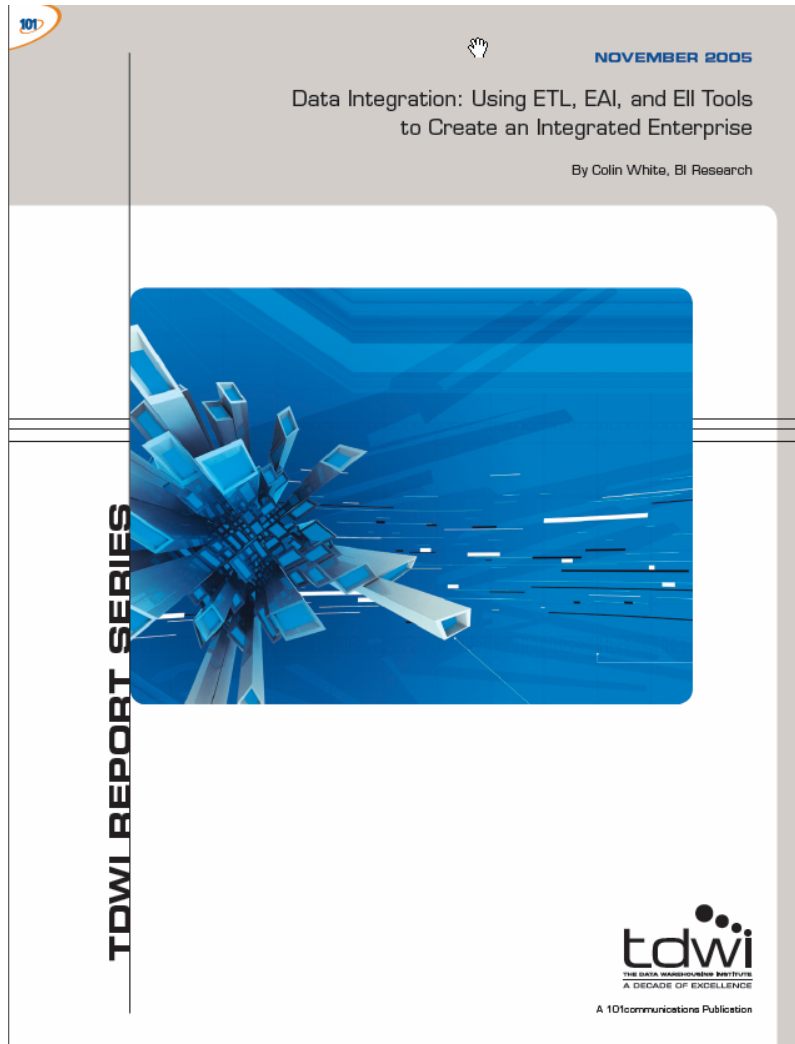




# Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise

**Colin White**  
**Founder, BI Research**  
**TDWI Webcast**  
**October 2005**

# TDWI Data Integration Study



**Business Objects**

**COLLABORATIVE**  
CONSULTING

**DATAFLUX**  
A SAS COMPANY

**DataMirror**

**IBM**

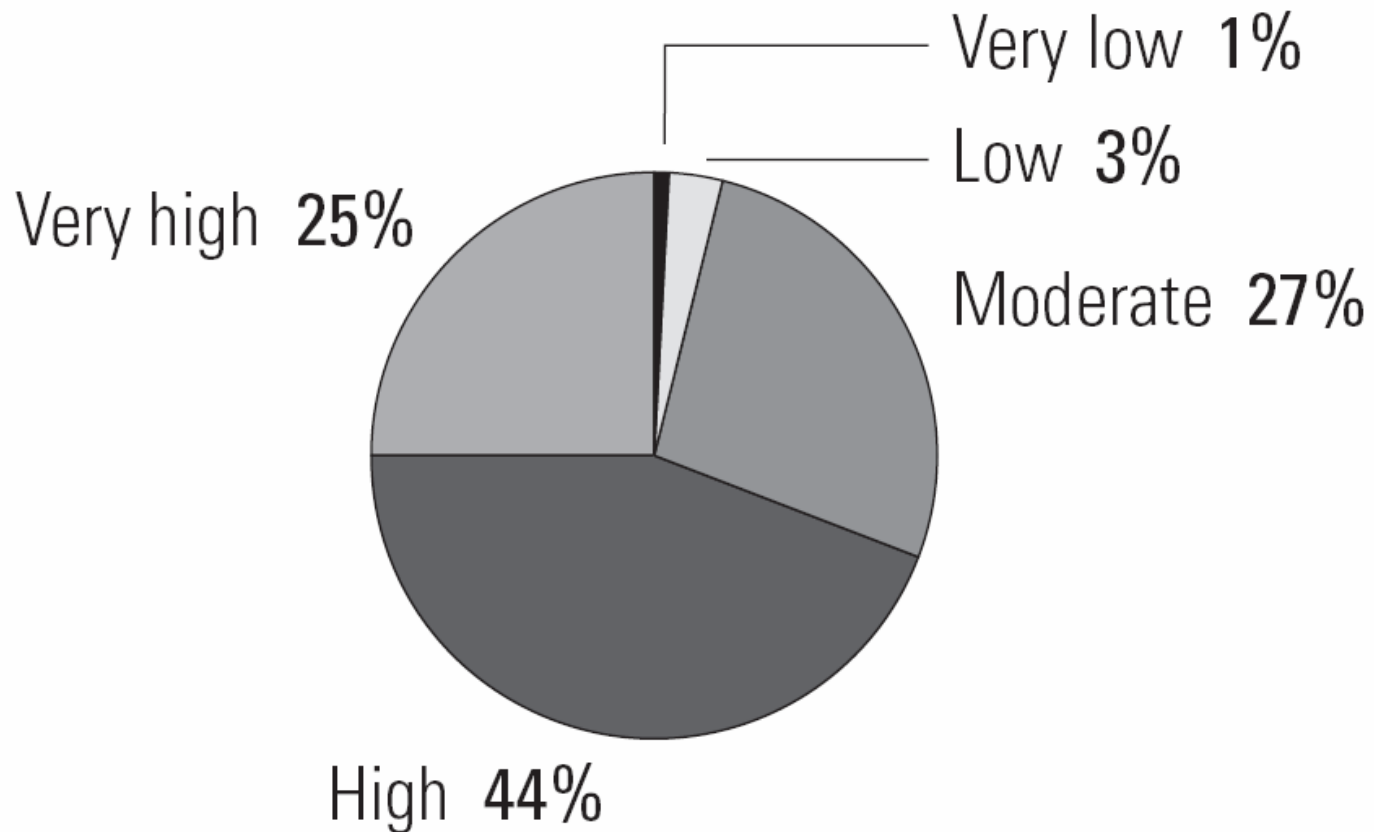
**INFORMATICA**  
Turning integration into insight.™

