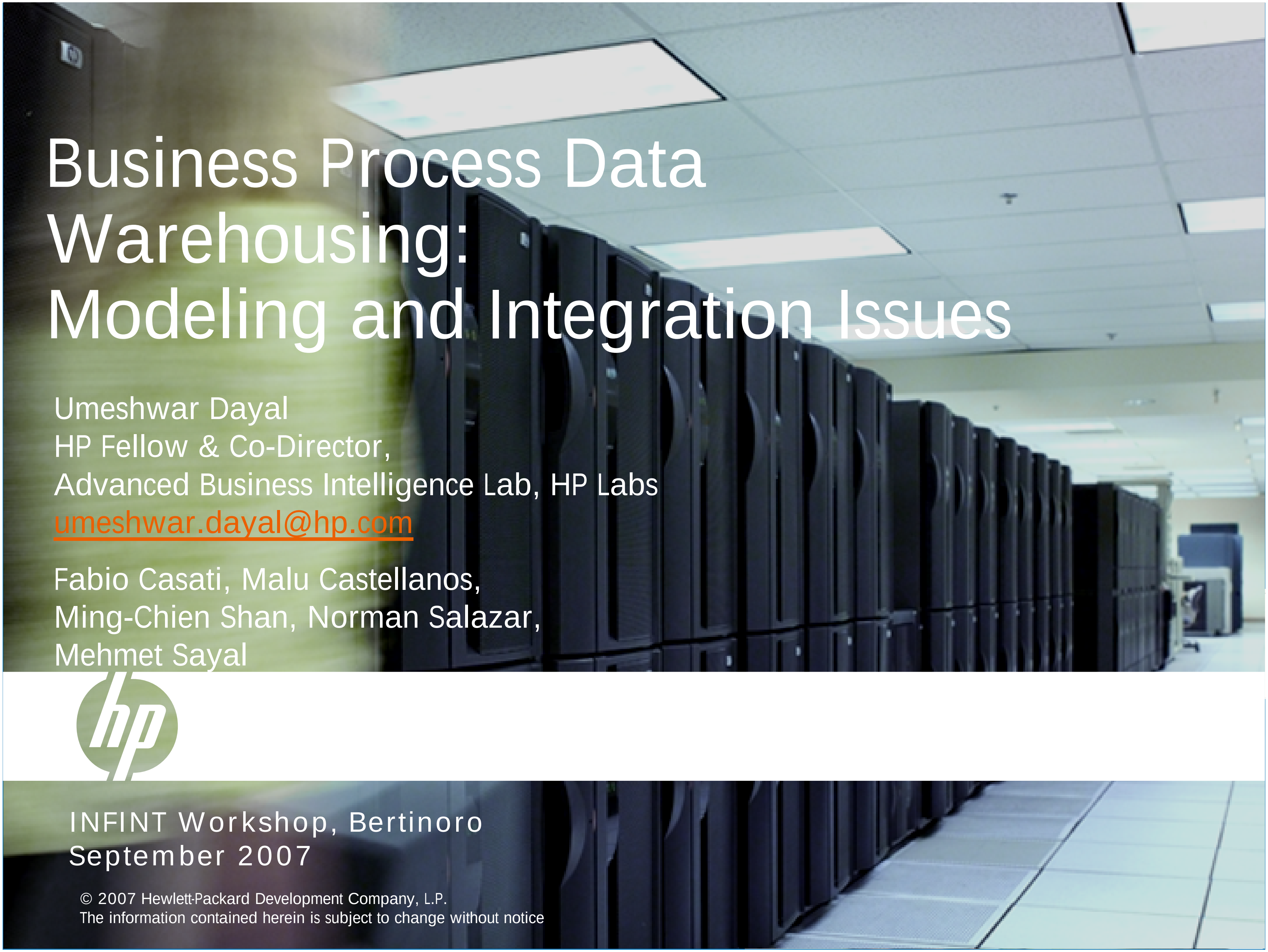


Business Process Data Warehousing: Modeling and Integration Issues

Umeshwar Dayal
HP Fellow & Co-Director,
Advanced Business Intelligence Lab, HP Labs
umeshwar.dayal@hp.com

Fabio Casati, Malu Castellanos,
Ming-Chien Shan, Norman Salazar,
Mehmet Sayal

INFINT Workshop, Bertinoro
September 2007

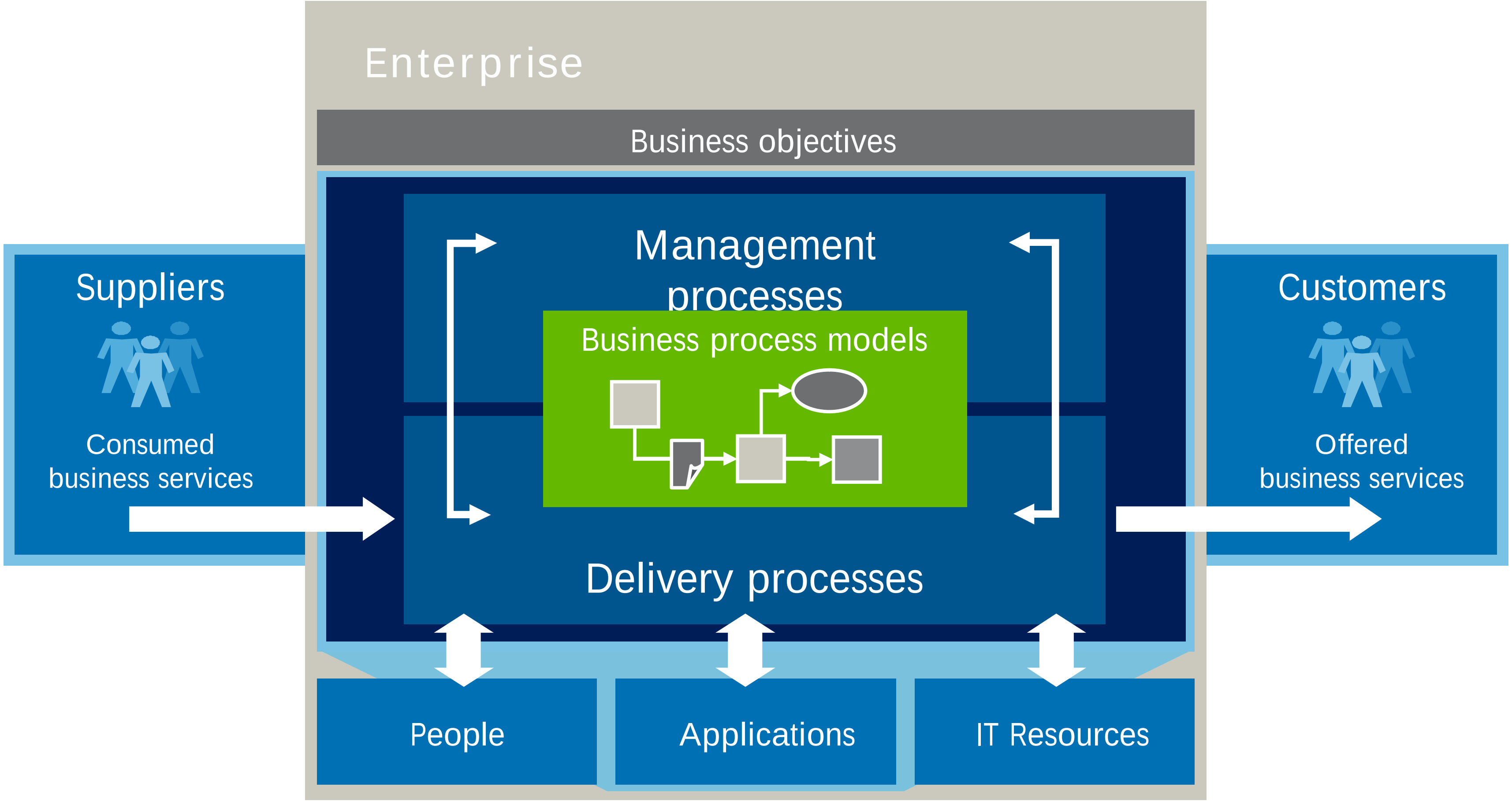
© 2007 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

Outline

- Context
  - Business Process Management
  - Business Process Intelligence
  - Relevance of Information Integration
- Process Modeling Issues
  - Process Views
  - Metrics Model
- Information Integration Issues
  - Generic Data Warehouse Schema
  - Abstraction Mechanisms
  - Generic ETL
  - Information Extraction from Semi-Structured Data
- Summary

© 2007 Hewlett-Packard Company

Business Processes Drive the Enterprise

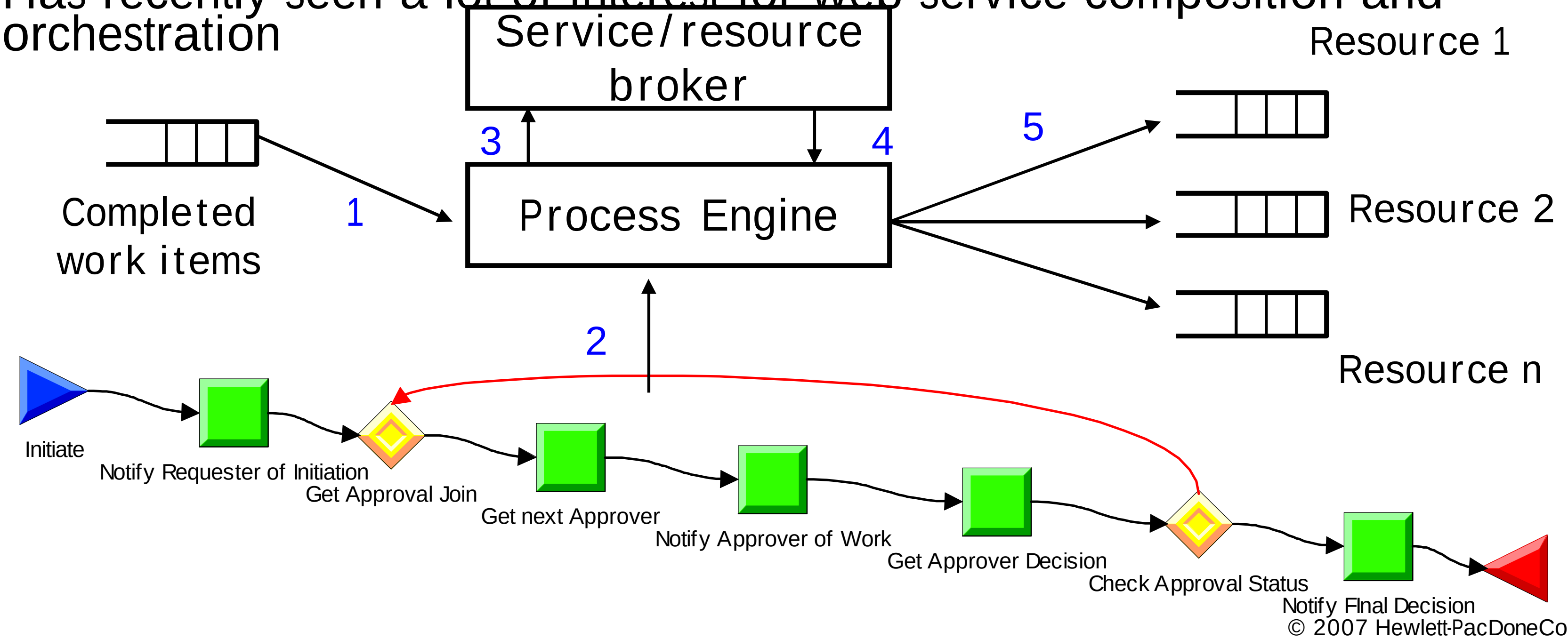


Supply chain processes: Procurement, Inventory, Logistics
Manufacturing processes: ERP, Product Lifecycle
Administrative processes: HR, Finance, Legal
Customer facing processes: support, CRM

© 2007 Hewlett-Packard Company

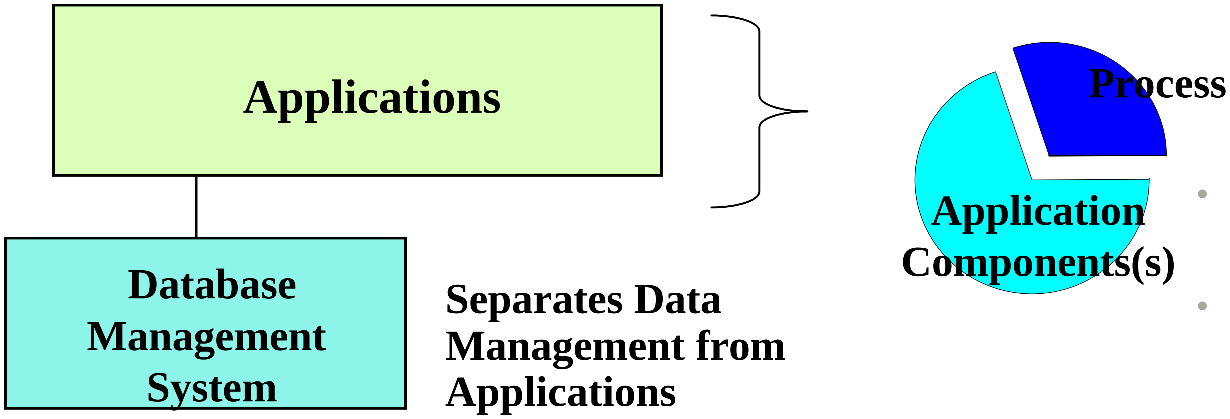
Business Process Management

- Support the definition, execution, and monitoring of operational business processes.
- Define and control the sequence, timing, and other dependencies among tasks, as well as enforce business rules and policies.
- Assign resources (humans, data, applications, services) to individual tasks.
- Monitor and track all aspects of the process execution for auditing, business reporting, and process optimization.
- Has recently seen a lot of interest for web service composition and orchestration



