Trends in Managing Data at the Petabyte Scale

Steve Kleiman
Sr. VP & CTO

Before we begin…

- **Disk reliability**
  - **SIGMETRICS ’07:**
    An Analysis of Latent Sector Errors in Disk Drives
    - **Lakshmi Bairavasundaram, Garth Goodson, Jiri Schindler, Shankar Pasupathy**
  - **Symposium on Reliability and Maintainability ’03,’04,’05**
    - **John Elerath and Sandeep Shah**

Petabyte Environments are Here!

- **2006 Q1-Q3 NAS+SAN Petabytes Shipped**

Vendor	PB
EMC	214
NetApp*	199
HP	180
IBM	92
Hitachi	64
Other	247
Total	996

*Current quarterly run rate >100PB
YoY Growth >100%

Source: IDC, Dec 2006

- **~25 NetApp customers with >1PB**
- **Largest: ~33PB**

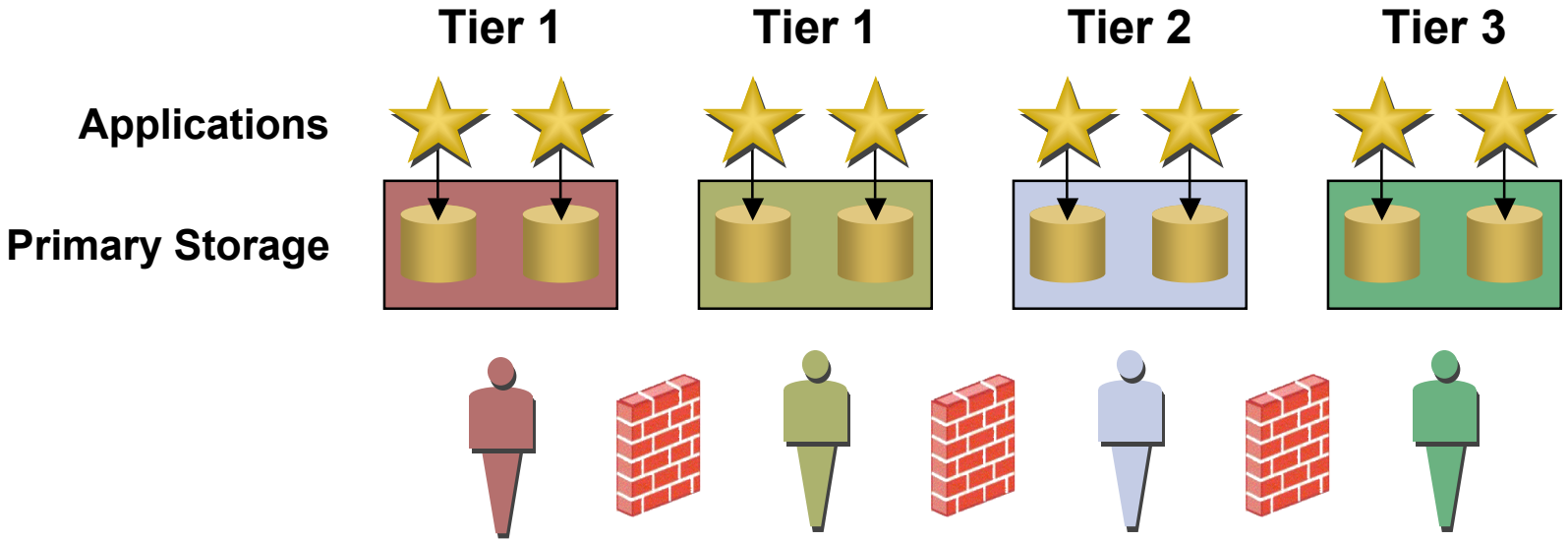
The Growing Burden of Data Ownership

- **Operational burdens**
  - Managing the data explosion
    - 50-100%
    - Unstructured, semi-structured, structured
  - Increasing dependence on data
    - Ensuring 100% availability
    - Protecting data from disasters
  - Rapidly deploying new applications
  - Global operations
    - Multiple data centers
    - Many remote offices
- **Financial burdens**
  - Controlling costs
    - Equipment, people, processes
    - Utilization

New/Hidden Burdens of Data Ownership

- **Legal burdens**
  - Complying with regulations
    - Discovery
    - Preventing unauthorized access
    - Retention
- **Social burdens**
  - Protecting your reputation
    - Disclosing data loss
- **Geo-political burdens**
  - Multiple cultures & legal systems