**SAP**

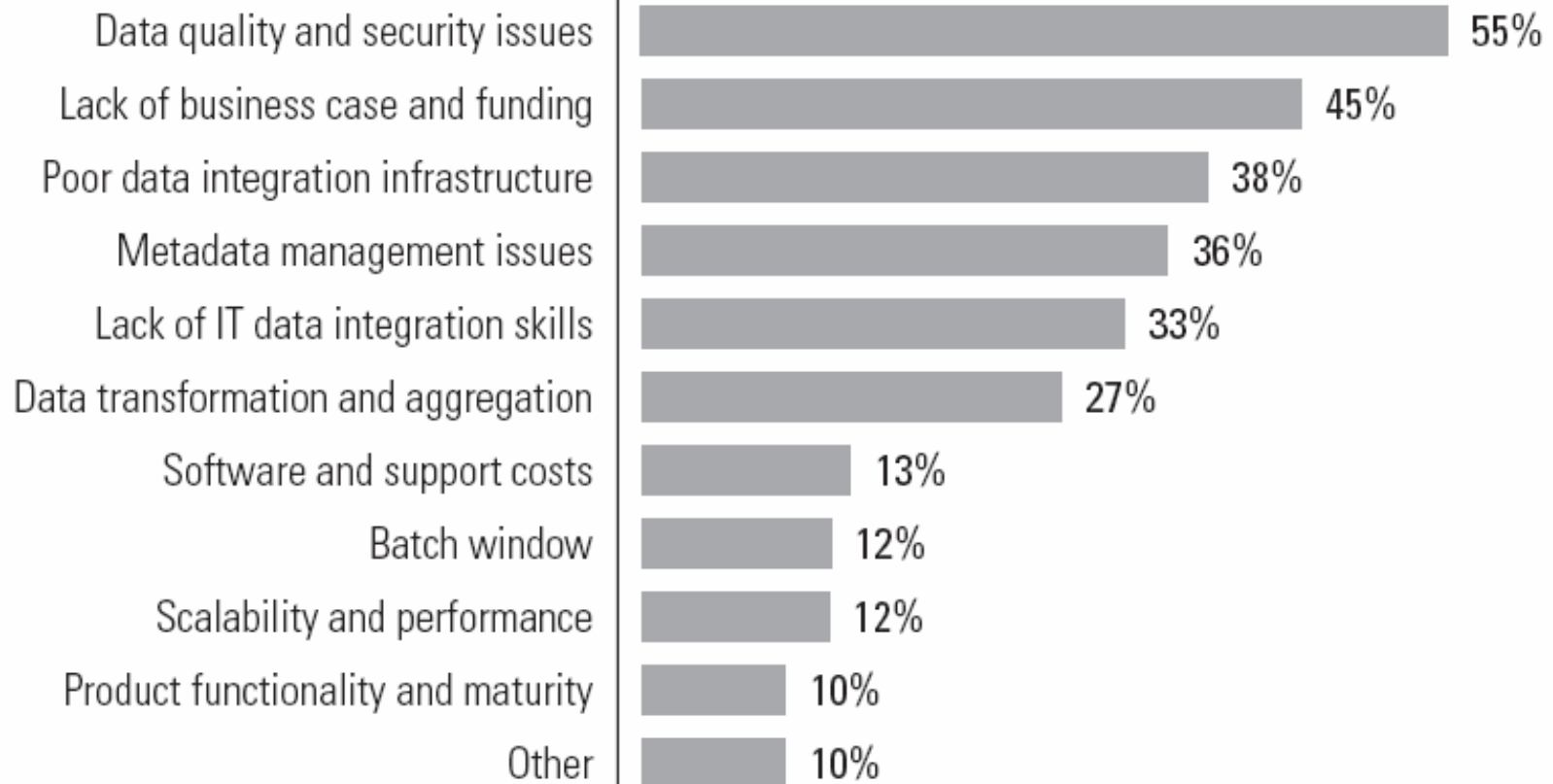
**Sunopsis**

**syncsort**

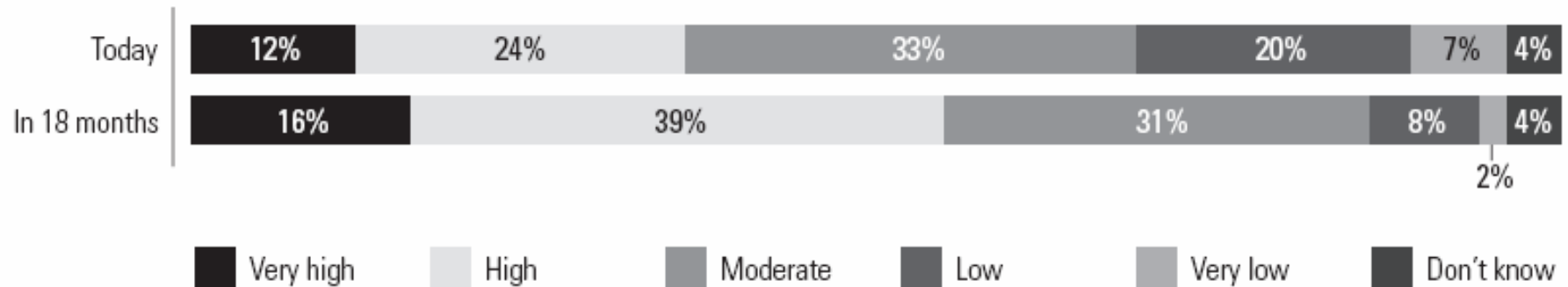
# Data Integration: Barrier to Application Development



# Top Three Data Integration Inhibitors



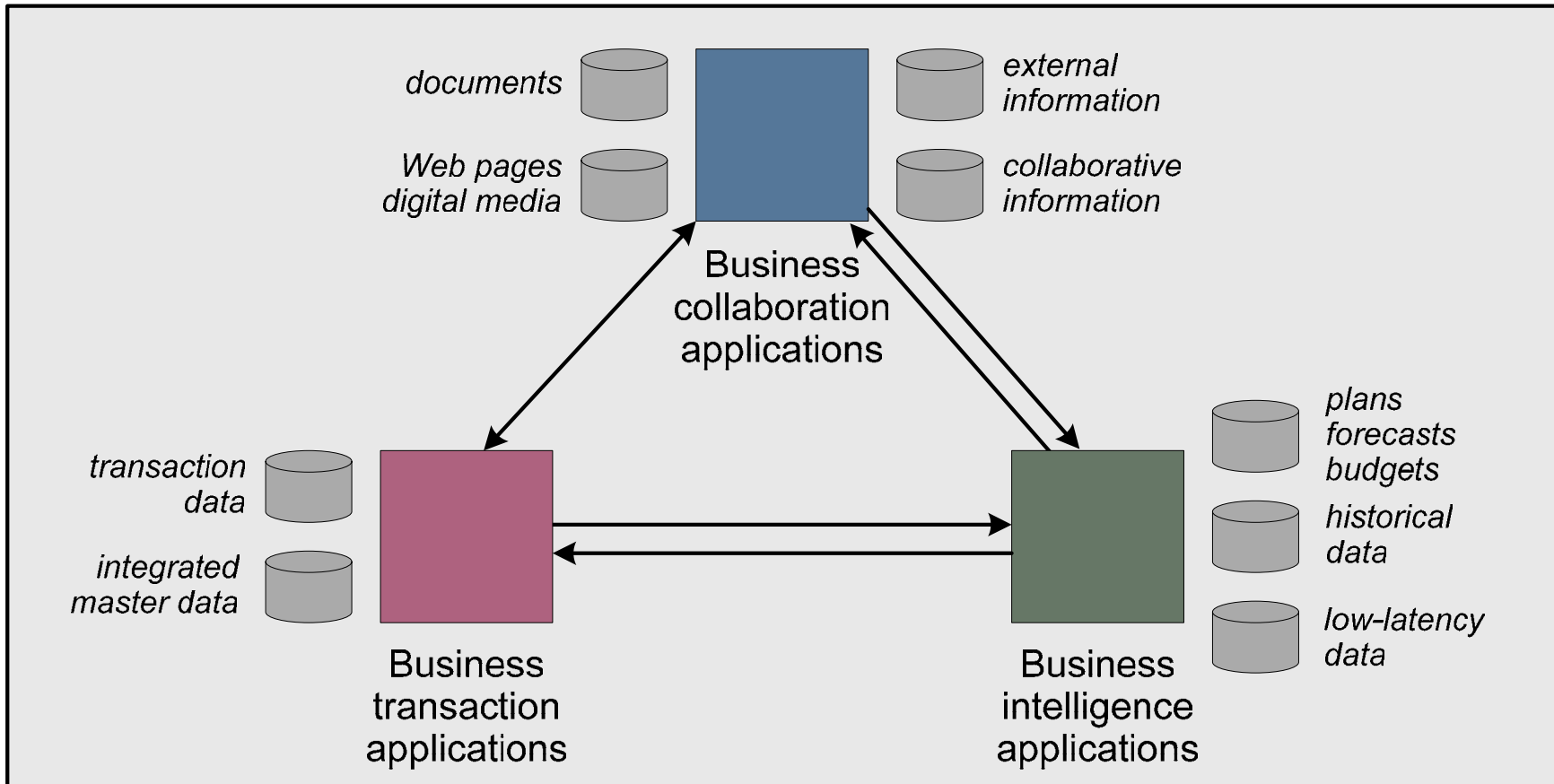
# Staffing and Budget for Data Integration