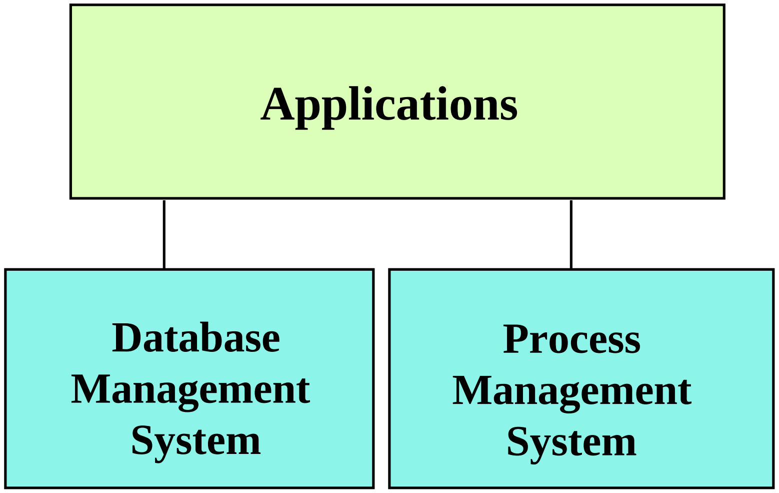
© 2007 Hewlett-PacDoneCompany,

Business Process Management Rationale



Separates Data Management from Applications

- Process logic is “hard wired”
- Need programmer to change process

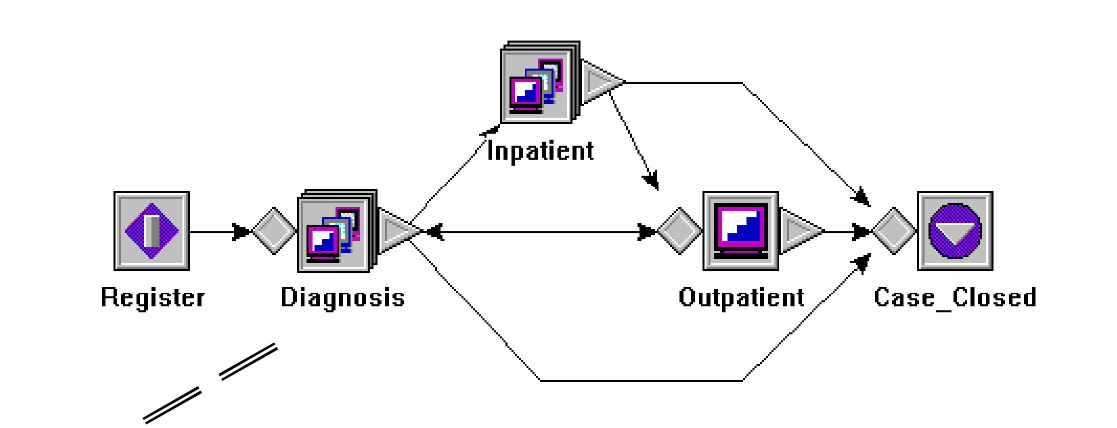


- Extract Rules for Process Flow and Resource Allocation from Applications and Data
- Flexible and modular development to minimize impact of changes in one area on other areas:
- Business Processes and Process Management become a corporate asset
  - a foundation for corporate business process reengineering and optimization

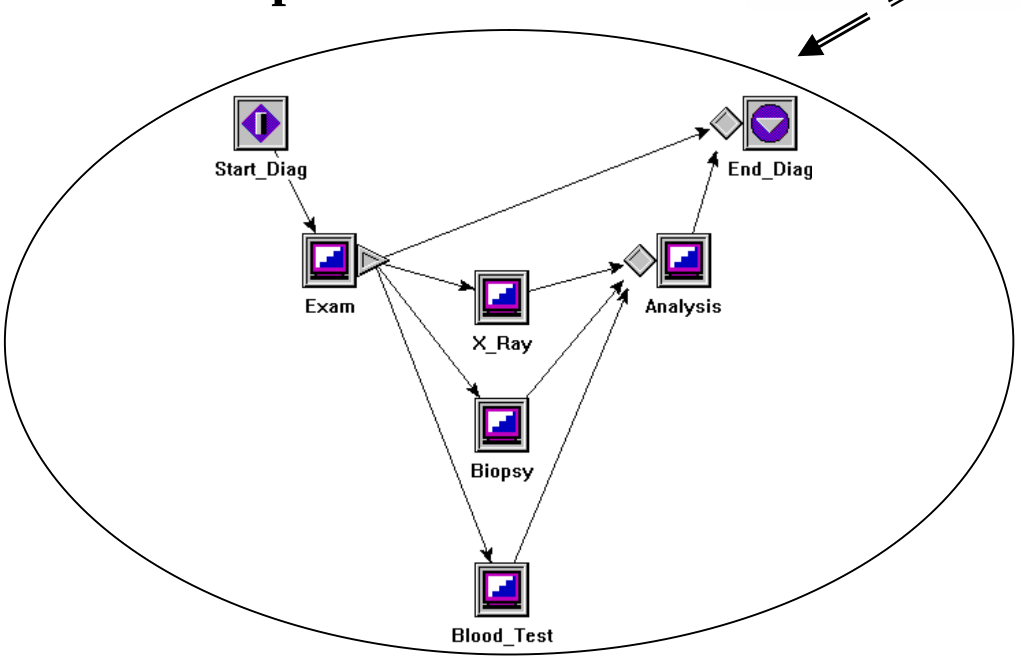
© 2007 Hewlett-Packard Company

Example of Business Process Definition

Main Process



Nested Sub process



Process definition captures:

•Tasks (activities): manual, automated
•Control Flow
•Data Flow
•Exception handling
•Escalation
•Event triggers
•Resource resolution
•Application integration

© 2007 Hewlett-Packard Company

Resurgence of Interest in BPM

- First wave (1980's - 1990's) was focused on Business Process Automation/ Execution using Workflow Management technology: this had little success
  - Heterogeneous components
  - Cost
  - Complexity of building the WF system
  - Lack of support for application development lifecycle
  - Lack of standardization

- Second wave (2000's) is focused on Business Process Improvement: now there are the conditions for BPM to succeed
  - Web services and Service-Oriented Architectures
  - Maturity in the basic middleware, Open source BPMs
  - Standardization (BPEL, RosettaNet)
  - Trend towards increased Outsourcing
    - Understand and model the customer's and provider's business processes
    - Define and formalize Service Level Agreements (SLAs)
    - Alert/Predict SLA violations
    - Audit: you are liable, need to track
    - Analysis/optimization: much larger emphasis on operational efficiency
  - Regulatory compliance

© 2007 Hewlett-Packard Company

Business Process Intelligence

- Goal: improve the quality and performance of intra- and inter-enterprise business processes
  - Internal quality, as perceived by the service provider (e.g., reduced operating costs, fewer exceptions)
  - External quality, as perceived by the service consumer (e.g., better quality of service, reduced cost of service)
- Approach: Apply business intelligence techniques (data warehousing, data mining, simulation and optimization) to data relevant to business process execution
  - Integrate data from many sources:
    - Process management system (workflow engine) logs
    - web service execution logs
    - application logs
    - audit logs
    - event logs
    - systems management data, resource utilization data,
    - financial data, other business and operational data, …
  - Use the data to analyze, understand, and optimize processes
    - Resource assignments
    - Reporting on performance and quality of resources, service providers
    - Load prediction and optimization
    - Exception understanding and prevention
    - Paths followed in the process graph

© 2007 Hewlett-Packard Company

Users want a crystal ball….

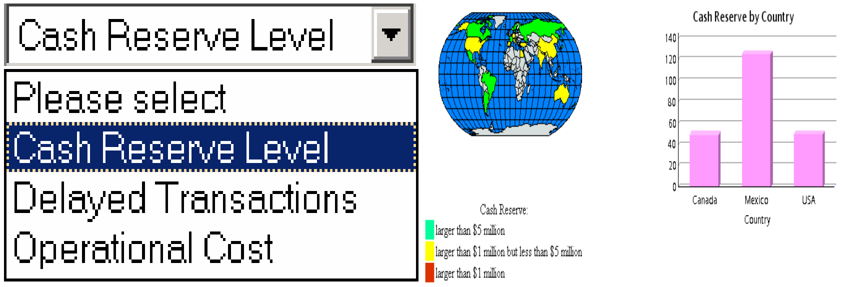
- How can I identify bottlenecks?
- What are the causes of missed SLAs (or other business performance metrics)?
- Can I predict my risk of missing SLAs (e.g., late payments)
- How much money do I save on electronic invoices versus paper ones? And how much time?
- How do the different payment methods compare in terms of cost and time?
- Can I predict my workload?
- What's my optimal resource plan? How many resources do I need to meet my SLAs (e.g., payment schedules)
- What's the disruption caused by unavailable resources?
- How do I “improve” my business process?
- What is the business impact of changing my IT infrastructure? My business process?

- Today this is difficult to do, requiring lots of custom design, system integration, and implementation effort.

© 2007 Hewlett-Packard Company

Business Process Intelligence
Embedding BI into Business Operations

Financial Controller



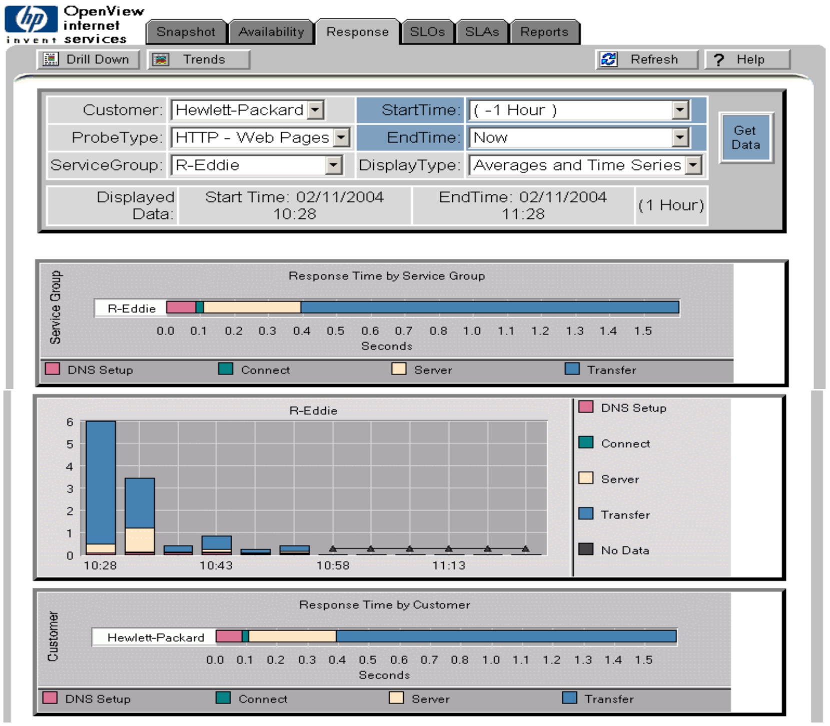
Cash Reserve Level

Please select
Cash Reserve Level
Delayed Transactions
Operational Cost

Today, it is extremely difficult to automatically discover these inter-relationships and their implications !

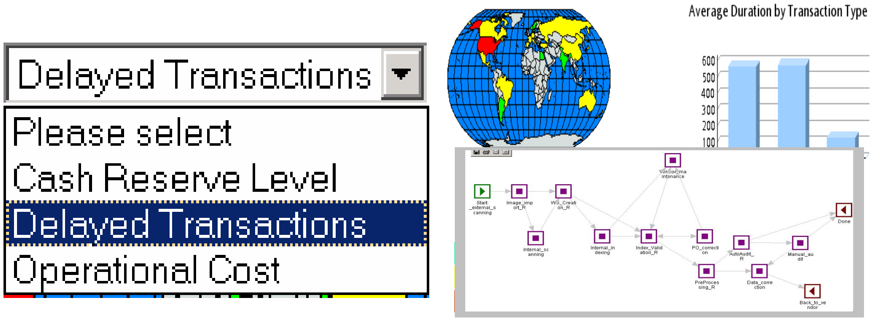
IT System Administrator

Name	**R-Eddie**
Type	IT resource
Status	Critical
Change time	2004-01-26 10:57:02.687
Duration	29 minutes 53 seconds



Customer: Hewlett-Packard
StartTime: (-1 Hour)
ProbeType: HTTP - Web Pages
EndTime: Now
ServiceGroup: R-Eddie
DisplayType: Averages and Time Series
Displayed Data: Start Time: 02/11/2004 10:28 EndTime: 02/11/2004 11:28 (1 Hour)

Operations Manager



Delayed Transactions

Please select
Cash Reserve Level
Delayed Transactions
Operational Cost

© 2007 Hewlett-Packard Company

Suddenly, a very crowded space

Buzzwords popping up all over the place

- Business Service Management
- Business Event Management
- Business Activity Monitoring
- Business Operations Management
- Business Performance Management
- Business Process Intelligence
- Operational Business Intelligence
- Real-time enterprise, Zero latency enterprise
- Executive dashboards

© 2007 Hewlett-Packard Company

Information Integration Approaches

- Federated Databases (virtualization, stateless)
- Data Warehouses (materialized, stateful)
- Operational Data Stores
- Master Data Management
- Active Data Warehouses (hybrid of DW and ODS)

