9th European TDWI Conference
Munich June 15th - 17th

Information Lifecycle Management

- Optimization of Large Data Warehouse Systems -

W2P

Agenda

- Drivers for Information Lifecycle Management (ILM)
- Definition of ILM
- Data Warehouse Challenges
- ILM in Data Warehousing
- Enterprise Data Warehousing

Agenda

- **Drivers for Information Lifecycle Management (ILM)**
- Definition of ILM
- Data Warehouse Challenges
- ILM in Data Warehousing
- Enterprise Data Warehousing

Challenges Facing the Information Lifecycle

- Data growth
  - Emails, attachments, Web sites, audio and video content, voice recordings …
  - Constantly increasing data volumes in BI
- Direct access capabilities
  - Predictable residence time for ERP data
  - Long-term direct accessibility appreciated, ad-hoc analysis needs
- Legal requirements
  - SEC, FDA, HIPPA, SOA, GDPdU, Basel II for ERP data
- Data value during lifecycle
  - Constantly decreasing in ERP environments
  - More long lasting in BI environments
- Costs
  - Personal costs, technology costs, process costs
- Technological innovations
  - ATA disks, blue laser, etc.
  - Write-once file system, NLS, etc.

The challenges cannot only be addressed by purchase of additional memory.

An effective administration of the data is necessary.

Information Lifecycle Management (ILM)

Everyone's talking about ILM

ILM is not only software

ILM is not only a tool

ILM is not only a product

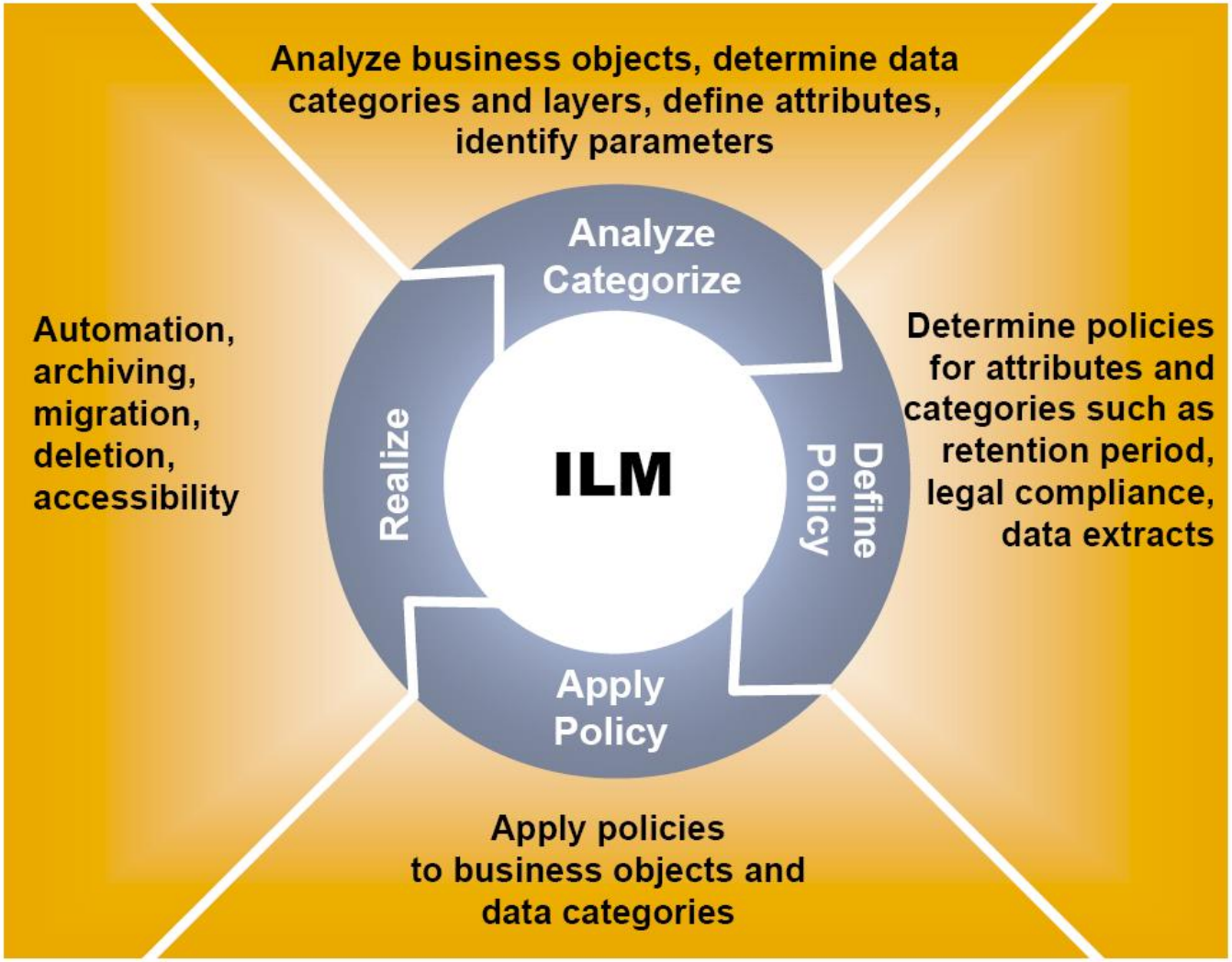
ILM is a storage management concept that actively manages all information objects during their entire lifecycle

ILM is a combination of processes and technologies with a goal of providing the right information at the right time in the right place with the lowest possible costs

Information Lifecycle Management

- ILM is a **combination of processes and technologies** whose goal it is to provide the right information at the right time at the right place with the lowest possible costs over the required life time of the data.
- The **new** and **added value** delivered by Information Lifecycle Management is **automation** and **completeness**!

Information Lifecycle Management – One View

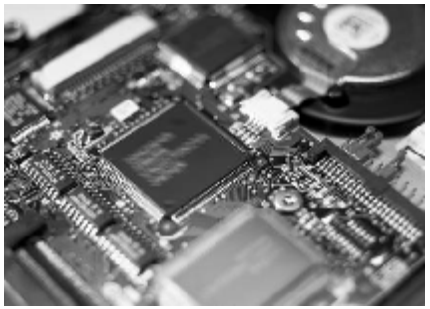


Drivers of ILM

Why ILM?

External Drivers	Internal Drivers
– Legal Retention Requirements – Product liability – Lawsuits (Legal holds, e-Discovery) – Tax Reporting, Audits – New technologies	– High costs for hardware and administration – Policies and service level agreements – Risk of litigation – Company-specific processes – System landscape harmonization/centralization – Mergers and acquisitions

Reduced Data Volumes



Legal Compliance

Reduced Risk

Reduced TCO

Number of Legal Requirements Keeps Increasing

Laws and regulations affecting ever more media and data types

Laws and regulations growing across industries and countries

Laws and regulations requiring data to be kept for longer periods

What is Driving ILM?

> With more-stringent corporate controls, an increasing number of companies are adding chief compliance officers to their boards.
>
> *- Computer Weekly*

> Doing business in compliance with applicable laws, rules, and regulations is deeply rooted in SAP history. To further demonstrate our ongoing commitment in this area [...] we are pleased to announce the establishement of a new Global Compliance Office.
>
> *- SAP Announcement, Feb 2007*

The Challenge

"With projected compounded annual growth rates for databases exceeding 125%, organizations face two basic options:

1) Continue to grow the infrastructure (e.g., server size, storage capacity)

OR

2) Develop processes [and architectures] to separate dormant [archive-ready] data from active data."

Meta Group Report

Databases on a Diet

The Challenge

"In the compliance age, the answer lies in any technology which meets all three of these criteria:

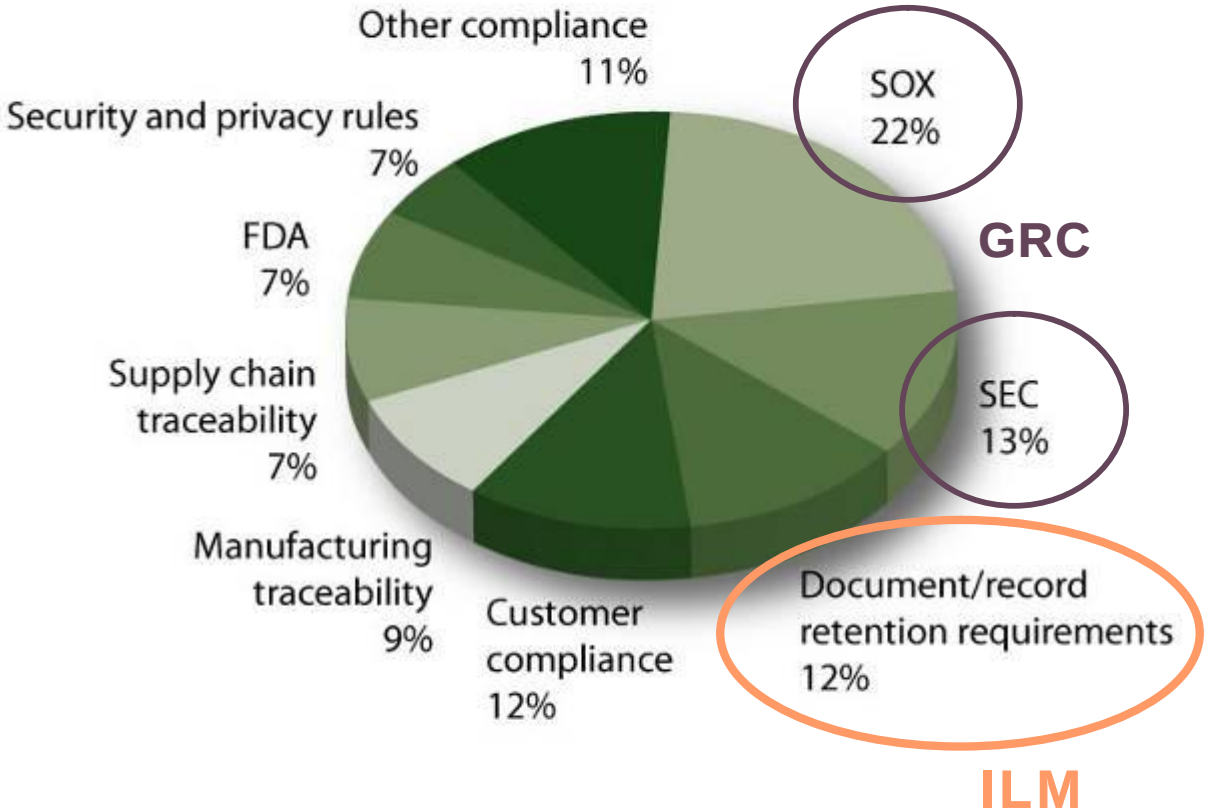
- Large Stored data volume
- Quick Availability
- Fast Query Response Time

and can do so within the seven-figure cost range"

SOX Journal 2005

Top 8 Programs by Spending

Figure 1: Compliance spend by major initiative



Source: AMR Research, 2006

Compliance spending in 2007 is estimated to be around $28B

Constantly Increasing Database Volumes

Meta Group

"70% of all corporate data is currently located in databases"

Gartner Group

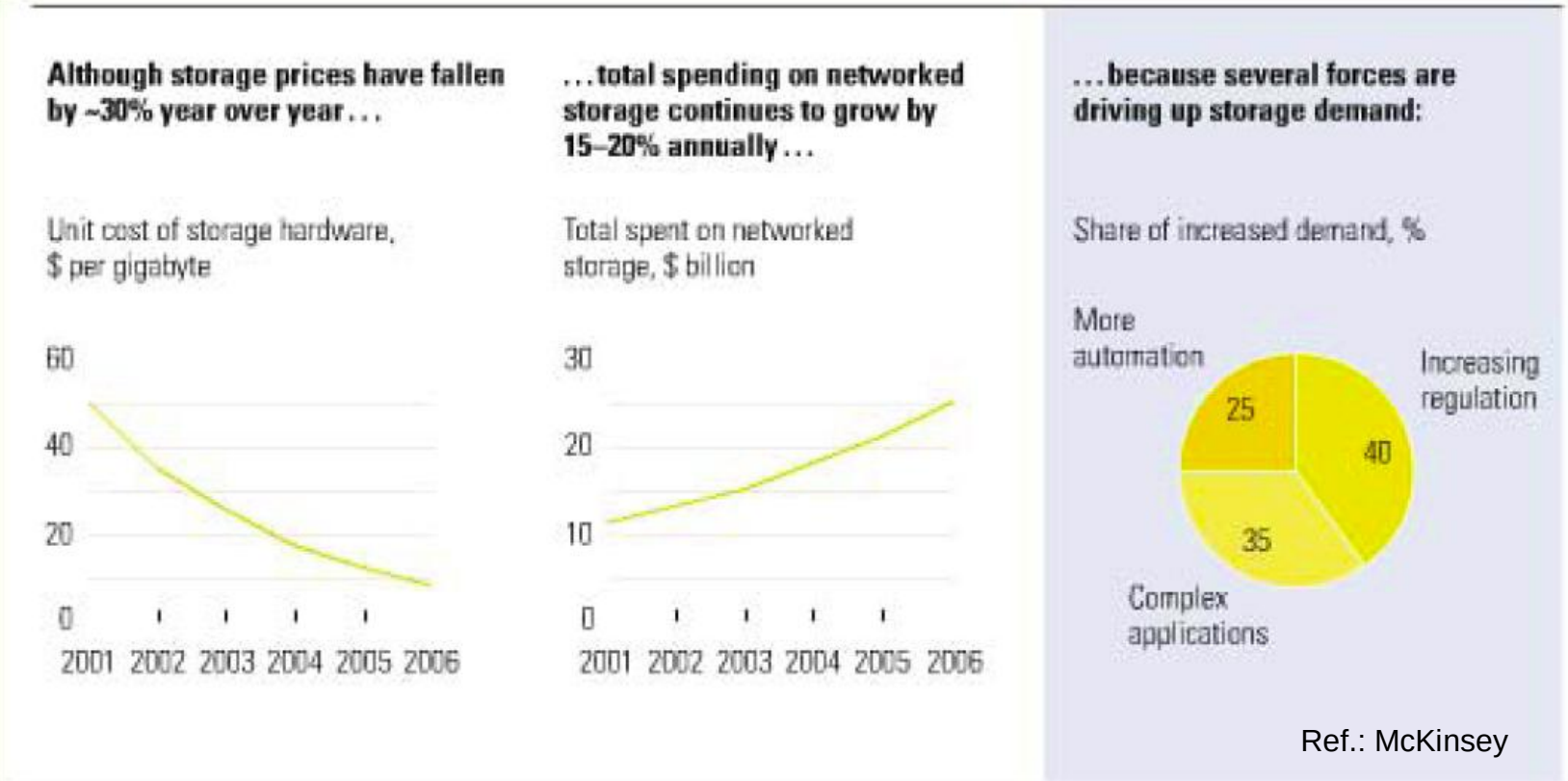
"Databases with multiple terabytes are already a reality – in the near future these will have hundreds of TBs"

"Business Applications – such as those from SAP – play a significant role in these growth in memory needs"

"Growth rate is 64%"

Total Coporate Spending on Storage …

… (disk drives, tape systems, specialized network gear, and the people and software to manage them) grows by 15 to 20 percent every year, even though the unit cost of storage drops by about 30 percent annually



Ref.: McKinsey

Agenda

- Drivers for Information Lifecycle Management (ILM)
- **Definition of ILM**
- Data Warehouse Challenges
- ILM in Data Warehousing
- Enterprise Data Warehousing

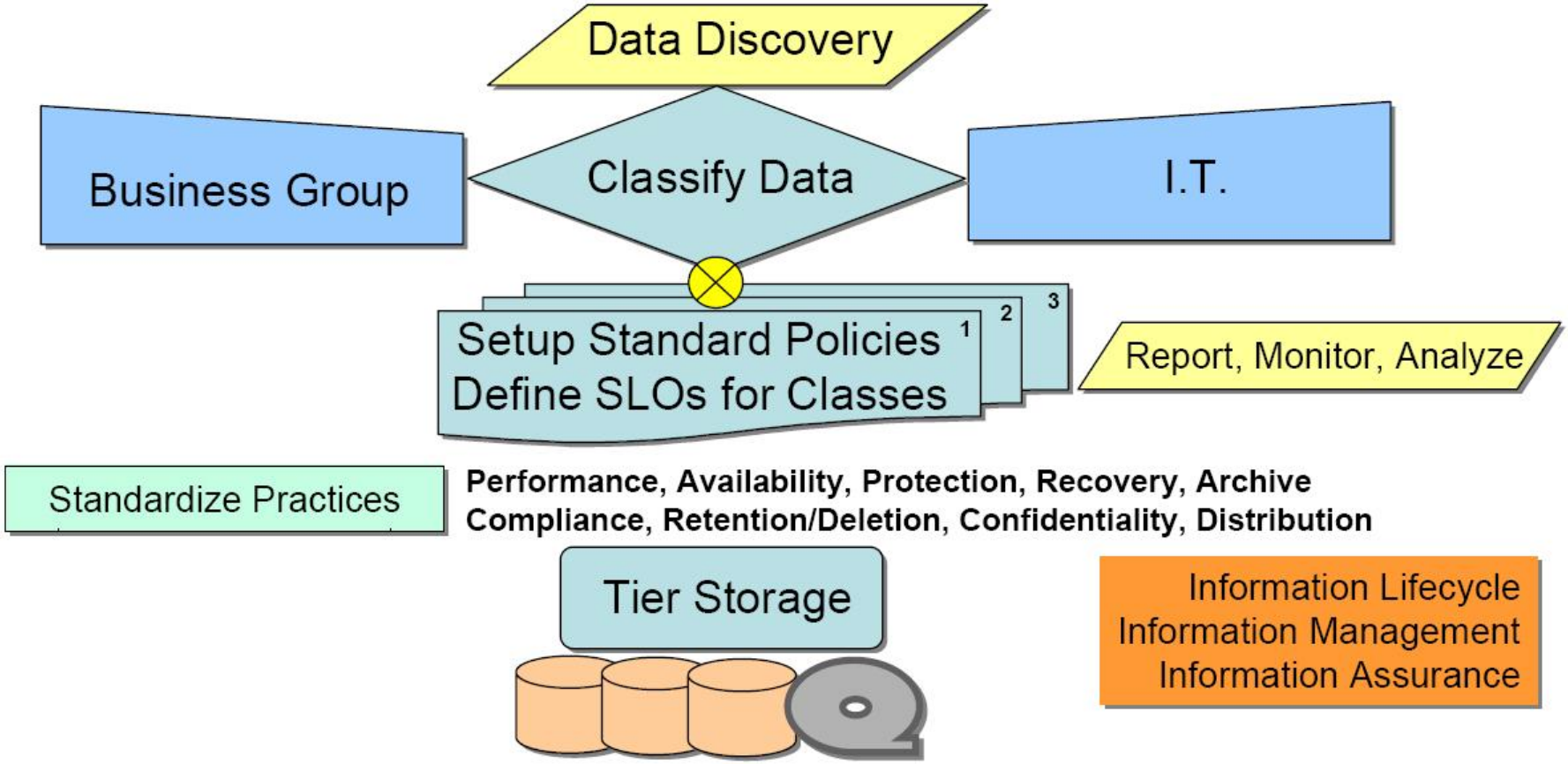
SNIA: Vision of ILM

- SNIA: Storage Networking Industry Association
  - not-for-profit organization
  - common goal: to set the pace of the industry by ensuring that storage networks become efficient, complete and trusted solutions across the IT community
- Data Management Forum
  - initiative of the SNIA
  - focused on building a community of I.T. professionals, integrators and vendors for the purposes of being the leading authority and resource on data management infrastructure and information lifecycle management
- Vision for Information Lifecycle Management
  - A new set of management practices based on aligning the business value of information to the most appropriate and cost effective infrastructure

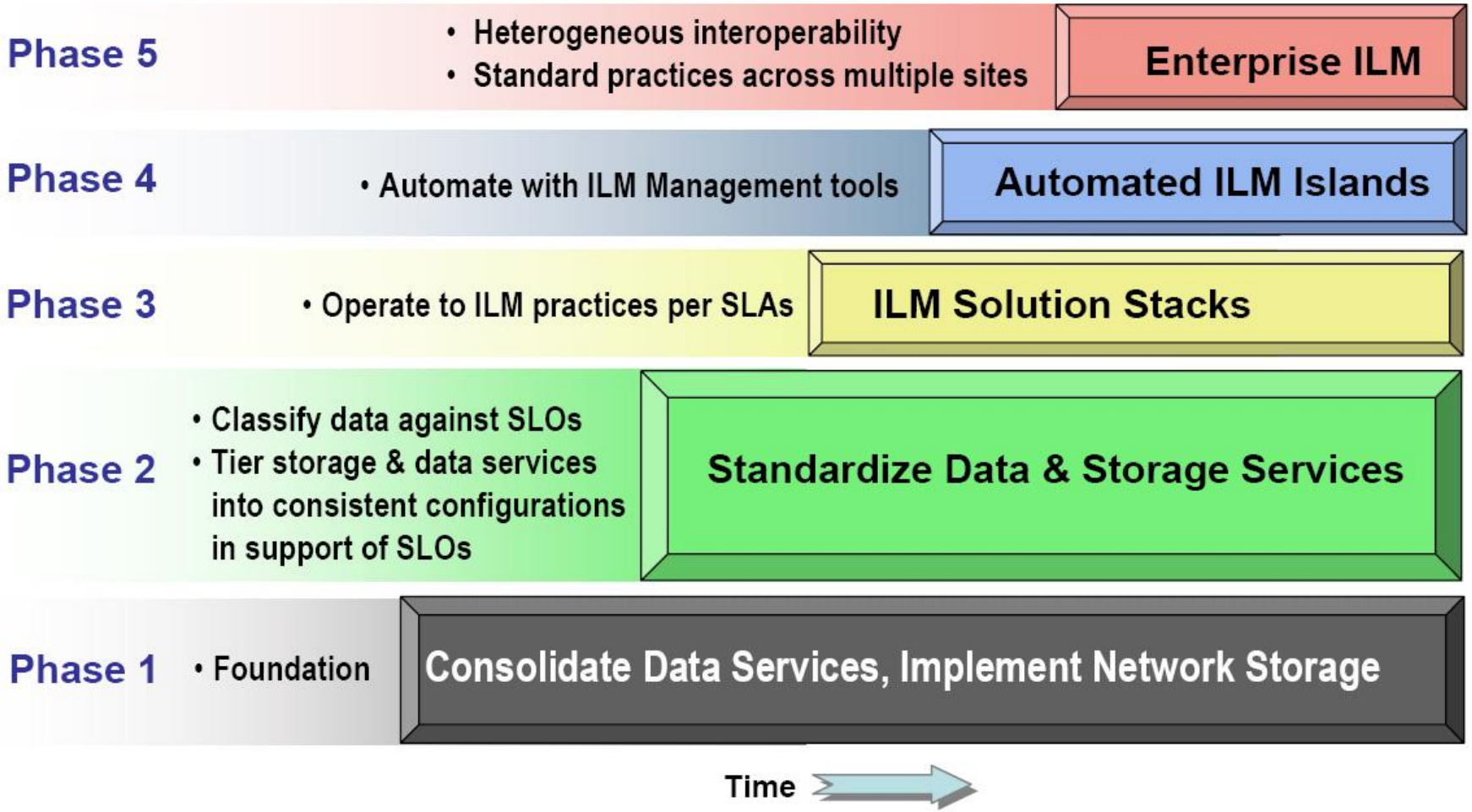
SNIA's Definition of ILM

Information Lifecycle Management is comprised of the policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost effective IT infrastructure from the time information is conceived through its final disposition. Information is aligned with business requirements through management policies and service levels associated with applications, metadata, and data.