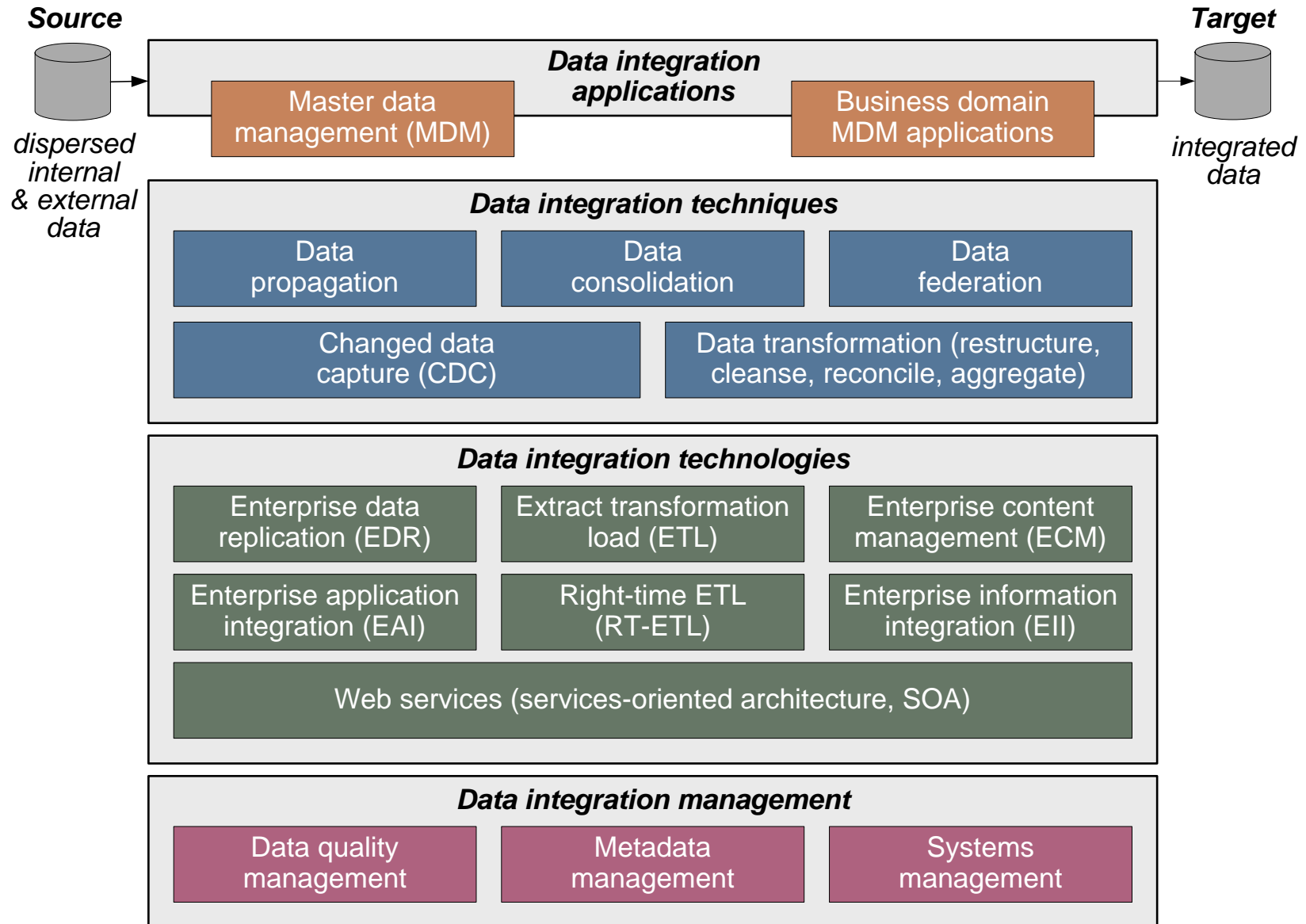
# Data Integration: A Definition

A framework of applications, products, techniques  
and technologies for providing a  
unified and consistent view  
of enterprise-wide business data

# Enterprise Business Data

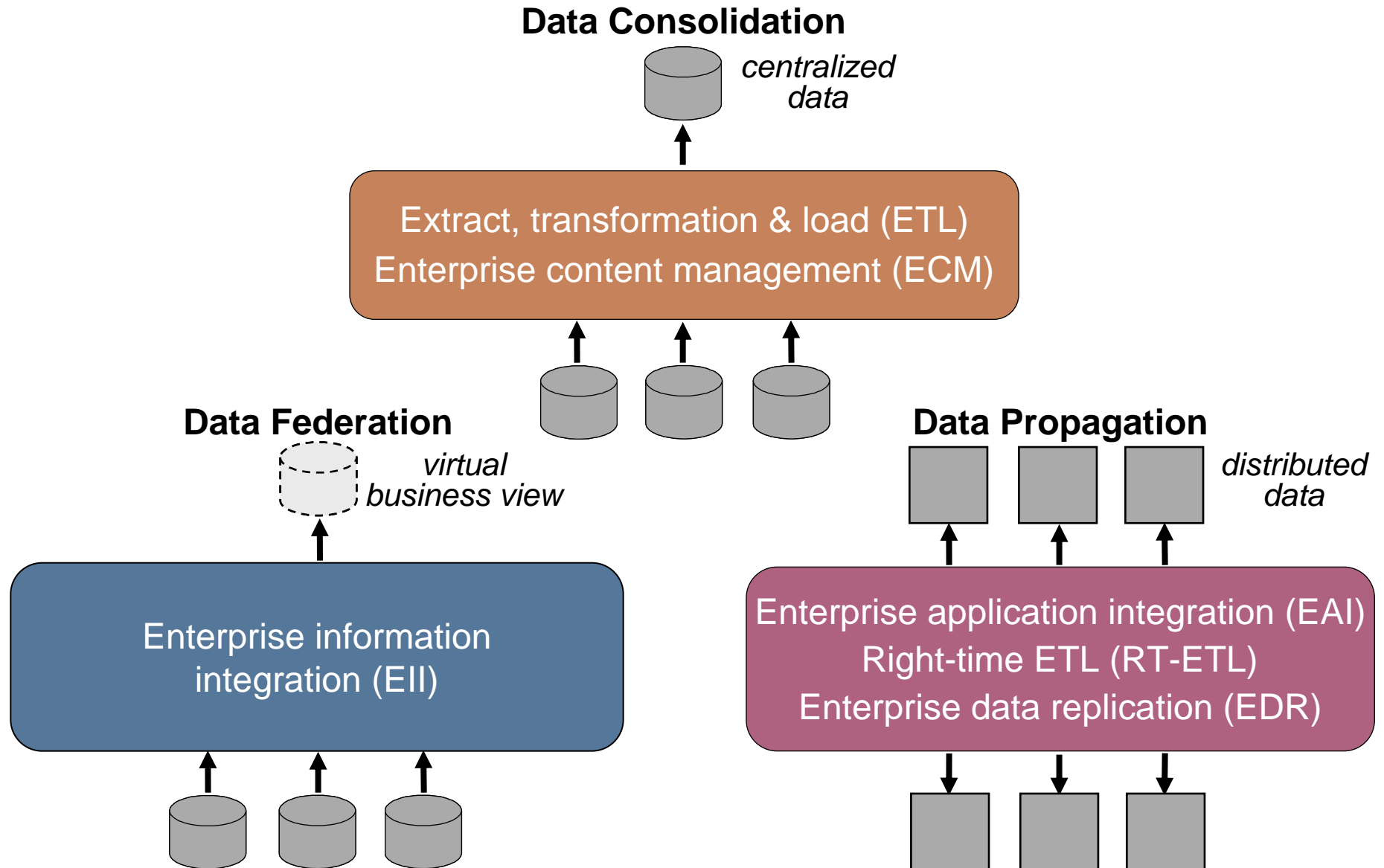


# Data Integration Architecture





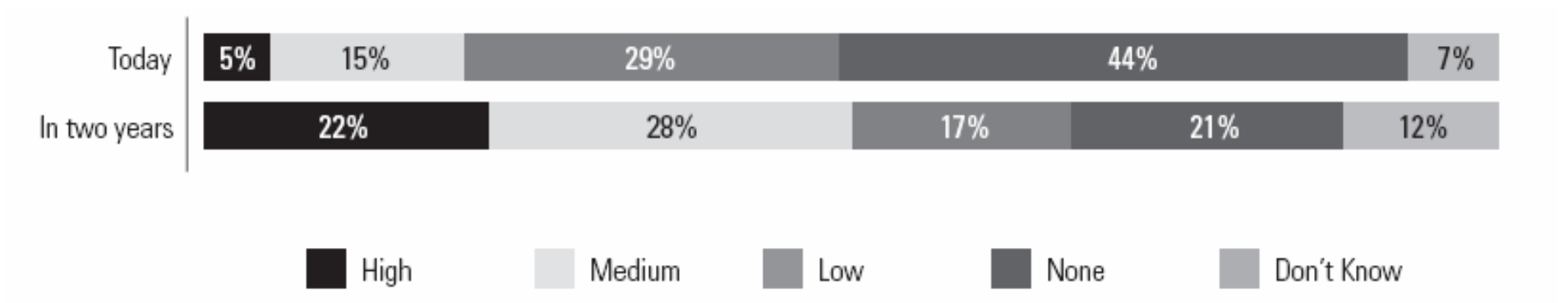
# Data Integration Techniques and Technologies