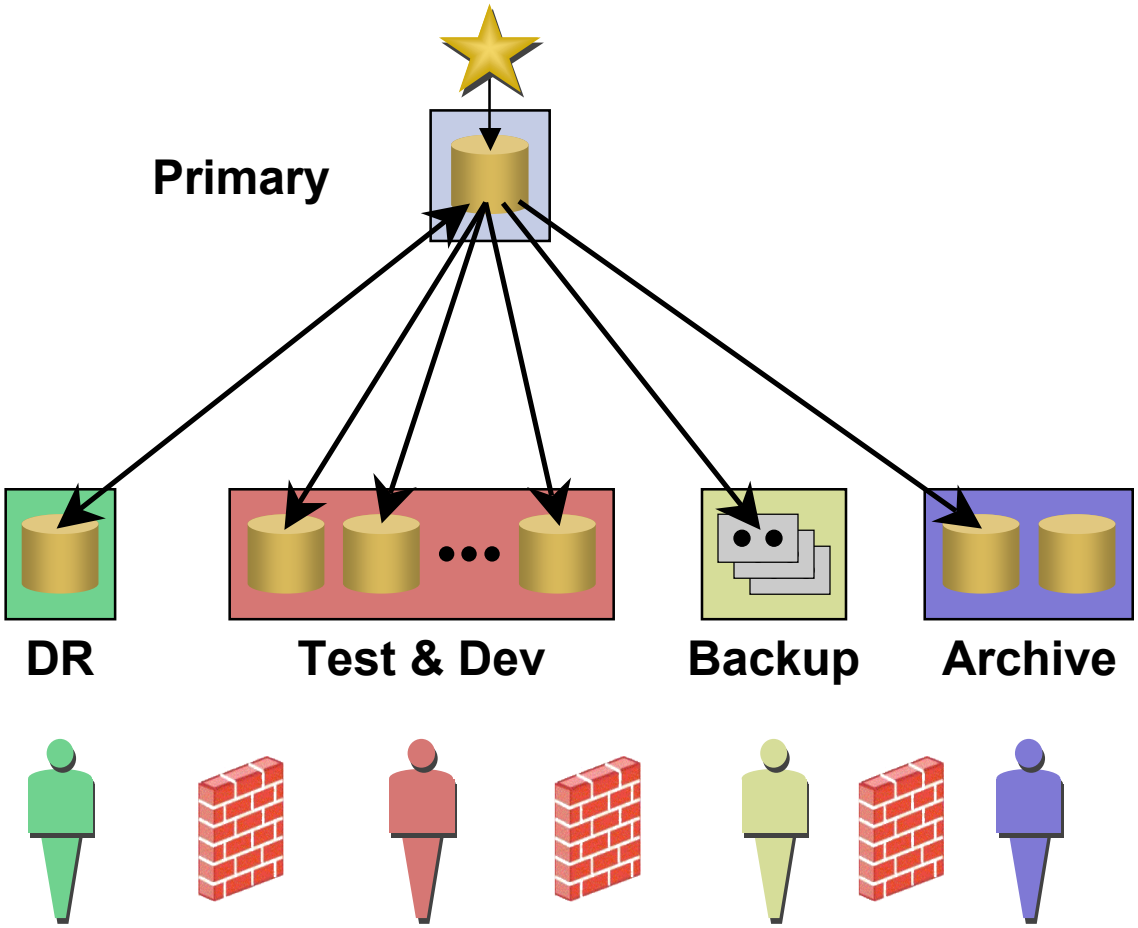
Traditional Infrastructure Build-out: Application-centric Silos



