Optimize All Data?
  - Mission-critical applications
  - May have regulatory issues for some data
  - Some data types not conducive to data reduction
  - Replicate incremental changes only, without other optimization

Data Protection: Transition for Big Data

- Backup to tape still common, but rarely used for "local" backups
  - More used for the "off-site" DR and/or for "long-term" requirements
  - Starting to see other data reduction technologies appear on tape drives
    - File system-like access technology
    - Deduplication
- Backup to Disk and other "fast" media will be most prevalent
  - Solid-state disk and various memory technologies
  - Will be less of an issue of RPO/RTO than pure speed and adaptability
- Multi-site / multi-media backups
  - Multi-site: N-way site replication
  - Multi-media: not constrained to one media type for data protection

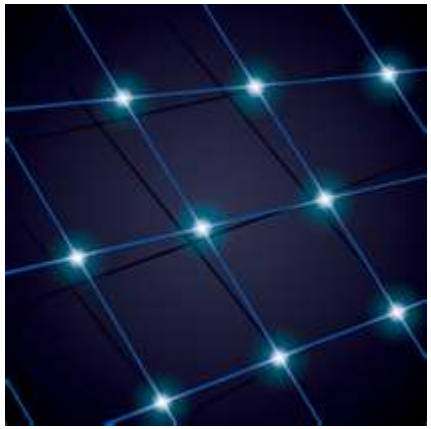
Data Protection: Transition for Big Data (Cont.)

- Backup to the "cloud"
  - Failover to and recovery from the cloud
  - Some of the technologies currently used and being developed:
    - Applications that write directly to the cloud
    - Storage-based software migration
    - On-ramps – devices that stage data and move to the cloud over time
    - Storage systems that make copies to cloud locations



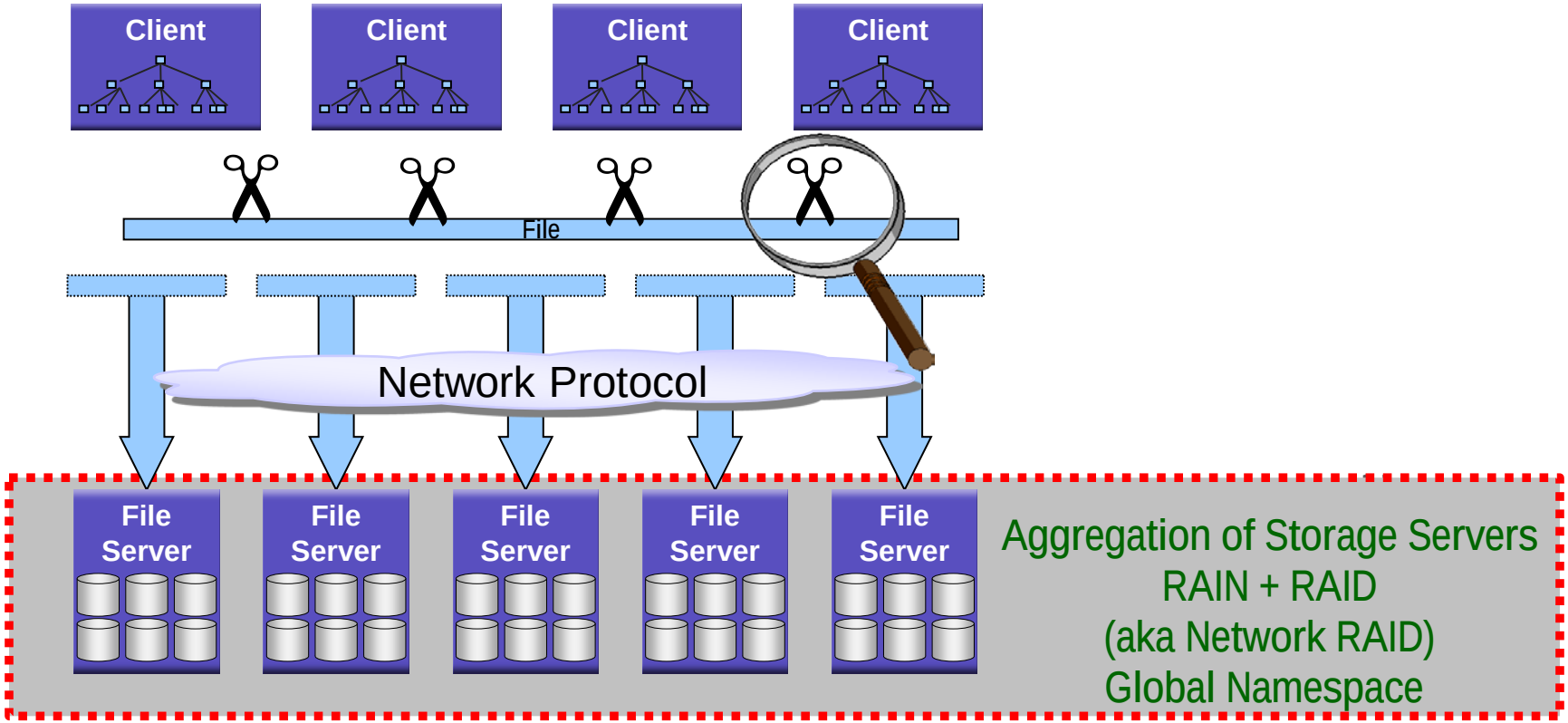
Data Protection: Transition for Big Data (Cont.)