# Data Integration Applications

- Applications use data integration techniques, technologies and tools to produce a packaged data integration solution
- Master data management (MDM) and customer master data management (CMD) are examples of data integration applications
- May use a combination of data integration techniques and technologies
- CDM application example:
  - Use data consolidation to build a single store of customer data – data latency varies based on business requirements
  - Use data federation to build a virtual view of customer data – data is current
  - Use data consolidation for shared data and data federation for application specific data
  - Use data propagation to synchronize name and address data across data stores

# MDM Usage in Organizations

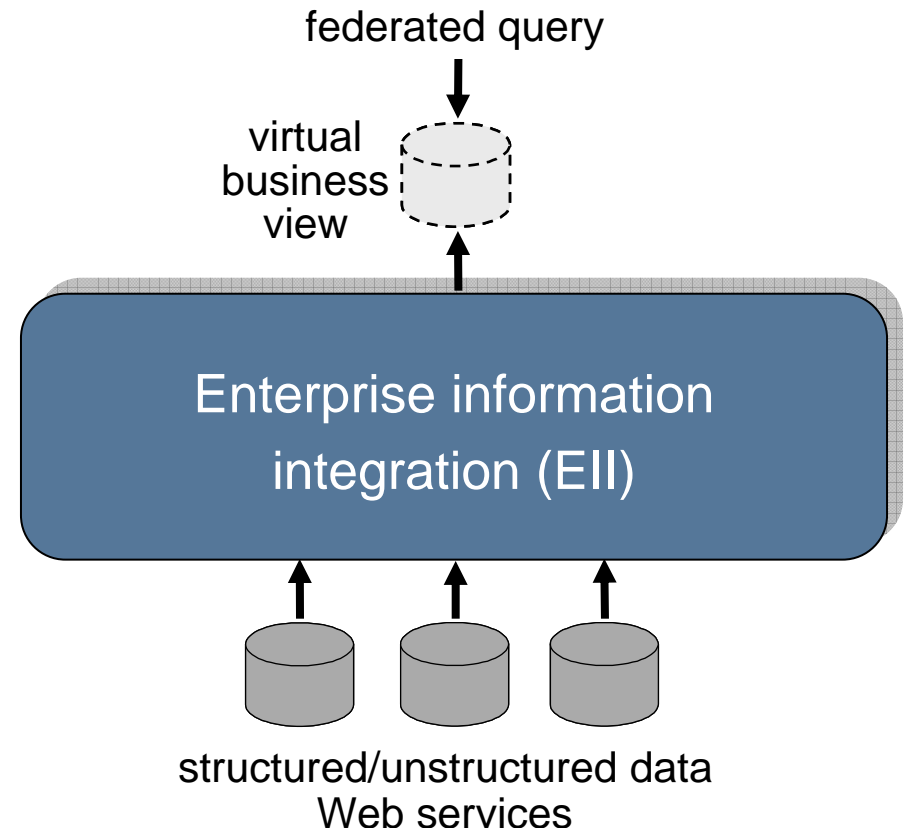


# Types of Data Integration Project

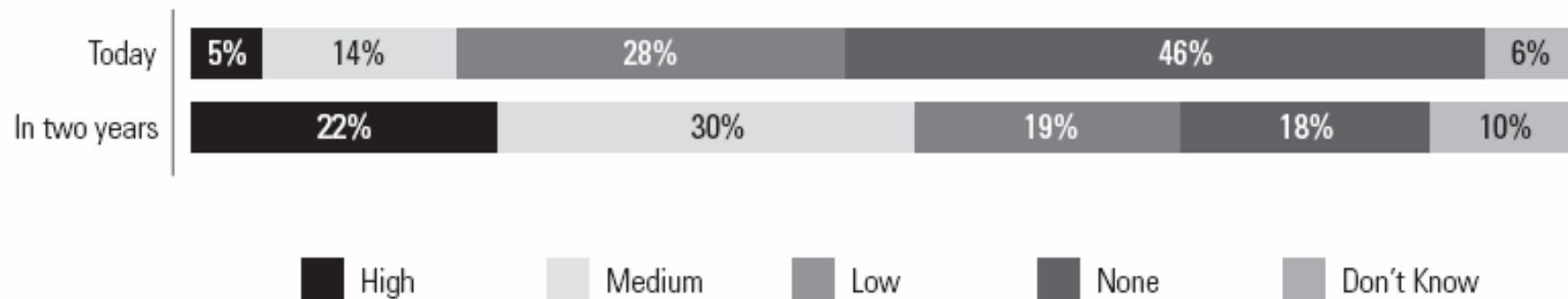


# Enterprise Information Integration (EII)

- EII provides *virtual* business views of dispersed data for demand-driven query access to live data by applications and business users
- EII can be used to access transaction, BI/DW and unstructured data
- EII offers a useful set of data integration technologies that enhance (rather than replace) the data warehousing and business intelligence environment



# EII Use in Organizations



# Using EII for Data Integration

## ■ Pluses

- Provides distributed query access to dispersed heterogeneous data without the need to extract or copy data
- Enables real-time access to live data (structured/unstructured)
- Supports dynamic business views

## ■ Minuses

- Not designed for complex queries against heterogeneous data
- Not suited for data with complex relationships or poor quality

## ■ Other thoughts

- Demand-driven (events can, however, drive federated queries)
- Products may have a query optimizer and support caching
- Data analysis and modeling similar to that required for ETL

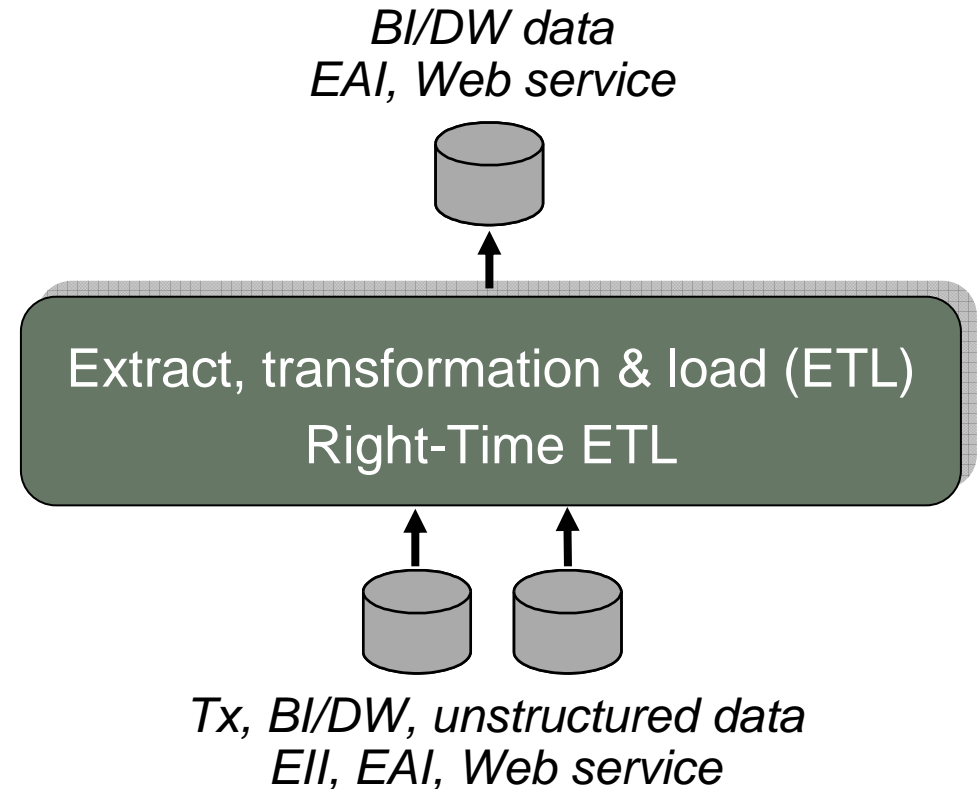
# When to Consider EII

- On-demand access to volatile data
- Direct write access to the source data (not all products support this)
- It is difficult to consolidate the original source data
- The cost of using federated queries is lower than the cost of data consolidation
- It is forbidden to make copies of the source data
- DW prototyping when the needs of users are not known in advance
- As temporary measure following a corporate acquisition

