IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

526 Data Warehousing

April 20, 2016

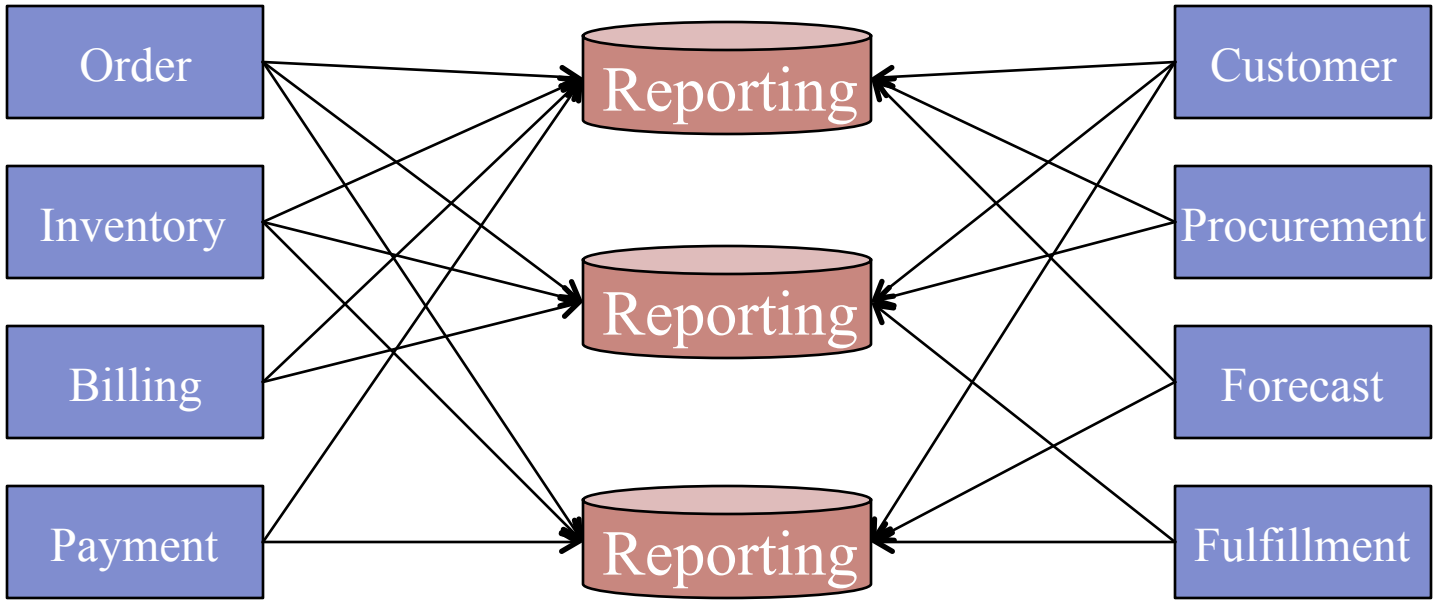
Week 14 Presentation

ITMD - 526

The Future of Data Warehousing

Reporting in 1970s and 1980s

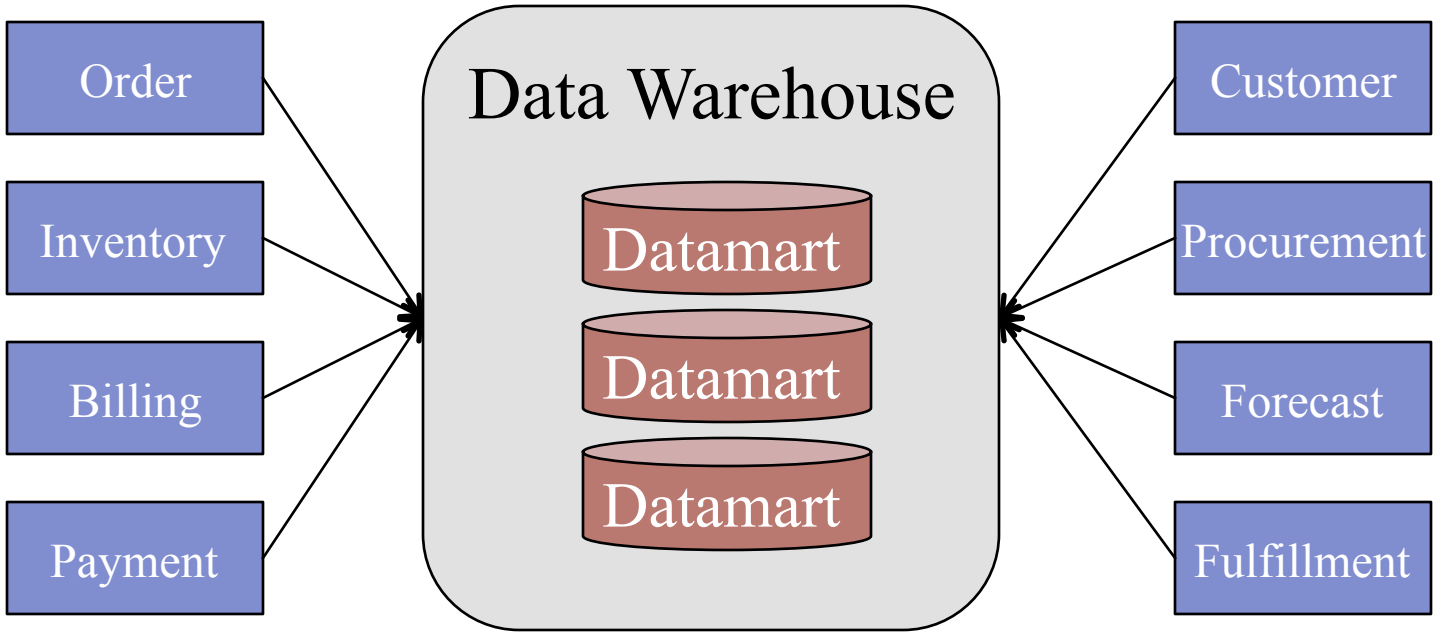
- Management Information Systems (MIS)
- Decision Support System (DSS)
- Executive Information Systems (EIS)
- On-Demand Extracts