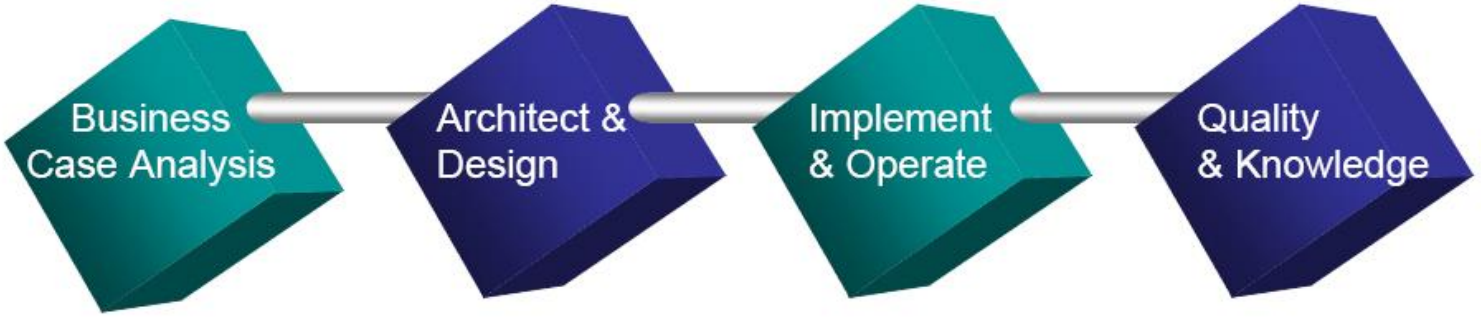
Implementing ILM according to SNIA



SNIA ILM Roadmap



Process-based Approach



Stage 1

- To provide Corporation with the necessary business and IT services
- Financial projections, financial metrics, and assessment of contingencies and risks
- **Engage stakeholders**: Gain support across the corporation to enter into an ILM driven initiative to reduce costs thru consolidation and operational improvements

Stage 2

- Business & IT Alignment Map
- Architecture Requirements Defined
- Operational Policy & Process Model
- Data Classification Policies defined
- SLA Validation & IT Services / Data Classification Menu
- Integration Roadmap

Stage 3

- Project Implementation and Integration
- Consolidation of Operational & System Objects
- Business Case Metrics Validation & TCO Model Completion
- Data Policy Model Validation
- Test, Measure & Validate

Ongoing

- Training & Skills Competency Model
- Professional Certification
- Quality Assurance Program
- Corporate & Market Communications Program

What is Data Classification?

- Organization of data and information into groups for management purposes.
  - Allows IT to create multiple service level offerings
  - Allows LOB to select services based on value of data
  - May use software to enable some of the process
- Represent corporate requirements:
  - Security officer: Secret, confidential, proprietary, …
  - Records Manager: retention time, …
  - Compliance officer (HIPAA, SOX, …): authorization, retention, …
- Represent LOB requirements:
  - Application performance, availability, recoverability, …
  - Staff response time, asset reporting, …
- IT Organization needs data classification:
  - Method to rationalize requirements into service level offerings

How is Data Classified?

Data Classification, Policies, SLOs are tightly coupled

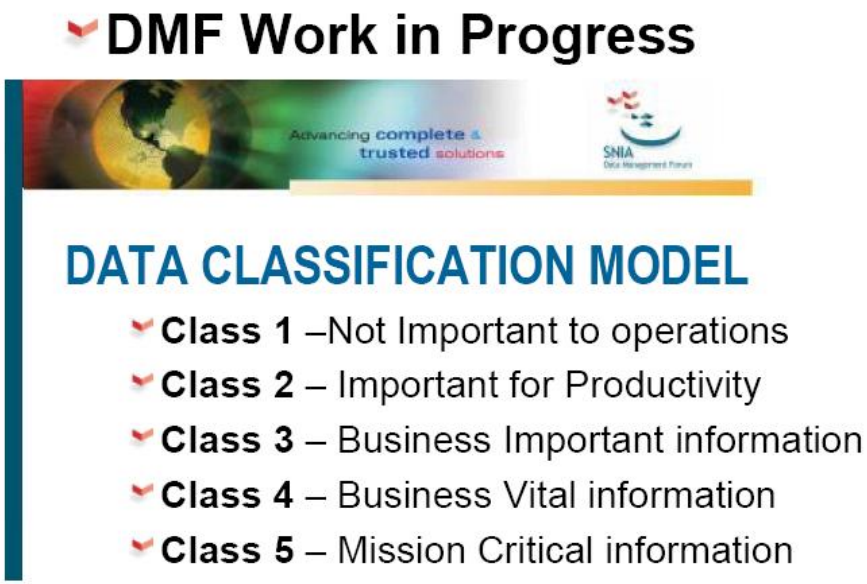
- Classify by application
  - All data from a specific App assigned same classification
  - Simple; good start; a first approximation
- Classify by groups of data
  - Production or Process data
  - LOB, Department, Owner, Customer, …
  - Compliance requirements by regulation type
- Classify by metadata
  - Time last accessed, date created, type of data, author, etc
- Classify by content
  - Content-filtering for compliance, grouping, risk classification
  - Security and Data classes can merge

Classification allows Classes of Service

- Define Service Level Objective framework
  - Class of infrastructure for performance & resiliency
  - Availability requirements (99.xxx%)
  - Data Protection & Recovery classes (RTO-RPO mins to days)
    - RTO: Recovery Time Objective, duration of time and a service level within which a business process must be restored after a disaster
    - RPO: Recovery Point Objective, time of data loss that is acceptable
  - Archival classes (online, tape, off-site, …)
  - Compliance classes (HIPAA, SOX, …)
  - Confidentiality (in the host, in the network, on storage, at rest…)
  - Others …
- Focus on what level of service is required for data
  - Not on how it is delivered
  - Technology changes, service levels don't
- Only create SLOs that are important to your business

Sample Class Models

- Security Classes:
  - CLASS-1 Public Information, CLASS-2 Internal Information, CLASS-3 Confidential Information, CLASS-4 Secret Information, CLASS-5 Hazardous Information
    Source: U.S. Gov, ISO 17799



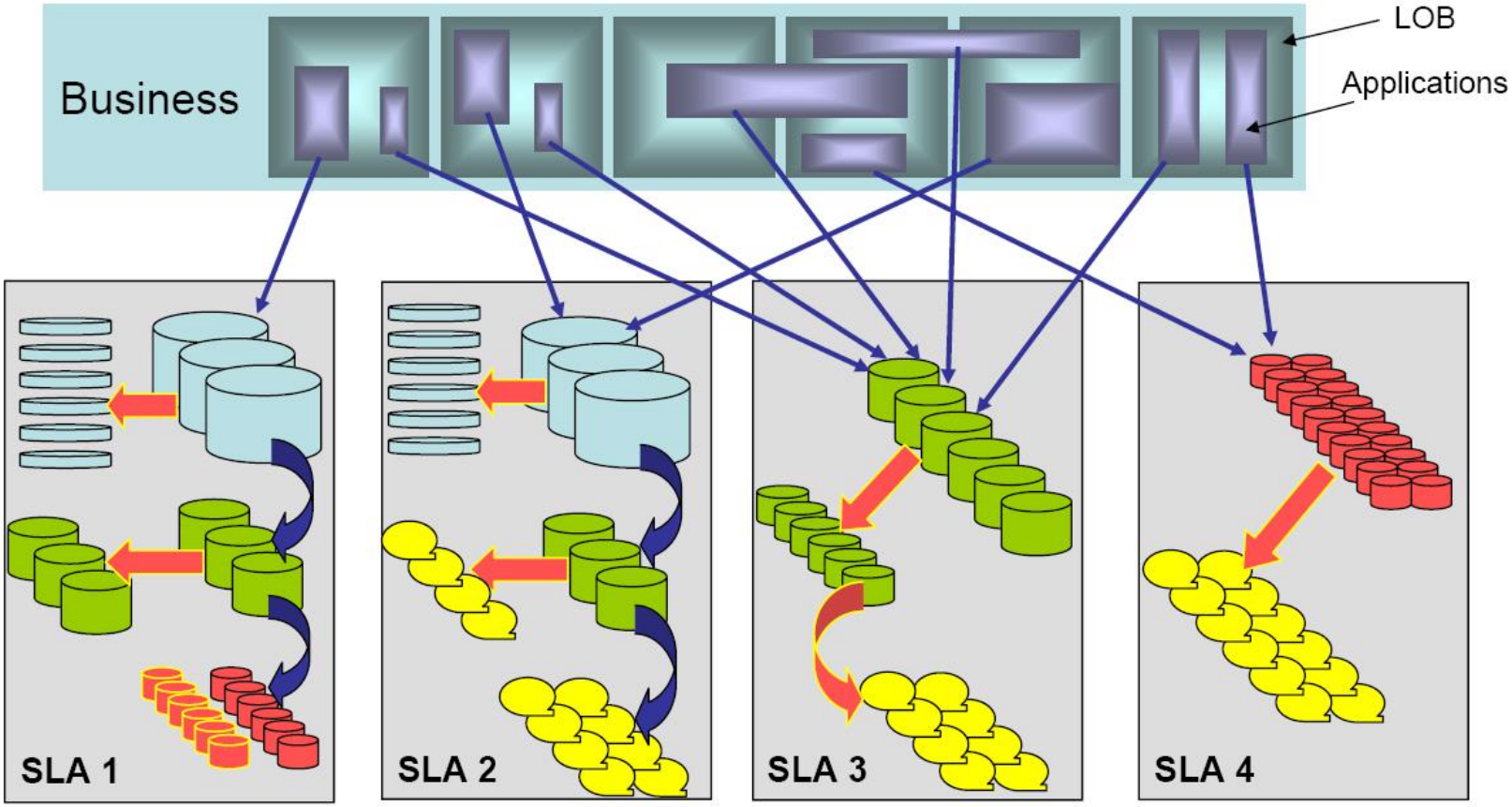
A possible Data Classification

- Critical Data
  - Needed for the critical applications
  - Loss of it represents catastrophe
- Essential Data
  - Needed for Daily Business
- Sensible Data
  - Daily Business Data that either can be reproduced quickly or that can be replaced by alternate data
- Non-Critical Data
  - Can be reproduced to low costs or duplicates exist

Acme Data Classification Menu

	Mission Critical	Business Important
Requirement	SLO Group	SLO Group
Performance	High Throughput	Medium throughput
Availability	99.999%	99%
Operational Recovery –RTO	<15 minutes	<2 hours
–RPO	<5 minutes	<8 hours
Disaster Recovery –RTO	<4 hrs	<1 week
–RPO	<5 minutes	<8 hours
Compliance (Retention)	30 yrs	7 yrs
Confidentiality	Class 4 – Secret	Class 3 - Confidential
Archive Inactive Data	120 days	180 days
Charge-back	$$$$	$$

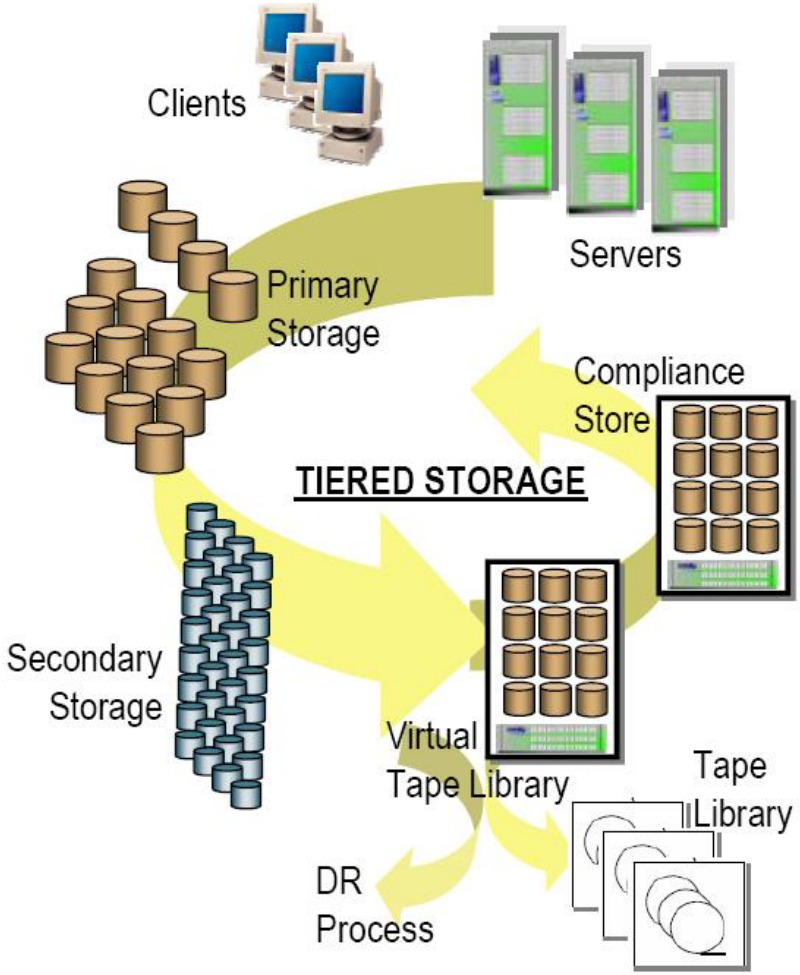
With Data Classification: Standard Configurations



à Simplified Management, more efficient, scalable

What is Tiered Storage?

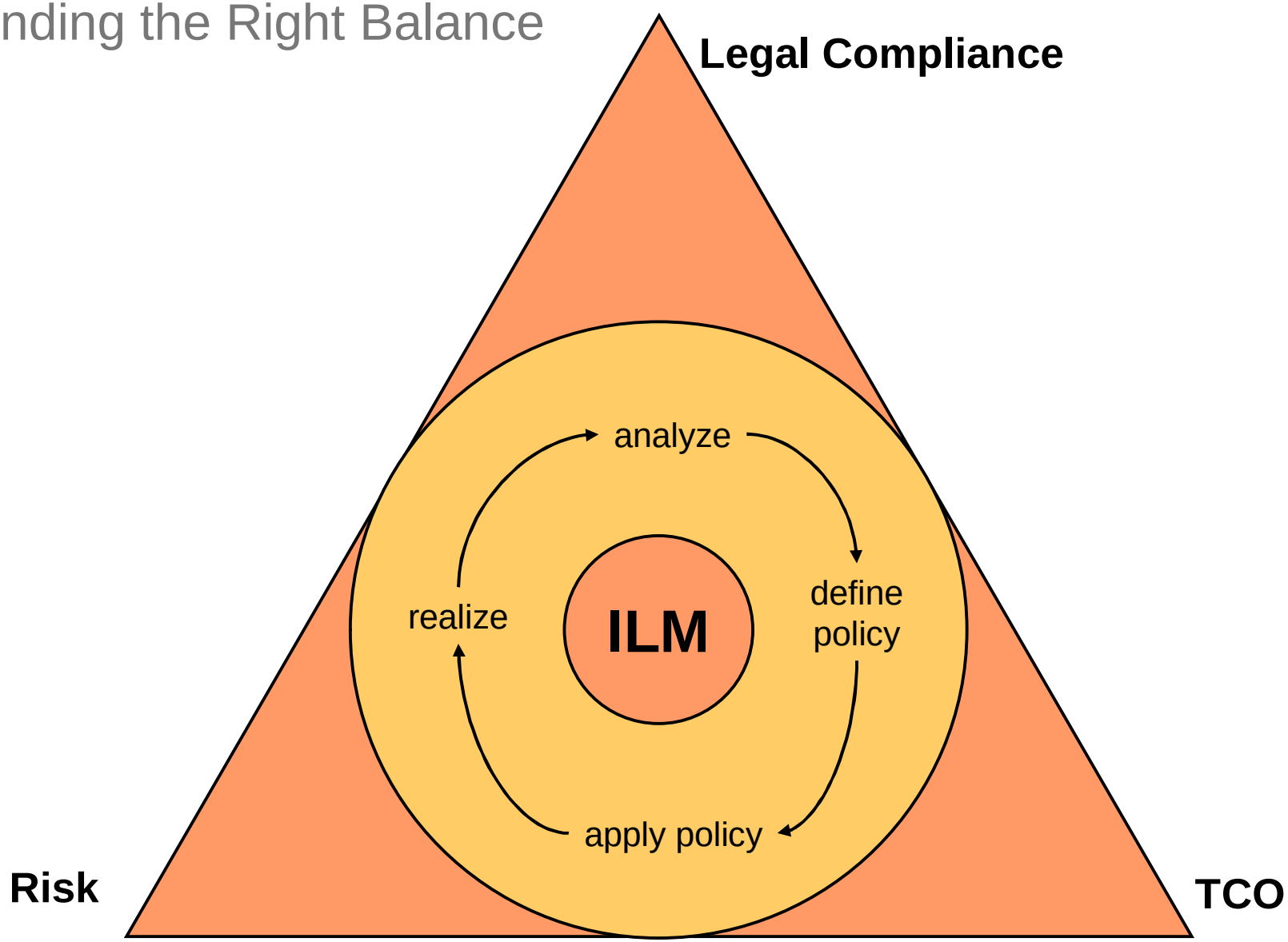
- Tiering means establishing a hierarchy of storage systems based on service requirements (performance, business continuity, security, protection, retention, compliance, etc.) and cost.
- Tiering storage requires some mechanism to place data:
  - Static: applications assigned to specific tiers
  - Staged: batched data movement (e.g. archive)
  - Dynamic: some active data mover (e.g. ILM policy services)



Legal Requirements for Data Extraction

- **Germany**: §§ 146, 147 of German Fiscal Code “GDPdU” enactment of 07/16/2001
- **Switzerland**: Decree by the Federal Department of Finance (FDF) on electronically transferred data and information of 01/30/2002 (“ElDI-V”)
- **Austria**: §§ 131 and 132 of the Austrian Fiscal Code (BAO)
- **France**: Contrôle Fiscal des Comptabilités Informatisées (CFCI)
- **UK**: Application of special check ABAPS by HM Customs & Excise
- **USA**: IRS Revenue Procedure 98-25

Finding the Right Balance

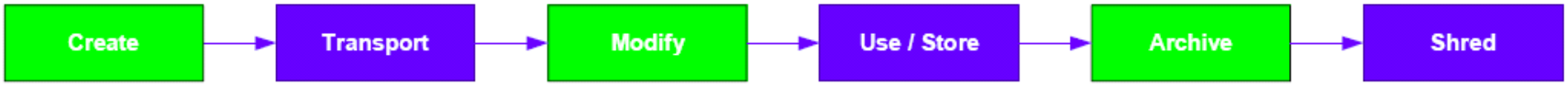


What Does an ILM Strategy Look Like in Practice?

- **Phase #1:**
  Analyze and Categorize — Get to know your data and how it behaves. Analyze data growth, determine what kinds of data you have in your system, and group or categorize that data according to different criteria and goals.
- **Phase #2:**
  Define Policy — Write out guidelines and rules for your data categories. These rules should be based on external and internal requirements for your data and should include retention schedules, archiving guidelines, destruction schedules, etc.
- **Phase #3:**
  Apply Policy — Map the policies you defined in the previous step to your data categories and prepare the system for implementation. This phase can include such tasks as customizing settings, setting up your storage system, and choosing the correct objects to archive.
- **Phase #4:**
  Realize — Turn your strategy into action. This is the phase in which you actually archive or delete data.

Classical Data Lifecycle

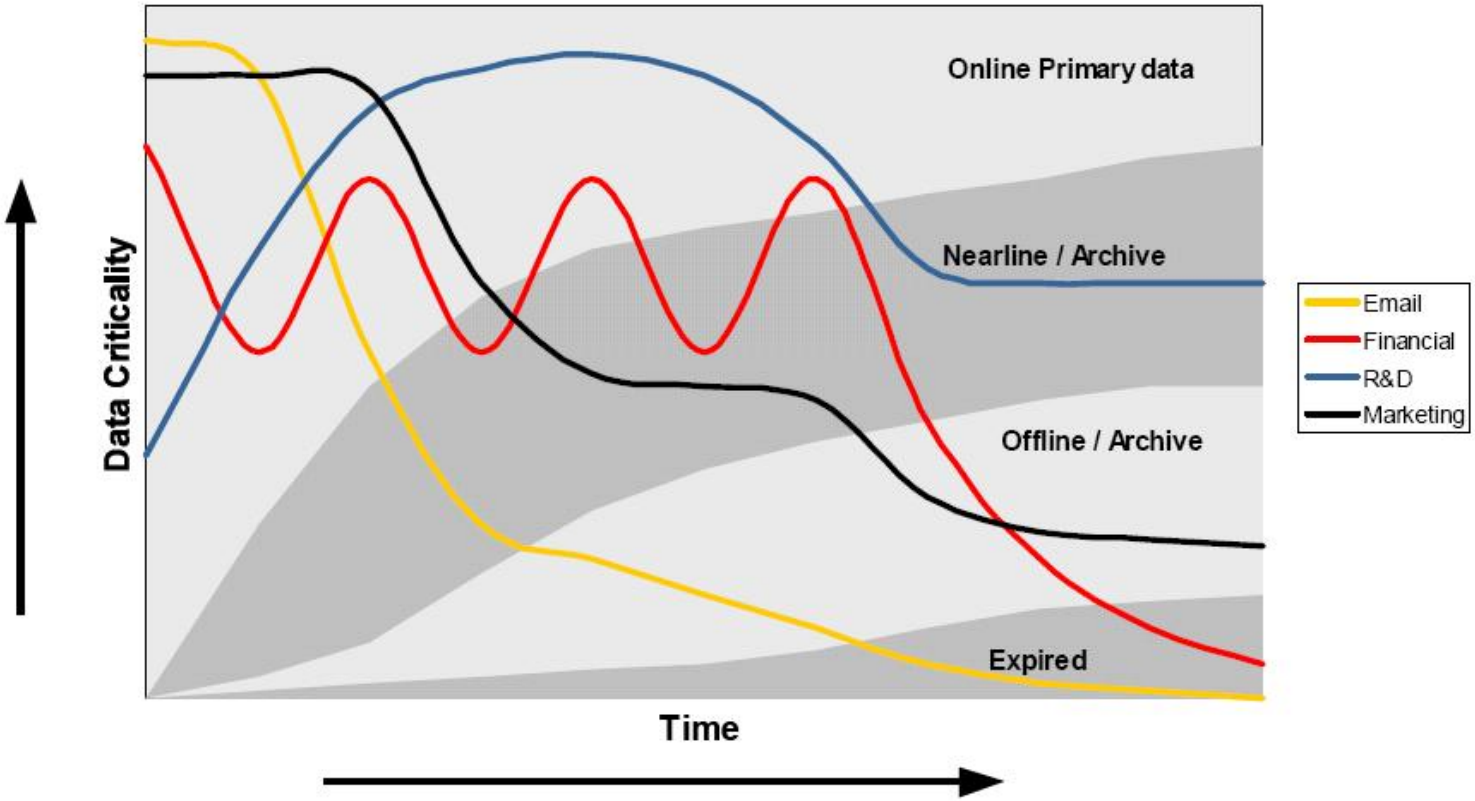
- Create
  - Annual Data Growth 30% worldwide
    (e.g. 2005 > 12 Exabyte = 12288 Pentabytes = 12582912 TB)
  - More than 90% of all data is stored on disk and tape
- Transport
  - 3.5 times more data transported than stored
- Modify
  - Only 10% of all data actually is modified
- Use & Store
  - After 30 days only 20% of the data will be accessed
- Archive
  - Often due to regulatory reasons
- Shred



Data Lifecycle Management Defined

- The process of managing business data throughout its lifecycle from conception until disposal across different storage media, within the constraints of the business process.

Data Criticality in Data Lifecycle Management



- some data is more critical than other data
- business criticality of data will change over time

Accessibility and Availability of Data: HSM and DLM

- A different class of storage does not only imply different price performance levels, but also different levels of protection, manageability, immutability, and so on
- One of the key differences between HSM (Hierarchical Storage Management) and DLM (Data Lifecycle Management):
  - HSM primarily focuses on optimizing data availability in a virtual online model (across a hierarchy of storage - typically disk and tape), whereas DLM also takes into consideration all other aspects of the data's lifecycle - including the protection levels, data retention, and destruction of data.