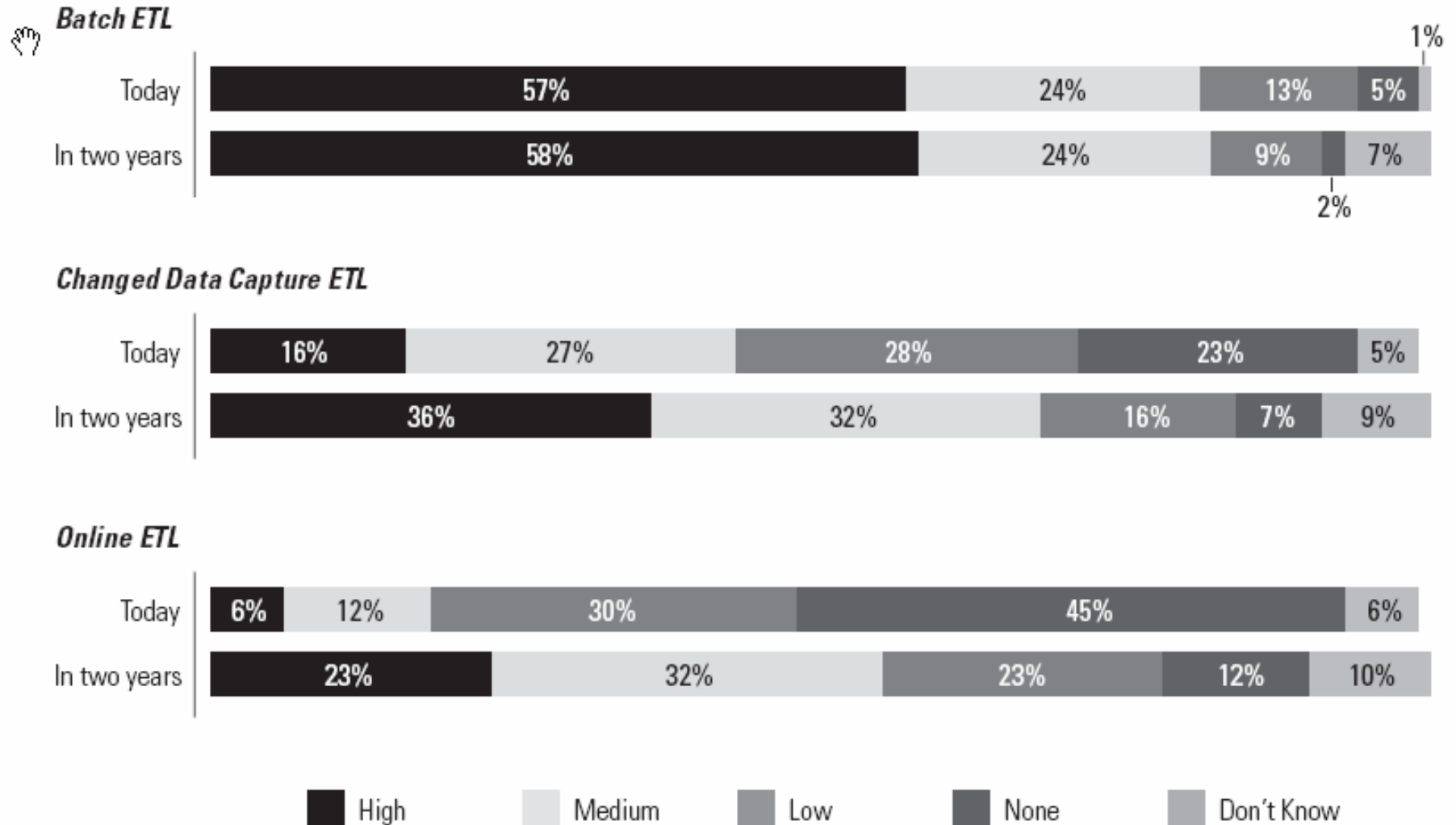


# Extract, Transformation and Load (ETL) Overview

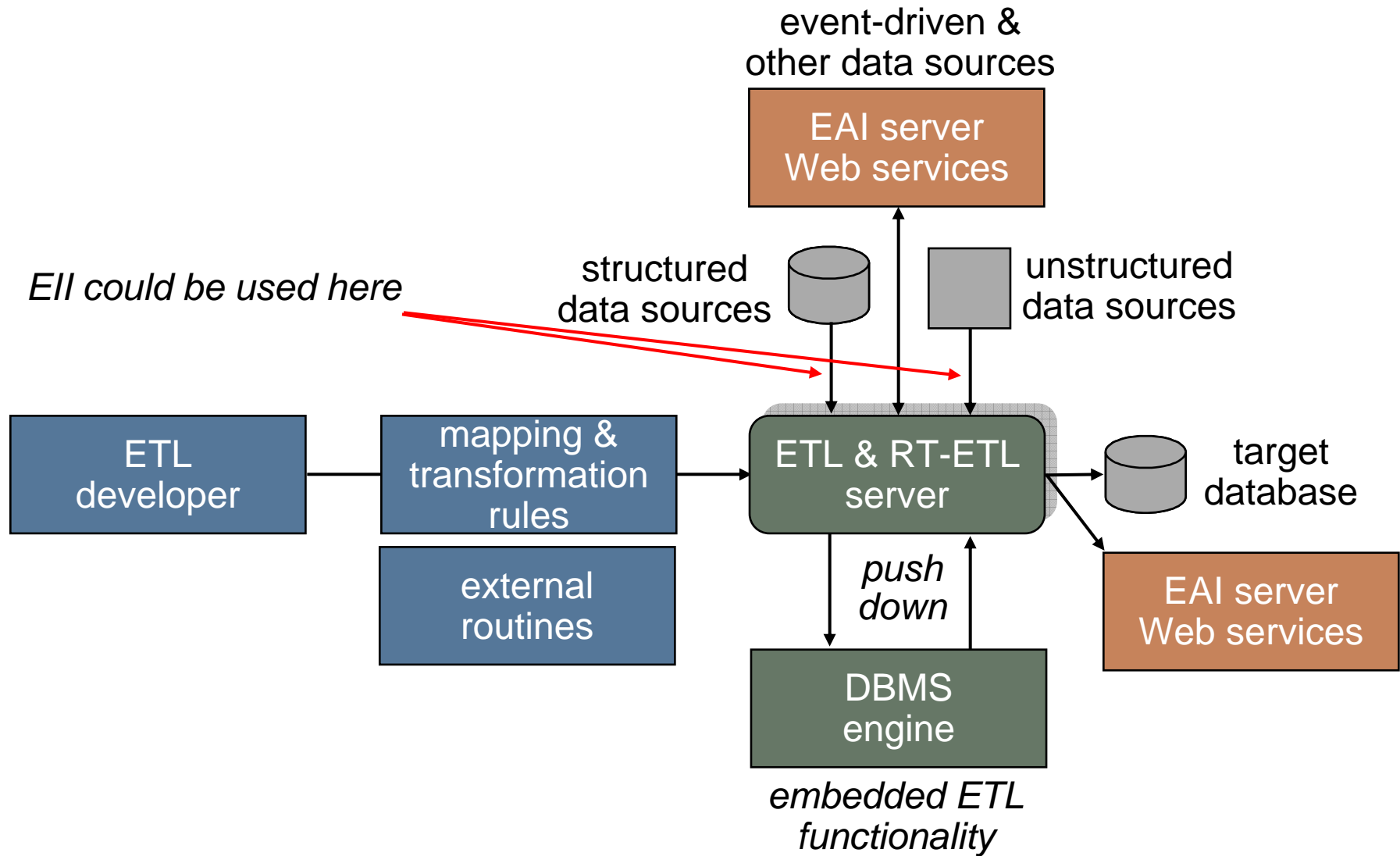
- Batch ETL
  - Extracts and transforms source data, and loads the results into a target data store
- Right-time ETL (RT-ETL)
  - Captures and transforms data events and uses the results to modify a target data store
  - Captures and transforms data events and routes the results to a message Q or Web service
  - Enables an ETL task to be called as a Web service
- Both EAI and EII servers can act as a source for ETL