The Future of Data Warehousing

Reporting in 1990s and 2000s

- Data Warehousing and Business Intelligence
  - Star Schema/OLAP
  - Dashboards
  - Visualizations



The Future of Data Warehousing

DW/BI Delivers Value

- Dimensional slice and dice of our data
- Mature tools from leading vendors
- Mature and proven modeling
- Years of best practices
- Many skilled BI and data warehousing professionals

The Future of Data Warehousing
Demands on Advanced Analytics

Analytics

- Discovery, Interpretation, and Communication of Meaningful Patterns in Data

(Reference: https://en.wikipedia.org/wiki/Analytics)

The Future of Data Warehousing

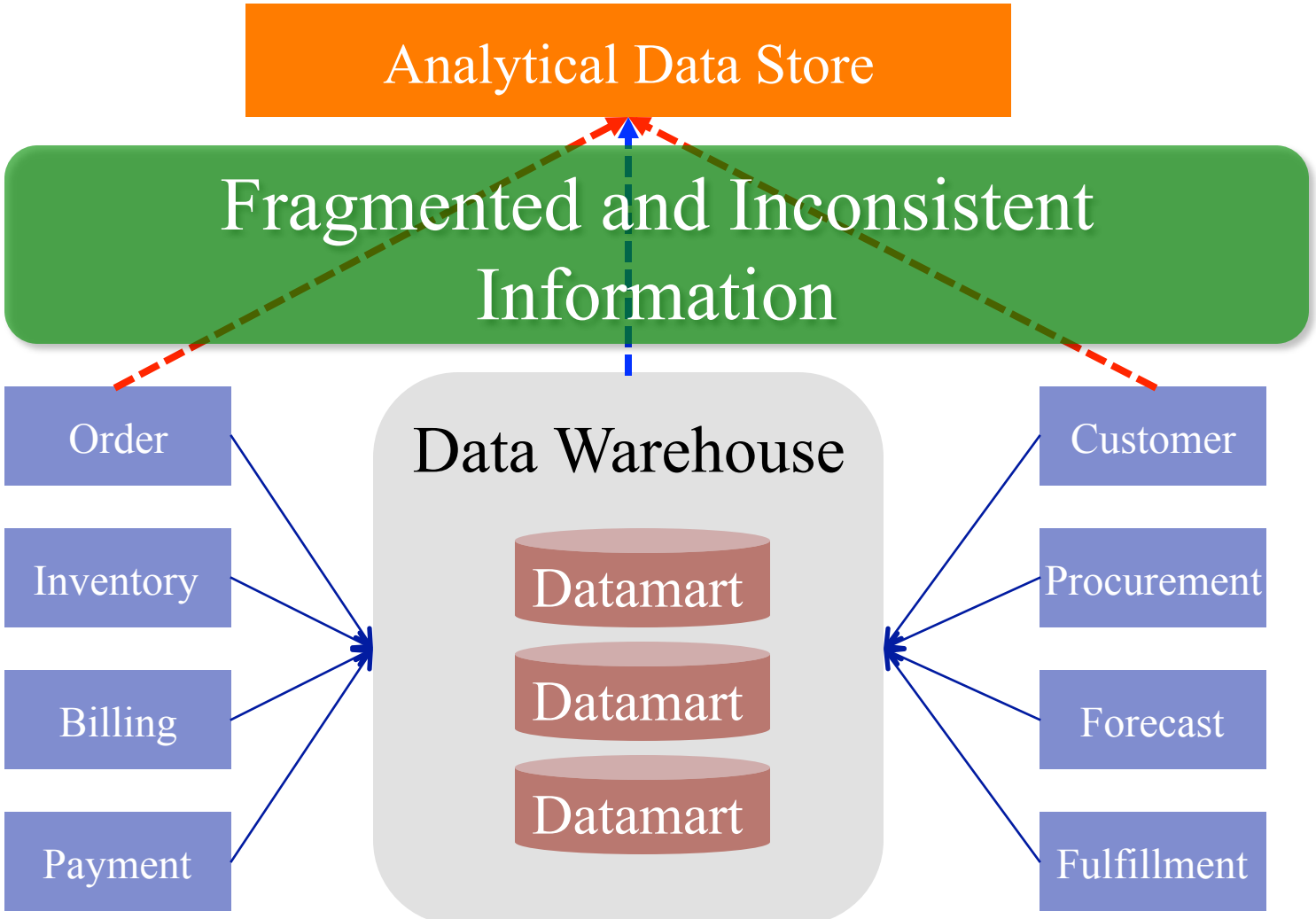
Three Levels of Analytics

ITMD - 526

Prescriptive	**Tell me what my options are?** **Tell me what I should do?**	• **Optimization** • **Simulation** • **MCDM/ Heuristics**
Predictive	**What else is most likely to happen?** **How else will it happen?** **How long will it continue to happen?**	• **Data/Text Mining** • **Forecasting** • **Statistical Analysis**
Descriptive	**What did happen, and why?** **What is happening, and why?** **Data Warehousing's Primary Focus**	• **Dashboards** • **Ad Hoc Reports** • **Standard Reports** • **Dimensional Modeling**

The Future of Data Warehousing

Current Approaches for Advanced Analytics



The Future of Data Warehousing
Traditional DW/BI Methodologies

Data First

Build data warehouse, then focus on data use

- Business Process
- Dimensions
- Fact
- Business Rules

Functionality Oriented

Focus on data use, then build data warehouse

- Reports, Dashboards
- Metrics
- KPIs
- Business Rules

The Future of Data Warehousing

DW/BI Methodologies – Reality Check

Regardless of the methodology, business rules are critical

- Deep Understanding on Data and Organization is Required
  - Organizational data
  - How organization will use data
  - How data is related to other data

… Before Designing and Building a Data Warehouse

The Future of Data Warehousing