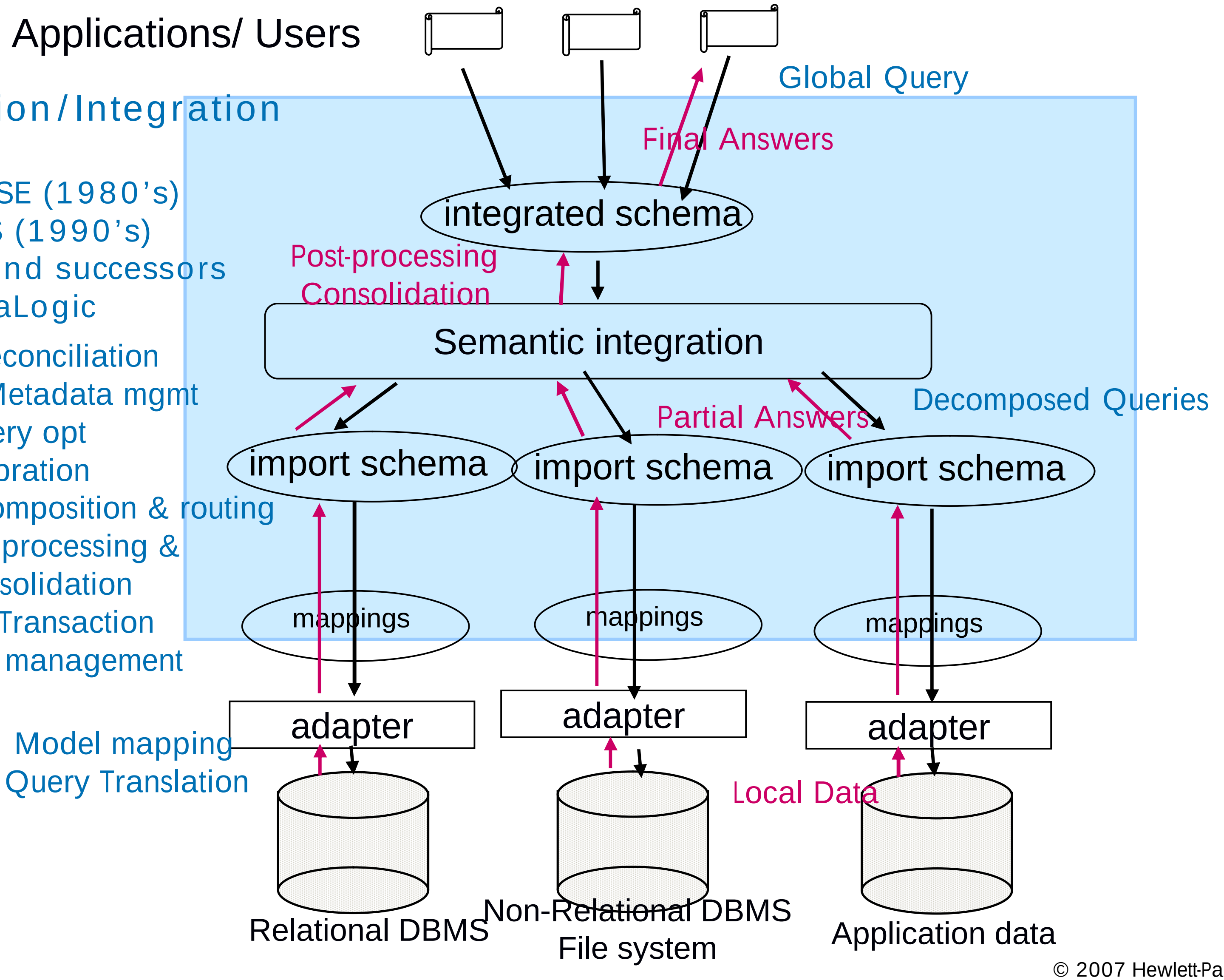
- For Business Process Intelligence, building a data warehouse (actually an active DW) is more appropriate
  - Need data from many sources (many of which are not databases)
  - Need historical data in addition to data about current process instances
  - Need complex transformations (e.g., map system events into abstract process progression)
  - Many reporting and analytic tools already work with data warehouses
- We built a research prototype, Business Cockpit, and tested it in several internal pilot solutions
- Several research challenges

© 2007 Hewlett-Packard Company

Federated databases: Data Virtualization

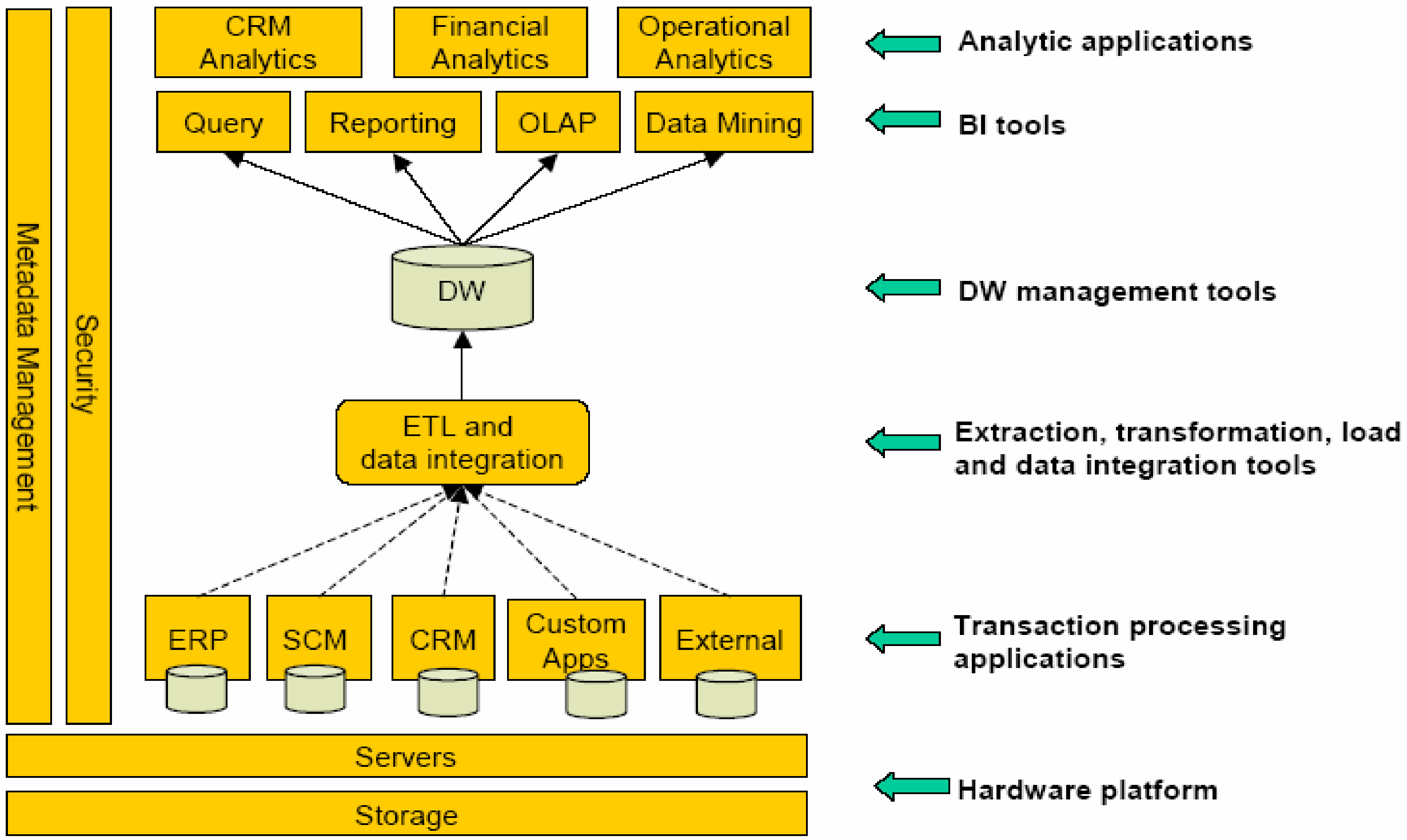
Federation/Integration Server
MULTIBASE (1980's)
PEGASUS (1990's)
GARLIC and successors
BEA AquaLogic

Semantic reconciliation
Catalog/ Metadata mgmt
Global Query opt
& Cost calibration
Query decomposition & routing
Query post-processing &
answer consolidation
Distributed Transaction
(Workflow) management



© 2007 Hewlett-Packard Company

Data Warehouse: Data Materialization

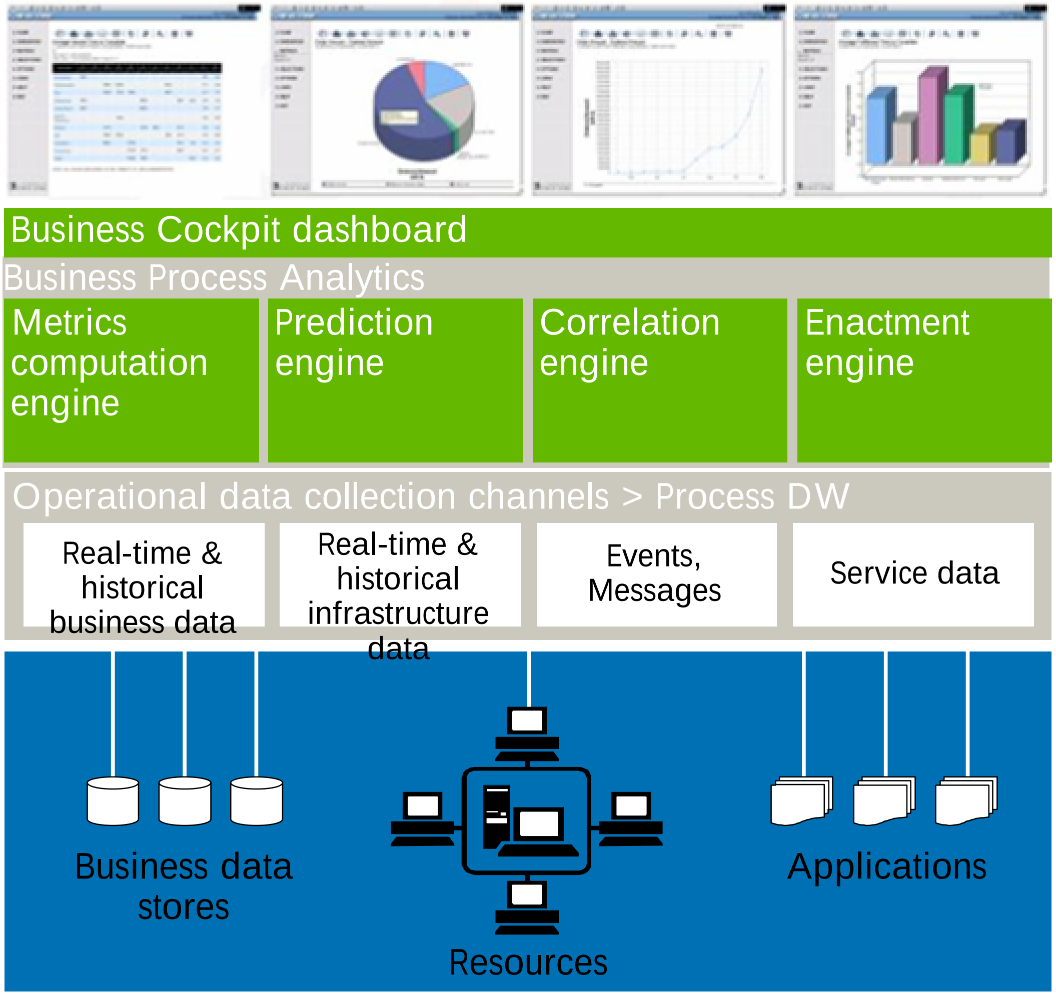


Source: IDC, 2005

© 2007 Hewlett-Packard Company

Business Cockpit: A Prototype Business Process Intelligence Solution

- Automatically correlate real-time and historical business process and IT systems data
- Customize / monitor / analyze user-defined metrics
- Identify critical factors influencing business metrics (e.g., SLA compliance)
- Assess and predict impact of events on business KPIs
- Allocate IT capacity to optimize business operations



Business Cockpit dashboard

Business Process Analytics

Metrics computation engine

Prediction engine

Correlation engine

Enactment engine

Operational data collection channels > Process DW

Real-time & historical business data

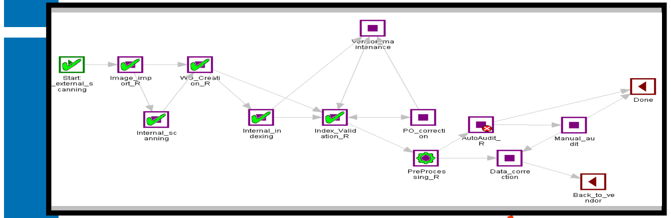
Real-time & historical infrastructure data

Events, Messages

Service data

© 2007 Hewlett-Packard Company

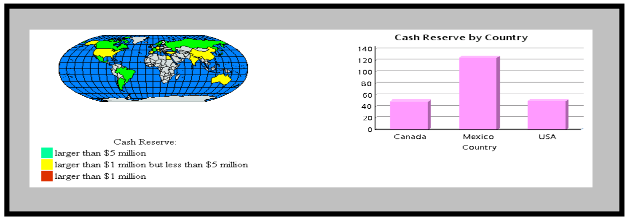
Business Process Intelligence Research Areas



Process modeling

Different views of the same process

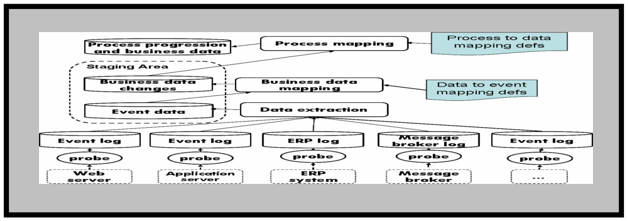
Mappings between process views



Metrics

Functions/Queries over process and related data

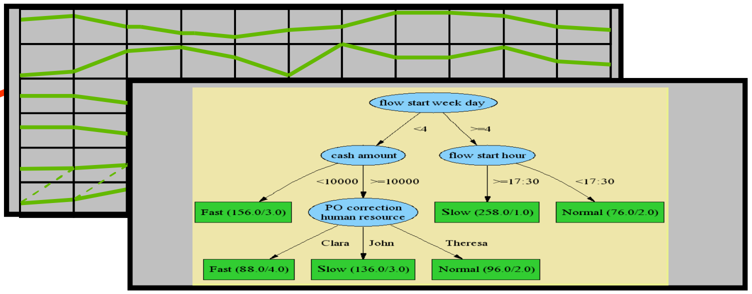
Mappings between views



Process Warehouse Design

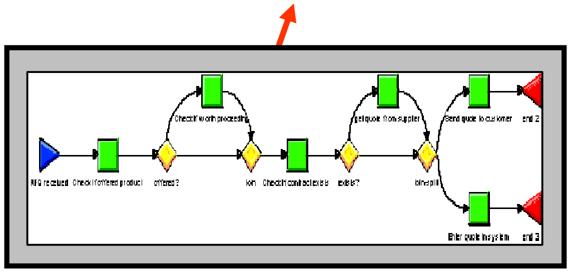
Generic and customizable schema and ETL suitable for any processes and data sources