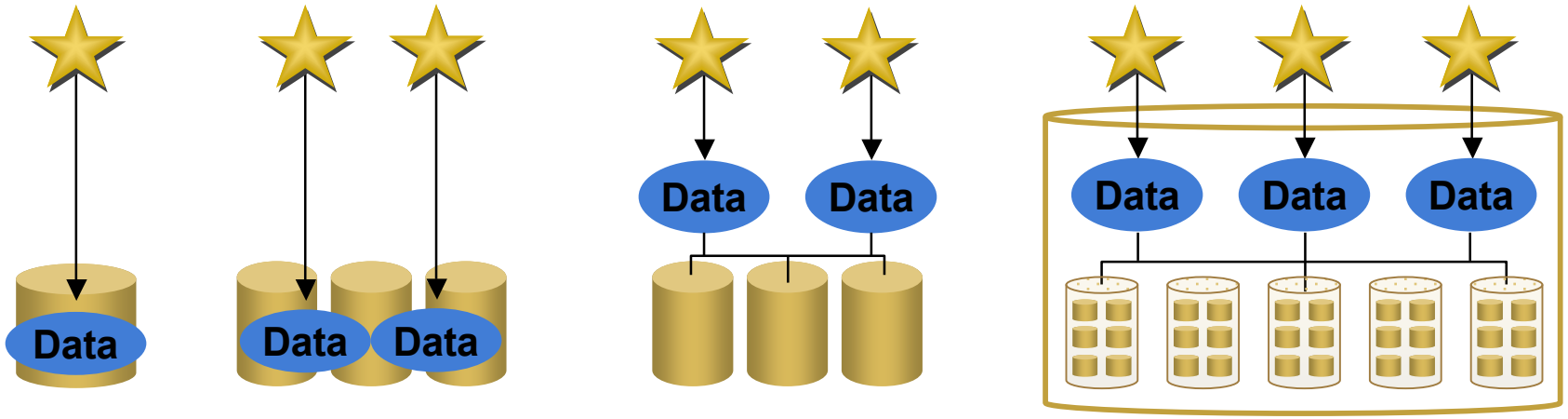
- Good Quality of Service

- Incompatible hardware
- Incompatible software
- Different processes
- Lots of experts
- Low utilization