DW/BI Methodologies – Reality Check

Disadvantages

- Data warehouse are built slowly
- Enhancing a data warehouse is also time consuming
- Often leads to separate, faster, and fragmented solutions
  - Kimball approach without Bus Architecture
  - Inmon's CIF approach bypassing normalized EDW
  - Varying definition, inconsistent reports

The Future of Data Warehousing

Real-Time Need for Data Warehousing

Data warehouse is not timely enough for advanced analytical needs

- Operational BI
  - Near real-time data warehouse
  - Possible fraudulent customer activity
  - Network intrusion
- Predictive Analytics
- Prescriptive Analytics

The Future of Data Warehousing

Analytics Gap in Traditional DW/BI

Need	Support
Unlimited source data breadth and depth	Difficult
Quickly add new data sets	Difficult
Real-time data and analysis	Difficult
Semi-structured and unstructured data	Difficult
Support predictive and prescriptive analytics	Difficult

The Future of Data Warehousing

Addressing Latency with Traditional DW/BI

- Increase the frequency of batch ETL feeds
- Switch to real-time messaging instead of ETL
- Build spinoff solutions (real-time data marts)

These are not natural for traditional data warehouse architecture

The Future of Data Warehousing

Big Data's Three Vs

Volume	• Source from hundreds or thousands of sources
Velocity	• Real-time data feeds • Ability to quickly add new data sources • High-velocity data ingestion (ELT instead of ETL)
Variety	• Structured • Semi-Structured • Unstructured

The Future of Data Warehousing

Big Data Paradigm

- Ingest and store now
  - Data Lake, Super-Sized Staging, Landing Zone, etc.
- Organize later
  - ELT instead of ETL
- New generation of tools & Ecosystem
- Beyond relational database rules

The Future of Data Warehousing

Hadoop

- One of Big Data technologies
- Breaks the VOLUME barrier of relational databases
- Support a VARIETY of data formats in same system
- Is Architected for high-VELOCITY data intake and usage