Business Cockpit



Correlation & Explanation

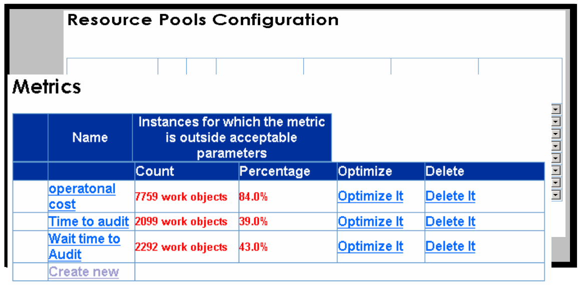
Data mining to learn relationships between metrics, and between metrics and events



Process Discovery

Mine event logs to learn "true" implemented process

Automatically suggest process restructuring


Resource Pools Configuration

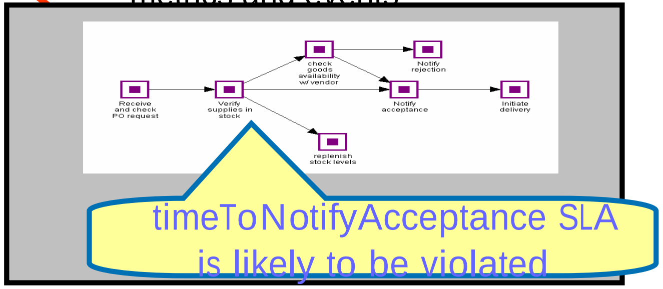
Metrics

	Name	Instances for which the metric is outside acceptable parameters			
		Count	Percentage	Optimize	Delete
	operatonal cost	7759 work objects	84.0%	Optimize It	Delete It
	Time to audit	2099 work objects	39.0%	Optimize It	Delete It
	Wait time to Audit	2292 work objects	43.0%	Optimize It	Delete It
	Create new				

Simulation and Optimization

Capacity management

Service/Resource selection



timeToNotifyAcceptance SLA is likely to be violated

Prediction

Time-series prediction

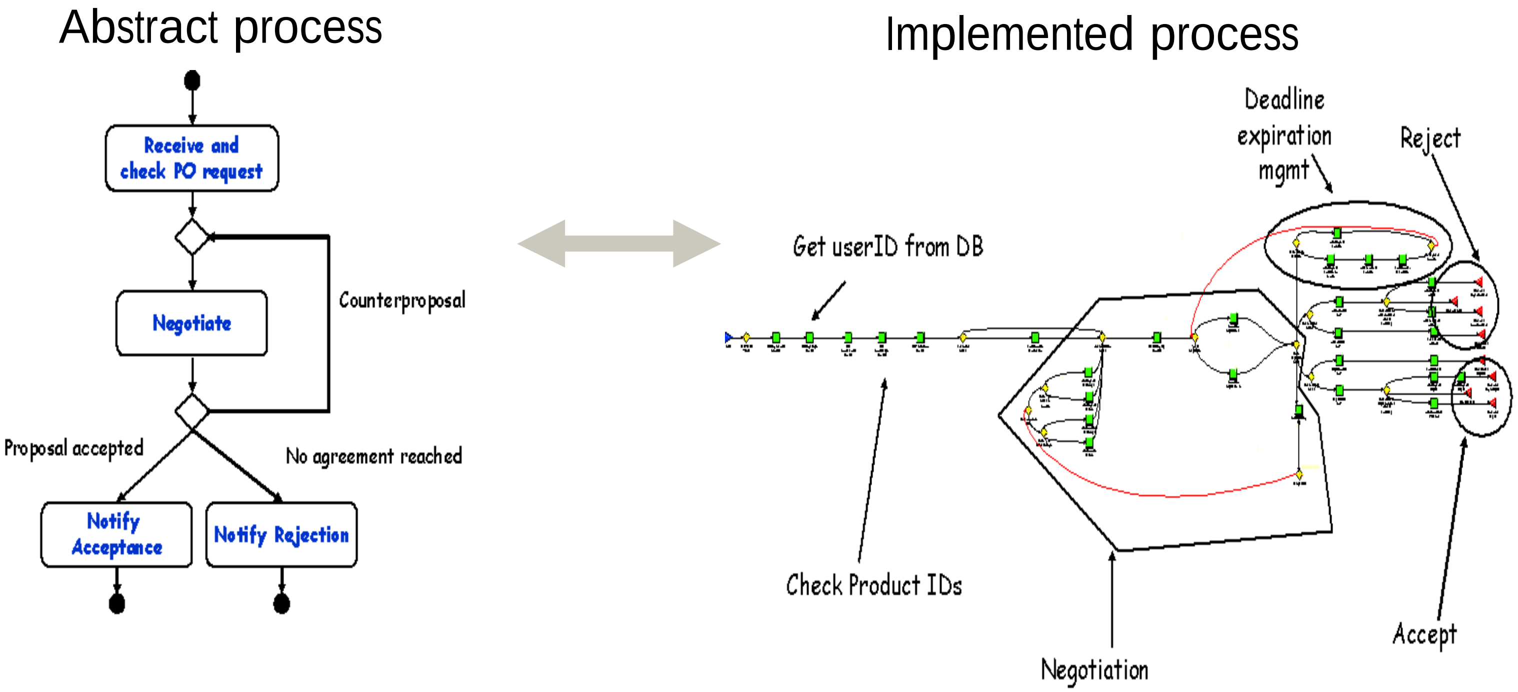
© 2007 Hewlett-Packard Company

Outline

- Context
  - Business Process Management
  - Business Process Intelligence
  - Relevance of Information Integration
- Process Modeling Issues
  - Process Views
  - Metrics Model
- Information Integration Issues
  - Generic Data Warehouse Schema
  - Abstraction Mechanisms
  - Generic ETL
  - Information Extraction from Semi-Structured Data
- Summary

© 2007 Hewlett-Packard Company

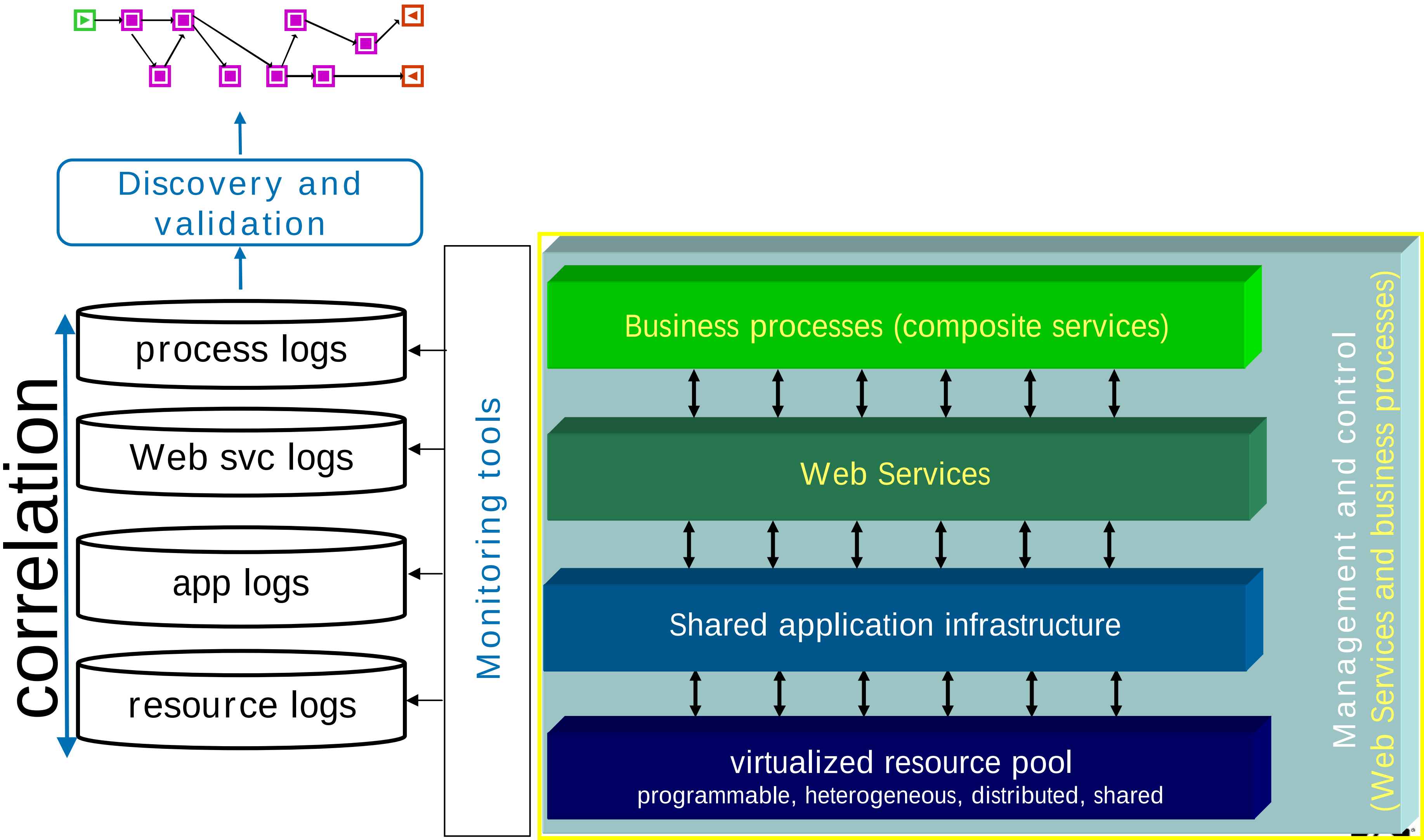
Process Modeling: Abstract versus Implemented Process Views



- How to map between these views? Which is the base model and which is the view?
- Abstract process models are usually constructed manually
- No good tools for model refinement and implementation; the implemented process is often not explicitly modeled
- Process discovery: data mining techniques to learn and validate the implemented process
- Construct mappings between abstract and implemented process views (schema level, instance level, data level)
- Process integration and views over multiple processes are wide open problems

© 2007 Hewlett-Packard Company

Process Discovery and Validation



Discovery and validation

process logs

Web svc logs

app logs

resource logs

correlation

Monitoring tools

Business processes (composite services)

Web Services

Shared application infrastructure

virtualized resource pool

programmable, heterogeneous, distributed, shared

Management and control

(Web Services and business processes)

© 2007 Hewlett-Packard Company

Log Abstraction

Trace: a set of time ordered log entries that refer to the same process execution

AsBsAeCsBeAsAeCe …

BsBeAsDsAeDs …

CsCeAsAeEsFsFeEe …

- Log: a set of traces
  - Log supported by transaction monitoring systems
- Problems: missing data, wrong timestamps, incomplete traces, correlation among log entries

© 2007 Hewlett-Packard Company

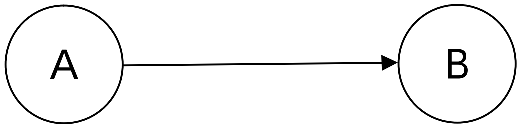
Discovery of Ordering Constraints

- Possible process flow structures
  - Sequences
  - Splits
    - XOR
    - AND
    - OR
  - Joins
    - XOR
    - AND
    - OR
  - Loops

© 2007 Hewlett-Packard Company

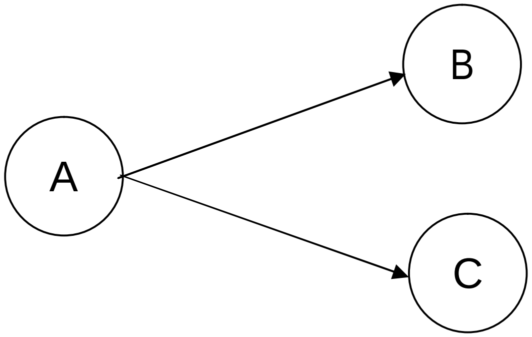
Mining frequent sequences (episodes) from logs

Sequence



AsAeBsBe

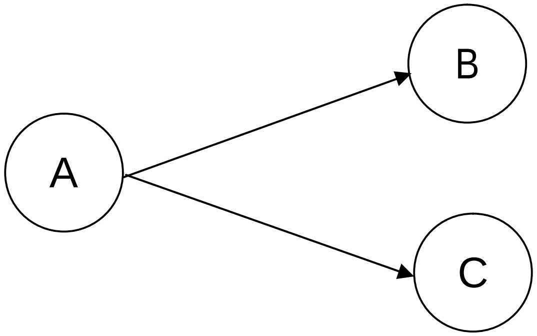
XOR-Split



AsAeBsBe

AsAeCsCe

AND-Split



AsAeBsCsBeCe

AsAeBsCsCeBe

AsAeCsBsBeCe

AsAeCsBsCeBe

© 2007 Hewlett-Packard Company

Model Discovery

- Progressive elimination
  - Probabilistic, iterative approach
  - Look at variance, average interval among activities
  - Look across traces
- It’s not just the flow
  - lots of thresholds, tuned through classification
- Tuning thresholds is difficult

© 2007 Hewlett-Packard Company

Monitoring Processes: Types of Metrics

- Process metrics
  - execution times, durations, volumes, paths taken, outcomes
  - correlation with “previous” step
- Resource metrics
  - Performance of human and system resources in executing steps.
  - Correlation between resource and process metrics: which resources statistically lead to successful or unsuccessful executions, or which resources have led to certain paths being taken, e,g., escalations or error handling
- Business metrics
  - Domain-specific metrics, e.g., order-to-cash, turn around time, cash reserve levels
  - Correlation of business data with process data, e.g., efficiency and quality of execution based on invoice type.
  - Correlation between business data and resources, e.g., number of invoices from a given center processed by a given employee