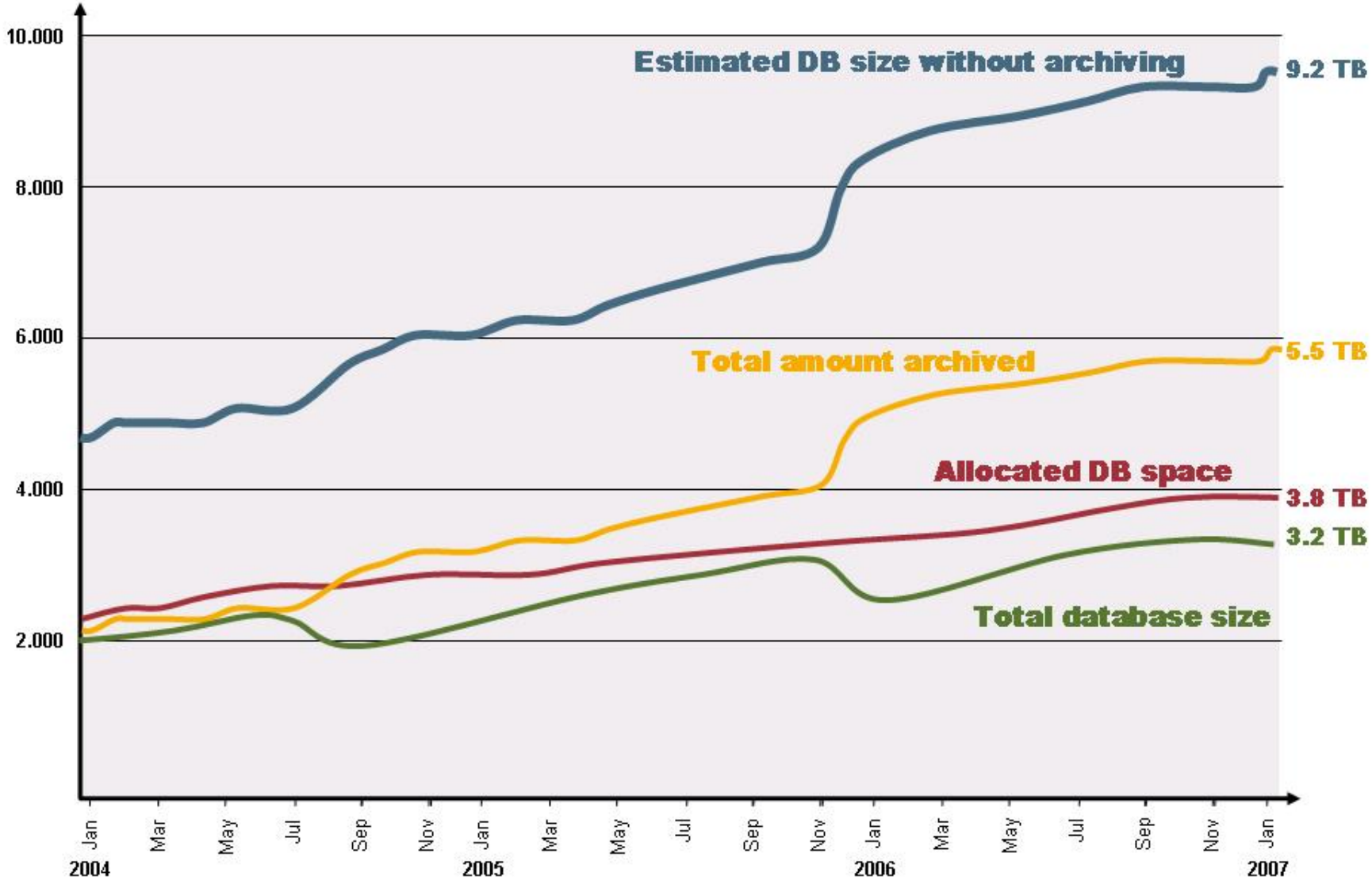
What's the Difference Between Data Management and ILM?

- How you handle data is called **data management**. To ensure high system performance, you must decide what to do with old or noncritical data. This is the main task of data archiving: to identify the data in the relational database that is no longer needed for business processes, and put it in a place where it can be stored very cheaply.
- However, ILM is not only about data, but all kinds of information. **Information lifecycle management** refers to the processes and technologies that come together to provide the right information at the right time in the right place, all at the lowest possible cost. ILM is about actively managing all information objects during their entire life cycle.
- In most cases DLM is a synonym for ILM

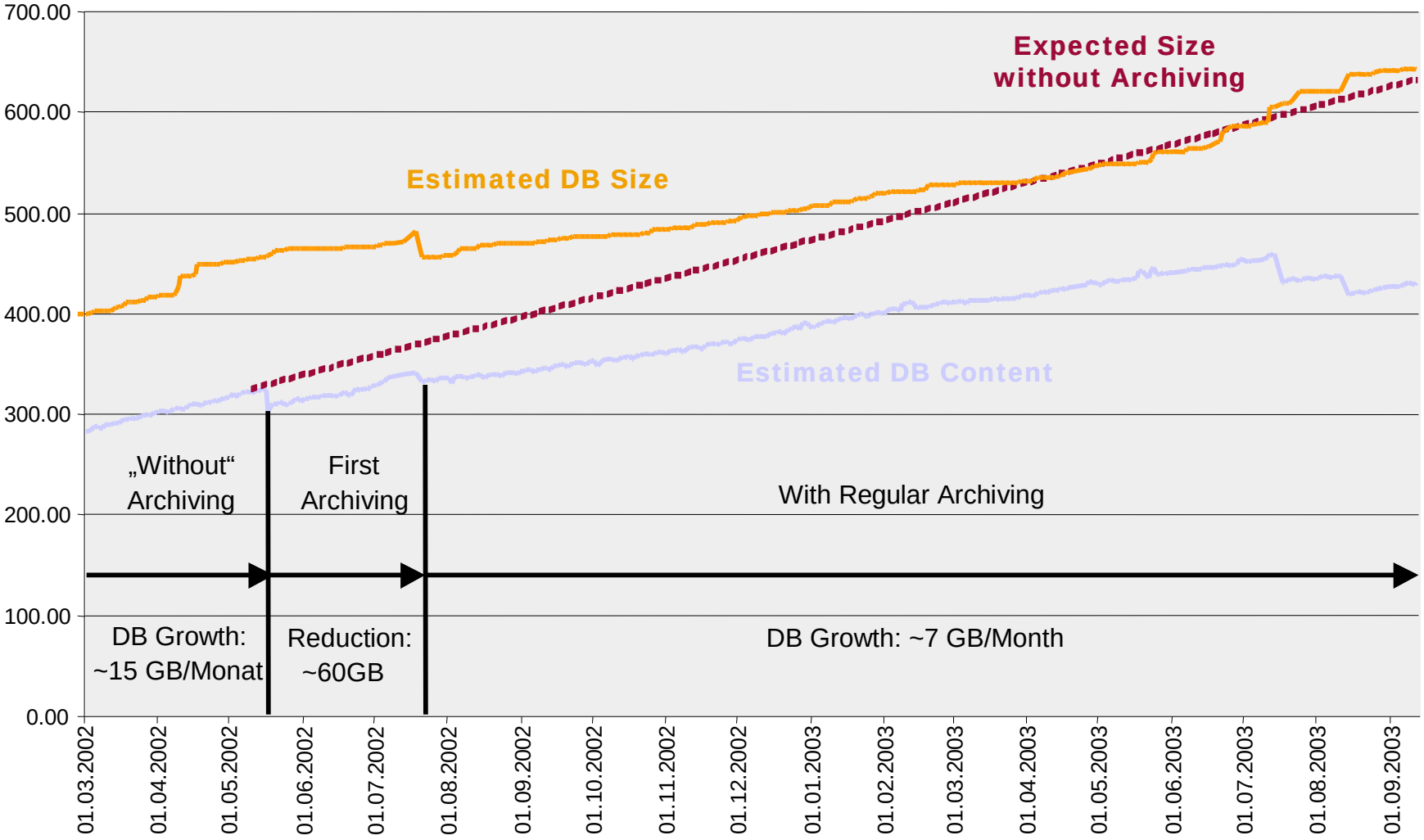
Benefits of ILM

- System Availability
  - Faster and easier upgrade to higher software releases. Shorter runtime for backup and recovery.
- Use of Resources
  - Reduced hardware costs for Disk, CPU, Memory as well as administration costs.
- Better Performance
  - Shorter response times in dialog mode for all employees.
- Legal Compliance
  - Meeting data retention requirements and setting up end-of-life scenarios.

Customer Example of Database Archiving



How Is Your Data Growth Looking?



Data Archiving/Data Management: Business Scenario and Benefits

- **System Availability:**
  - Faster and easier upgrade to higher software releases
  - Shorter runtime for backup and recovery

- **Use of Resources:**
  - Reduced hardware costs for Disk, CPU, Memory
  - Lower administration spending

- **Performance:**
  - Shorter response times in dialog mode for all employees

- **Legal Compliance:**
  - Meeting data retention requirements through archiving
  - Setting up end-of-life scenarios

Data Volume Management

Agenda

- Drivers for Information Lifecycle Management (ILM)
- Definition of ILM
- **Data Warehouse Challenges**
- ILM in Data Warehousing
- Enterprise Data Warehousing

Typical Data Warehouse Problems

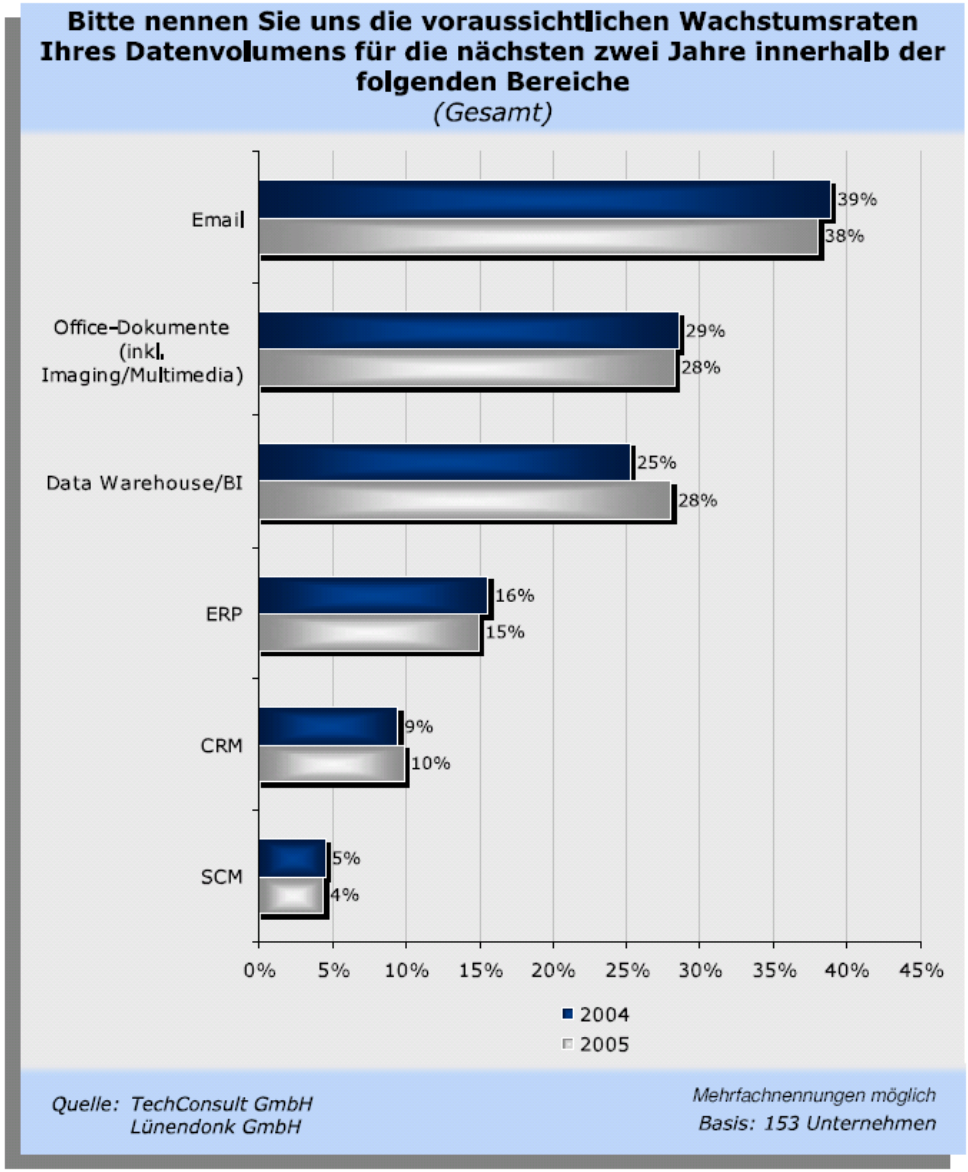
- End-User Challenges
  - Making timely, informed business decisions
    - Users cannot wait for historical data to be restored
    - Transparent access to data for regular reporting and ad-hoc analysis
- IT Management Challenges
  - Meeting end-user data demand while managing cost
    - High costs of adding/managing online disk storage
    - High costs of backup and recovery – especially when data is accessed infrequently
    - Data protection and availability

As Data Explodes….

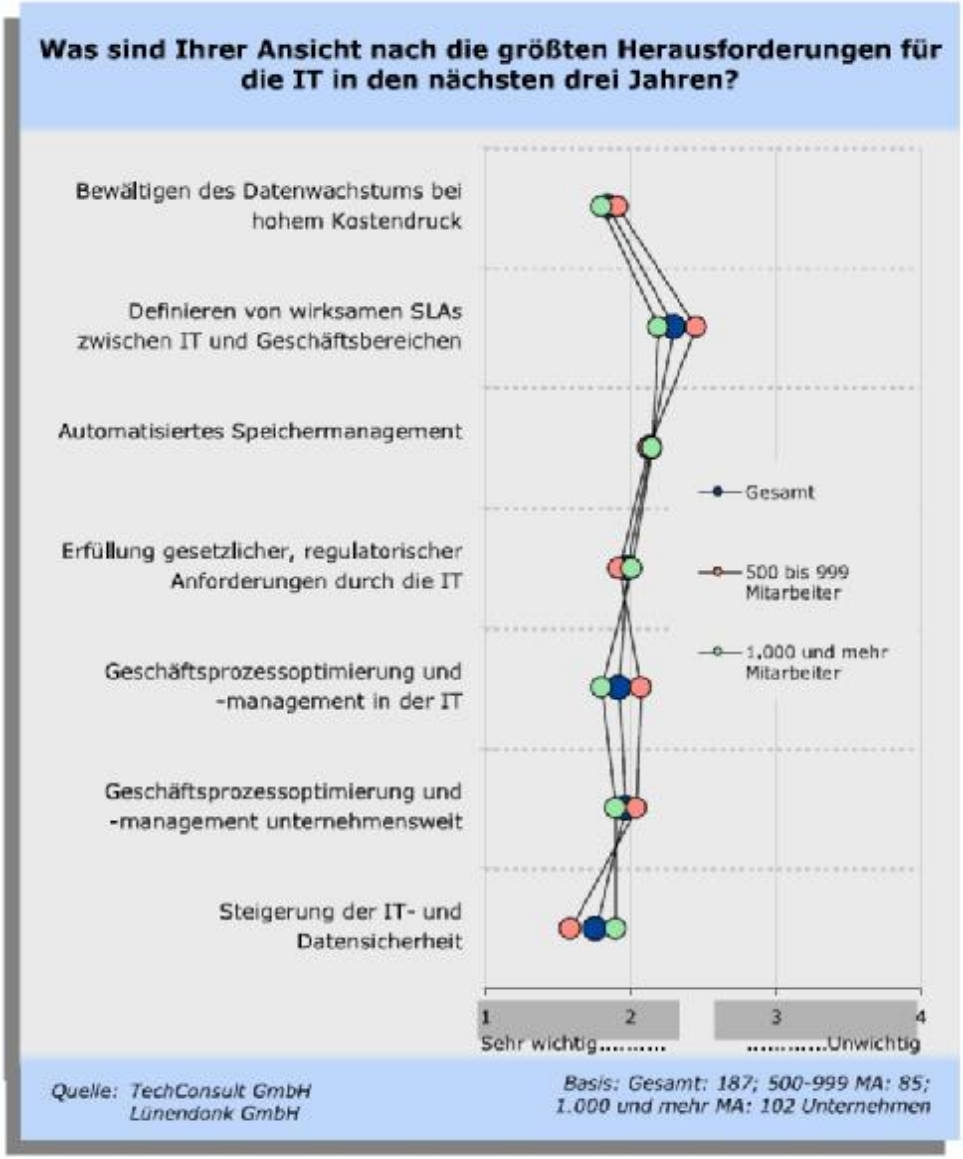
- Unprecedented data growth – *"Our warehouse was already 5TB when it went live!"*
  - Driven by business growth - more transactions, more customers, more everything
  - Driven by need to keep new types of data – IM files, web logs, RFID
  - Driven by user demands - more in-depth and on-demand analysis/reporting
  - Driven by regulatory mandates - e.g. SOX, Basel II, Data Protection Act
  - Driven by reluctance to purge data – "just in case"

- Data warehouse architecture is under stress
  - All the data you ever wanted AND high performance are incongruent goals
  - Warehouses use Aggregated Data for "Standard" Reporting
  - Increasing demand for access to more granular data, longer time periods for analysis
  - MORE users with MORE demands
  - Do more but SPEND LESS

Data Growth



IT Challenges

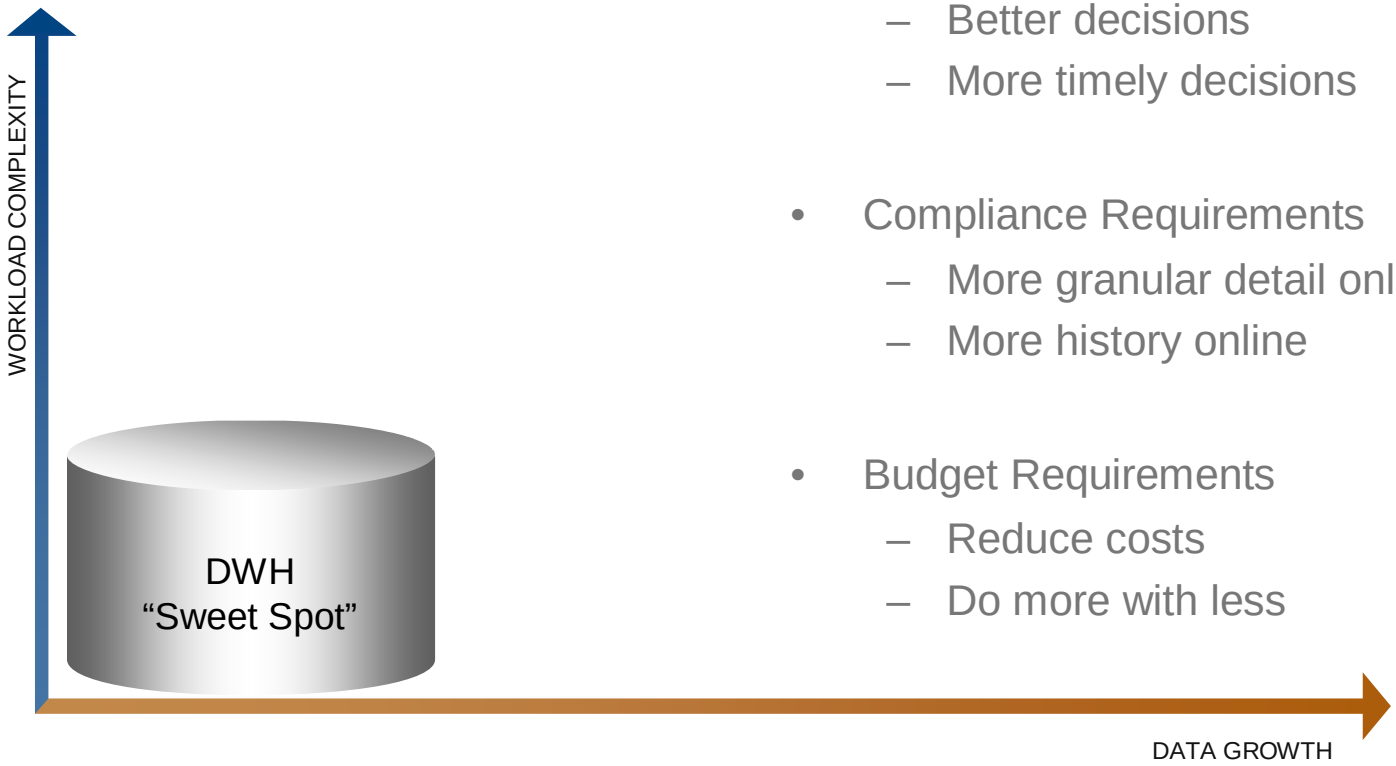


Challenges....

- “We Can’t Meet our Batch Windows”
  - Monthly / Daily Preparation of Revised KPI’s & Reporting
  - Backing Up Data
  - Rebuilding Warehouse Data
- “Our Costs are Spiraling”
  - Storage Hardware / Replication
  - Processors to Handle Storage
  - Floor Space / Power / Air Conditioning
  - Data Administration
- “The Targets Keep Changing”
  - New Business Directions
  - Special Project Demands
  - External / Internal Audit Responsiveness

Existing Warehouses Under Stress

- Increasing workload complexity – adhoc reporting, faster response, more users
- Increasing data growth – compliance, consolidation, detailed analysis



- Competitive Requirements
  - Better decisions
  - More timely decisions
- Compliance Requirements
  - More granular detail online
  - More history online
- Budget Requirements
  - Reduce costs
  - Do more with less

Data Growth Challenges

- Complex system administration required for the online relational database
  - Stress of database processes
    - Loads
    - Queries
    - Indexing
    - Deletions
    - Aggregations
    - Reorganizations
    - Data Remodeling
  - Longer backup/recovery times
  - Expensive and complex Disaster Recovery
- Increased total cost of ownership
  - New hardware
  - More storage
  - Database consultancy time for online relational database tuning

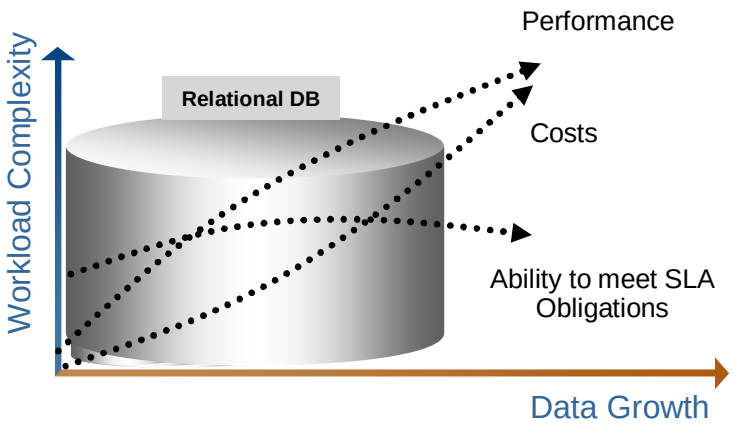
Challenges for Data Warehousing

§ **Explosive data growth & increased performance requirements**

- Corporate expansion and increased sales – more transactions, more customers, etc.
- New data types, e.g. RFID, IM, logs (transaction logs, web logs, system logs)
- Increased user expectations, e.g. for more detailed analyses for longer periods
- Data Remodeling
- More ad hoc reporting
- New legal regulations such as SOX, Basel II
- Centralization and consolidation of data warehouse systems
- “Controlled" redundancy within the EDW

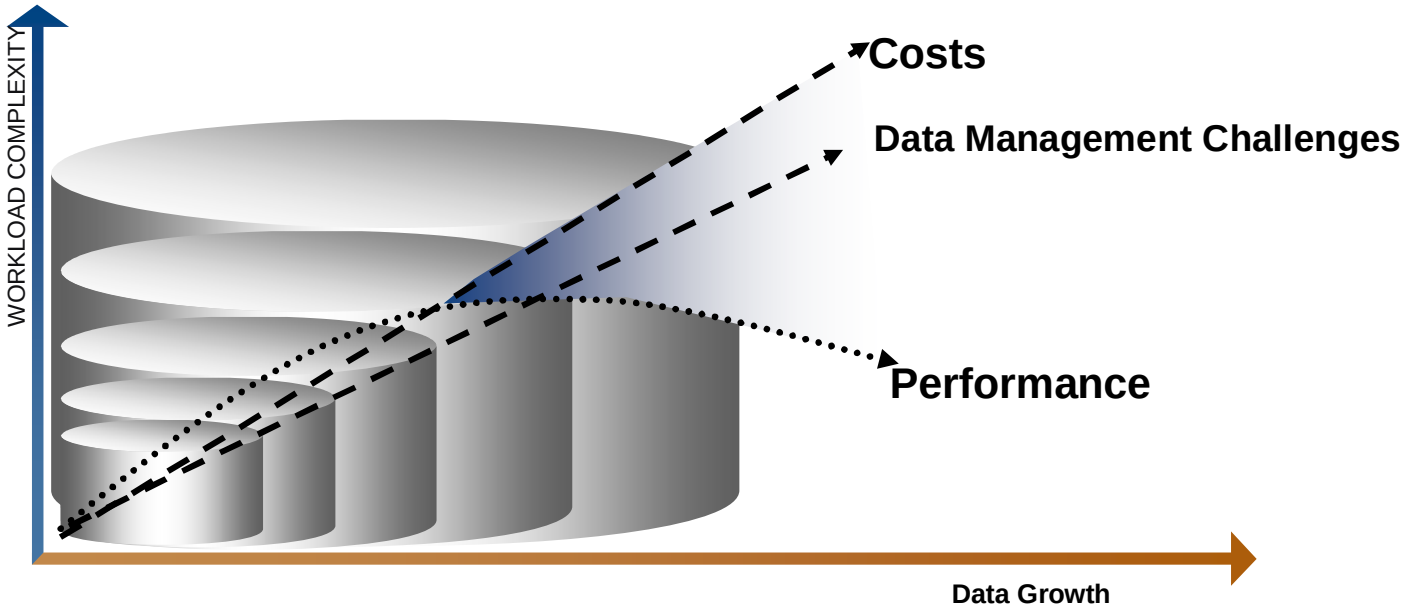
è Data warehouse management challenges

- è Decreased performance
- è increased TCO
- è Increased complexity
- è Failure to provide required levels of service



Result: Missed Service Levels

- Performance Can’t Keep Pace
- “Batch Windows” for Data Preparation Unmanageable



WHAT ARE THE OPTIONS????