The Future of Data Warehousing

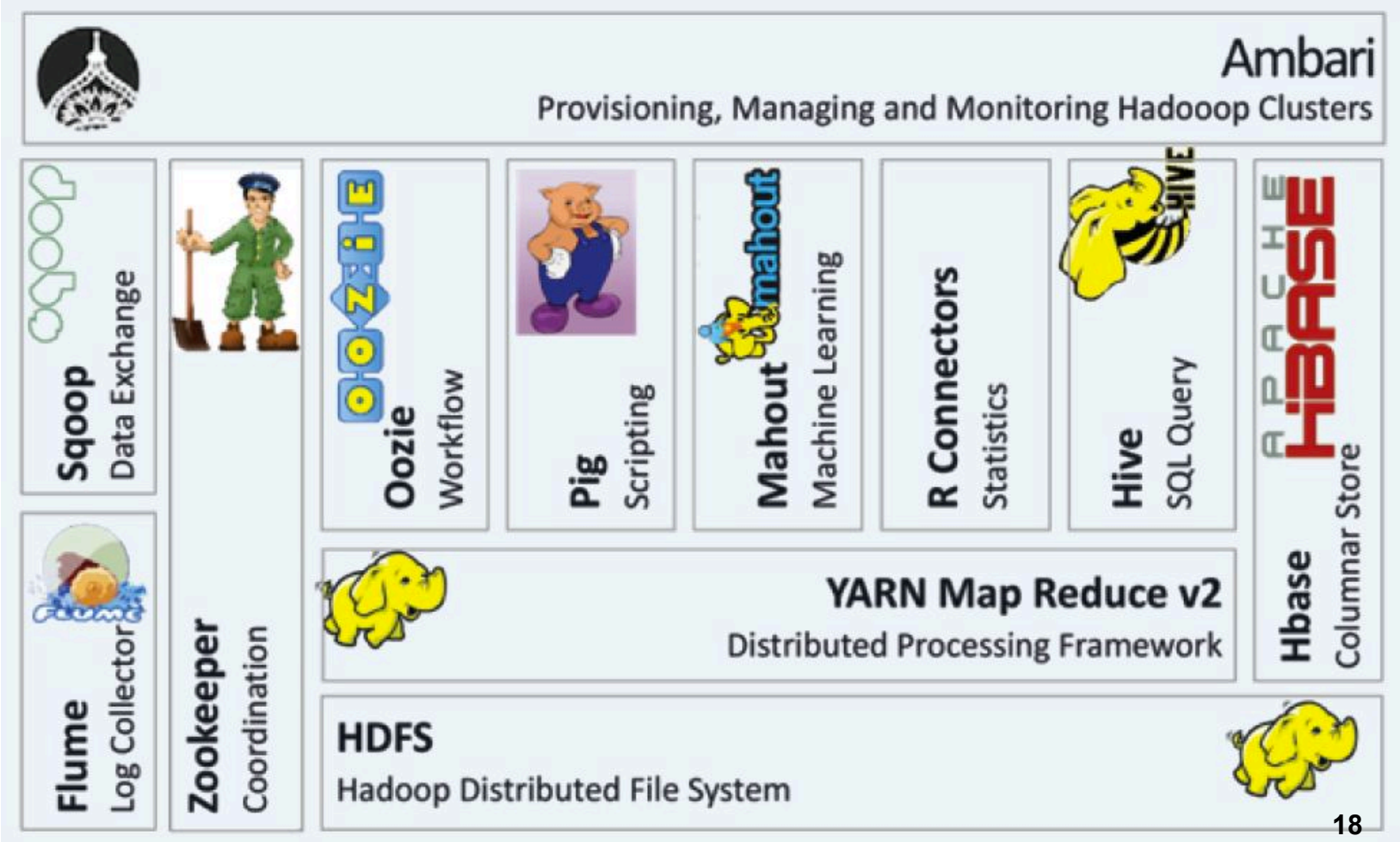
HDFS: The Core of Hadoop

HDFS: The Core of Hadoop

- Hadoop Distributed File System
- Distributes and manages data across many different commodity server
- Similar to DDBMS