It's Not Just the Primary Storage



Separation of Data from Physical Containers



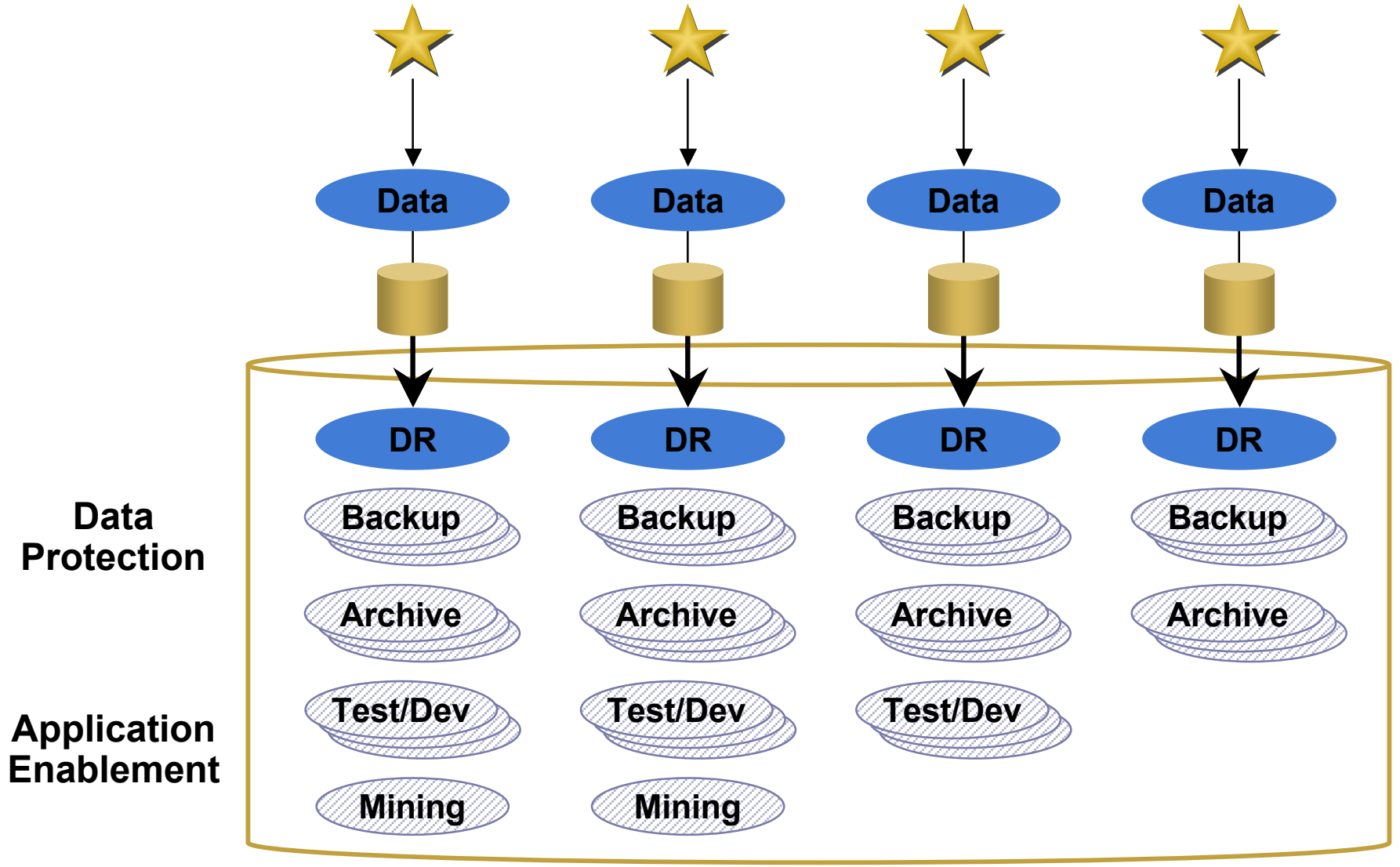