Agenda

- Drivers for Information Lifecycle Management (ILM)
- Definition of ILM
- Data Warehouse Challenges
- **ILM in Data Warehousing**
- Enterprise Data Warehousing

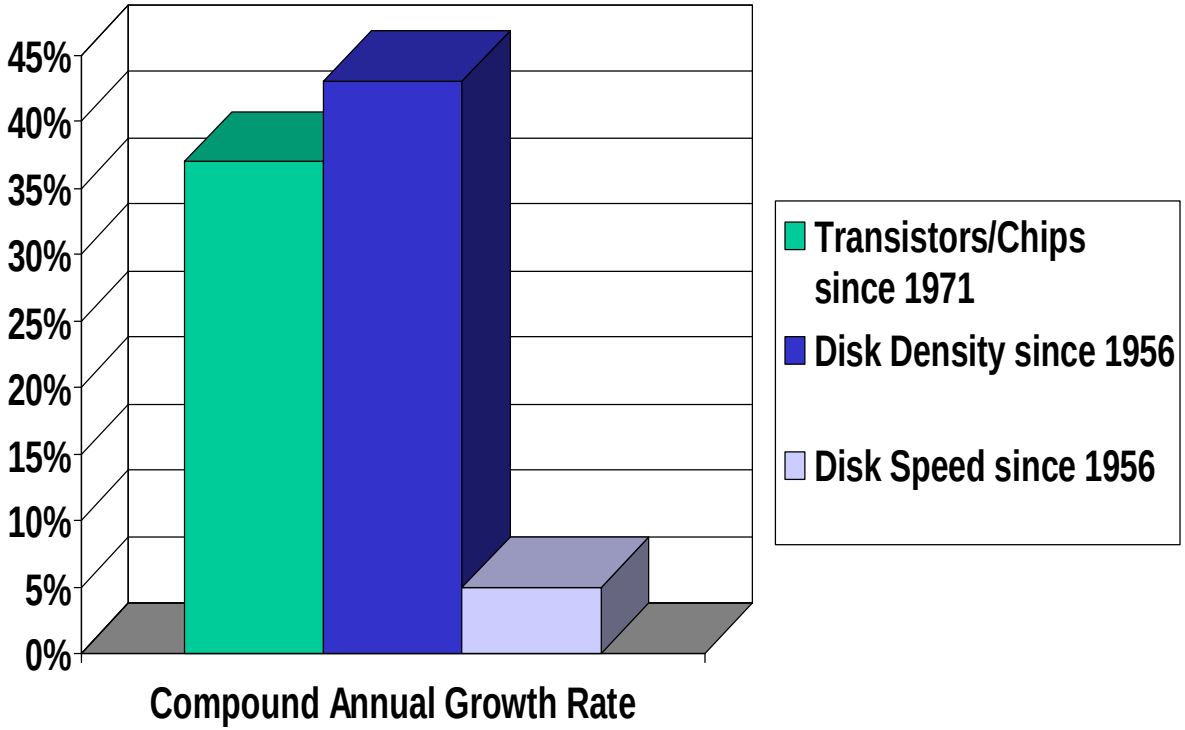
Options

- Traditional Solution: Increase the Hardware Landscape
  - Adding processing power
  - Adding Memory
  - Adding Storage capacity
- Data Model Optimization
  - Implement Summary Table
  - Use Materialized View
  - Table Partitioning
  - In Database Compression

Moore's Law, Kryder's Law, and a huge exception

Growth factors:

- ***Transistors/chip*:**
  >100,000 since 1971
- ***Disk density:***
  >100,000,000 since 1956
- ***Disk speed:***
  12.5 since 1956



The disk speed barrier dominates everything!

Source: Monash Research

Traditional Solution – Not the Answer

- Data volumes are growing faster than the price/performance ratios of disk storage technology.
- Fast disks are still expensive
- Data stored in production environments requires failover and backup technology
- For every dollar a company spends on data storage devices, an estimated additional $5 to $7 is required to manage those devices over the lifetime of the equipment

• è Total costs > $ 100.000 per TB per year

è More importantly, large volumes of data have adverse effects on system responsiveness, in areas such as:

- Ø Data loading performance
- Ø Performance of change runs, rollups
- Ø Backup and recovery times
- Ø Migration and upgrade times.

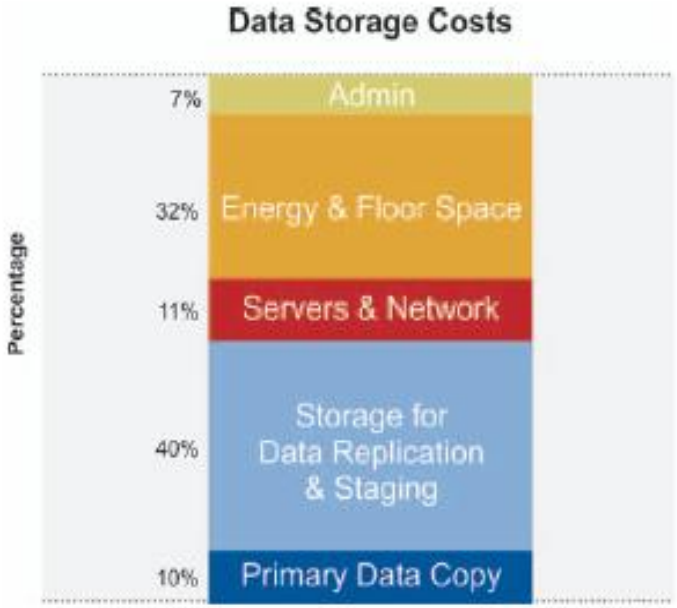
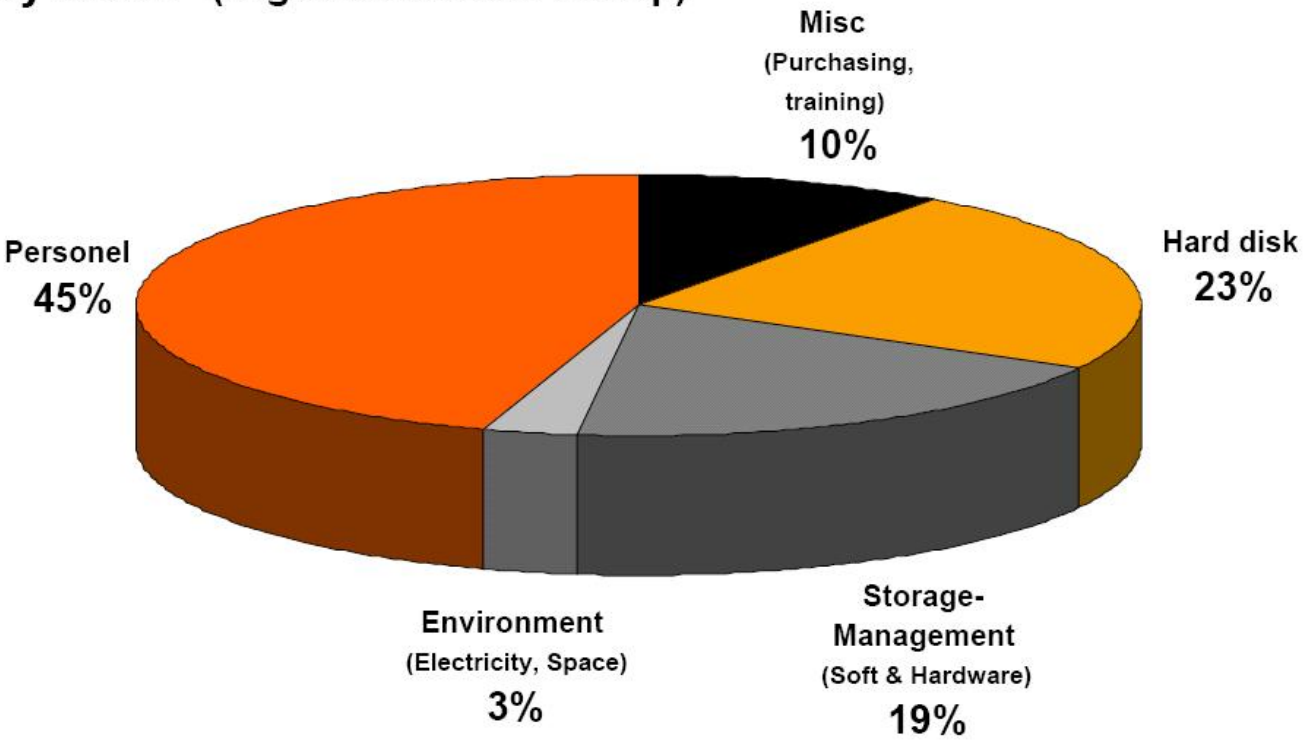


Figure 1: Breakdown of Data Storage Costs

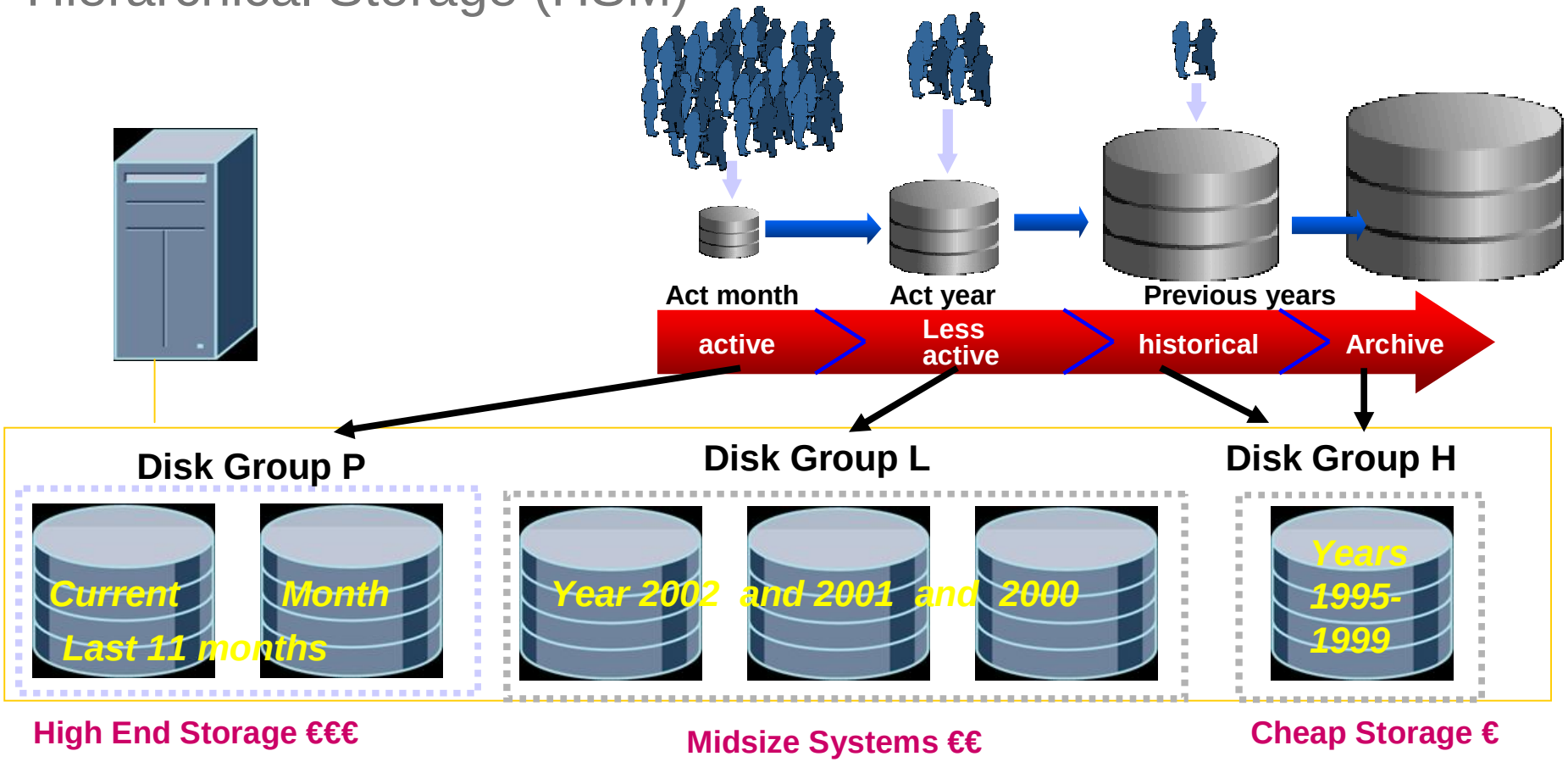
Distribution of Memory Costs

"The costs for data media don't even make up a quarter of memory costs" (Giga Information Group)



"The administrative expense for 1 terabyte of memory is five to seven times as high as the memory cost itself" (Dataquest/Gartner)

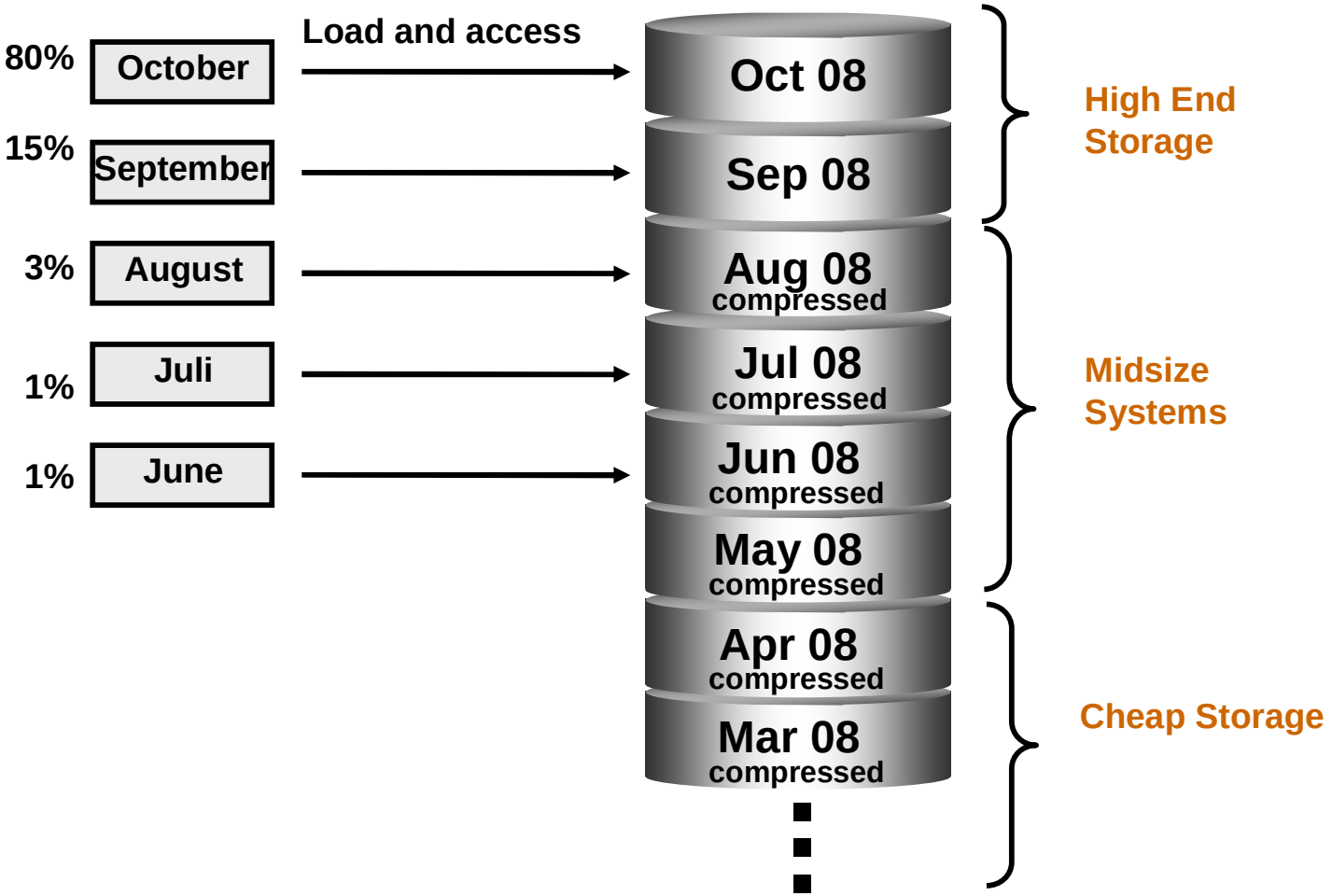
Hierarchical Storage (HSM)



- Use High End Storage for frequently used data
- Use Low End Storage for infrequently used data

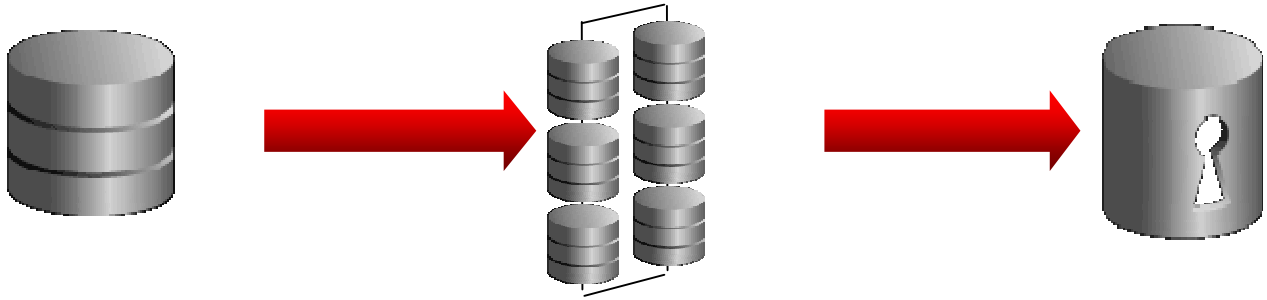
Disadvantage: the data weight is still fully present

Table Partitioning and database compression enables ILM



Example with 1 TB storage

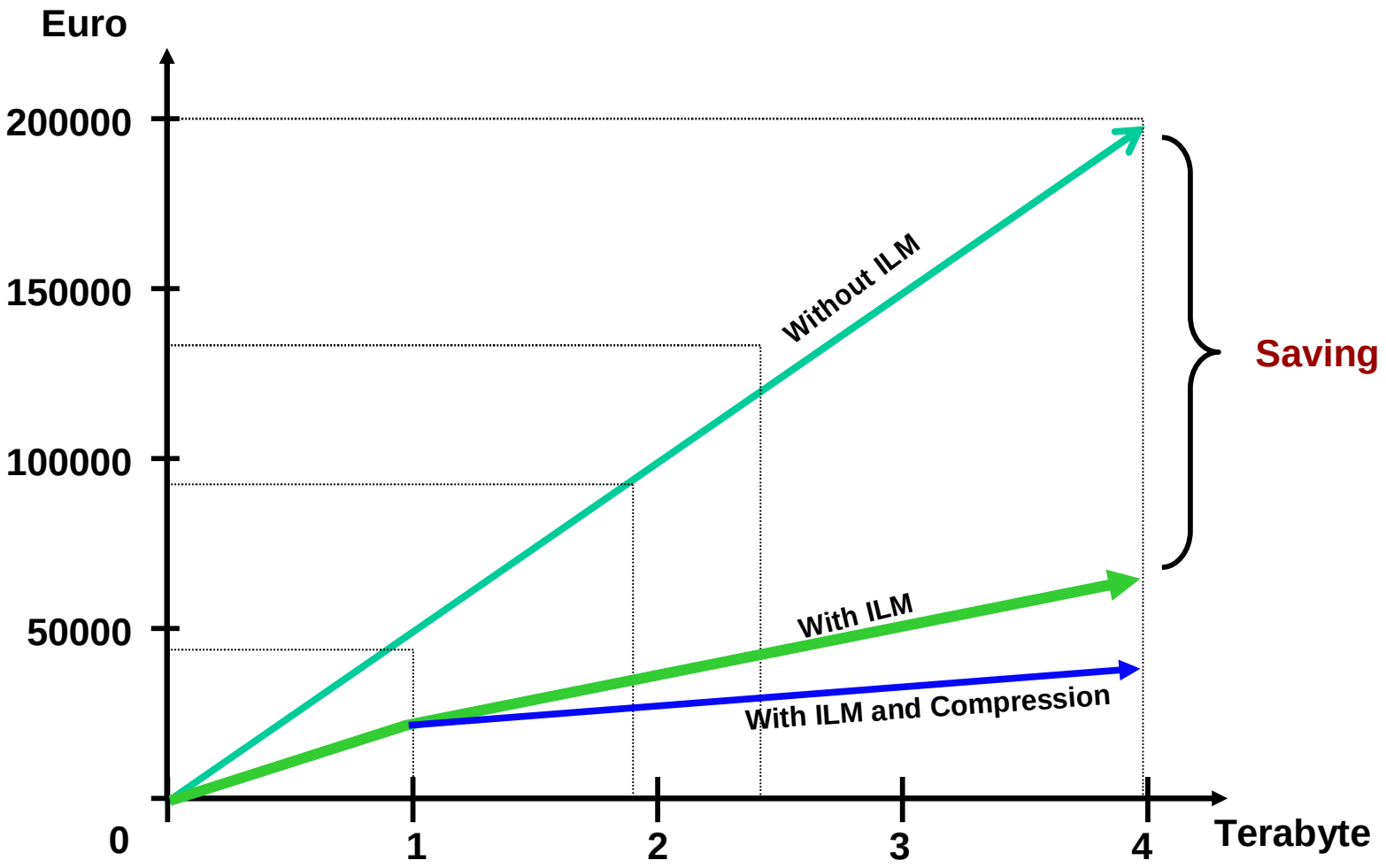
High End Storage	Midsize System	Cheap Storage
40€ / GB	**8€ / GB**	**1€ / GB**
5% active	35% less active	60% historical
~40*1000*0,05	**~8*1000*0,35**	**~1*1000*0,60**
2000€	**2800€**	**600€**



Storage Saving in the example

Without ILM	With ILM	With ILM and Compression (Factor 2,5 = footprint 40%)
40€*1000	2000€ 2800€ 600€	2000€ / 2,5 2800€ / 2,5 600€ / 2,5
40.000€	5.200€ + Software	2.160€ + Software

Cost Saving for larger warehouses



Analyst View

Delivering the Data Warehouse Program

The big theme for data warehousing in 2008 is the increased demand for more data, in more places and doing so with an evolutionary approach. Data warehouses are mission-critical and integrated into operations. Failure to support business process change and inflexibility will not be tolerated. The problem is how to take the years of effort and funds invested previously and leverage them into a modern data warehouse because "rip and replace" strategies are not acceptable.

Source: Gartner
Key Issues For Delivering a Data Warehouse Project, 2008
Publication Date: 18 April 2008

Bill Inmon‘s View

“Indeed, leaving infrequently accessed data on disk storage greatly <u>HURTS</u> performance.
… Data warehouse performance is hurt because mixing infrequently used data with actively used data is like adding lots of cholesterol into the blood stream.”

Information Lifecycle Management
for Data Warehousing:
Matching Technology to Reality
An Introduction to
SAND Searchable Archive
By W.H. Inmon
Copyright ©2005
SAND Technology.