ITMD - 526

The Future of Data Warehousing
Hadoop Ecosystem



The Future of Data Warehousing

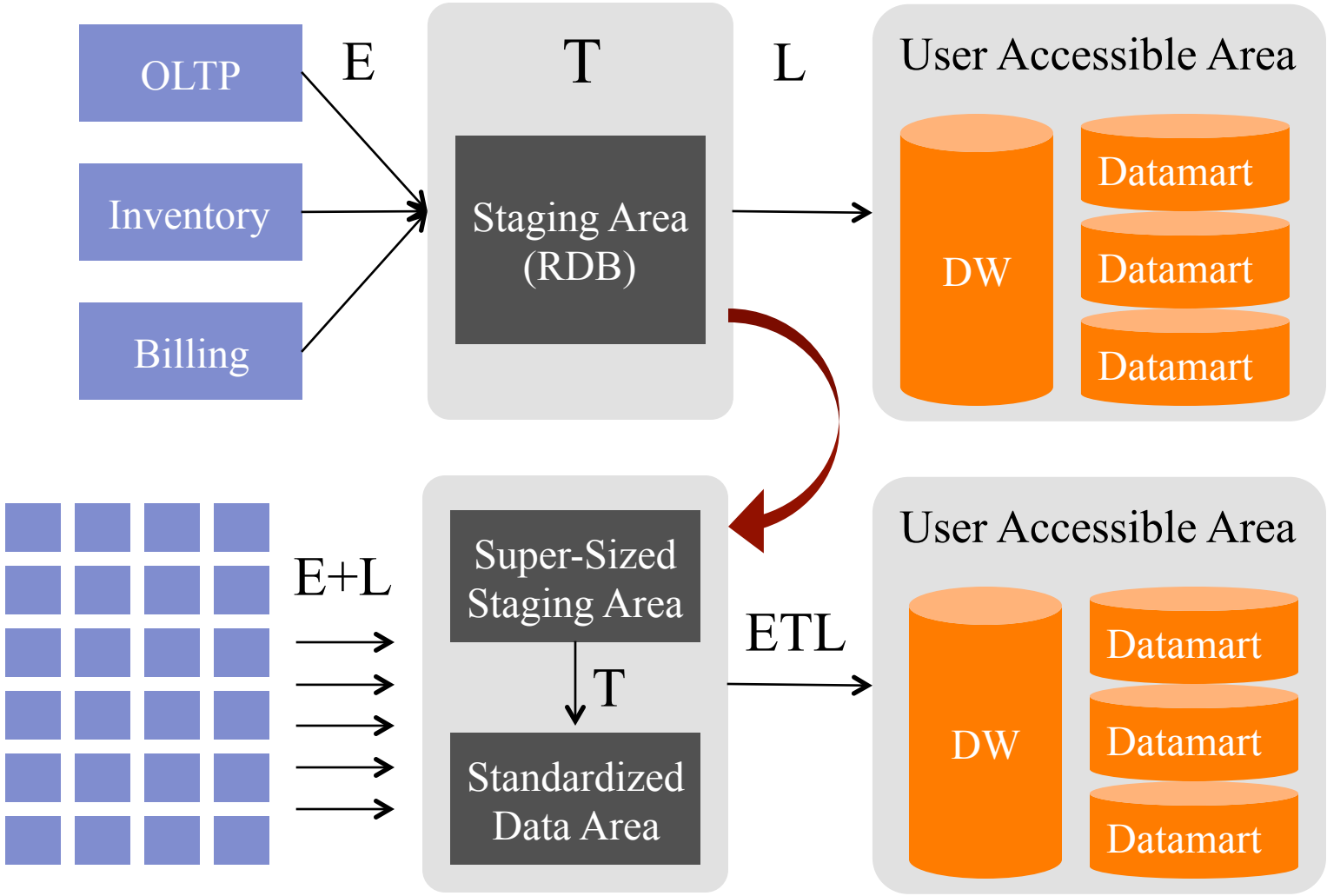
Two Hadoop DW Approaches

Super-Sized Staging Area

Non-Relational EDW

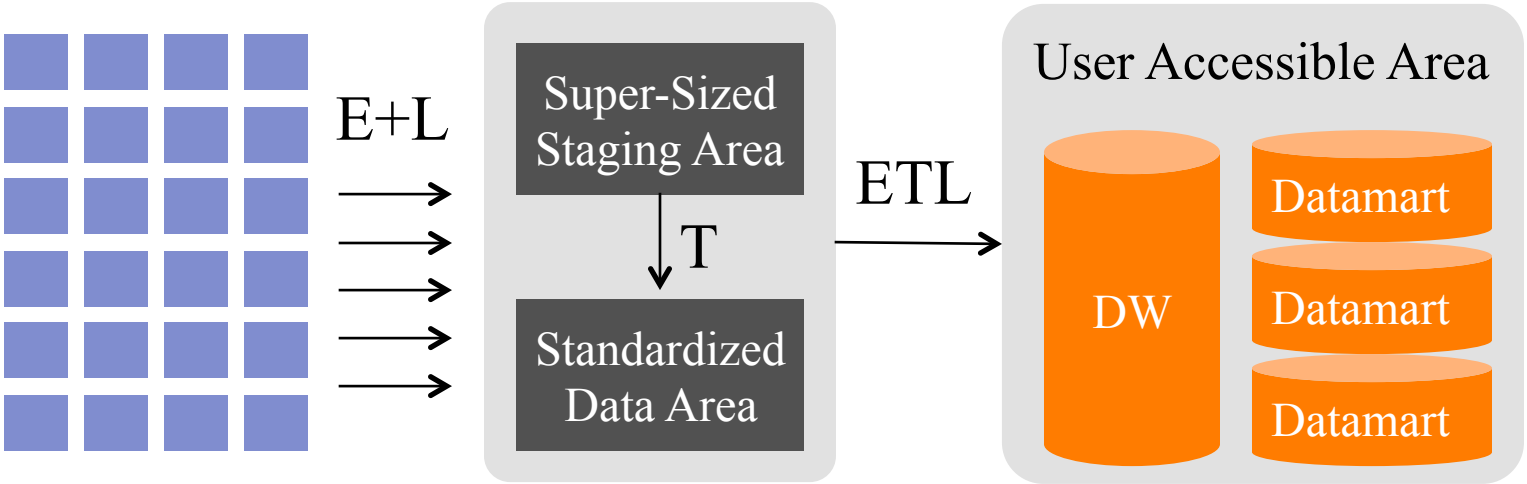
ITMD - 526

The Future of Data Warehousing
Super-Sized Staging Area



ITMD - 526

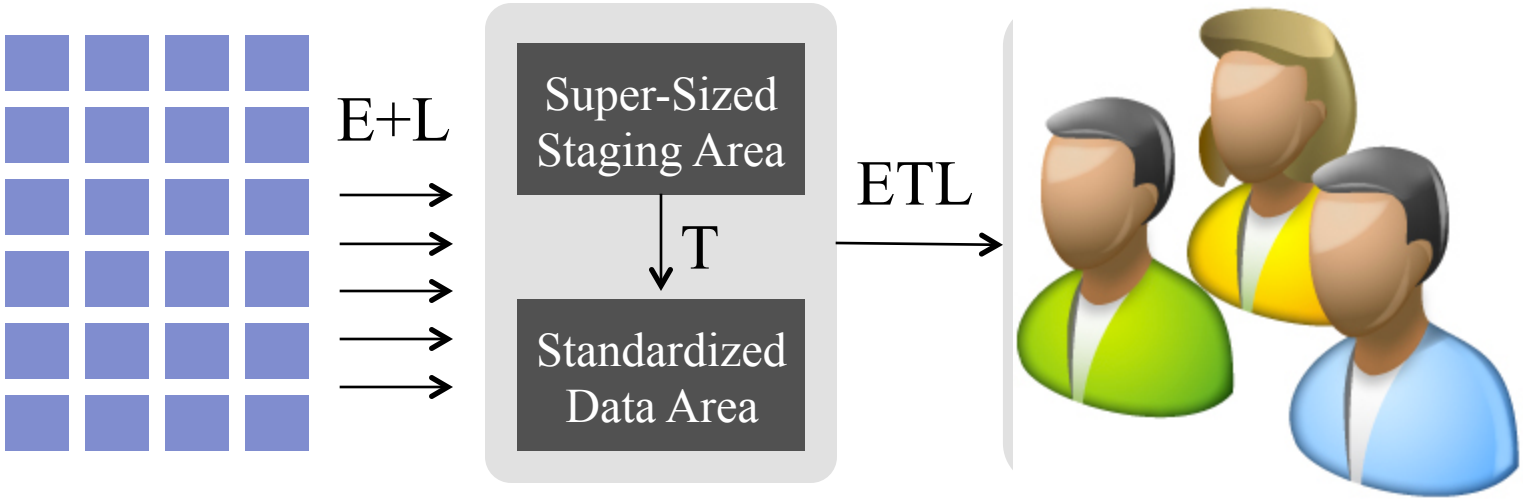
The Future of Data Warehousing
Super-Sized Staging Area



Need	Support
Unlimited source data breadth and depth	Good
Quickly add new data sets	Good
Real-time data and analysis	Partial
Semi-structured and unstructured data	Partial
Support predictive and prescriptive analytics	Difficult

ITMD - 526

The Future of Data Warehousing
Non-Relational EDW



Need	Support
Unlimited source data breadth and depth	Good
Quickly add new data sets	Good
Real-time data and analysis	Good
Semi-structured and unstructured data	Good
Support predictive and prescriptive analytics	Good
Data Governance & Security	Difficult

The Future of Data Warehousing
ETL vs. ELT

	ETL (Traditional Approach)	ELT (Big Data Approach)
Source Data	Some	All
Data Transfer	Primarily batch	Bulk and Streaming
Data Cleaning	Before load	Deferred
Master Data Standardization	Before load	Deferred

The Future of Data Warehousing

ELT in Big Data

- Extraction in Big Data
  - Theoretically all source data can be extracted
  - No need for deep requirements to drive extraction
  - Mixture of bulk load and streaming
- Loading in Big Data
  - Direct pipeline to extraction process
  - Load the data that has been extracted
  - No cleaning or norming before loading