© 2007 Hewlett-Packard Company

Conventional approach to defining and computing metrics

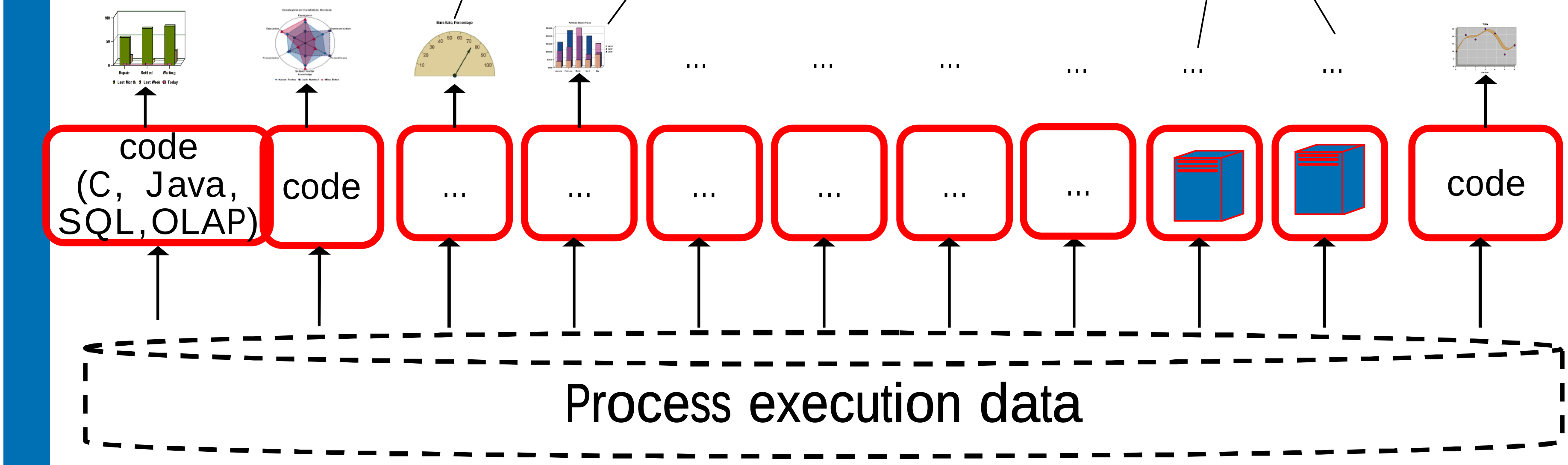
average **execution quality** by process

total **'time to acceptance' SLA violations** by customer

predictions and explanations on **SLA violations**

total **execution quality** by day

details on total **'time to acceptance' SLA violations**



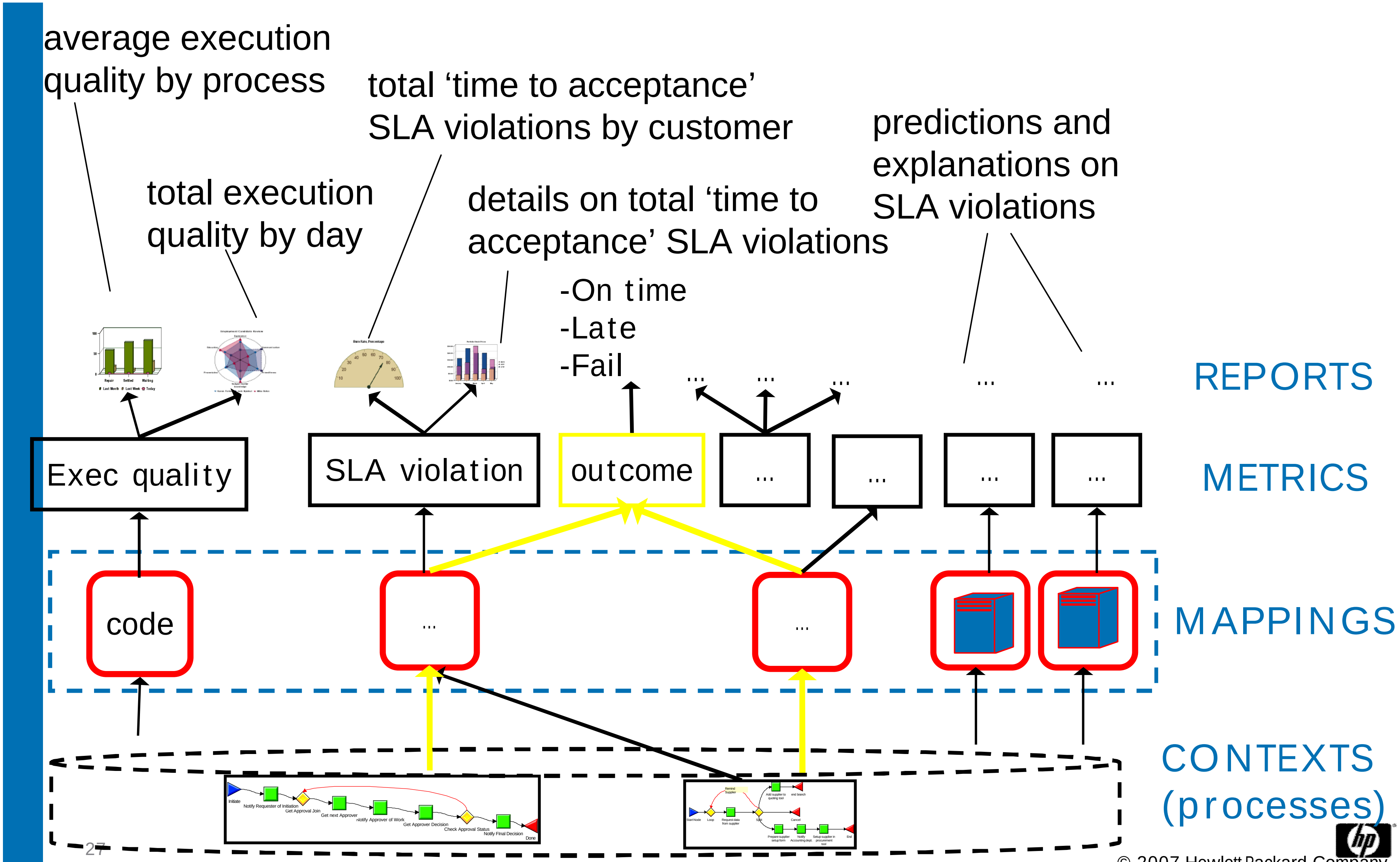
© 2007 Hewlett-Packard Company

Problems

- Long development time
  - Want to define 100 reports? Write 100 queries, test, deploy
- Poor performance
  - Want to view the reports? Run 100 queries
  - Real time? Concurrent users?
- Poor functionality
  - Limited support for drill-downs
  - No support for polymorphism
- Not robust to changes
  - Want a little different perspective? Write code, test, deploy
- No support for analytic features
  - Root-cause analysis, predictions, impact,…

© 2007 Hewlett-Packard Company

Our Approach: Provide Modeling Abstractions



average execution quality by process

total execution quality by day

total 'time to acceptance' SLA violations by customer

details on total 'time to acceptance' SLA violations

predictions and explanations on SLA violations

-On time
-Late
-Fail

... ...

REPORTS

Exec quality

SLA violation

outcome

... ...

METRICS

code

... ...

MAPPINGS

CONTEXTS (processes)

© 2007 Hewlett-Packard Company

Modeling Abstractions

- Domain Model for Processes
  - Entities (e.g., “process”, “event”, “action”, “service”)
  - Attributes, relationships, aggregations,..
  - Benefits:
    - Allows definition, computation, analysis, monitoring of “things”
    - Enable easy and quick “verticalization”
- Metric Model
  - Mappings: Functions that compute values from raw data
  - Metrics: Measurable properties of an entity (e.g., transaction value by type) defined by mappings
  - Benefits:
    - Polymorphism (different definitions for different contexts)
    - Minimize number of functions needed
    - Reuse: share definitions across metrics, contexts, data models
- Reporting / Analytic Model
  - Once a metric has been defined, lots of report types are immediately available without requiring coding
    - Domain-specific aggregations
    - Temporal aggregations
    - Analytics on metrics: correlations, predictions, explanations, root cause analyses, etc.
  - Benefits:
    - Rich reporting of generic measures
    - Flexibility: Enabling aggregations of “something” by “something else

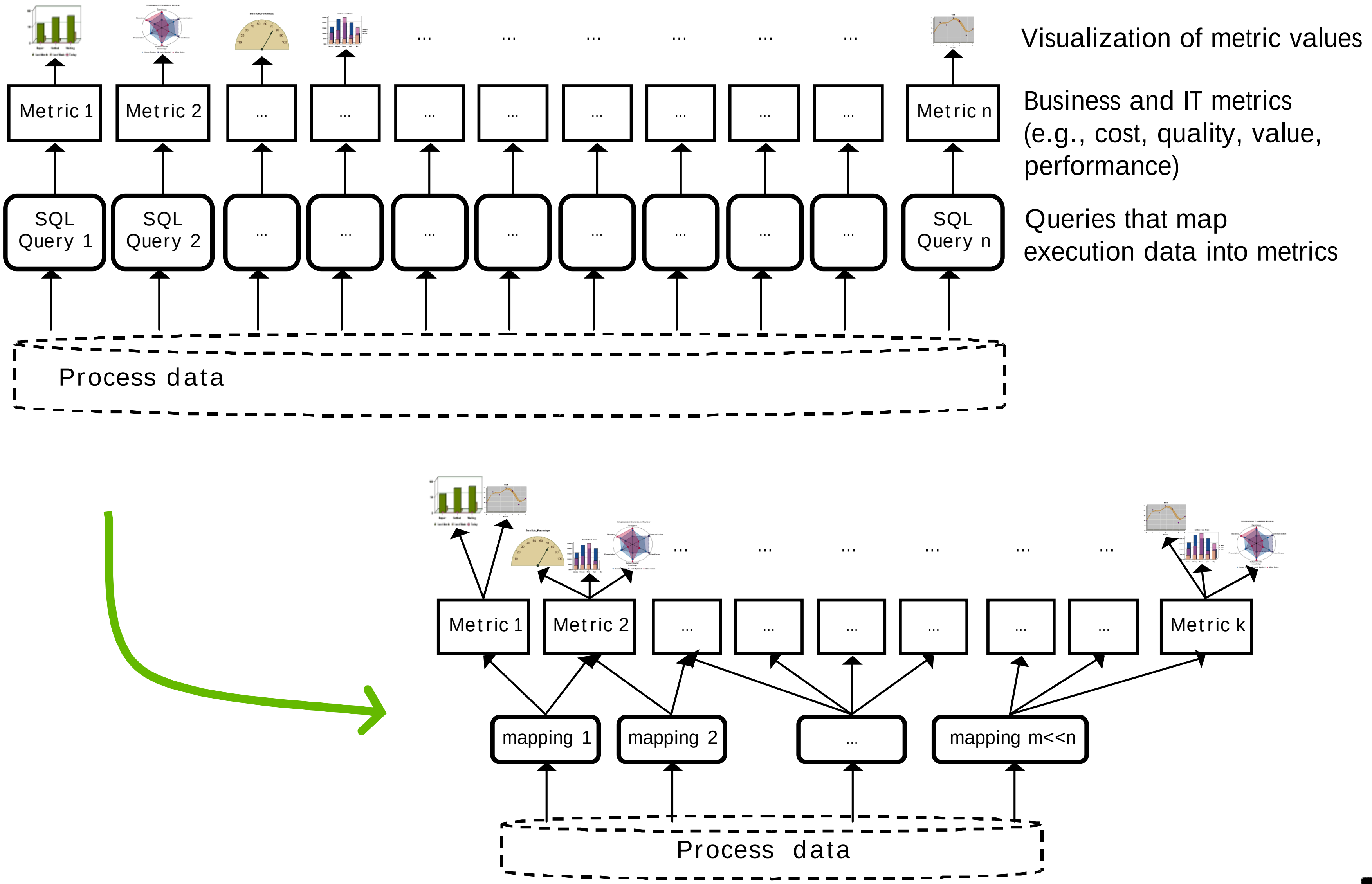
© 2007 Hewlett-Packard Company

Adding Semantics:
Behaviors, Contexts, Taxonomies

- Associate semantics with process instances
- Map execution data into categorical metrics
- Examples:
  - Instances lasting more than avg+stdev are slow
  - Orders >$1000 are large orders
- Many behavior templates are predefined
  - T1: “instances in which node N was executed”
- Users define behaviors by instantiating a template (filling forms)
  - T1 (N=“notify acceptance”) maps to accepted
- New templates can be defined (e.g., in SQL)
- Contexts define the instances to which the template should be applied
  - Processes of group supply chain where suppliers are in group “IT suppliers”
- Define Taxonomies based on behaviors
  - Duration taxonomy: acceptable, fast, slow, unacceptable.