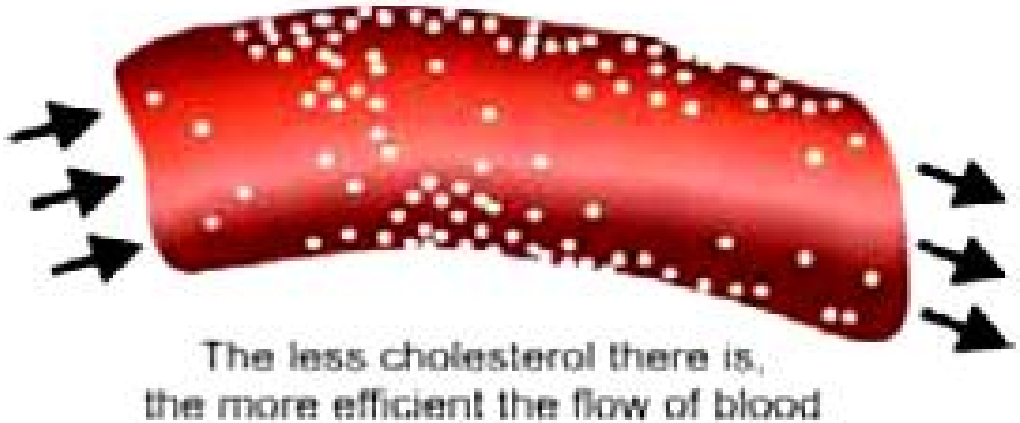
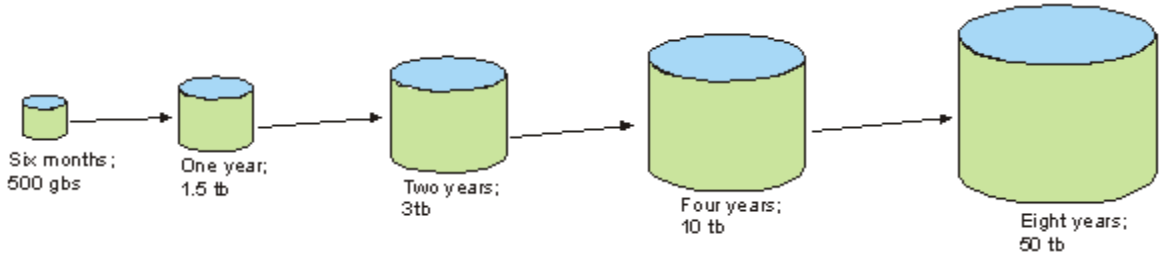


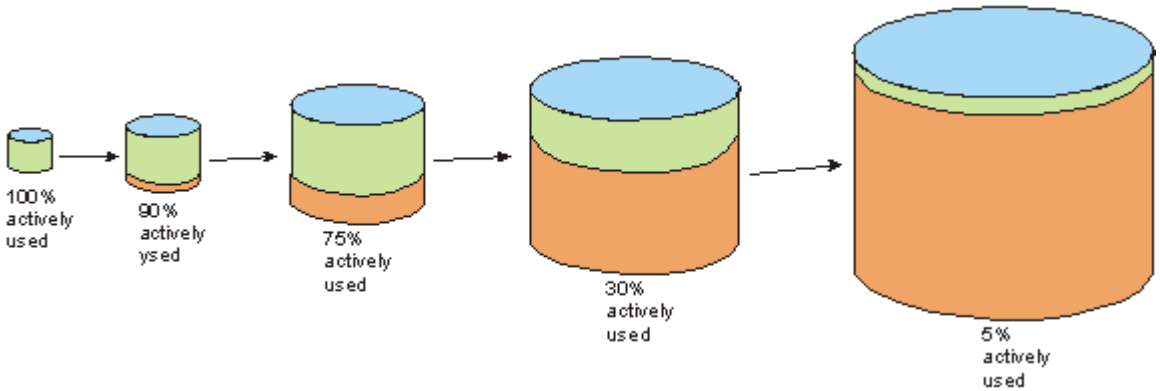
Figure 8: Cholesterol’s Effect on Circulation

Data Access vs Data Growth

- Typical Data Growth



- Typical Data Access vs Data Growth



- As data grows in volume, the probability of access of data changes dramatically

What is Data Aging?

- Data warehousing is a very powerful concept for creating a unified and consistent view of the business
- In a data warehousing environment, it is typical that:
  - Data is amassed and analyzed at an increasing rate
  - As time progresses, companies face the dilemma of storing more and more historical data
  - Over time, data tends to lose its “day-to-day” relevance and is therefore accessed less frequently
  - The costs associated with maintaining historical data are high
- Data aging is a strategy for managing data over time, balancing data access requirements with TCO
- Each data aging strategy is uniquely determined by the customer’s data and the business value of accessing the data
- Need: solution that provides alternatives for the typical “cost vs. business data availability” conundrum

Motivation for a Data Aging Strategy: Benefits

- Performance
  - Faster data load times
  - Faster query execution times
- Cost
  - Storage costs: High availability, high IO disks, etc.
  - Resource and Administration overhead
    - System: CPU, Memory, etc.
    - Headcount: Number of full-time employees, etc.
  - Control of system growth
- Availability
  - Data availability – faster rollups, change runs, etc.
  - System availability – less downtime for backups, upgrades, etc.

Information Lifecycle management (ILM) - again

Everyone's talking about ILM

ILM is not only software

ILM is not only a tool

ILM is not only a product

ILM is a storage management concept that actively manages all information objects during their entire lifecycle

ILM is a combination of processes and technologies with a goal of providing the right information at the right time in the right place with the lowest possible costs

SNIA's Definition of ILM - again

"Information Lifecycle Management is comprised of the policies, processes, practices, and tools used to align the business value of information with the most appropriate and cost effective IT infrastructure from the time information is conceived through its final disposition. Information is aligned with business requirements through management policies and service levels associated with applications, metadata, and data."

SNIA: Storage Networking Industry Association

Business Intelligence and ILM/DLM

- Data Lifecycle Management in general doesn't make a difference between OLTP data and DSS data
- But this differentiation is crucial for Business Intelligence
- Data stored in BI systems
  - impact the company value
  - Should not only be classified by storage costs and regulatory reasons
  - Additional BI specific classification criteria are needed
  - All layers (Data Warehouse, Data Mart) should be considered
    - Source Data can be critical, essential, sensible and non-critical
    - Core Data Warehouse Data is essential (can be reproduced)

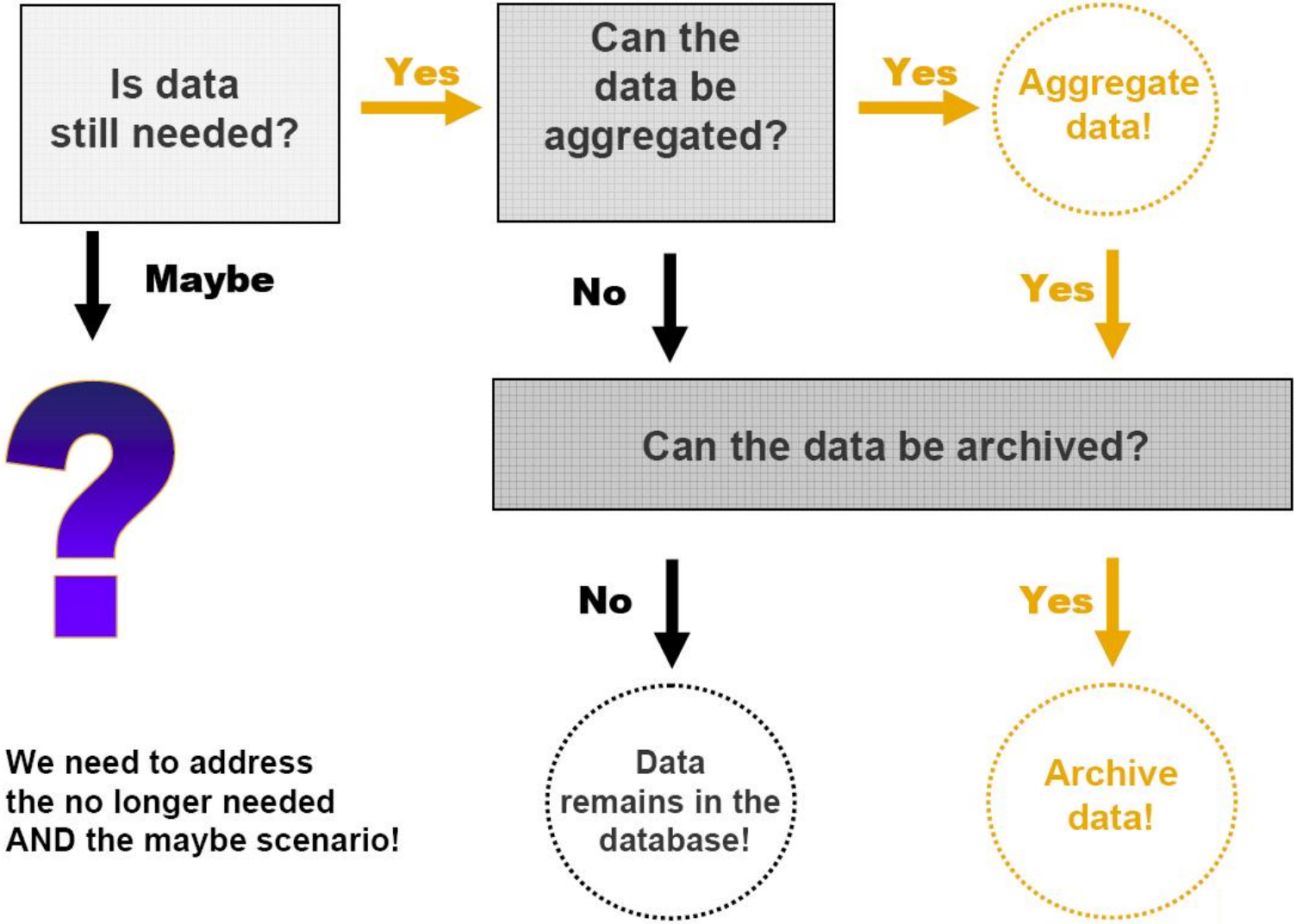
Business Intelligence Data Classification

- All layers (Data Warehouse, Data Mart) should be considered
  - Source Data can be critical, essential, sensible and non-critical
  - Core Data Warehouse Data can be essential
    - Probably can be reproduced from source data
  - Data Mart Data is critical
    - Loss of it impacts daily business to an extent
    - Needed for decision support
    - Loss of it impacts basis for business decisions

A possible Data Classification including Data Warehousing

- Critical Data
  - Needed for the critical applications
  - Loss of it represents catastrophe
- **Business Decision Data**
  - **Data needed for corporate management and planning**
- Essential Data
  - Needed for Daily Business
- Sensible Data
  - Daily Business Data that either can be reproduced quickly or that can be replaced by alternate data
- Non-Critical Data
  - Can be reproduced to low costs or duplicates exist

How to Avoid High Data Volumes in a DW Environment?



We need to address the no longer needed AND the maybe scenario!

Data Aging Strategy Implementation – Initial Steps

- Data aging is a strategy for managing data over time to balance data access requirements with TCO
  - Each data aging strategy is uniquely determined by the customer's data and the business value of accessing the data
- Classification according to business value or frequency of usage

Data Model Design and Strategies: Definition

Include data aging early on in the blueprinting phase

- Data retention should be determined during requirements gathering
  - Determine retention for all data layers including transactional and master data
    - Transactional data should be evaluated from both a Source and an individual Target perspective
  - Evaluate legal reporting requirements
  - Evaluate regulatory reporting and retention requirements
  - Future business data analysis requirements should also be considered
- Observation: In Data Warehouses typically data retention of three to five years
- Data model sizing should be included in overall Data Warehouse capacity planning
  - Data volume and growth should be determined
  - Data “change” activity profiles should also be determined
    - Frequency of data deletion and data updates should be included
  - Data Warehouse capacity plans and TCO should be revisited regularly!

Data Model Design and Strategies: Definition (cont.)

- Data warehouse data retention should be integrated with your OLTP system's archiving plans
  - Data warehouses normally retain more historical data than operational transaction processing systems
  - Do you need to archive Data Warehouse data if your ERP system also archives it?
  - Does your OLTP archiving strategy limit future data warehouse developments?
- Keys to a successful data aging strategy development:
  - Define data retention for all data
  - Profile your data activity and access
  - Determine the capacity impact of ongoing data storage
  - If possible, determine a cost model for your data storage/access
  - Choose and implement technology that does not limit your business

Wikipedia definition of Nearline

Nearline storage (where the word "nearline" is a contraction of **near**-on**line**) is a term used in computer science to describe an intermediate type of data storage that represents a compromise between online storage (supporting frequent, very rapid access to data) and offline storage/archiving (used for backups or long-term storage, with infrequent access to data.

ILM in Data Warehousing

1. Online
   - Data persistent in the database
   - Data modeling aspects important
   - Use multiple layers to control data growth
   - Frequent cleanup necessary
2. Near-line
   - Near-line Storage (NLS)
   - Set up proper nearline concept (archiving policy)
   - Transparent access for reporting
3. Offline
   - Classic archiving
   - Very cheap storage medium can be used
   - No access for reporting

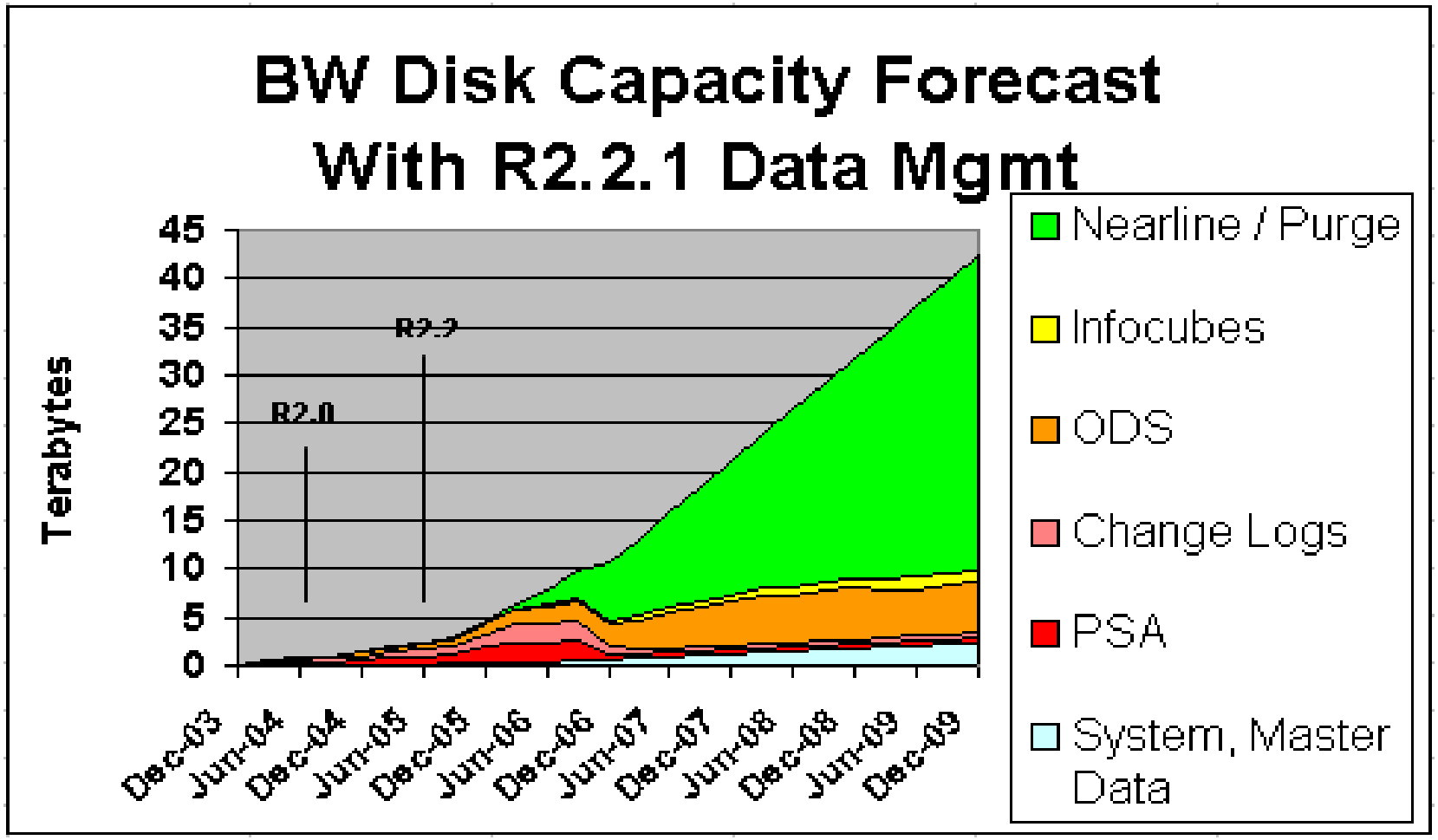
Information Lifecycle Management

- Right-sized (right-priced) data storage approach
- Base storage on age or frequency of access
- Move data to the next level after a specified retention period

	Online Database Storage	Near Line Storage	Data Archiving
Frequently read/updated data	ü		
Infrequently read data	ü	ü	
Very rarely read data	ü	ü	ü

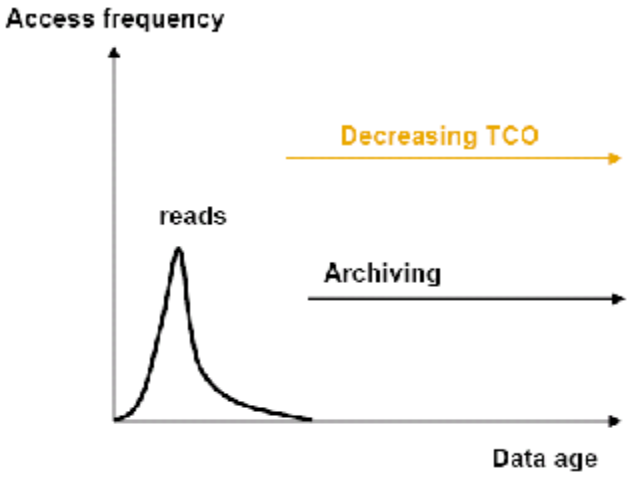
- Nearline in context of Information Life Cycle Management (ILM) :
  - Keep a “skinny”, responsive relational database
  - Keep all data accessible and usable over time
  - Satisfy analytic and legal requirements
  - Control data storage budget
  - Ensure system availability according SLA obligations (happy users!)

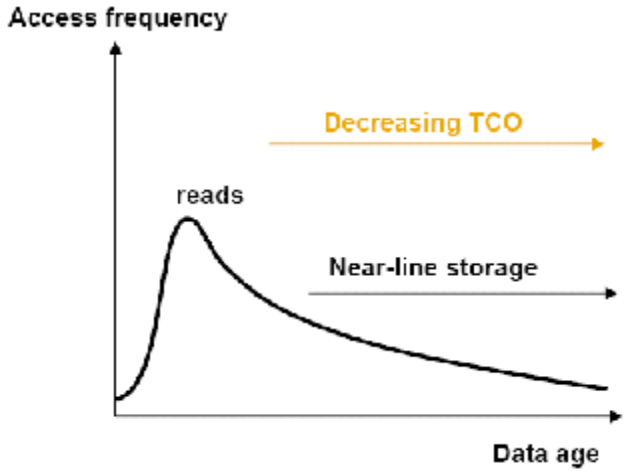
Example: SAP BW Forecast with Data Management Strategy



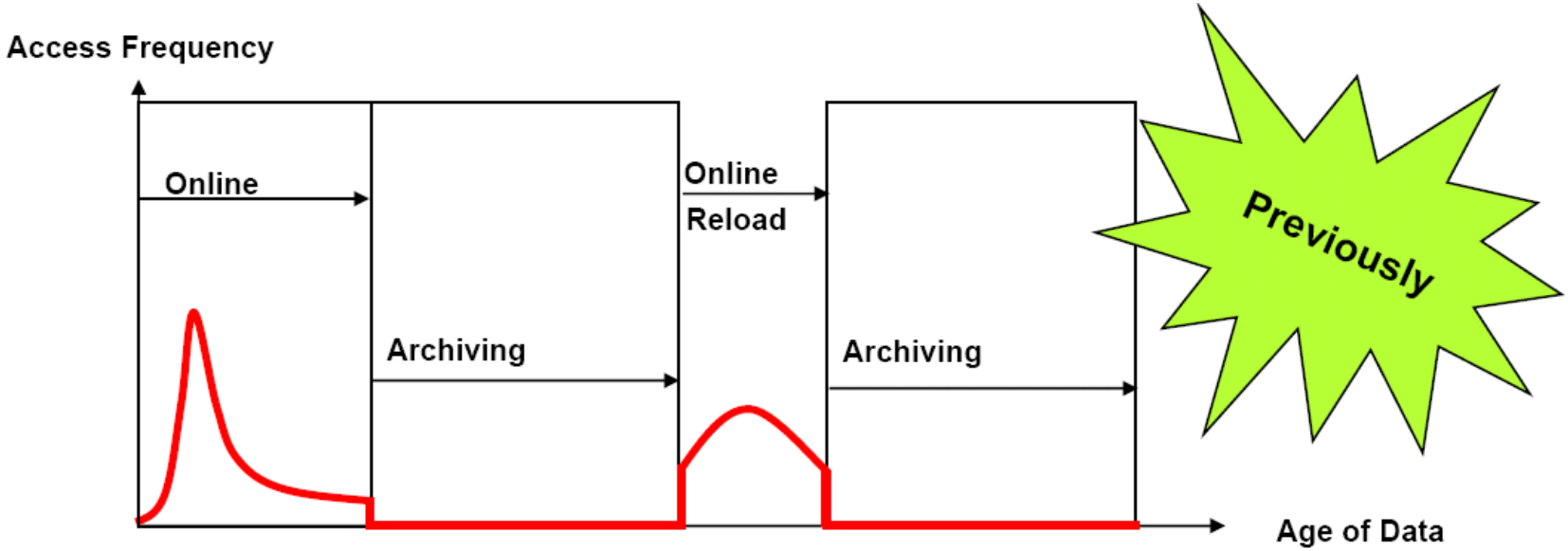
Where Is Archiving and Near-line Storage Applicable?

- Archiving
  - For analysis, archived data must be reloaded first again into the DW database
  - Reduction in costs of data retention on alternative media
- Near-line storage
  - Direct access to data in alternative storage media for queries
  - Performance and data retention costs to access aged data can be minimized



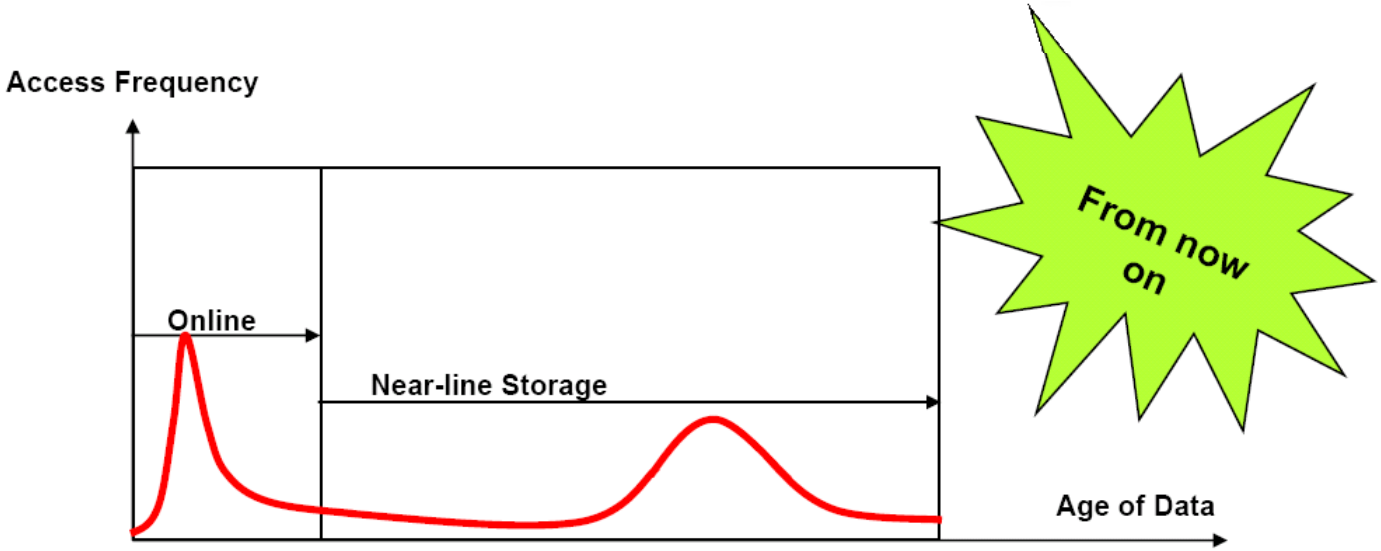


Classic Archiving



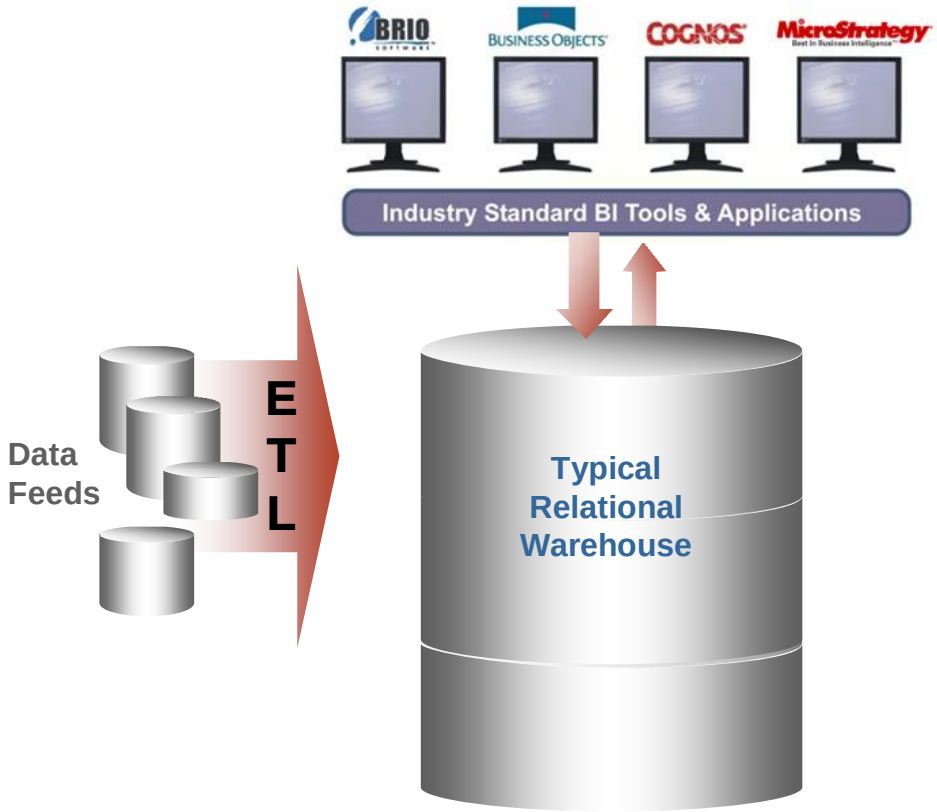
- **Cost reduction due to storing data on alternative storage media**
- **Data Warehouse Query access ineffective**
- **Archived data must be reloaded into the Data Warehouse for analysis purposes**