The Future of Data Warehousing

ELT in Big Data (cont'd)

- Transformation in Big Data
  - Transformation as needed
  - Data mining can still use "dirty" data
  - Downstream relational EDW? Transformation then
  - Big Data (Hadoop) serving as EDW? Set aside "clean data" area

The Future of Data Warehousing

ETL & ELT :Different Goals

Data Warehousing ETL	Big Data ELT
Clean, standardized data closely aligned with specific requirements	As much data as possible, as quickly as possible, even without specific needs identified

The Future of Data Warehousing

Assessing Analytics Status

	Complexity	Quality	Timeliness	Support
Descriptive Analytics	Good	Good	Challenge	Good
Predictive Analytics	Challenge	Challenge	Challenge	Challenge
Prescriptive	Challenge	Challenge	Challenge	Challenge

- Issue with high latency?
  - Greater need for data velocity
  - MPP (Massively Parallel Process), HDFS, etc.
- Weak predictive and prescriptive analytics?
  - Greater need for ETL with Big Data
- Mature statistical assets with compromised/fragmented data?
  - More focus on DW/BI and data governance

The Future of Data Warehousing

Reasons Not to Invest in Big Data

- Analytical needs overwhelmingly descriptive (DW/BI)
- Data needs highly compartmentalized and fragmented
- Reports seldom used, especially by executives
  - Decisions often made by force of personality
- Adversarial culture to "what the data says"
  - To overcome this, the primary focus should be in building enthusiasm and interests through pure analytical packages (SAS, SPSS, R, etc.), not Big Data