# ETL Usage in Organizations



# ETL Today



# Using ETL for Data Integration

## ■ Pluses

- ETL: schedule-driven consolidation of data into a target data store
- RT-ETL: event-driven propagation of data into a target data store or EAI destination (e.g., message Q, Web service)
- Strong data transformation (restructuring, content cleaning, reconciliation and aggregation) capabilities

## ■ Minuses

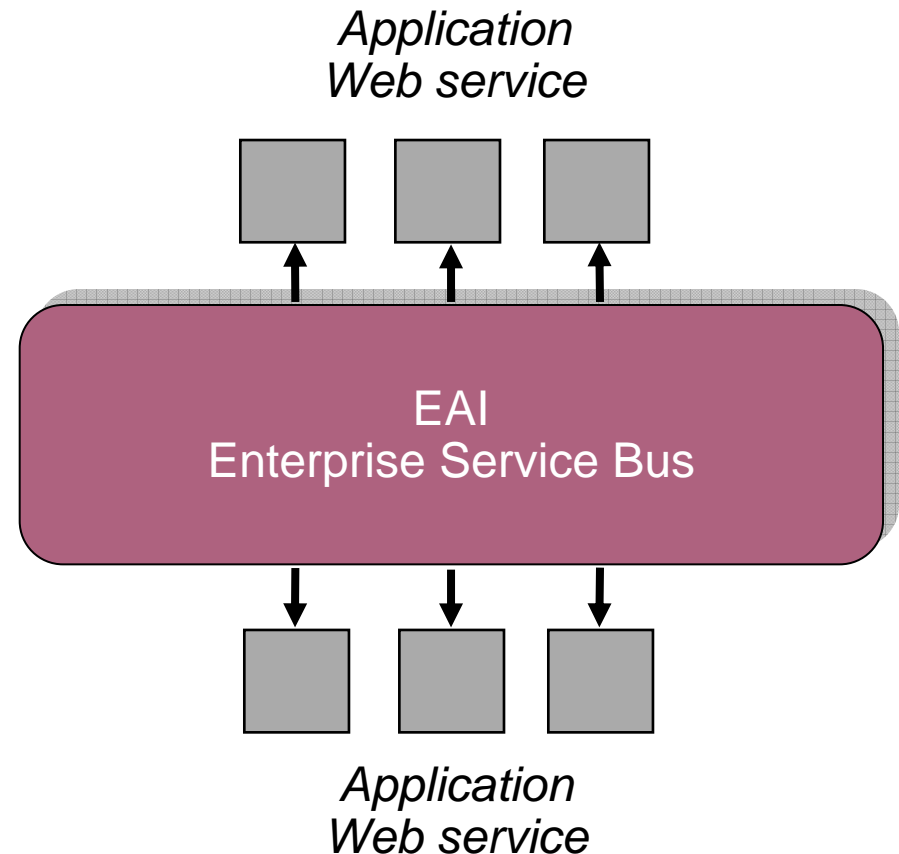
- Target data store not synchronized with source data stores
- Resources required to create and maintain target data store
- Usually the target data store contains snapshots of source data, rather than a complete historical record of source data changes

# When to Use ETL

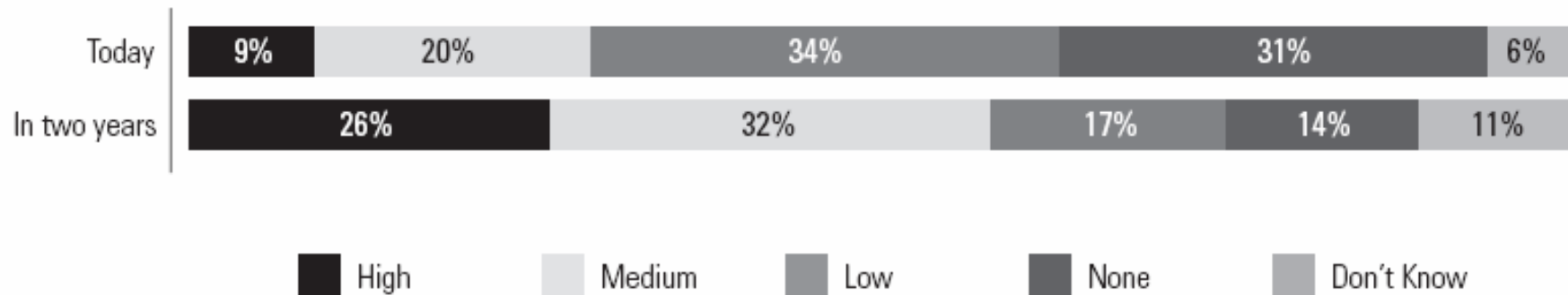
- Read-only access to reasonable stable data
- Users need historical or trend data
- Data access performance and availability are key requirements
- User needs are repeatable and can be predicted in advance
- Data transformation is complex
- Conclusions:
  - Both EII and ETL have a role to play in data integration
  - Organizations need an ETL/EII product set with shared metadata
  - EII can act as a source for ETL
  - EII may need to be deployed on a platform where ETL is not required, e.g., in a business portal

# Enterprise Application Integration (EAI) Overview

- EAI software may be offered a separate product or as a feature of a complete *application platform*
- Key EAI technologies for data integration are:
  - Messaging middleware (JMS, IBM WebSphere, Microsoft MSMQ, etc.)
  - Web services
- Additional features offered by an application platform may include:
  - Application server and enterprise portal
  - Business activity monitoring
  - Other integration-related software: business process management, master data management, EII, etc.



# EAI Usage for Data Integration in Organizations



# Using EAI for Data Integration

## ■ Pluses

- Enables applications to exchange transactions, messages and data
- Designed to propagate small amounts of “data” from one application system to another (synchronous or asynchronous)
- Supports real-time or near-real-time processing
- Good process/transaction semantics, message transformation and guaranteed message delivery
- Has evolved to support an enterprise service bus (ESB) and a services-oriented architecture

## ■ Minuses

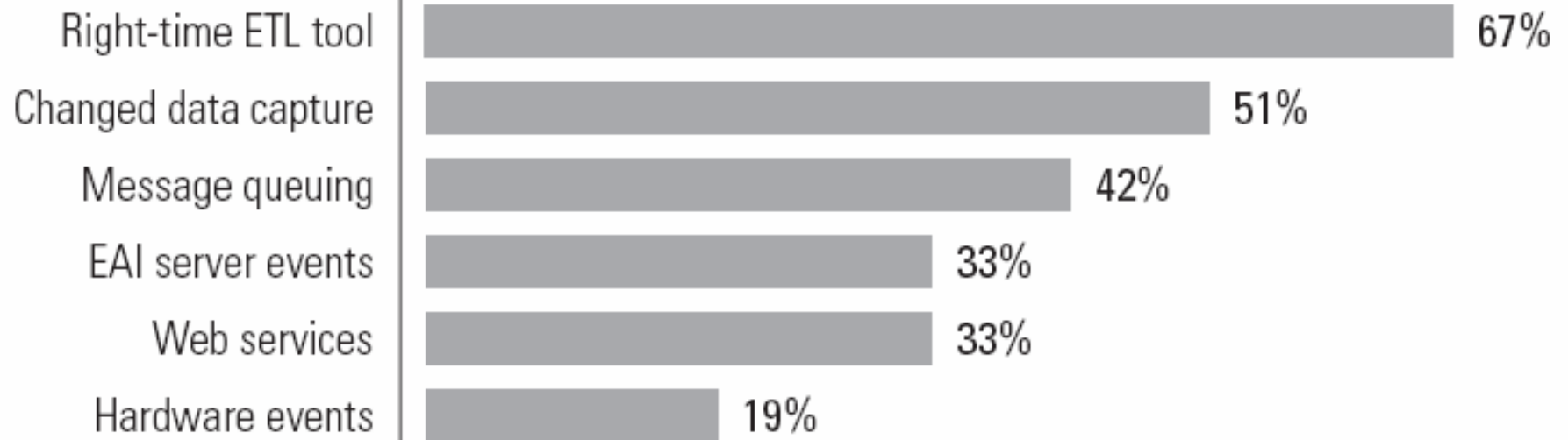
- Cannot handle complex data/metadata structures
- Products have a business process/application development focus and approach