Near-line Storage Strategy



- direct access to archived data in various storage media
- Availability of historic data while reducing costs
- Physical decoupling of frequently, less frequently, or rarely used data
- Reloading of data only necessary in exceptional cases

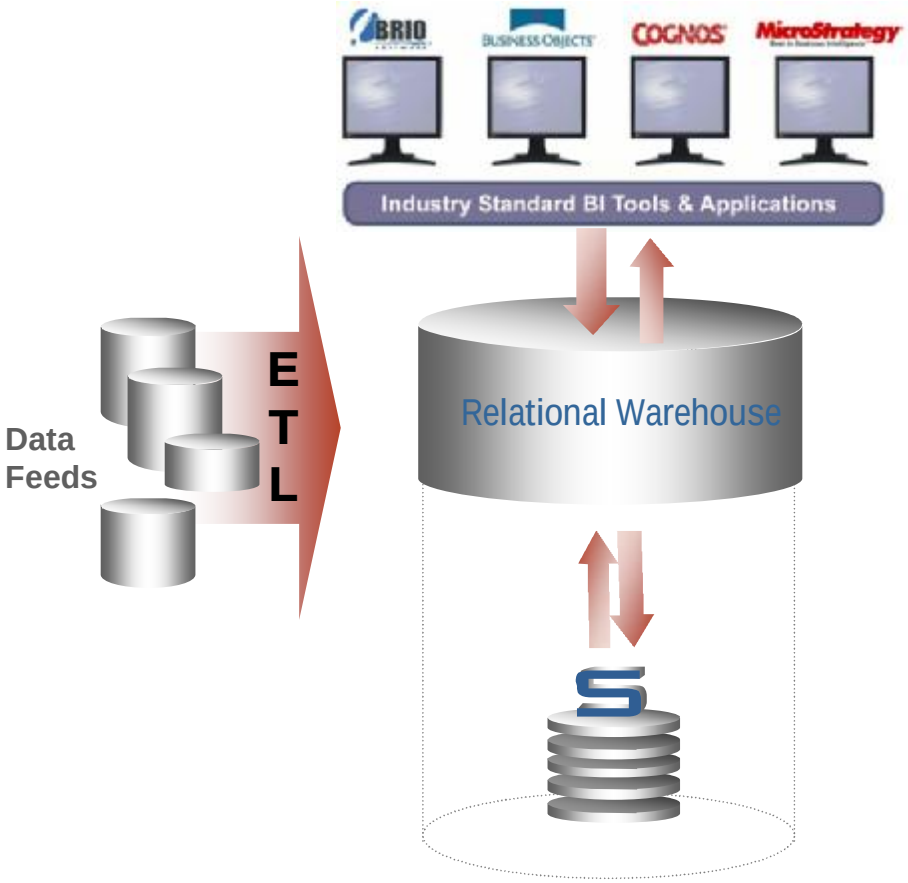
Growing Data Warehouse



Relational Warehouse:

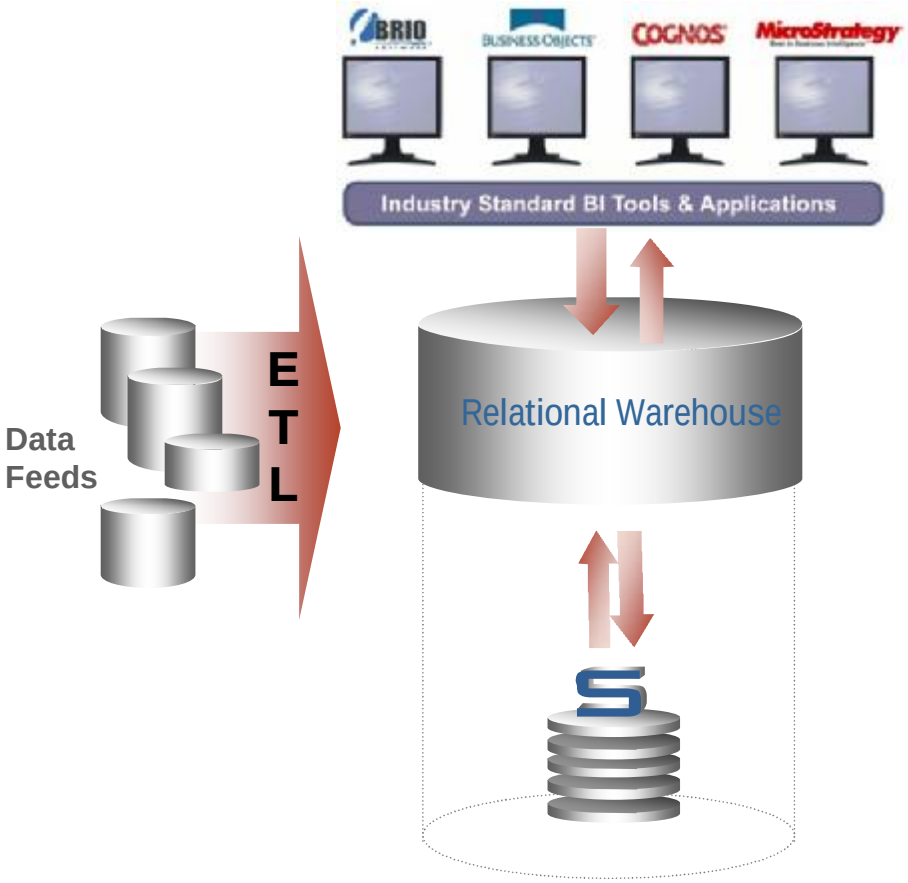
- Large and getting larger
- Slow and getting slower
- Due for yet another infrastructure upgrade

DW - ILM Nearline



- **Relational Warehouse** houses current data
  - Improved performance against reduced footprint size
  - Reduced index load
  - Reduced Management complexity
- **Nearline** holds secondary data
  - Efficient storage
  - More potential data
  - Use in place or restore on-demand

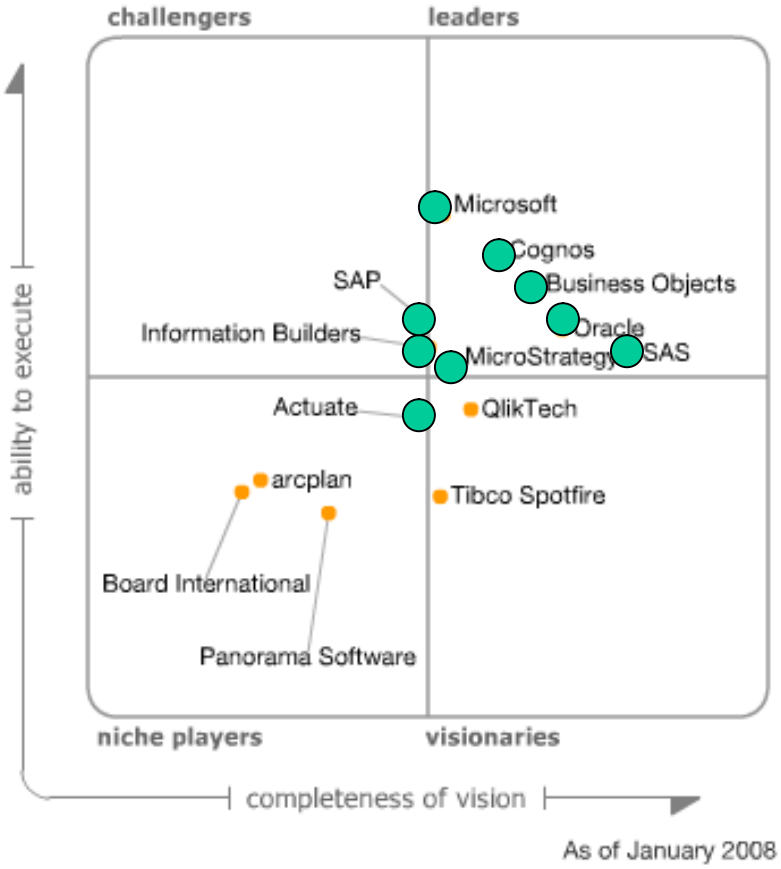
DW - ILM Nearline



- Communication
  - Based on Database Federation
  - A single layer for communication
  - One common connection for your user community
- Integration
  - NLS acts as just another data source with an ODBC/JDBC interface.
  - Transparency achieved using ALIAS & VIEWS
  - Preserve existing report and ETL process
- Data Management
  - Data Movement based on business migration rule: 80/20 rule
  - Automatic End Of Life data Management

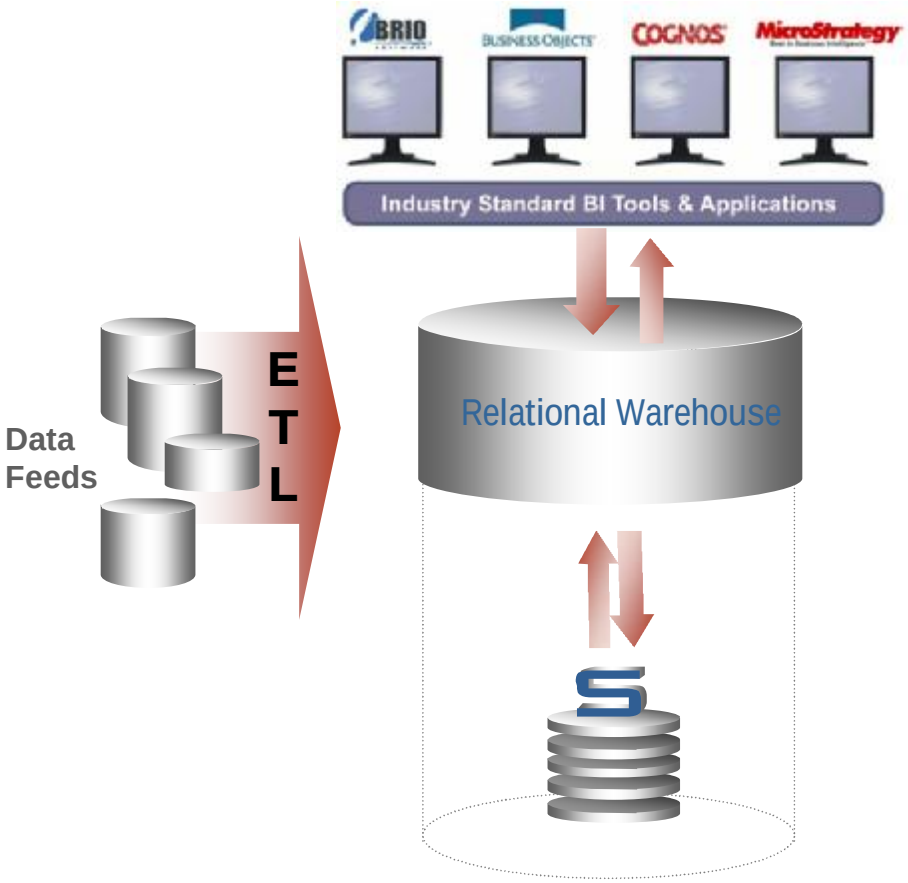
How this fit in the BI Market

Gartner – Magic Quadrant for Business Intelligence Platforms 2008



BI Solution friendly to ILM Nearline

„Extending“ the Relational Database



Database Extension:

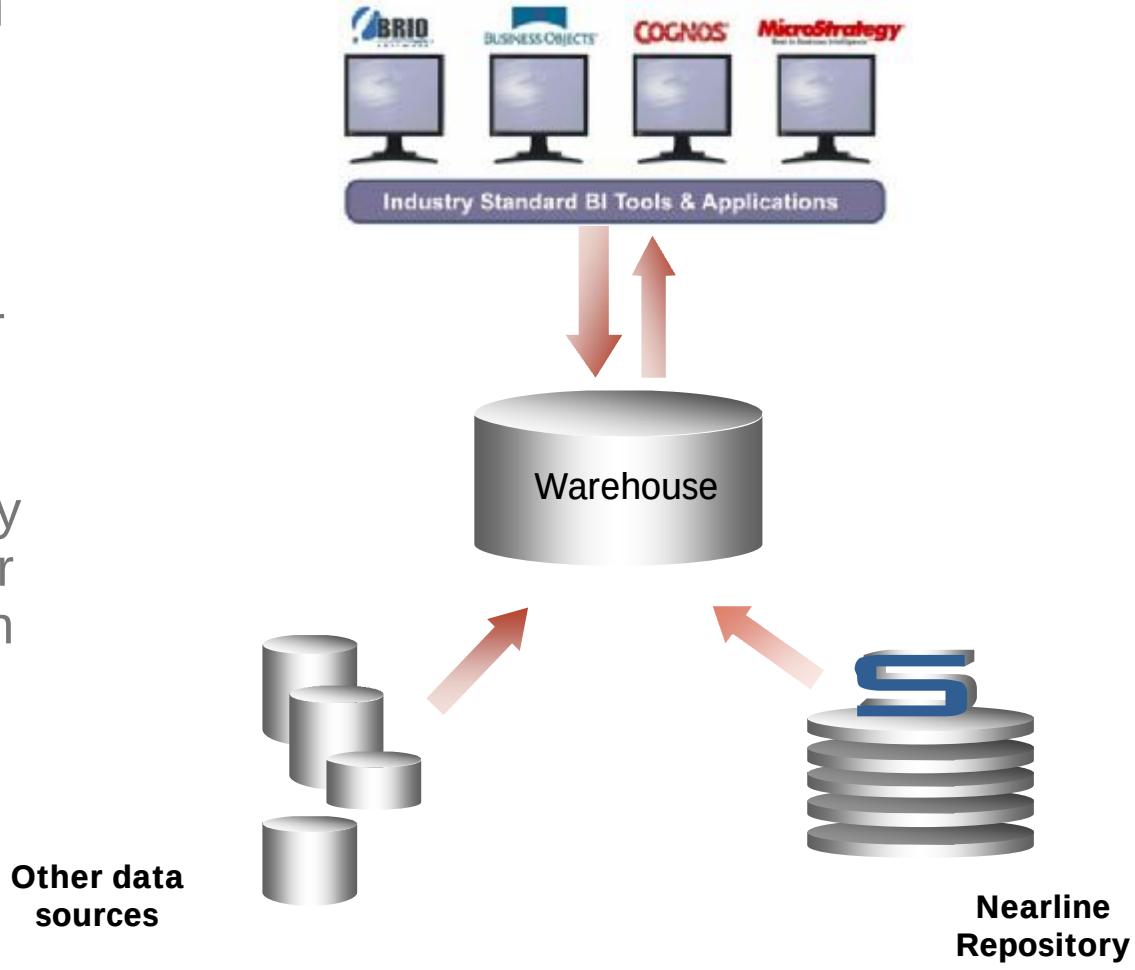
- **Relational Warehouse**

  houses current data
  - Improved performance against reduced footprint size
  - Reduced index load
  - Reduced Management complexity
- **Nearline Repository**

  holds secondary data
  - Efficient storage
  - More potential data
  - Use in place or restore on-demand

Database Extension

- A single layer for communication
- One common connection for your user community
- Nearline Repository acts as just another data source with an ODBC/JDBC interface.
- Transparency achieved using ALIAS & VIEWS



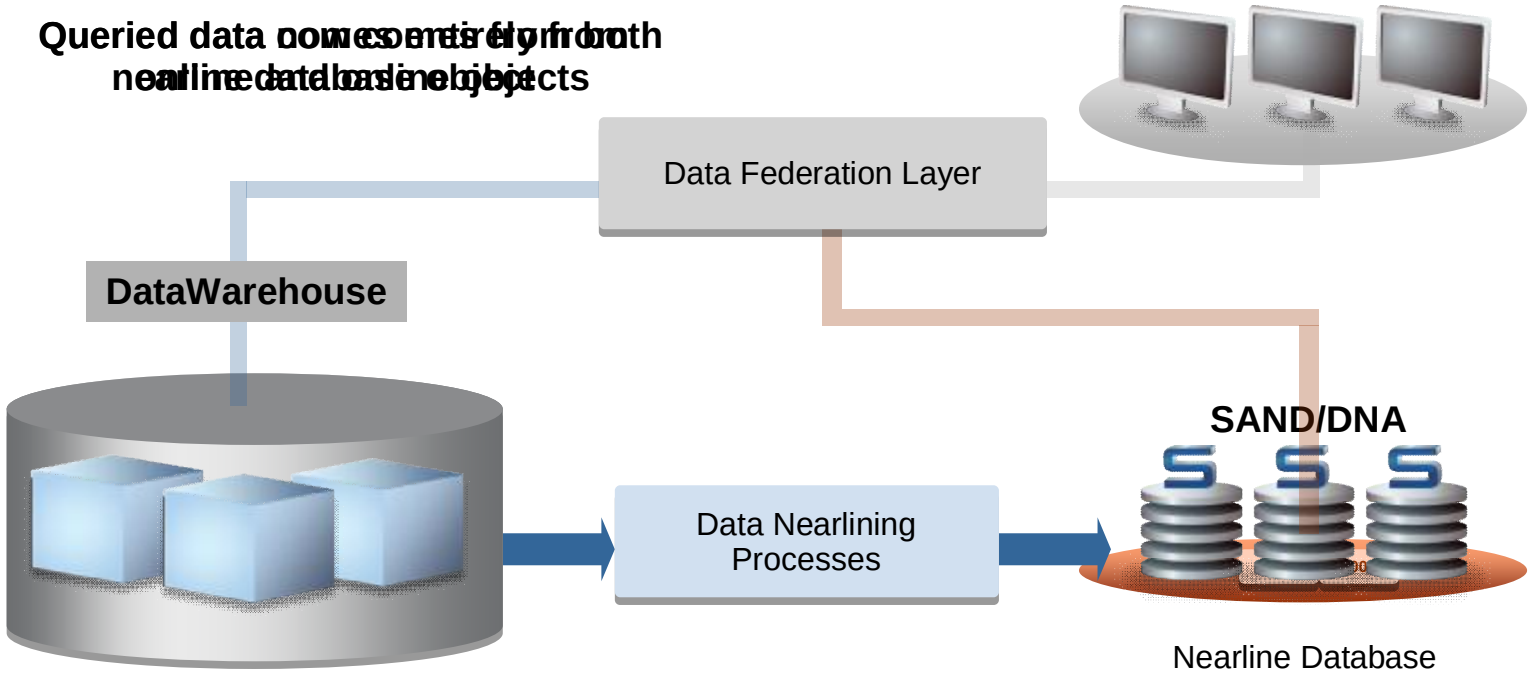
Database Extention Access - Transparent for End Users

1. User executes a query

3. User executes the same query

End-user access via certified front-ends

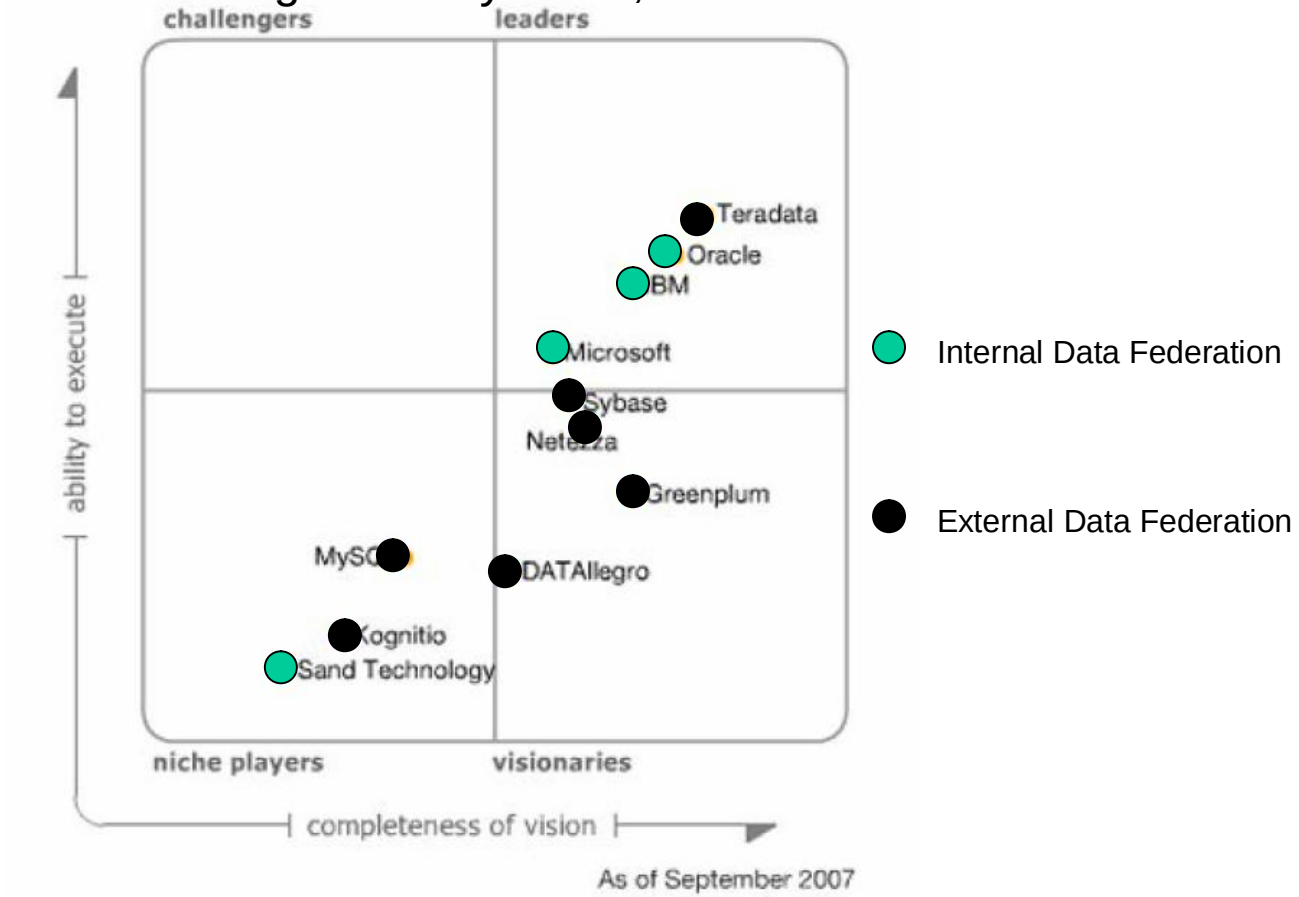
Queried data comes transparently from both nearline and online objects



2. Part of the online database object is sent to nearline

RDBMS Nearline Integration

Gartner – Magic Quadrant for Data Warehouse Database Management Systems, 2007



Source: Gartner (September 2007)

Benefits of a Fundamental ILM Strategy for BI

- Increase Volume
  - Manage and use even larger amounts of information more effectively
  - Information available for any time frame for ad-hoc analyses and rebuilds
- Reduce Resource Consumption
  - Reduction of hardware costs for hard drive on the BI side
  - Main memory and CPU as well as costs for system administration
- Increase Availability
  - Reduced backup and recovery times
  - Intelligent data access
- Optimize Performance
  - Speed up loading processes in Data Warehouse
  - Improve Analytical Query Performance

Benefits of ILM and Nearline Storage for Data Warehouses

- Reduced TCO
  - Reduction of hardware storage costs for main Enterprise Data Warehouse
  - Lower memory and CPU requirements, and reduced system administration costs
  - Older data moved to nearline yet remain accessible

Benefits of ILM and Nearline Storage for Data Warehouses

- Ability to efficiently meet Service Level Agreements
  - Reduced backup and recovery times
  - Faster data loading processes in Enterprise Data Warehouse
  - Faster query processing for reporting and analytics

Benefits of ILM and Nearline Storage for Data Warehouses

- Ability to efficiently meet Service Level Agreements
  - More historical data with greater granularity available for Business Intelligence activities
  - Information available for any time frame for ad-hoc analyses and rebuilding of Key Performance Indicators (KPI)
  - Easy data retention for regulatory compliance
  - Framework to ease the process of Data Reconciliation and Data Audit

Improved SLA

- Full Backup to tape takes 10 hours
  - 80 % Inactive – reduces to 2 hours!
- Full Recovery takes 15 hours
  - 80 % Inactive – reduces to 3 hours!