The Future of Data Warehousing

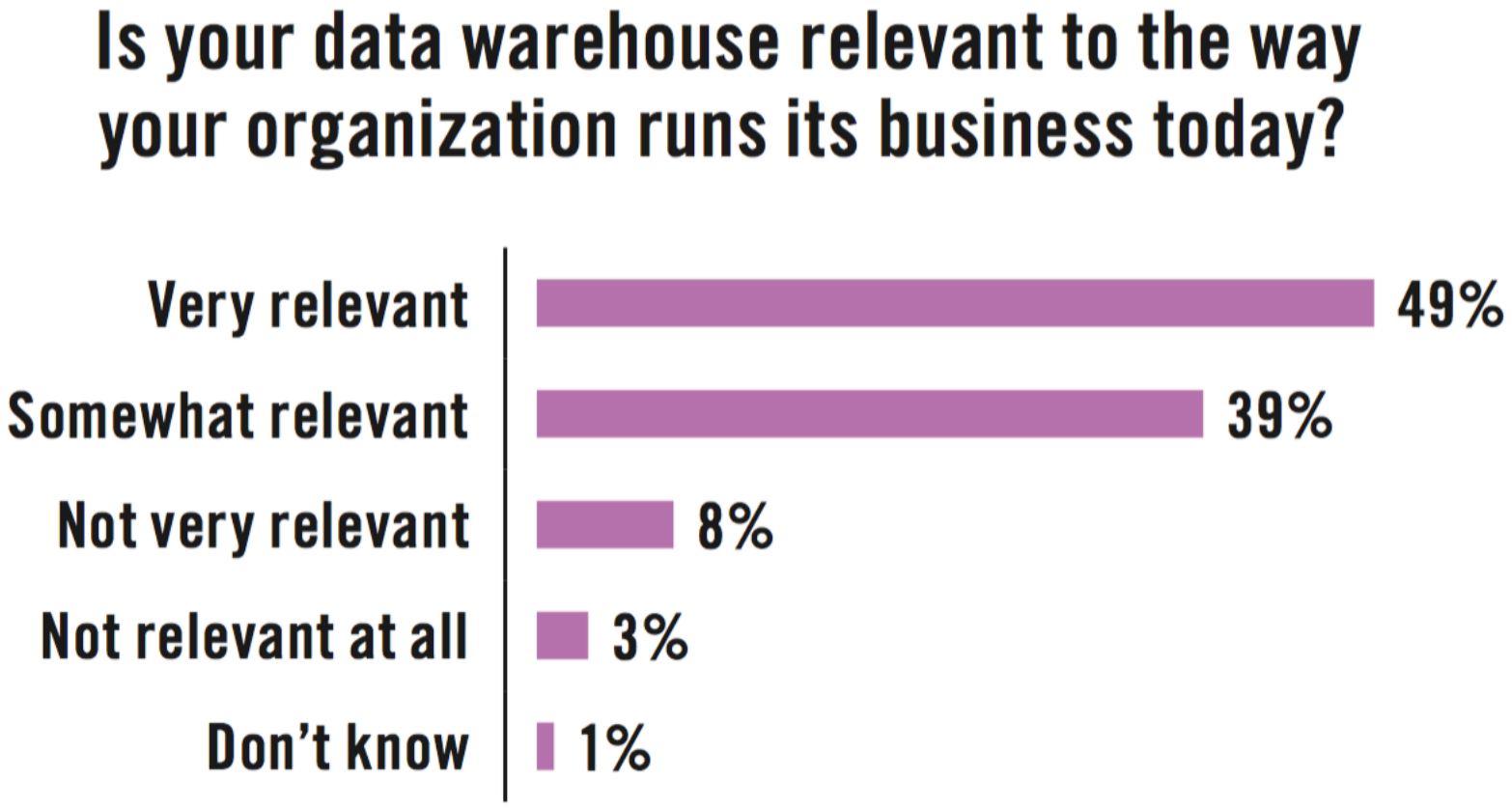
Separate Hype from Reality

Be Cautious

- Separate buzz-word-hype from reality
- Make sure to educated and well informed decisions

The Future of Data Warehousing

Data Warehouse As-Is

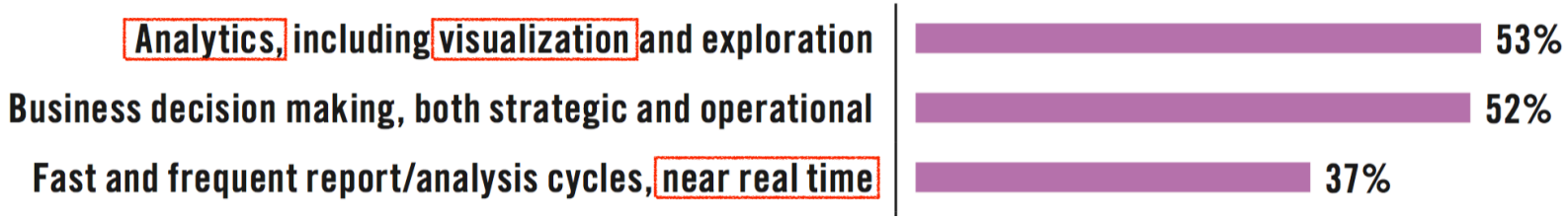


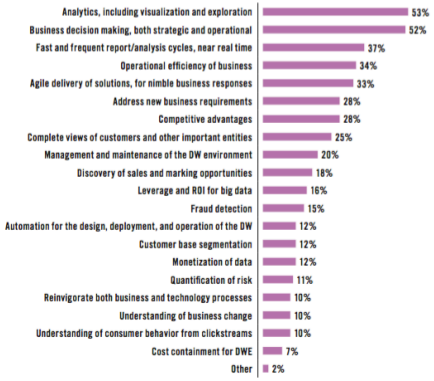
Source: Data Warehouse Modernization in the Age of Big Data

The Future of Data Warehousing

Data Warehouse To-Be

What are the top business and technology tasks that would benefit if your organization implemented the forms of data warehouse modernization you are contemplating? Select one to five answers.