© 2007 Hewlett-Packard Company

Sharing mapping functions



Visualization of metric values

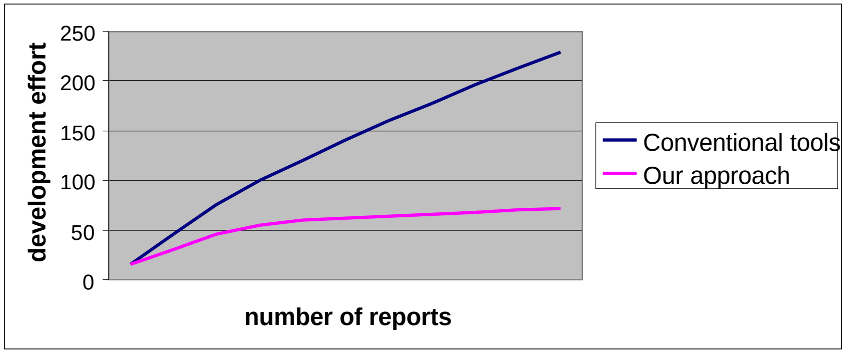
Business and IT metrics (e.g., cost, quality, value, performance)

Queries that map execution data into metrics

© 2007 Hewlett-Packard Company

Development Time: Automatic Query Generation

- Benefits
  - Make things simple, minimize user errors
  - Automatic query generation



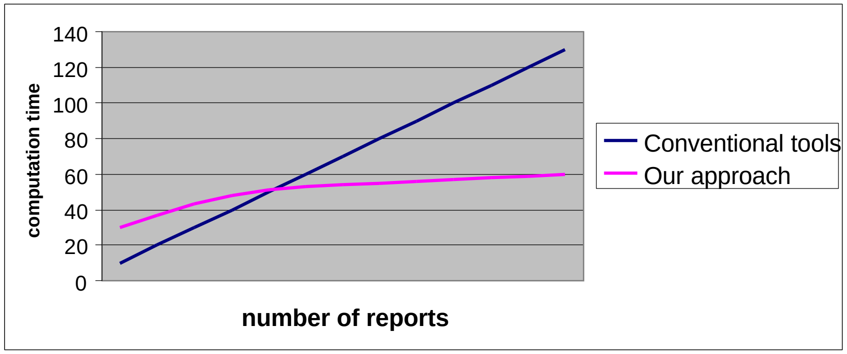
```
SELECT EI.EVENT_INST_UUID
FROM EVENT_INSTANCE EI
WHERE (ei.time_completed-ei.time_triggered)< $THRESHOLD$
```

```
SELECT MTE.METRIC_ID, mte.id, EI.EVENT_INST_UUID,
mte.metric_class_id, NULL,NULL, NULL, SYSDATE
FROM active_METERS MTE, CONTEXTS CTX, EVENT_INSTANCE EI,
ENTITIES E, EVENT EV,
EXTENDED_METER_PARS emp
WHERE MTE.MAPPING_ID=5
AND E.EXTENDED_NAME='Event Data'
AND EI.EVENT_ID=EV.ID AND CTX.METER_ID=MTE.ID AND
emp.meter_id=mte.id
AND (ei.time_completed-ei.time_triggered)< emp.PAR1
AND  ((EV.NAME=CTX.EVENT_NAME OR CTX.EVENT_NAME IS NULL)  )
```

© 2007 Hewlett-Packard Company

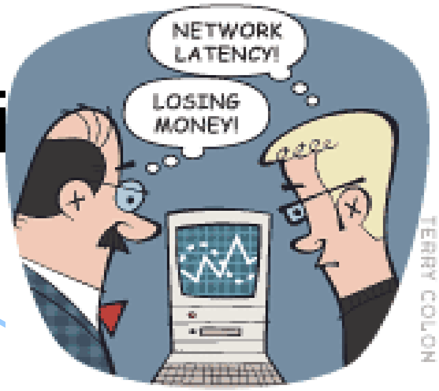
Runtime: Multi-query processing

- Benefits
  - Minimize the number of queries to be executed
    - One mapping can compute several metrics
  - Enable real-time computation and optimizations
    - Smart refresh
    - Take advantage of shared mappings
  - Streaming data, continuous queries



© 2007 Hewlett-Packard Company

Correlations between Metrics: Root cause & business impact analysi



- Learn relationships between IT data/events and business metrics

LOB	Detect no orders placed on any suppliers in the last hour. $$$$$$$$	Why ?	Implies
Business Solutions	PO Collaboration solution is not generating orders.	Why ?	Implies
Applications & Services	I2 business application received no PO requests in the last hour.	Why ?	Implies
Middleware	EAI bus transmitted no requests to i2.	Why ?	Implies
	EAI / SAP adapter is down.	Why ?	Implies
Systems Storage, Devices, Servers	Integration server is down for the last hour.	!!!!	!!!!

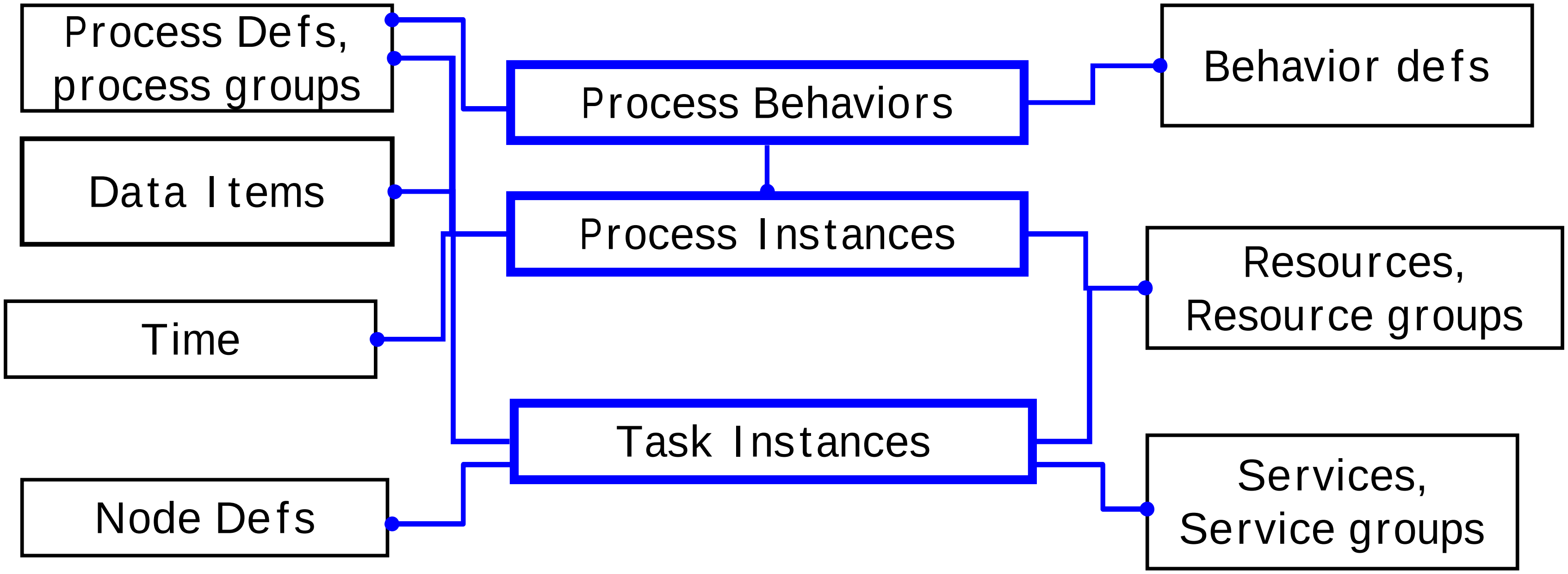
© 2007 Hewlett-Packard Company

Outline

- Context
  - Business Process Management
  - Business Process Intelligence
  - Relevance of Information Integration
- Modeling Issues
  - Process Views
  - Metrics Model
- Information Integration Issues
  - Generic Data Warehouse Schema
  - Abstraction Mechanisms
  - Generic ETL
  - Information Extraction from Semi-Structured Data
- Summary

© 2007 Hewlett-Packard Company

Process Data Warehouse Design: Constellation Schema (almost)



Must be generic for domains, processes, resources, etc., and yet easily customizable and extensible to new process types, data sources, metrics, report types

Several tricky modeling issues

Also, challenges in how to deal with real-time data, event streams, text, etc.

© 2007 Hewlett-Packard Company

Generic Process Warehouse Model

- Challenges for a generic model
  - Multi-level instance data
    - Step level facts, process instance level facts, data-related facts
    - Facts may have to be self-correlated
  - Business data complexities
    - Different from process to process
    - Complex structures
    - Can change at every step during the process
    - à representation hard to generalize
  - Process and step executions go through a lifecycle
    - Step status changes (created, activated, completed, etc --> process events mark progression); number of states can be unlimited (suspend / reactivate cycles)
    - Different systems supporting the execution may have different lifecycle phases

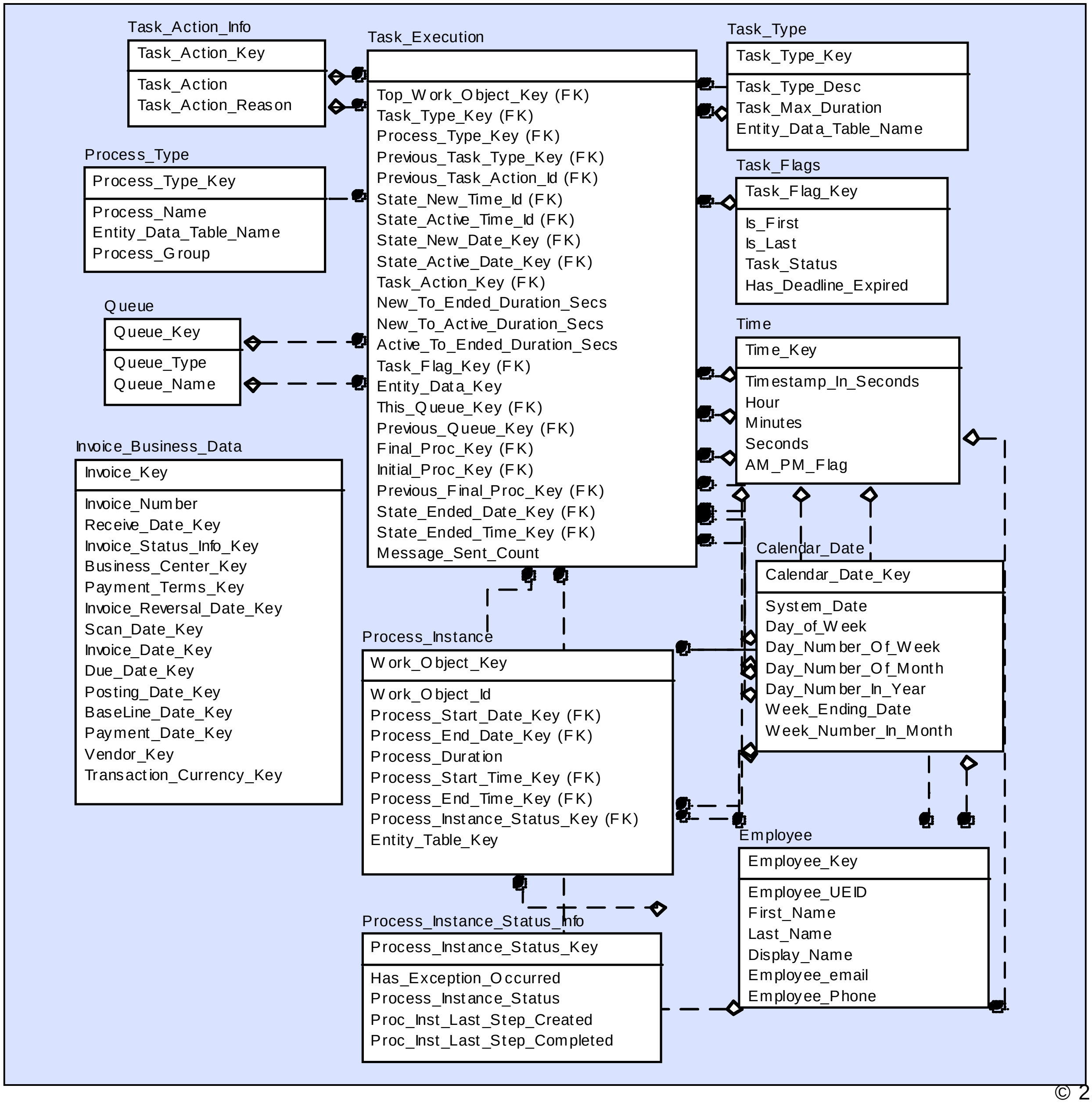
© 2007 Hewlett-Packard Company

Main elements of the generic warehouse model

- Single granularity for steps (rather than at the level of status changes)
- Single fact table for any step of any process
  - Enables analyses across processes
  - Includes aggregation of most common step event measures
- Correlation with previous step data handled via additional columns
- Separate business data tables for each process type
- Blind links to handle step/process correlation with business data

© 2007 Hewlett-Packard Company

Process warehouse model



Task_Action_Info
Task_Action_Key
Task_Action
Task_Action_Reason

Process_Type
Process_Type_Key
Process_Name
Entity_Data_Table_Name
Process_Group

Queue
Queue_Key
Queue_Type
Queue_Name

Invoice_Business_Data
Invoice_Key
Invoice_Number
Receive_Date_Key
Invoice_Status_Info_Key
Business_Center_Key
Payment_Terms_Key
Invoice_Reversal_Date_Key
Scan_Date_Key
Invoice_Date_Key
Due_Date_Key
Posting_Date_Key
BaseLine_Date_Key
Payment_Date_Key
Vendor_Key
Transaction_Currency_Key

Task_Execution
Top_Work_Object_Key (FK)
Task_Type_Key (FK)
Process_Type_Key (FK)
Previous_Task_Type_Key (FK)
Previous_Task_Action_Id (FK)
State_New_Time_Id (FK)
State_Active_Time_Id (FK)
State_New_Date_Key (FK)
State_Active_Date_Key (FK)
Task_Action_Key (FK)
New_To_Ended_Duration_Secs
New_To_Active_Duration_Secs
Active_To_Ended_Duration_Secs
Task_Flag_Key (FK)
Entity_Data_Key
This_Queue_Key (FK)
Previous_Queue_Key (FK)
Final_Proc_Key (FK)
Initial_Proc_Key (FK)
Previous_Final_Proc_Key (FK)
State_Ended_Date_Key (FK)
State_Ended_Time_Key (FK)
Message_Sent_Count

Process_Instance
Work_Object_Key
Work_Object_Id
Process_Start_Date_Key (FK)
Process_End_Date_Key (FK)
Process_Duration
Process_Start_Time_Key (FK)
Process_End_Time_Key (FK)
Process_Instance_Status_Key (FK)
Entity_Table_Key

Process_Instance_Status_Info
Process_Instance_Status_Key
Has_Exception_Occurred
Process_Instance_Status
Proc_Inst_Last_Step_Created
Proc_Inst_Last_Step_Completed

Task_Type
Task_Type_Key
Task_Type_Desc
Task_Max_Duration
Entity_Data_Table_Name

Task_Flags
Task_Flag_Key
Is_First
Is_Last
Task_Status
Has_Deadline_Expired

Time
Time_Key
Timestamp_In_Seconds
Hour
Minutes
Seconds
AM_PM_Flag

Calendar_Date
Calendar_Date_Key
System_Date
Day_of_Week
Day_Number_Of_Week
Day_Number_Of_Month
Day_Number_In_Year
Week_Ending_Date
Week_Number_In_Month

Employee
Employee_Key
Employee_UEID
First_Name
Last_Name
Display_Name
Employee_email
Employee_Phone