Reduced Storage Acquisition Costs

- 5 TB on high-end storage @ $50 per GB
  - Cost = $250,000
- If 80% of data is Inactive
  - Migrate 4TB data to low-end disk $5 per GB
  - Saving = $180,000

Reduced tape Cost

- 5 TB Full Backup to tape @ 1$/GB = $5,000
  - Full Backup every week with 6 months retention
  - Total Annual tape cost: $130,000
- Migrate 80% of inactive data
  - Reduces total annual tape cost to $26,000
- Saving of $104,000

ILM & Data Aging Strategy in Data Warehousing: Key Points

1 The "Healthy" System

Don't start thinking about data archiving when your system is about to crash!

2 Timely Planning

Proactive preparation for sustainable system performance

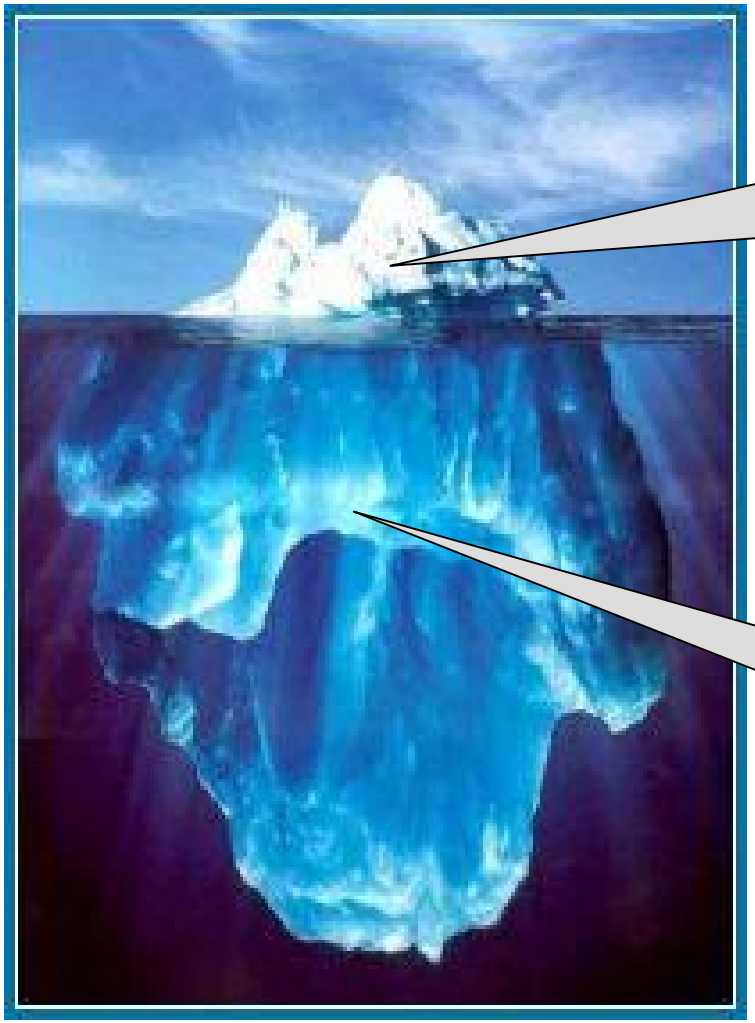
3 Interdisciplinary Process

Data archiving requires a large amount of coordination between IT and those who are responsible for applications

Agenda

- Drivers for Information Lifecycle Management (ILM)
- Definition of ILM
- Data Warehouse Challenges
- ILM in Data Warehousing
- **Enterprise Data Warehousing**

Information as Corporate Asset – We Do not Know What we not Know...



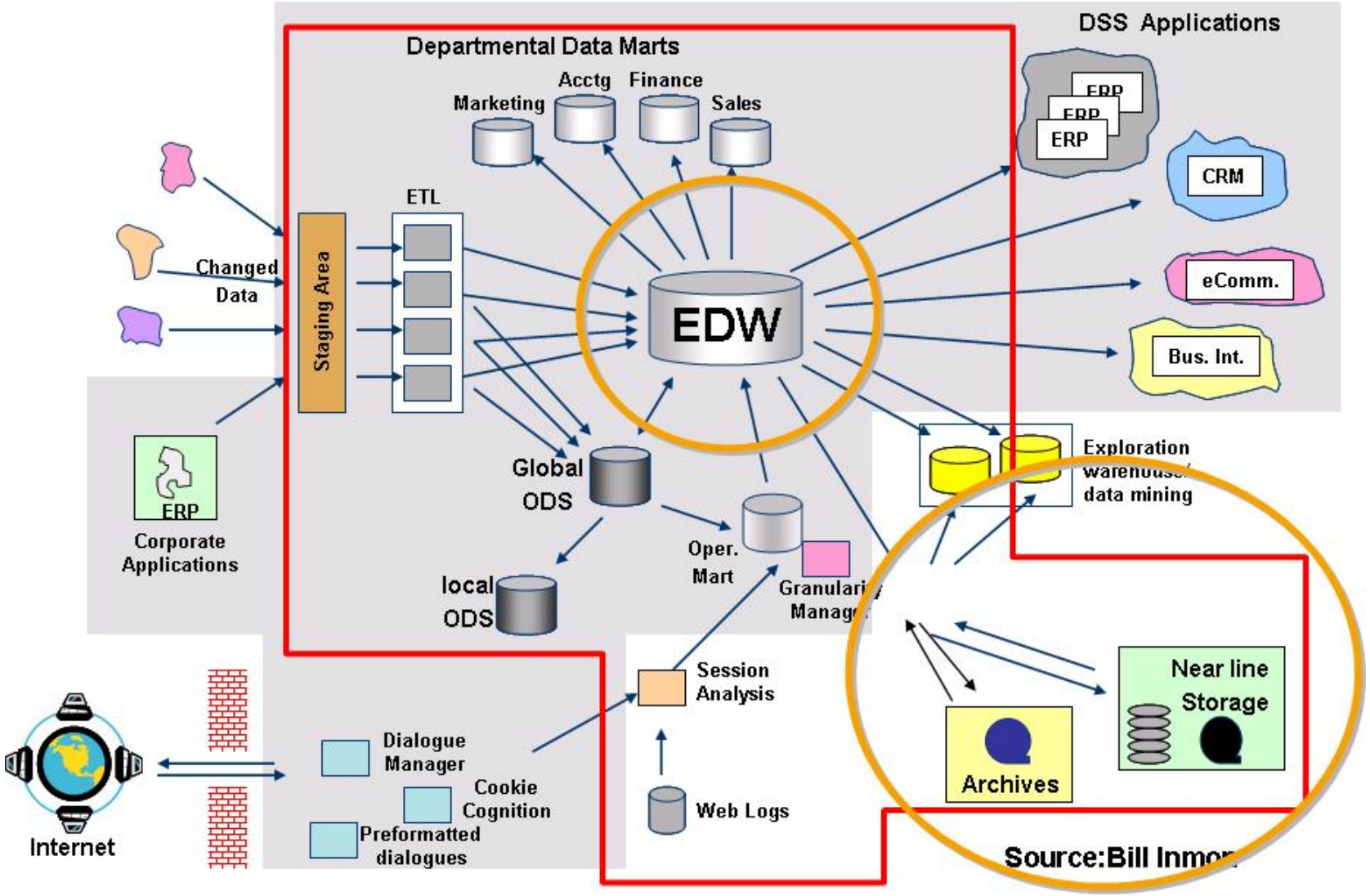
The Known

Current BI implementations are set up to answer known requirements

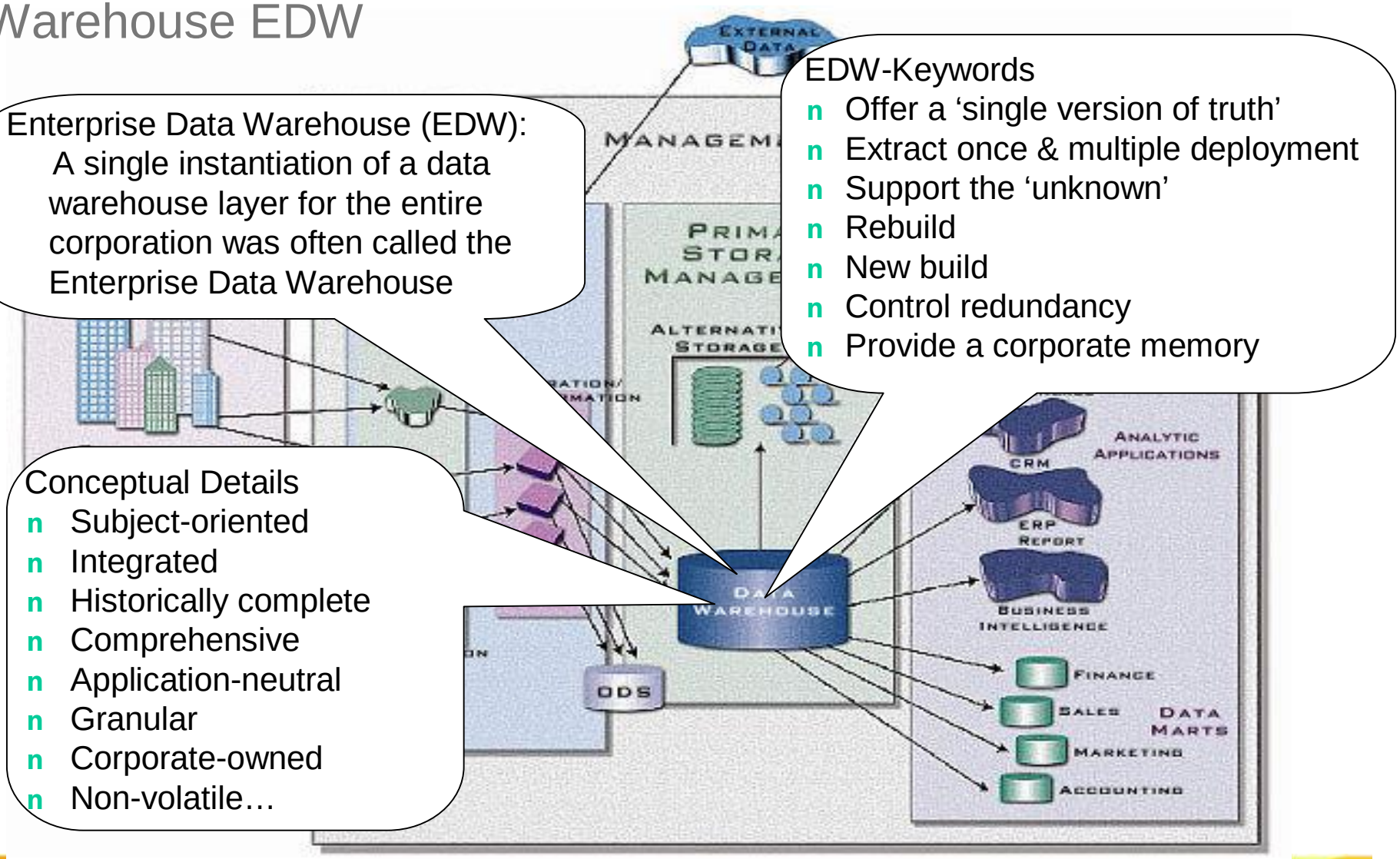
The Unknown

Little or nothing is done to be prepared for unpredictable future information needs

Bill Inmon's: Enterprise Data Warehousing Concept



Bill Inmon's Corporate Information Factory & Enterprise Data Warehouse EDW



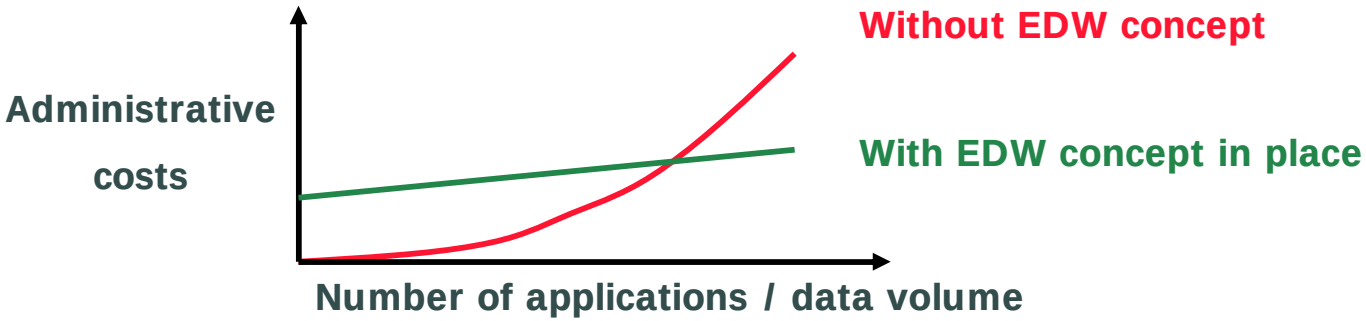
Enterprise Data Warehouse (EDW):
A single instantiation of a data warehouse layer for the entire corporation was often called the Enterprise Data Warehouse

EDW-Keywords
- Offer a 'single version of truth'
- Extract once & multiple deployment
- Support the 'unknown'
- Rebuild
- New build
- Control redundancy
- Provide a corporate memory

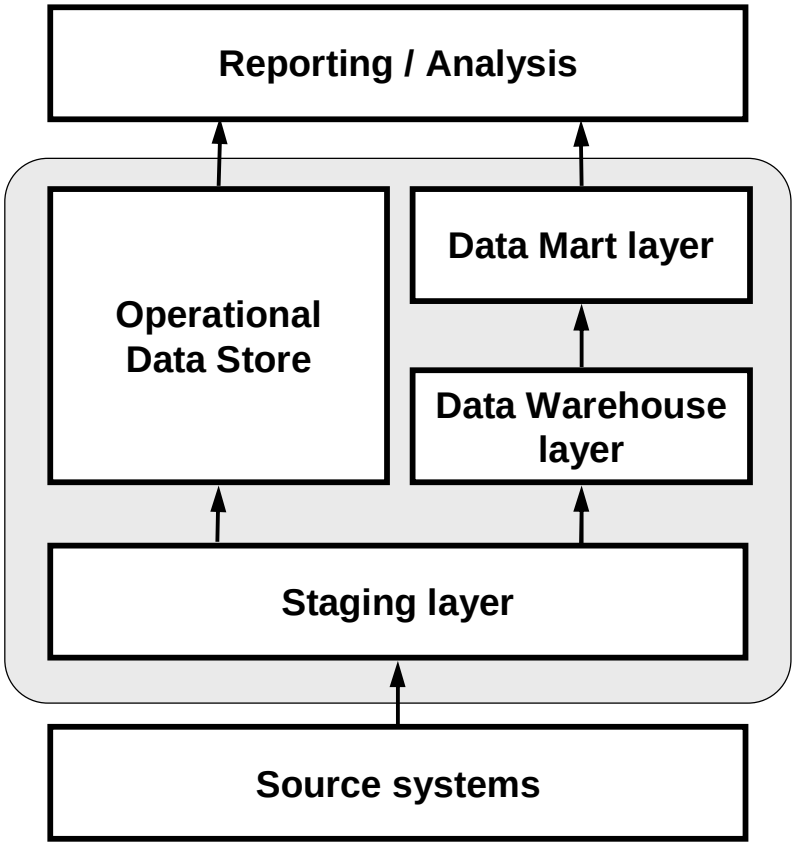
Conceptual Details
- Subject-oriented
- Integrated
- Historically complete
- Comprehensive
- Application-neutral
- Granular
- Corporate-owned
- Non-volatile…

Motivation of EDW concept – *Anticipating the unknown*

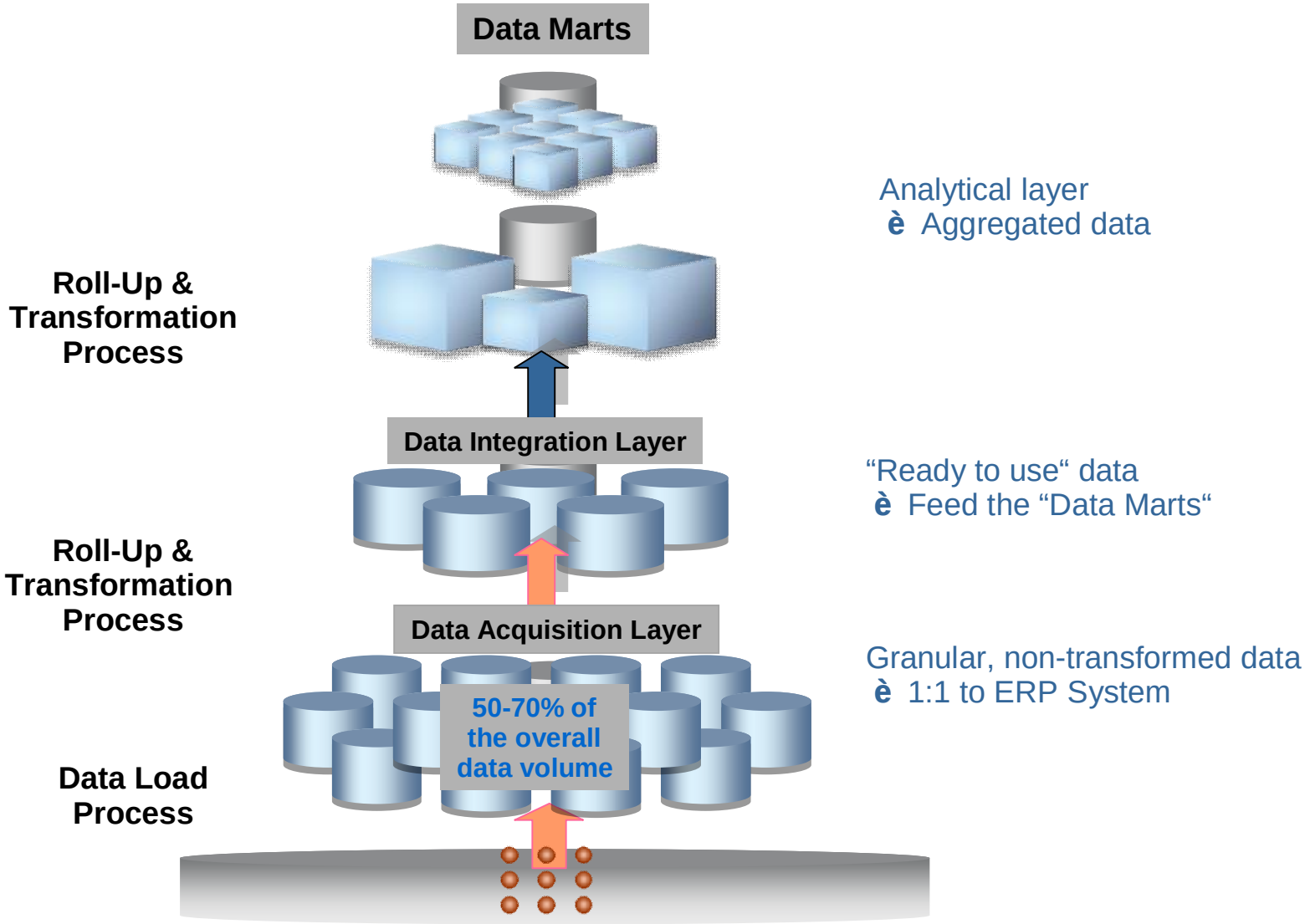
- Data growth
- Increasing number of applications
- Resulting in
  - Increasing administrative costs
  - Higher risk of breakdown of applications
  - Risk of total breakdown



Conceptional Multi-Layer-Architecture



Enterprise Data Warehousing: Data Management



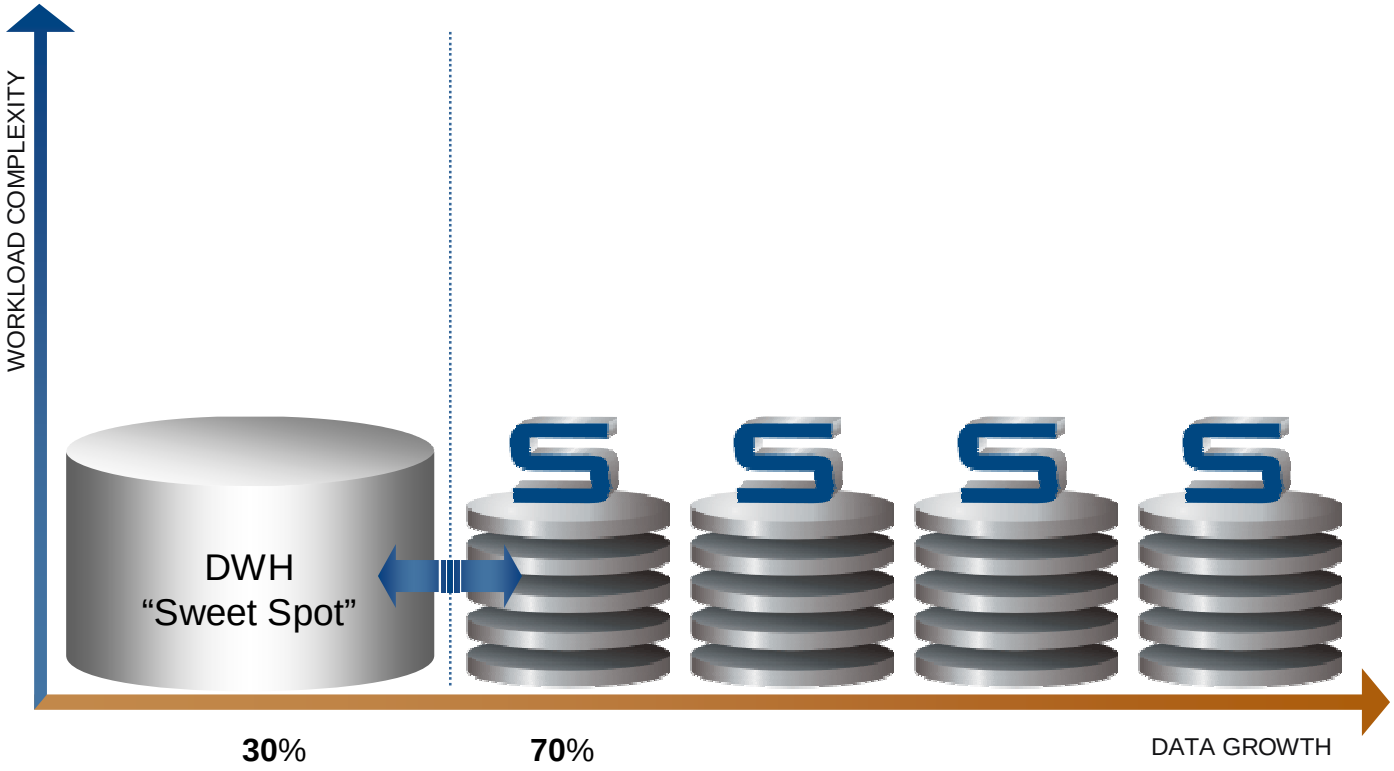
Analytical layer
è Aggregated data

"Ready to use" data
è Feed the "Data Marts"