- "RAIN" technology (Redundant Array of Independent Nodes)
  - Allows for data to be copied to one or more nodes within a cluster
  - Sometimes negates the need for traditional RAID and/or local Backup
    - Especially if the Nodes are in remote locations – for DR purposes



Data Protection: Transition for Big Data (Cont.)

- Distributed / Parallel / Global file system technologies
  - Allows for geographically dispersed data on a file system
  - Can create policies to make "copies" data for multi-site data redundancy



Summary

- Many different technologies are used to cope with protecting big data
  - Compression
  - Deduplication
  - WAN accelerators
  - Faster storage technologies (Solid State Disks, Flash/Memory, etc.)
- But data reduction technologies are only one aspect of dealing with the challenge of protecting Big Data…

Summary (Cont.)

- "Content-aware" and "application-aware" data protection schemes are becoming the mainstay
  - Potential for greater data reduction with more knowledge of specific data structures within respective data types
  - Potential for greater "flexibility" in how data is protected
    - Data protection can become much more streamlined as we are more "aware" of data types, data change-rate models, data access models, etc.
  - Greater "adaptability" in how data is protected
    - New "hybrid" data types are being generated
    - Advanced data protection schemes can be developed to better accommodate big data requirements

Summary (Cont.)

- Much higher performance data protection required for big data
  - Near instantaneous backups
  - Much faster restores (in some cases, near instantaneous)
- More data protection automation is required
  - Less people will be required to manage
  - Gone will be the days of reviewing “failed backup jobs” logs
- Advanced data protection “flexibility”
  - Need to support multi-vendor, multi-application environments

Summary (Cont.)

- Growing requirement for abiding by regulations
  - In US: SEC 17a-4, HIPAA, etc.
  - Outside of US: EU Data Protection Directive, UK Data Protection Act, etc.
  - Automation will be a large part of meeting these requirements
- High Reliability / Resiliency
  - SLAs becoming more stringent (cannot afford to lose any access to data)
  - Relying on a standard "backup" is a thing of the past: a new paradigm is emerging…
- Need to support protecting data to the "cloud"
  - Data feeds coming from other sources via public, private & hybrid clouds
  - Also need support for recovery "to" the cloud (not just "backup" to…)

Summary (Cont.)

- Make sure the Cloud service provider has the appropriate data protection in place for your data
  - Based on SLAs that meet your business requirements
- Check on support for the ability to “upload” data to a Cloud service provider, to then allow for further data analytics by third-parties

Attribution & Feedback

The SNIA Education Committee thanks the following individuals for their contributions to this Tutorial:

Authorship History

Original Author: DPCO Committee, 8/2012

Updates:

DPCO Committee, 9/2012
DPCO Committee, 2/2013
DPCO Committee, 5/2013
DPCO Committee, 8/2013
DPCO Committee, 3/2014
DPCO Committee, 8/2014
DPCO Committee, 1/2015
DPCO Committee, 4/2016

Additional Contributors

Ashar Baig
David A. Chapa
Kevin Dudak
Mike Dutch
Larry Freeman
Eric Hibbard
David G. Hill
Tom McNeal
Gene Nagle
Ronald Pagani
Molly Rector
Thomas Rivera
Tom Sas
Gideon Senderov
Paul Talbut
SW Worth

Please send any questions or comments regarding this SNIA Tutorial to ***tracktutorials@snia.org***