© 2007 Hewlett-Packard Company

Mapping events to abstract processes

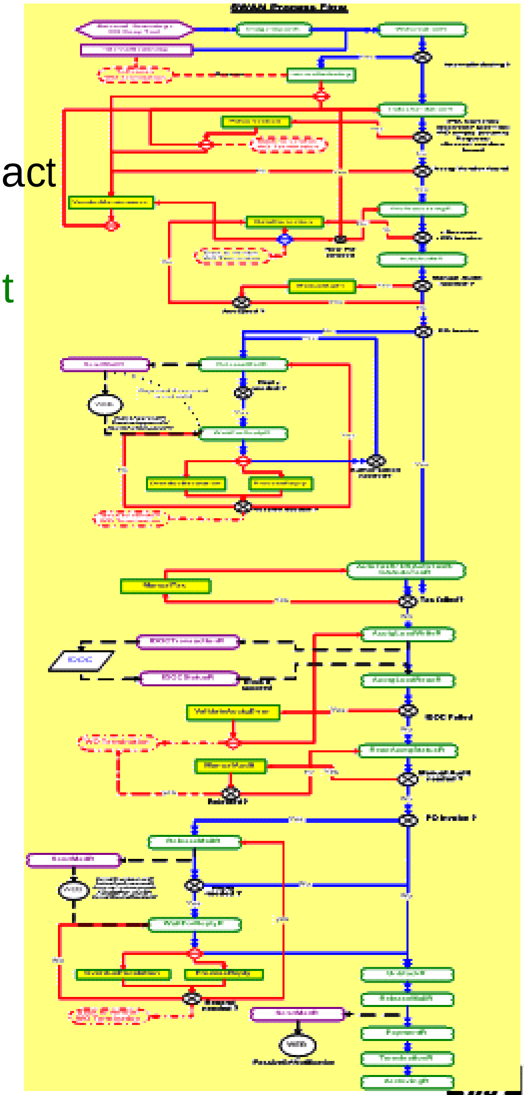
- Two facets to provide abstracted process representations
  - A way to model the abstraction
    - Describe the high level process
    - Describe how its progression maps to underlying IT events
  - ETL mechanism to load warehouse with abstracted process execution data

© 2007 Hewlett-Packard Company

Map Events to Abstract Process Progression

- Typically, events signal status changes in steps of the implemented process
- Have to specify or learn abstract process progression
- Mappings between monitored events and start/completion of abstract process steps, data relevant to the abstract process, …

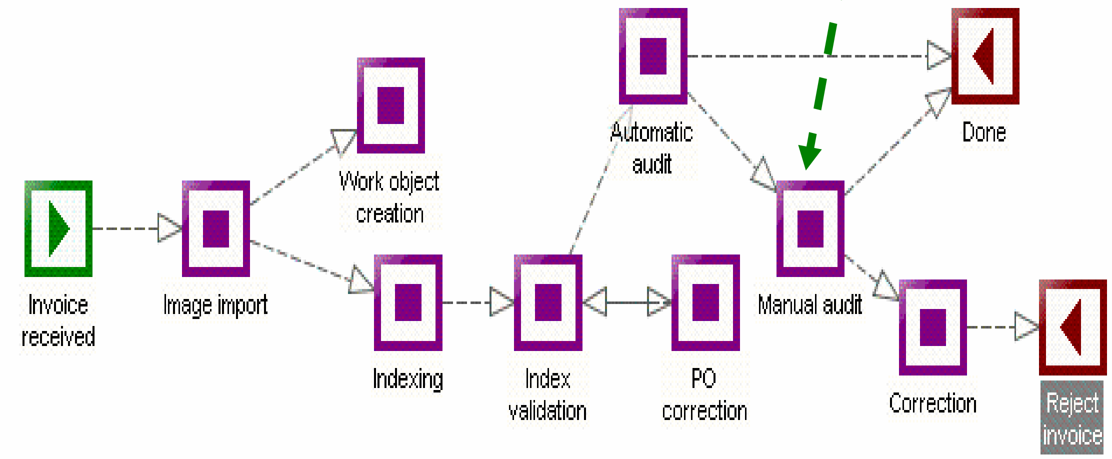
Implemented process



Audit Msg
Invoice_ID=123
Amount=$100
response=OK

Event

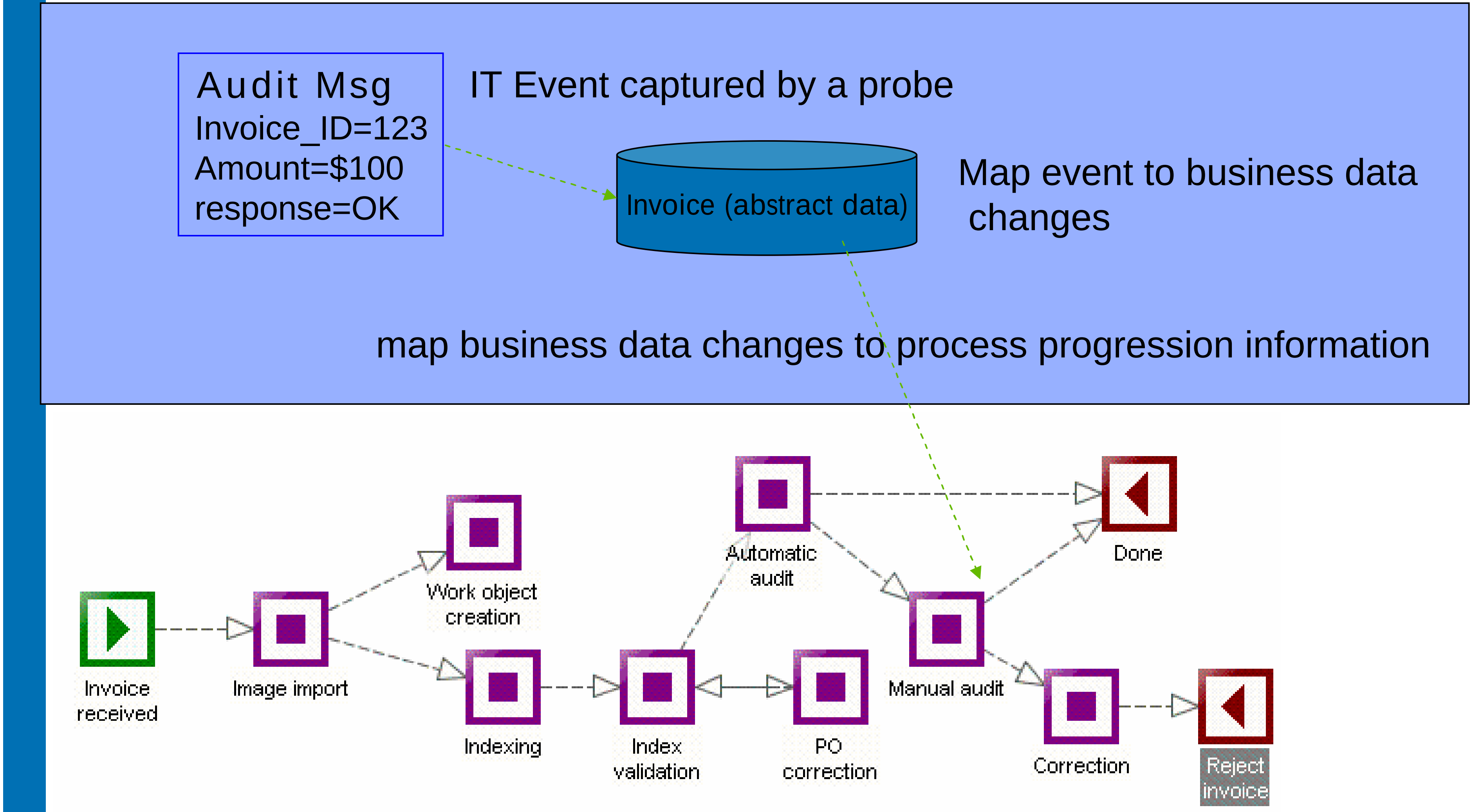
Process data changes map to progression information



Abstract invoice payment process

© 2007 Hewlett-Packard Company

Two-phased mapping



Audit Msg
Invoice_ID=123
Amount=$100
response=OK

IT Event captured by a probe

Invoice (abstract data)

Map event to business data changes

map business data changes to process progression information

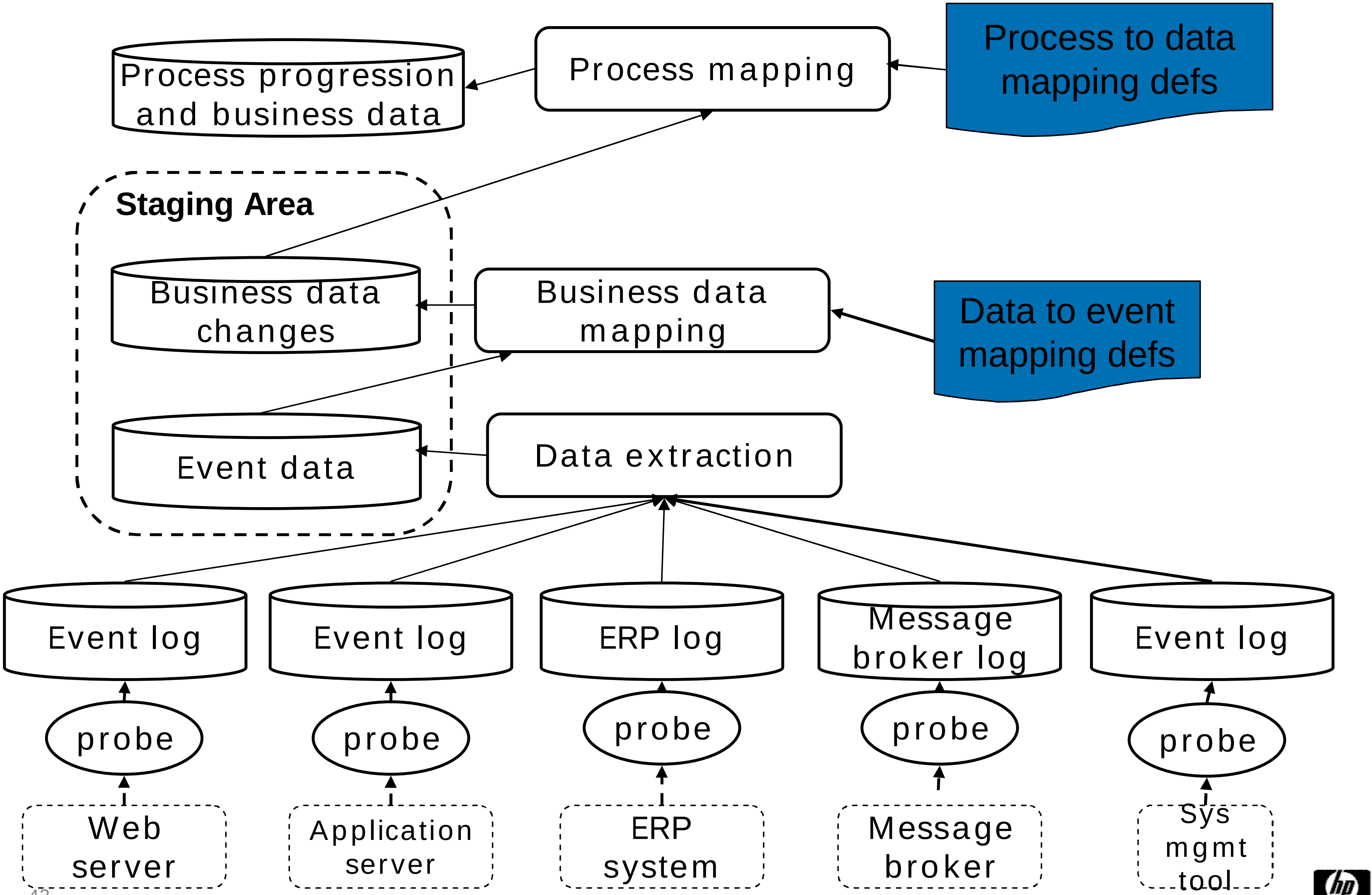
© 2007 Hewlett-Packard Company

Why the indirection?

- Many different events may cause the same change to a business data item
- Same business data can be used to support and mark progression of instances of different process types
- In practice, for abstract processes the progression often depends on business data changes
- Benefits
  - Reduces specification & maintenance effort
  - Specs are more robust to changes in the info sources (event specs updated but no need for business data or progression info)

© 2007 Hewlett-Packard Company

Extraction & abstraction of process data



© 2007 Hewlett-Packard Company

Loading process data

- Modeling specs used by the ETL to map across levels of abstraction
    - IT events captured with probes and logged with timestamps
    - ETL reads event tables in logs and orders them by time
    - Events are mapped to business data changes
    - Business data changes are ‘replayed’ in order and relevant changes are detected for computing process progression
    - Process progression creates records for the step execution data which are loaded into the warehouse

© 2007 Hewlett-Packard Company

ETL generation

- Automates staging area creation & maintenance
- Automates generation of executable transformation scripts
  - Indirection of mappings from IT events to process progression à Two-phased transformation
    - Phase 1: IT events mapped to business data changes
    - Phase 2: business data changes mapped to process progression

© 2007 Hewlett-Packard Company

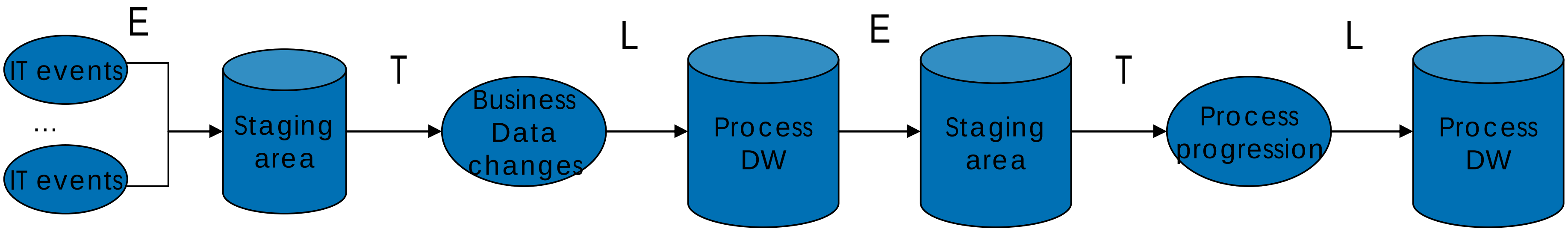
Staging area

- Three types of tables
    - Landing tables
        - Buffering of extracted IT events data
        - Checks for errors in the extraction
        - Refreshes at every cycle
    - Image tables
        - Keep an image of the IT events records extracted since the first extraction
        - Input to first transformation phase
    - Comparisons between landing & image tables
        - To detect duplicates
        - Determine manipulation operation (I, d, u)
    - Intermediate tables
        - Output of first transformation phase
        - Business data changes
        - Input to second transformation phase