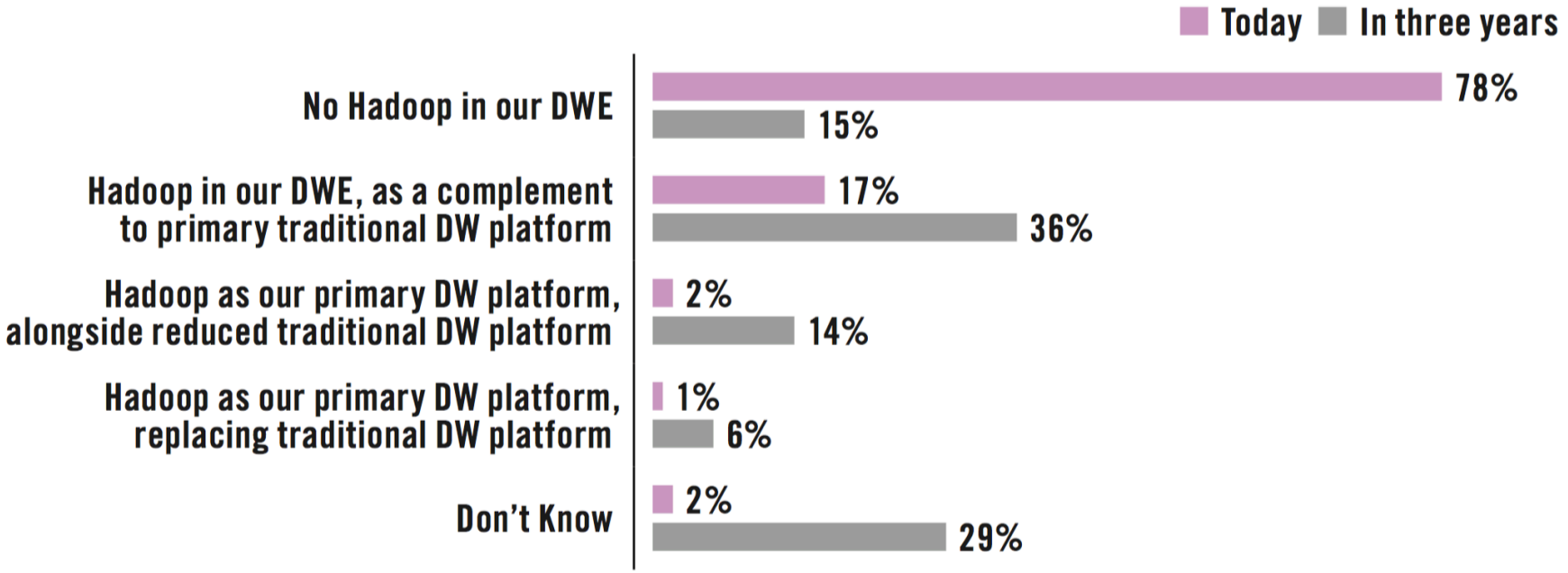




The Future of Data Warehousing

Data Warehouse with Hadoop

What is the role that Hadoop plays in your extended data warehouse environment (DWE) today? What about in three years?



The Future of Data Warehousing

DW Modernization/Optimization

- More practical approach for DW/BI
- Focus on off-loading pain points to new technologies while keeping the traditional DWE (Data Warehouse Environment) as the backbone of the enterprise reporting and analytics system
- DW/BI is mission-critical and so can't afford failure
  - Start with simple cloud based file storage

The Future of Data Warehousing (cont'd)

DW Modernization/Optimization

- Expect to see lots of permutations of Hadoop and its ecosystem components with data warehouse, business intelligence, predictive analytics and data visualization technologies.
- Be prepared to see these specialty areas more unified, rationalized and seamlessly combined.

The Future of Data Warehousing (cont'd)

How to Prepare for Big Data Era?

- Data Warehouse and Business Intelligence
  - Dimensional Modeling
- Statistics/Analytics tools
  - R, Matlab, SPSS, SAS, etc.
- Visualization tools
  - Tableau, BIME, QlickView, etc.
- Hadoop Ecosystem
  - Hive, Pig, Spark (higher level language abstraction for analysts)
- Cloud Solutions
  - Amazon Web Services (EC2, S3, Aurora, Redshift, Machine Learning, Hadoop, Kinesis, etc.)