Granular, non-transformed data
è 1:1 to ERP System

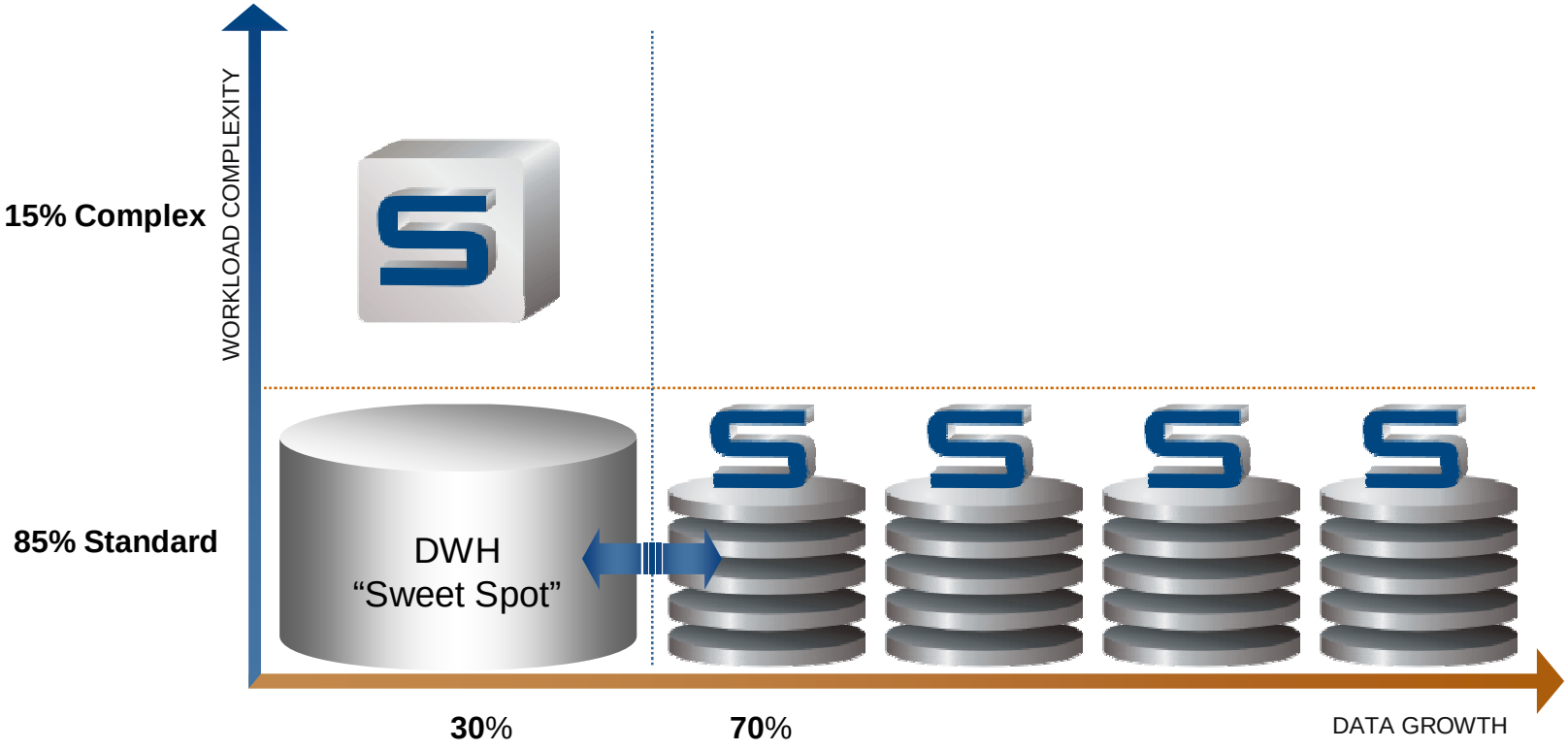
Data Growth

- Most Standard Reporting uses less than 30% of Enterprise Data
- Other Data Required “Just-in-Time” to Meet SLA’s

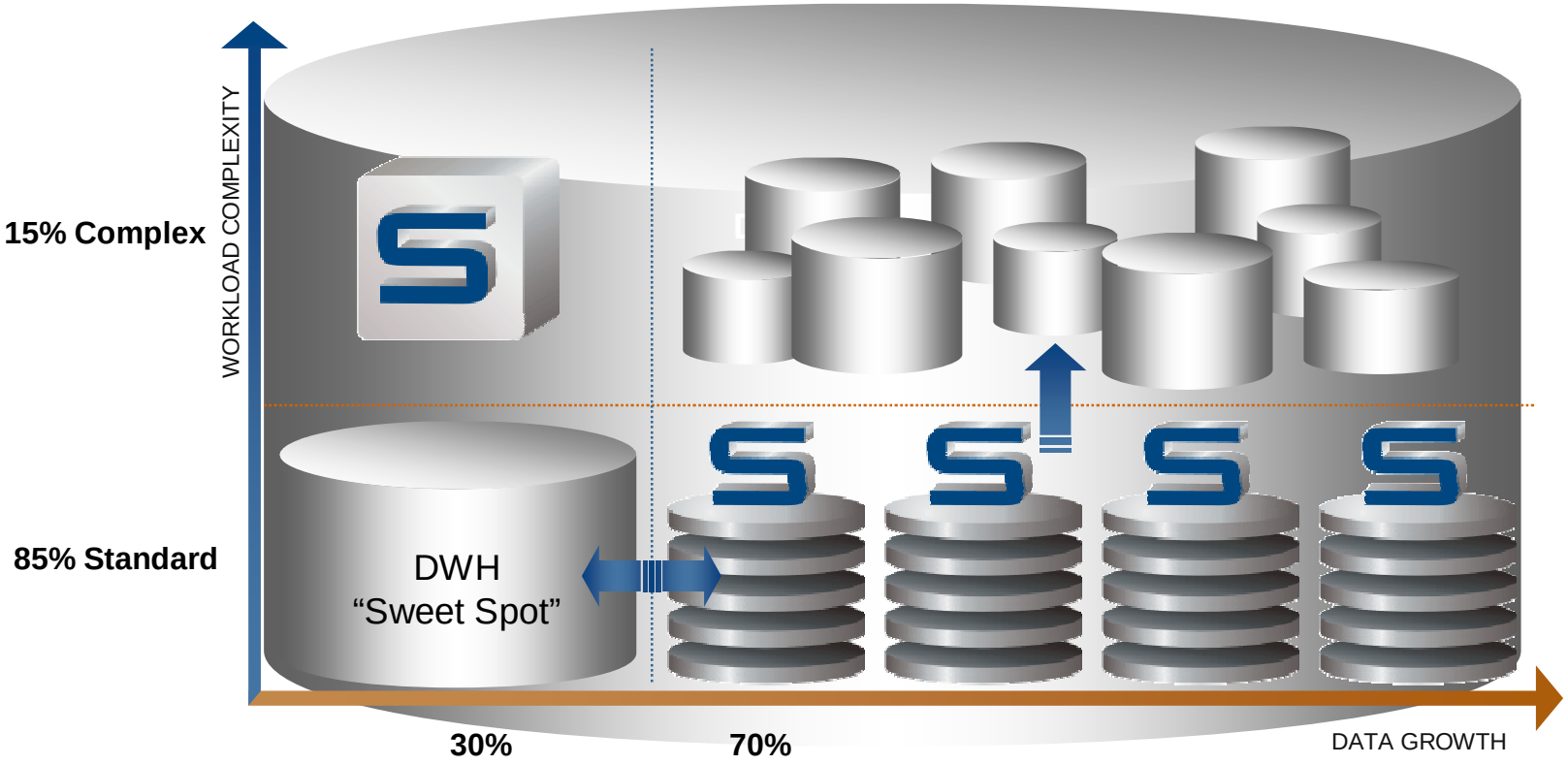


Workload Complexity

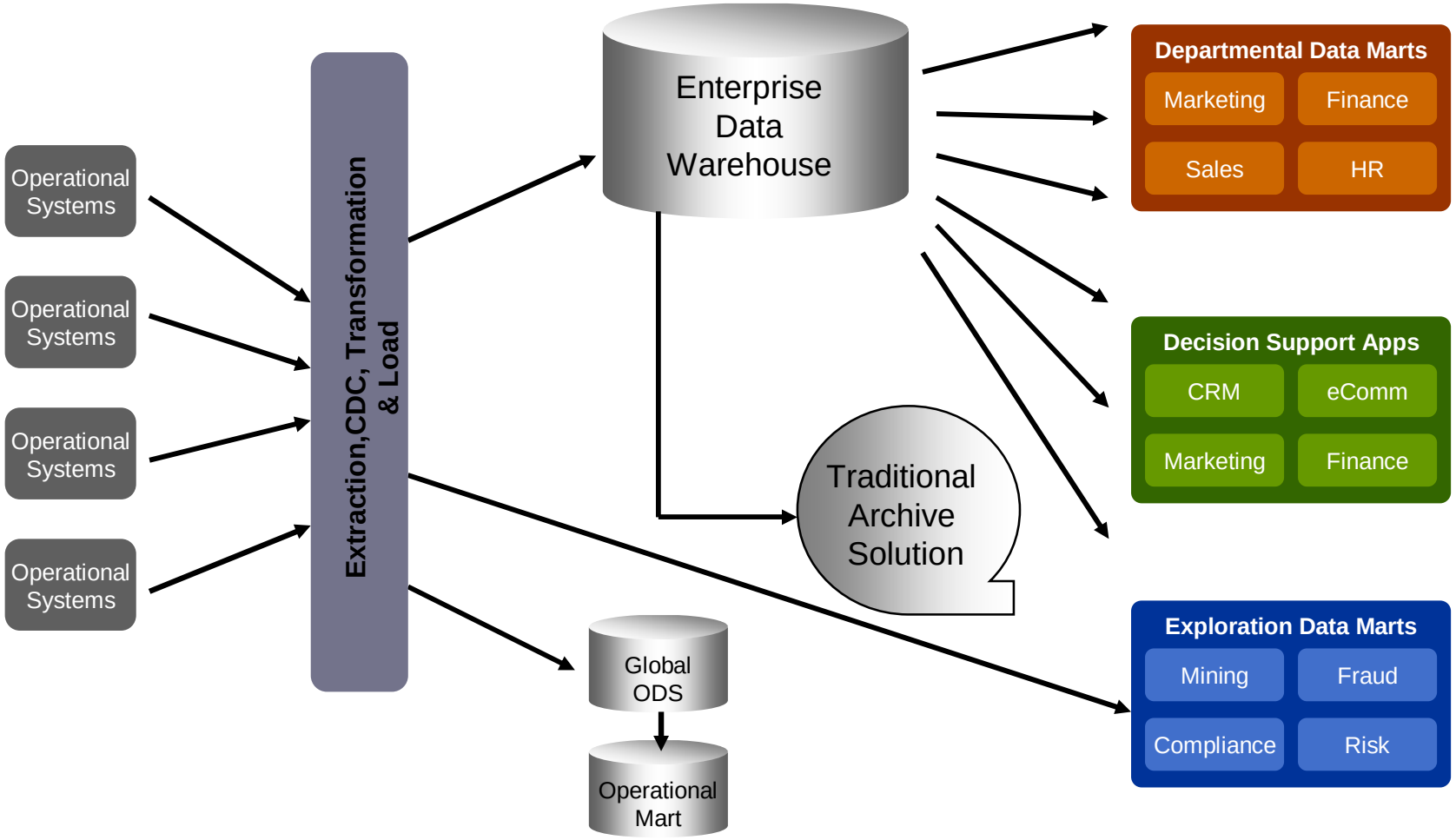
- Most Standard User require less complex analytical function and performance and use predictive queries
- The complex Analytic Users required a pure ad hoc environment specialized for their needs



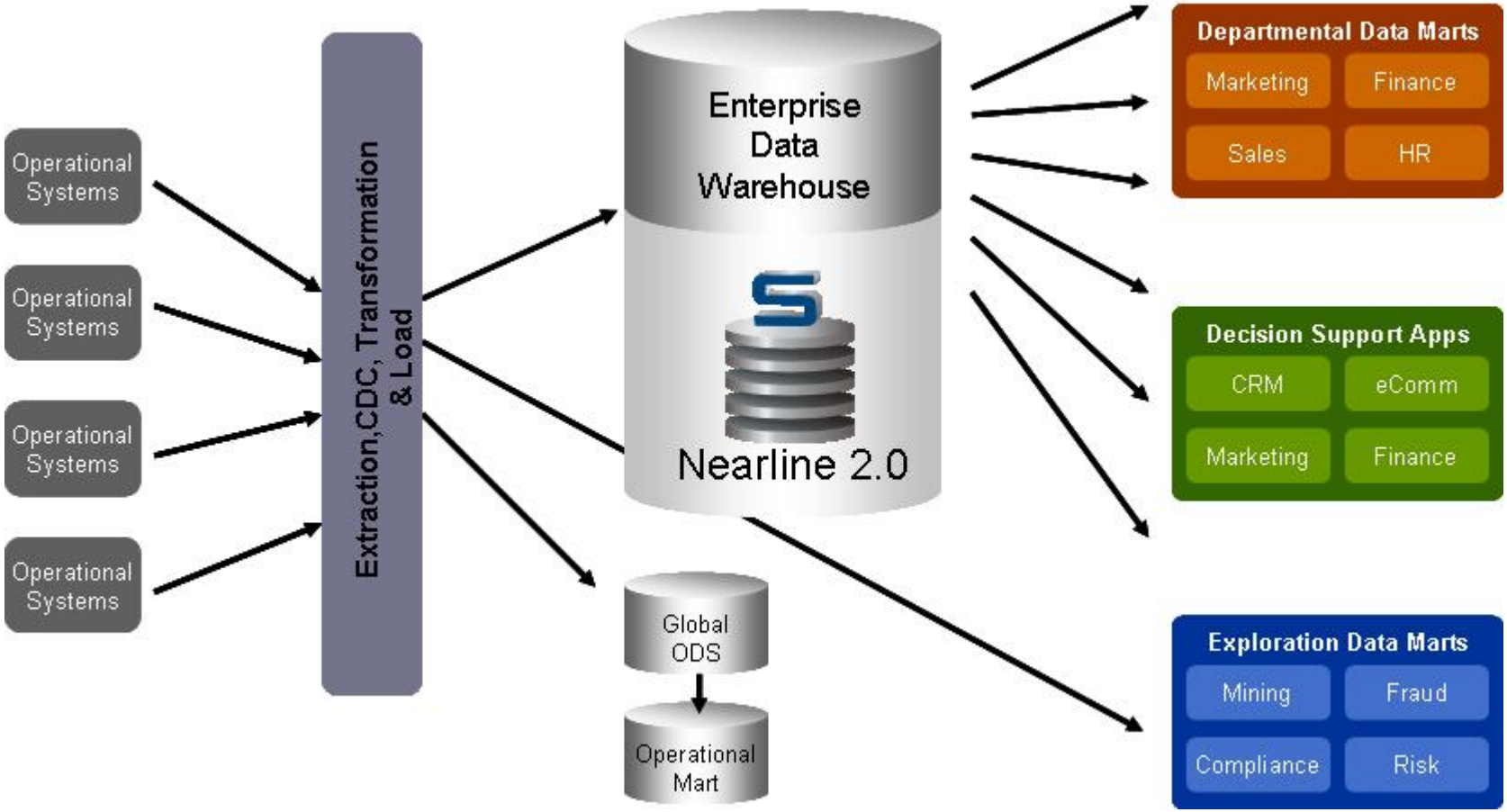
Exploration Warehouse – Data Marts on Demand



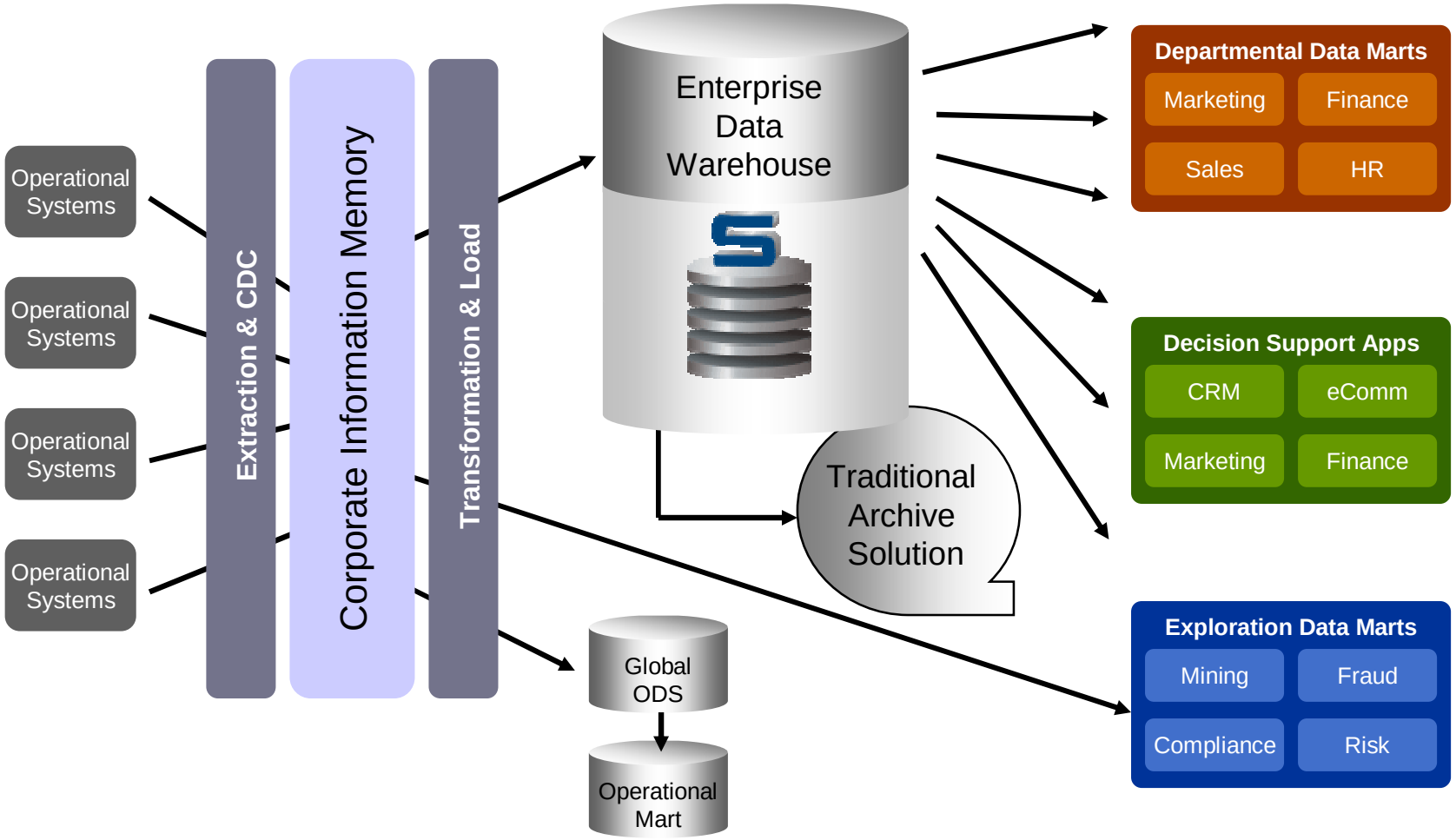
Corporate Information Factory



Tactical Approach



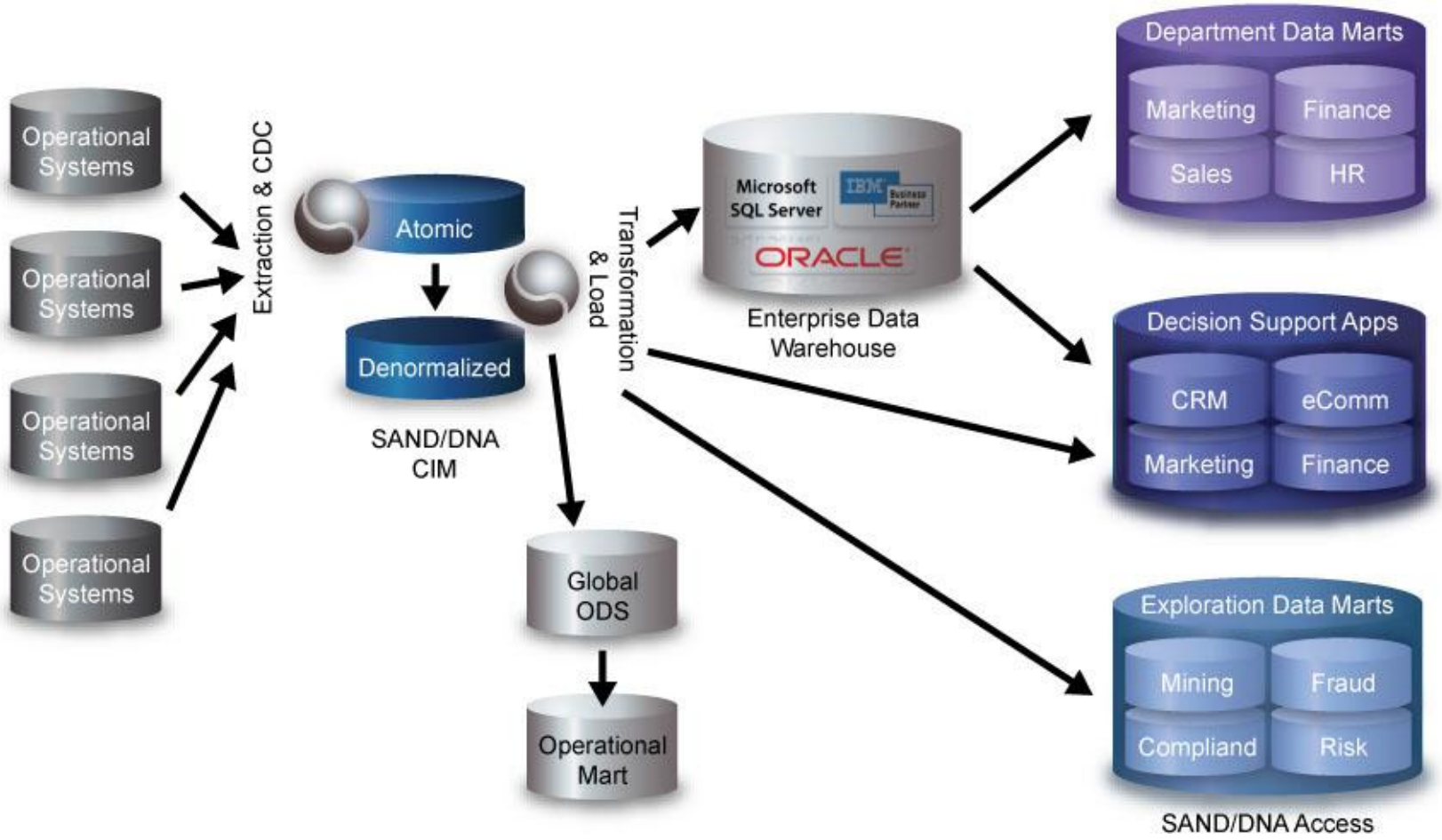
Strategic Approach: Data Centric Architecture



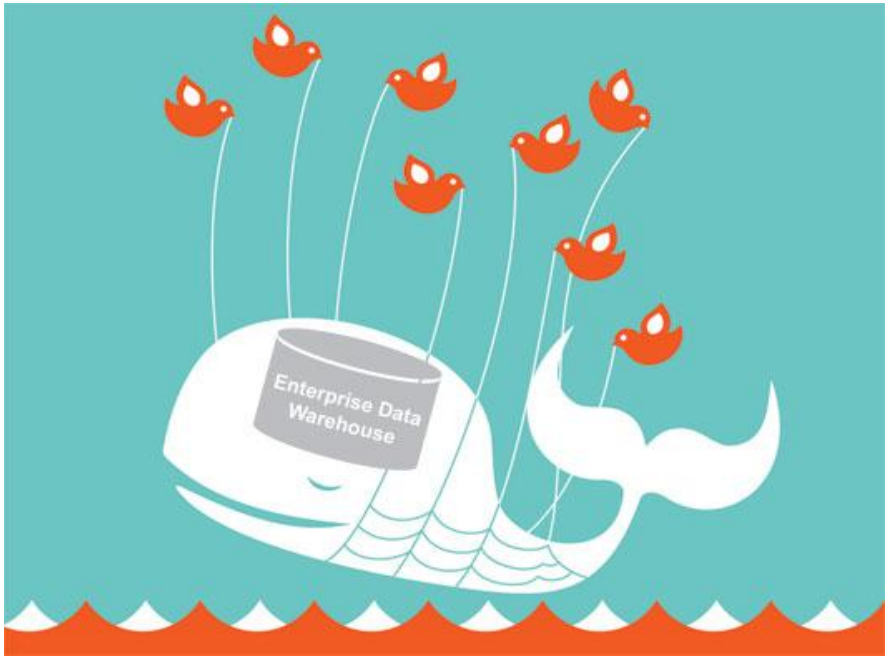
Corporate Information Memory

- Is extension of the Corporate Information Factory
- Is a new data architecture component required to fulfill the audit requirement of new regulation: the “Just in Time” data accessibility/traceability
- Delivers a different SLA and a TCO reduction
- Has the following assumptions:
  - Some Data is used on a regular basis and required very high performance access
  - Generic Ad Hoc Analytic environment has to be enable to deliver real business intelligence on unplanned event
  - Some Data should be keep into their original form without any transformation for audit
  - Some Data should be keep around without any specific requirements outside audit required by new regulation: CDR, syslog, application log, weblog & XML log
  - Some historical data should be keep for specific ad hoc reporting needs, but no more the application infra-structure: Application Sun Setting, Mergers/Acquisitions

Putting the EDW back together again



Reliability



- Immutability of data
- Meet Service Level Agreements
- Control TCO
- Fully auditable
- Enable Full disaster recovery (remote site(s))

Availability

- Enable “power users” without impacting standard reporting
- Full, effective disaster recovery
- Flexible, massively scalable approach

Serviceability

- Determine if “Atomic Level Data” has been tampered with
- Create data-centric model to support multiple business views of data
- Leverage existing infrastructure investments
- Enable Ad-hoc data investigation
- Easily add new functionality