DAS

Networked Storage

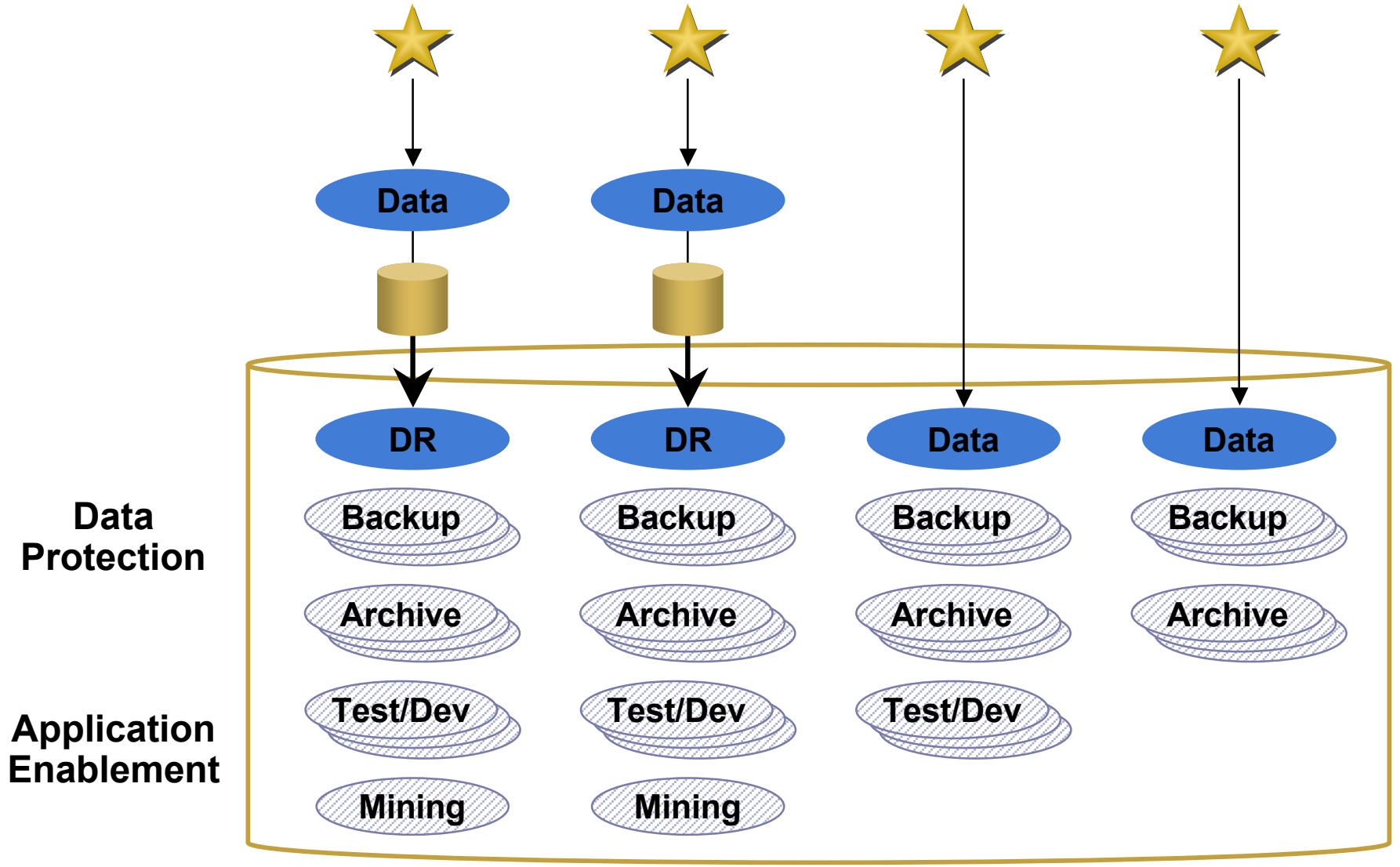
Snapshots
Clones
Thin-provisioning
Data mirroring