# Thoughts About EAI

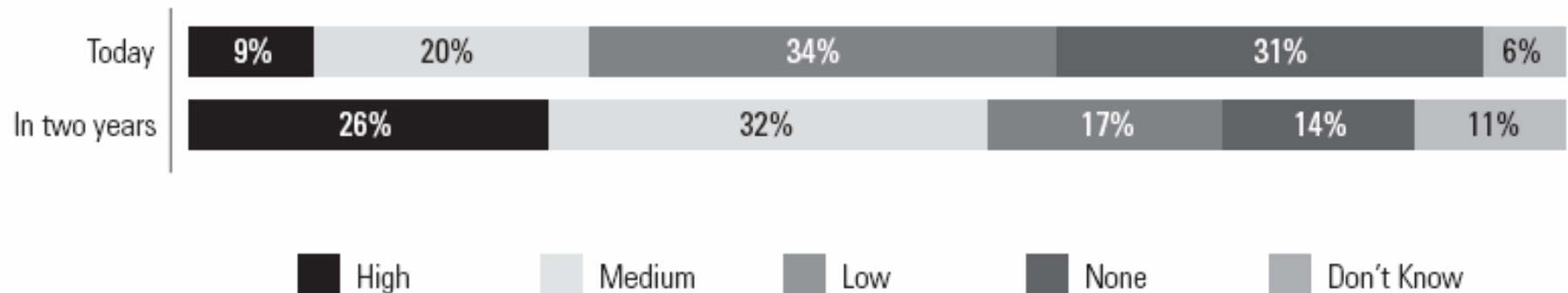
- EAI and ETL are not competing technologies – even though some vendors would have you believe otherwise!
- EAI provides access to a wide range of enterprise data
  - An EAI-to-ETL interface can give ETL access to this data
  - This interface can provide access to real-time events and can also be a target for RT-ETL
  - The interconnection can be done using a message Q, Web services, etc.
- Current EAI-ETL interfaces are often immature – many companies are using EAI to create files for input to ETL
- EAI applications can use ETL as service for data transformation, etc.

# Data Source for Right-Time Data Integration

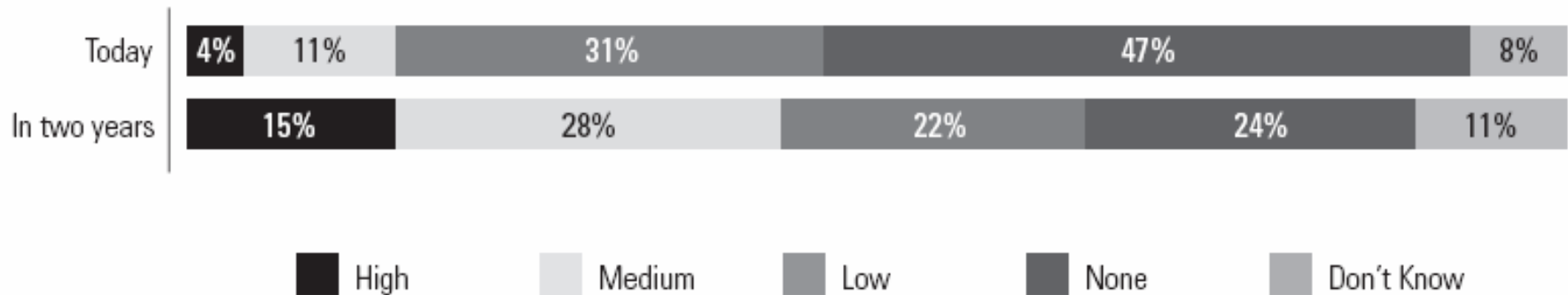


# Usage of Other Data Integration Technologies in Organizations

## Enterprise Data Replication



## Enterprise Content Management



# Data Quality Considerations

- Two key data quality tasks in data integration
  - Analysis of the source data
  - Cleansing of poor quality data
- Source data analysis
  - Determines source data contents and quality – data profiling tools can be used here
- Source data cleansing
  - Done by inserting a data transformation process in the data integration workflow
  - Transformation involves data restructuring, data cleansing, data reconciliation and data aggregation
  - The data cleansing and reconciliation steps are where data quality issues are addressed most often
  - Data cleansing may be handled by a data integration tool, a third-party product, or both

# Selecting the Right Solution: Application Variables

Source  
data  
type

*structured*  
*semi-structured*  
*unstructured*  
*packaged application*  
*EAI/Web service*  
*metadata*

Source  
data  
organization

*homogeneous/heterogeneous*  
*centralized/distributed*  
*federated/dispersed*

Data  
scale

*number of data sources*  
*data store size*  
*data store volatility*

Source data  
transformation

*restructuring*  
*reconciliation*  
*cleansing*  
*aggregation*

Target data  
currency & access

*real-time*  
*near-real-time*  
*point-in-time*  
*read-only/read-write*

Data  
integration  
technique & mode

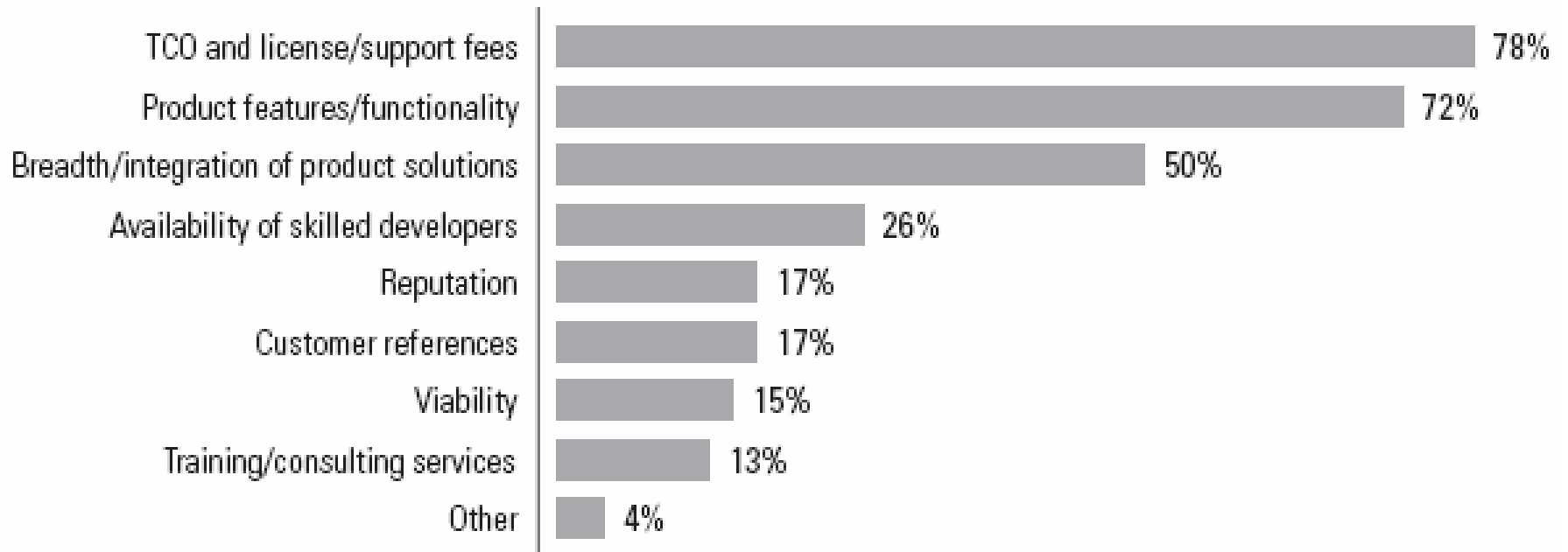
*consolidation*  
*federation*  
*propagation*  
*changed data capture*  
*synchronous/asynchronous*  
*on-demand pull/event push*