Global Namespace
Scale-out
Multiple Tiers

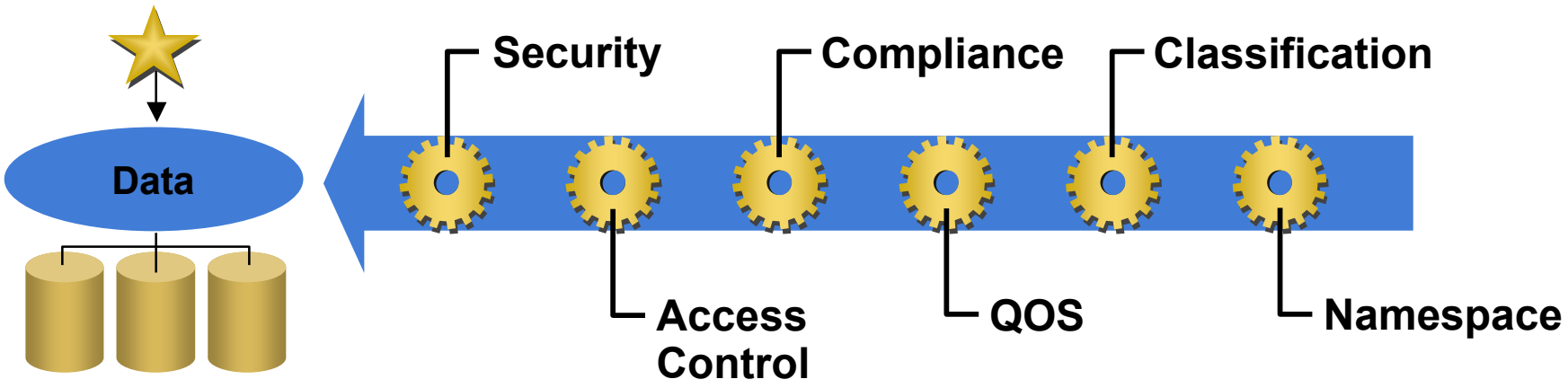
Unified Protection & Enablement Environment



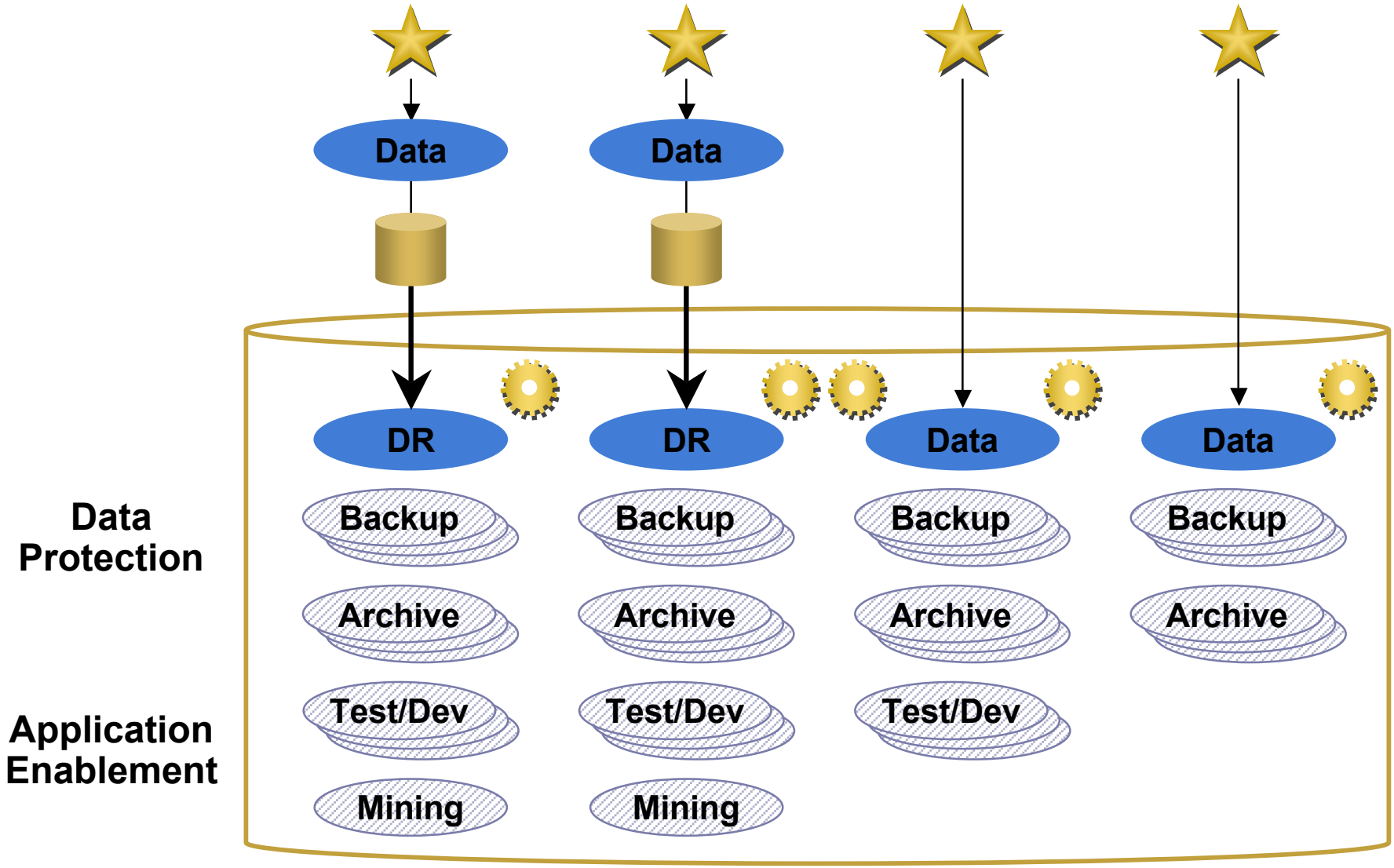
Unified Protection & Enablement Environment



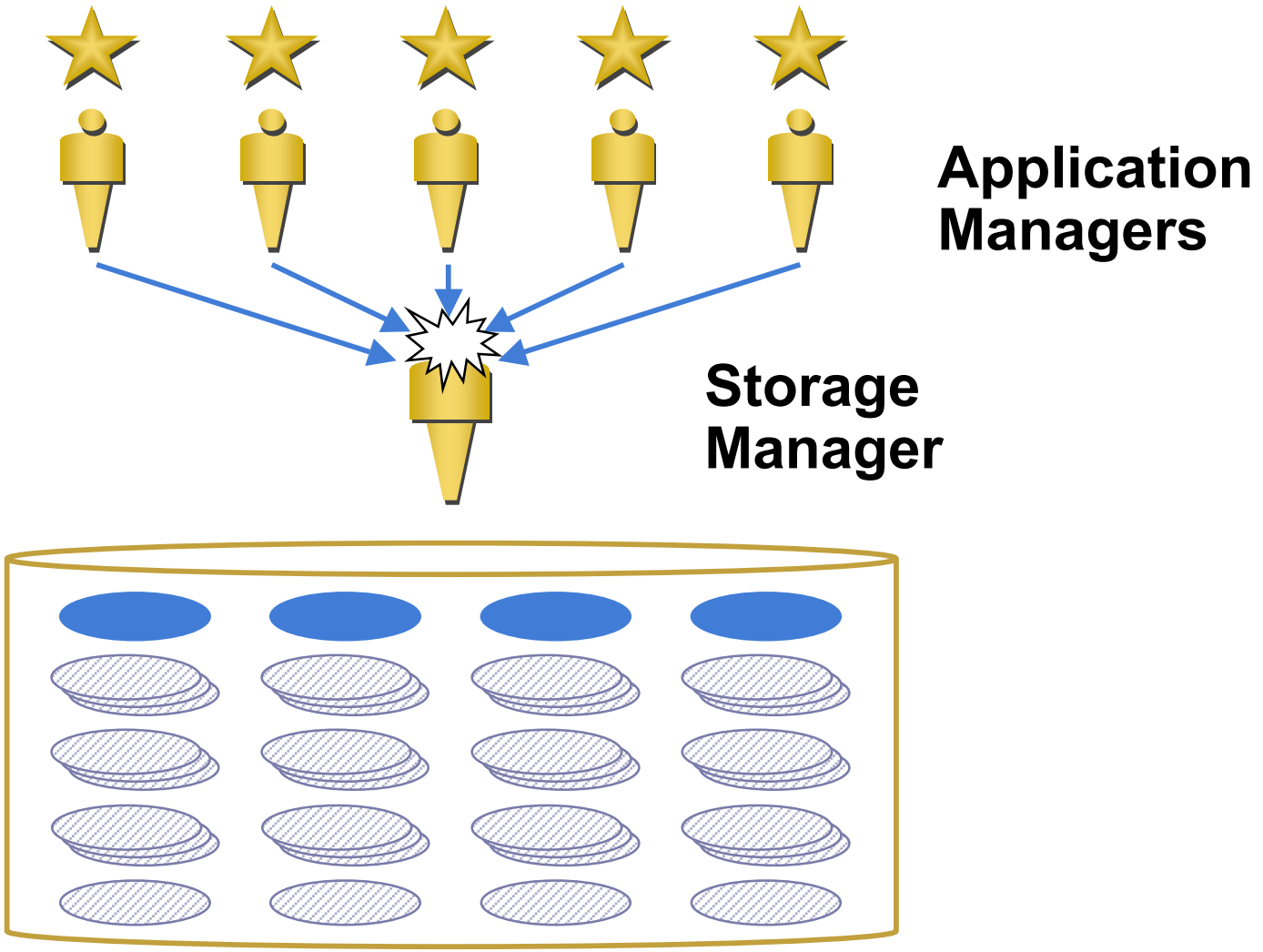
Non-copy Data Properties



Unified Protection & Enablement Environment



The Storage Admin's Challenge



Managing The Copies

1 Oracle database:

- 17 tables on Primary
- + 17 tables on remote DR site
- + 17 mirror relationships between primary and DR
- + 17 tables on secondary dev & test
- + 17 mirror relationships between primary and secondary
- + Backups
- + Archive copies

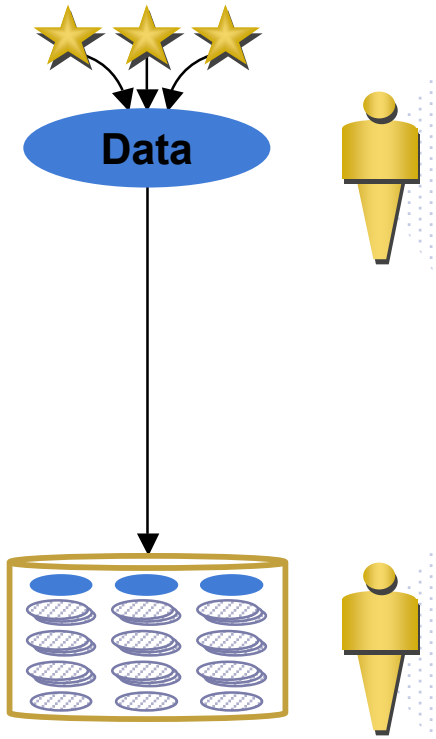
Or

- 1 Dataset

What's a "Dataset"?

- **Dataset: A collection of data meaningful to the user or data administrator having similar properties**
  - **A set of database tables**
  - **A home directory**
  - **A server root LUN**
- **Datasets have properties**
  - **Redundancy, Disaster recovery**
  - **Compliance, Saved versions**
  - **QOS**
  - **Security, Access control**
  - **???**
- **Datasets can span storage servers**
  - **A higher level of abstraction allows automation**

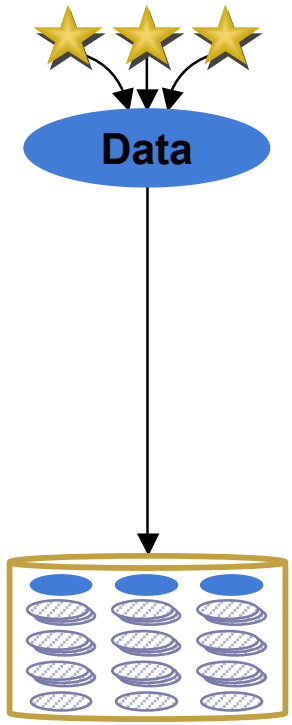
Simplification Through Integrated Data Management



- Application admins set data properties
- Properties assigned to logical sets of data
- Properties define business requirements for data

- Storage admins create & manage processes
- Processes deliver on data requirements
- Automation & service delivery become possible