Data  
integration  
technology

*ETL & RT-ETL*  
*EAI, EDR*  
*EII, ECM*

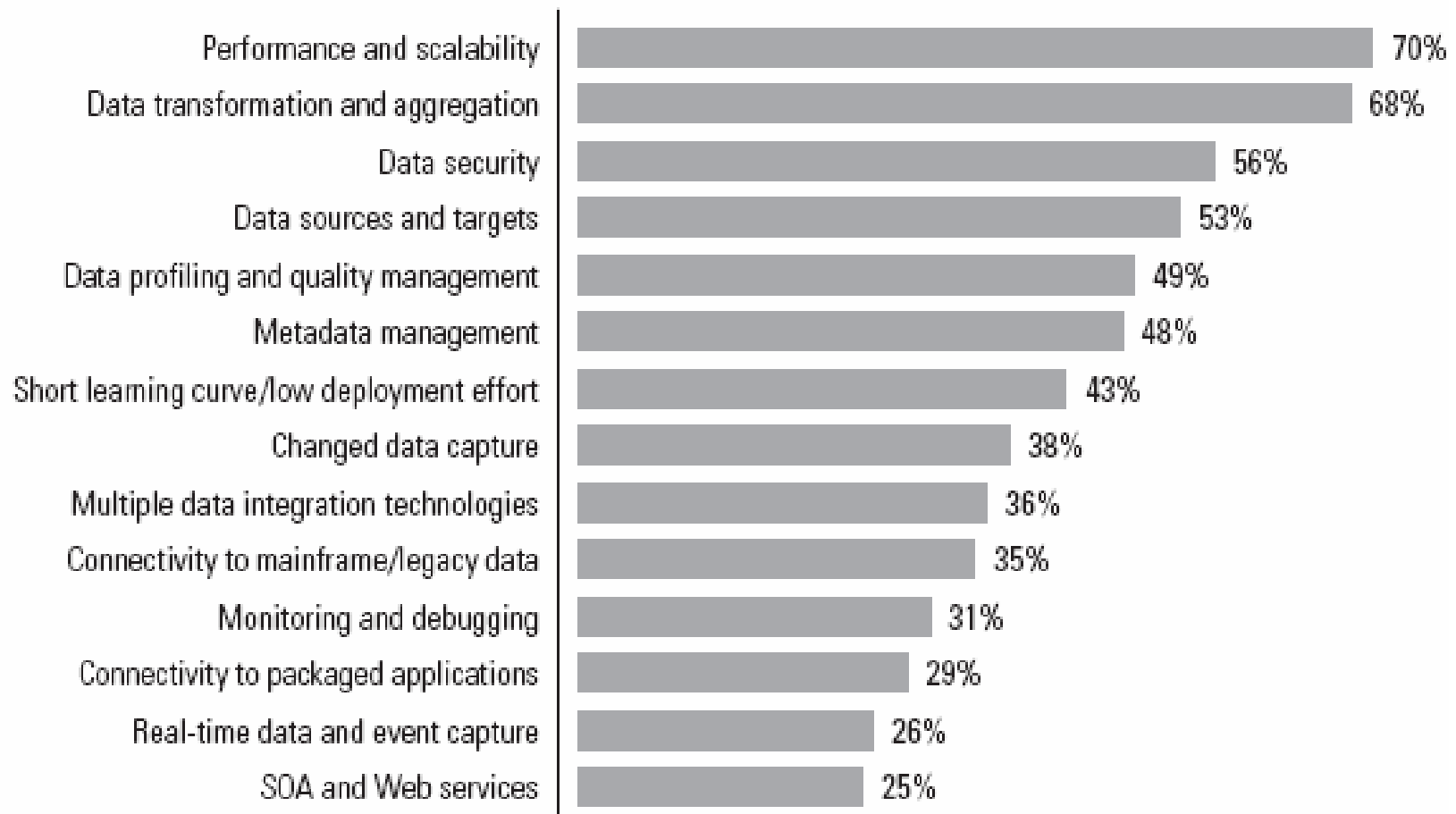
# Criteria Used to Choose a Data Integration Vendor - 1

## Main Criteria for Choosing a Data Integration Vendor

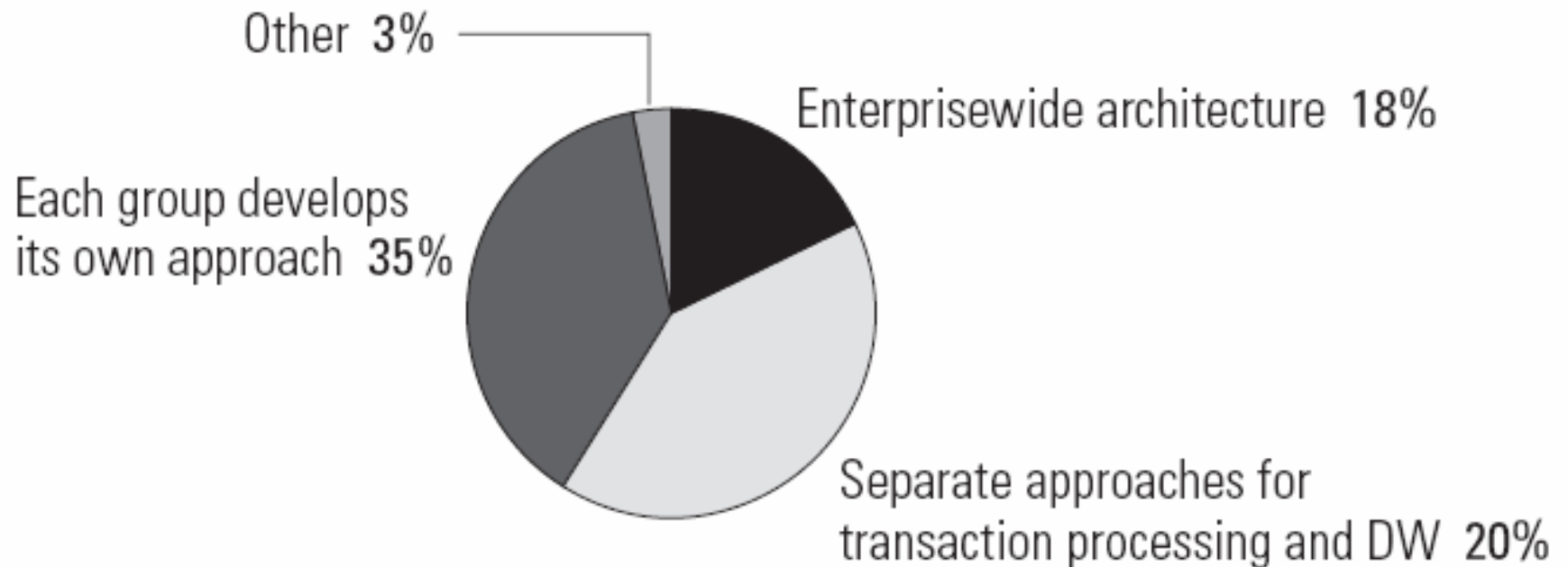


# Criteria Used to Choose a Data Integration Vendor - 2

## Most Important Data Integration Features



# Developing a Data Integration Strategy





# The Integration Competency Center (ICC)

“Our survey showed that 11 percent of respondents have a data integration competency center, and 28 percent plan to have one.”

“Follow-on interviews indicated that while most companies had created a competency center to provide integration services and support to line-of-business development groups, many also created the center to eliminate the political battles and turf wars over which IT department should deploy new integration applications.”

“Often, these centers not only cover data integration, but also the complete field of enterprise integration and business intelligence.”

# Data Integration: TDWI Study Conclusions

- *Large organizations* are moving toward building an enterprise-wide data integration architecture
  - Have a multitude of data stores and large amounts of legacy data
  - Focus on buying an integrated product set and are interested in leading edge data integration technologies
  - Buy high-performance best-of-breed products to handle the integration of large amounts of data
  - More likely to have a data integration competency center.
  
- *Medium-sized companies* are focused on data integration solely from a BI viewpoint
  - Evaluate products from the perspective of how well they will integrate with the organization's BI tools and applications
  - Often have less legacy data
  - Less interested in leading-edge approaches such as right-time data and Web services