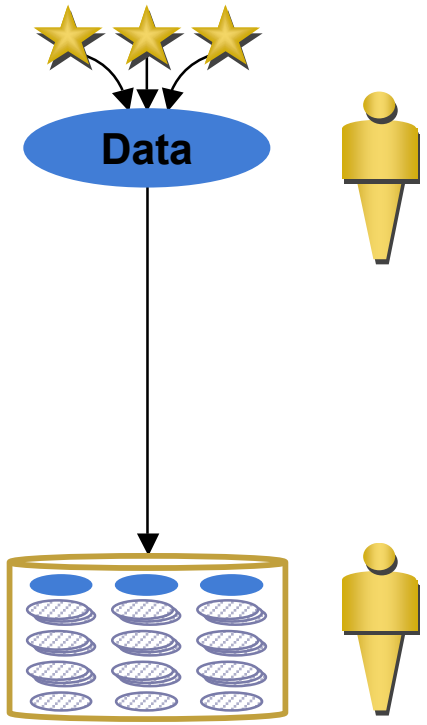
Simplification Through Integrated Data Management



- Recovery time objective: 0 sec
- Applicable regulations: SEC-17A
- Security level: high

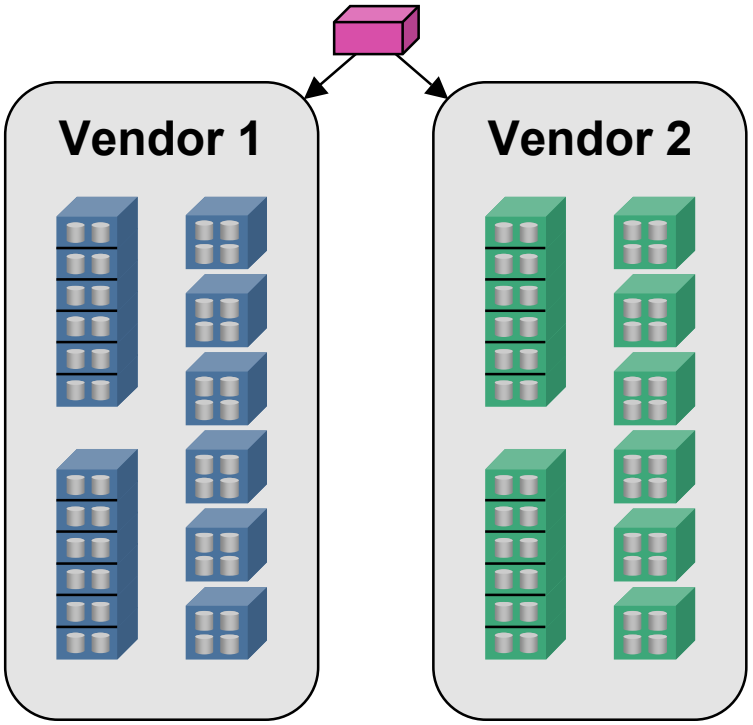
- Low RTO: use synchronous mirroring
- SEC-17A: enable SnapLock; delete after 7 years
- Hi Security: enable encryption

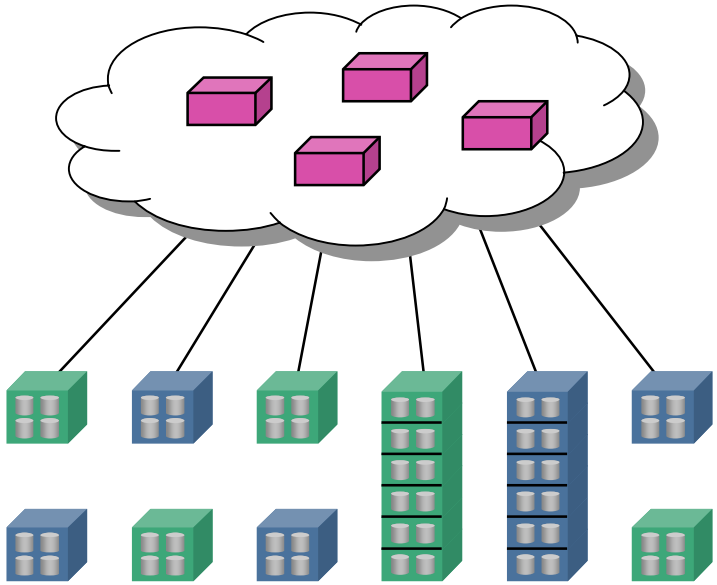
Simplification Through Integrated Data Management



- Right decisions are made by the right people
- Easier to change and automate
  - Goal: Automate 80% of workflow
- Data properties can remain constant while processes adapt to new technologies

“Two Worlds” vs. Storage Virtualization Architecture





Long Term Trends

- Unification of capabilities in a single storage infrastructure
- Property-based dataset management adopted for simplification and automation

It's starting to happen now

- Unified model
- Scale-out & Grid
- Value-added copies
- Virtualization
- Data sets & properties
- Heterogeneous replication

Summary

It’s good to be in storage!