© 2007 Hewlett-Packard Company

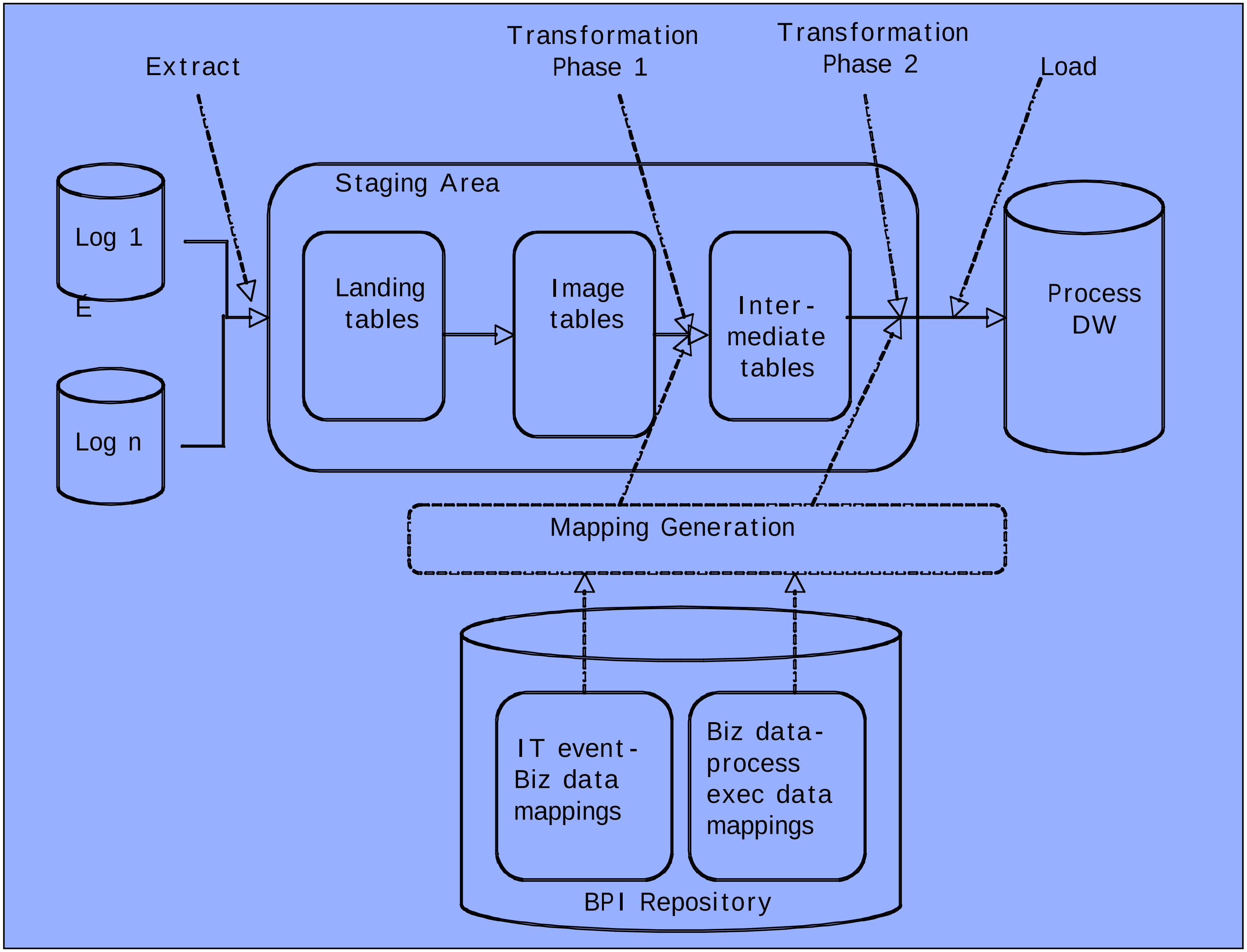
Intermediate tables

- Alternative design: 2 separate ETL processes but …
  - Inefficient
    - Extraction and staging of business data changes
    - Additional tables to keep all business data changes to mark process progression
      - DW only stores the last version of a business data instance



© 2007 Hewlett-Packard Company

ETL Transformation phases



© 2007 Hewlett-Packard Company

Executable mapping generation

- How to execute the transformations?
  - Agnostic to underlying tool
  - Modeling: declarative mappings
  - Mapping Generator derives prescriptive mappings
    - Two phases
      - Prescriptive logical mappings
        - Canonical language to express executable semantics (pseudo-SQL)
      - Prescriptive executable mappings
        - Specific translators (or manually)
      - Orthogonal to the two transformation phases

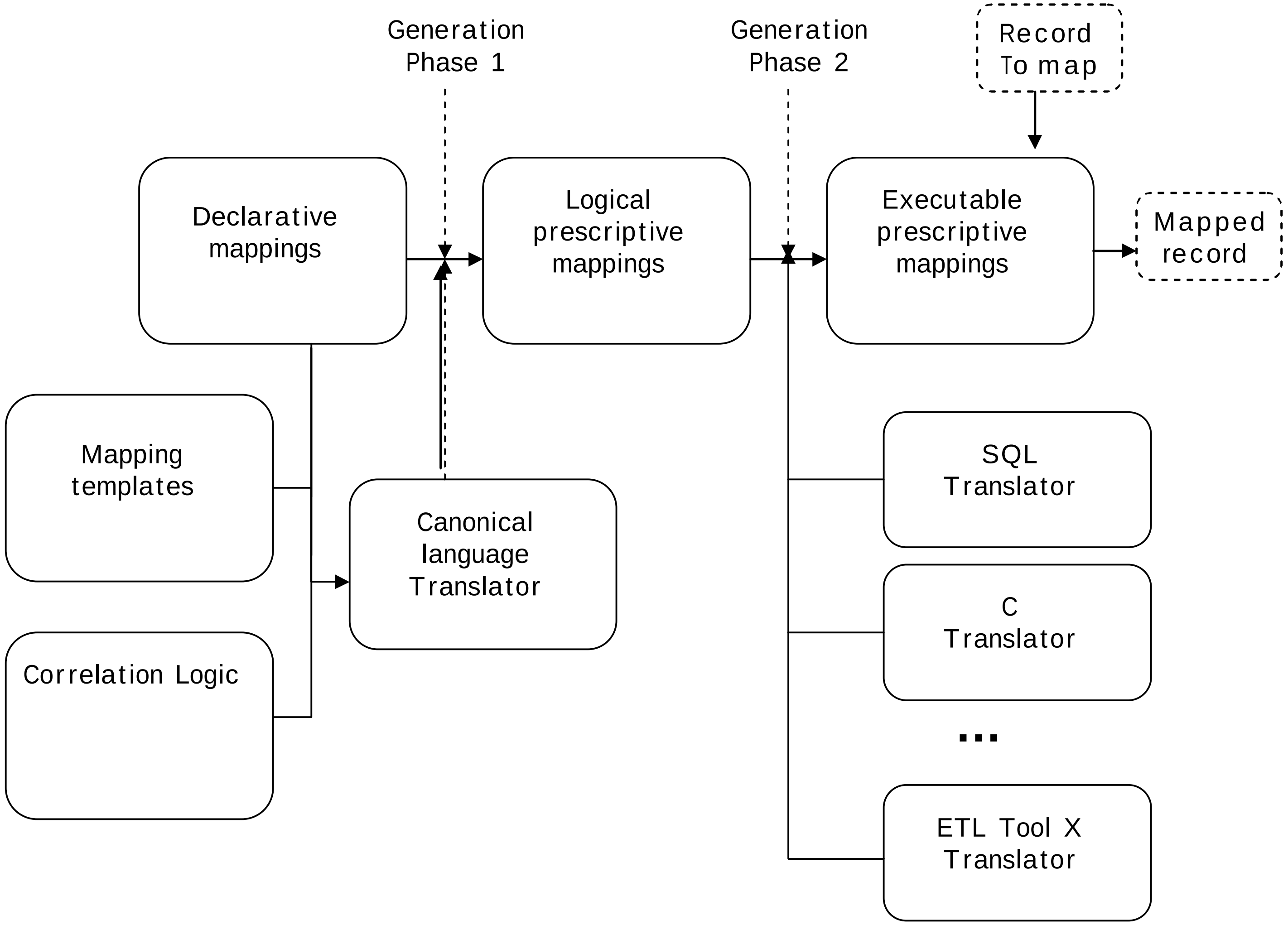
© 2007 Hewlett-Packard Company

Mapping Generator

- Core: mapping templates
  - Parameterized logical scripts in canonical language
    - Capture executable semantics
      - Factor out commonalities of mapping between the layers of abstraction
      - Exploits DW semantics
      - Captures other correspondences not specified by the declarative mapping (e.g., duration)
    - Parameters: event-, business entity-, process step-related
    - Templates instantiated by declarative mappings
    - Different template types (e.g., bizEntity_to_endStep)
    - Not executable
    - Canonical language translator

© 2007 Hewlett-Packard Company

Mapping generation phases



© 2007 Hewlett-Packard Company

Outline

- Context
  - Business Process Management
  - Business Process Intelligence
  - Relevance of Information Integration
- Process Modeling Issues
  - Process Views
  - Metrics Model
- Information Integration Issues
  - Generic Data Warehouse Schema
  - Abstraction Mechanisms
  - Generic ETL
  - Information Extraction from Semi-Structured Data
- Summary

© 2007 Hewlett-Packard Company

Information Extraction from Less Structured Data

- Information relevant to business processes may be contained in less structured data
- Many enterprises have a large corpus of contracts, customer service logs, reviews, etc.
- For example, the enterprise must be able to respond to events that might affect existing contractual relationships
- Critical information remains buried in text, e.g., key parameters of Service Level Agreements
- Need to incorporate this information into Business Process Intelligence solution

© 2007 Hewlett-Packard Company

Our approach

- Automatically identify “facts” in text documents
- Based on the use of
  - two object models
  - information extraction techniques
- Identified facts can be
  - automatically tagged with XML
  - extracted into the business process data warehouse as additional business data
- Extracted information becomes readily available for BPI: metrics definitions, queries, reports, analyses

© 2007 Hewlett-Packard Company

Document templates

- Common practice: set of free text templates for different contract types
  - For each type à one or more templates
  - E.g.,
    - Long term agreement (LTA)
    - Corporate purchase agreement (CPA)
- Contract templates organized in clauses
  - Each clause: specific kind of factual information
  - E.g.,
    - Term clause

© 2007 Hewlett-Packard Company

CPA term clause template

This CPA will be a [TERM] Agreement for the period [START DATE] to [EXPIRATION DATE] inclusive. Both parties agree to meet prior to [MM/DD/YY] to consider an extension of [##] year(s). In like manner, both parties shall meet prior to [MONTH/DAY OF EXPIRATION DATE] of each year to consider future extensions.

© 2007 Hewlett-Packard Company

Template instantiation

- Relevant fact types in templates: attributes
- Clause template instantiation
  - Value assignment to attributes (facts)
- Contract
  - Combination of clause template instantiations
  - Some variations
    - In the text
    - In the order of clauses

© 2007 Hewlett-Packard Company

Instance of CPA term clause

This CPA will be a one year Agreement for the period 05/01/03 to 05/01/04 inclusive. Both parties agree to meet prior to 04/01/04 to consider an extension of one year. In like manner, both parties shall meet prior to 05/01 of each year to consider future extensions.

© 2007 Hewlett-Packard Company

Document annotation

- Learning rules to automatically identify facts requires a training set
  - Subset of contract collection annotated with tags for relevant facts
- Two kinds of tags
  - Attribute tags
  - Structural tags

© 2007 Hewlett-Packard Company

Attribute tags in term clause instance

This CPA will be a <TERM> one year </TERM> Agreement for the period <START_DATE> 05/01/03 </START_ DATE> to <EXPIRATION_DATE> 05/01/04 </EXPIRATION_DATE> inclusive. Both parties agree to meet prior to <IMMEDIATE_EXTENSION_MEET_DATE> 04/01/04 </IMMEDIATE_EXTENSION_ MEET_DATE> to consider an extension of <EXTENSION_PERIOD> one </EXTENSION_PERIOD> year. In like manner, both parties shall meet prior to <FUTURE_EXTENSION_MEET_DATE> 05/01 </FUTURE_EXTENSION_MEET_DATE > of each year to consider future extensions.

© 2007 Hewlett-Packard Company

Structural tags for term clause

<TERM_CLAUSE> This CPA will be a <TERM> one year </TERM> Agreement for the period <START_DATE> 05/01/03 </START_DATE> to <EXPIRATION_DATE> 05/01/04 </EXPIRATION_DATE> inclusive. Both parties agree to meet prior to <IMMEDIATE_EXTENSION_MEET_DATE> 04/01/04 </IMMEDIATE_EXTENSION_ MEET_DATE> to consider an extension of <EXTENSION_PERIOD> one </EXTENSION_PERIOD> year. In like manner, both parties shall meet prior to <FUTURE_EXTENSION_MEET_DATE> 05/01 </FUTURE_EXTENSION_MEET_DATE > of each year to consider future extensions *</TERM_CLAUSE>*

© 2007 Hewlett-Packard Company

Object models

- Objective: guide annotation task
- XML-based
- Two kinds
  - Document object model (DOM)
    - Specifies structural components
      - Sections & clauses
      - Order can vary
  - Facts object model (FOM)
    - Specified relevant attributes
      - E.g., contract expiration date

© 2007 Hewlett-Packard Company

Document Object Model

```
<DOM>
        <id>  7002 </id>
        <contract type>
                LTA
        </contract type>
                 …
        <section>
                <name> Shipment and Delivery </name>
                <clause> prospective failure </clause>
                <clause> untimely shipment </clause>
        </section>
        <section>
                <name> Term </name>
                <clause> term </clause>
        </section>
                …
</DOM>
```

© 2007 Hewlett-Packard Company

Facts Object Model

```
<FOM>
                <id> 235 </id>
                <contract type>
                         LTA
                </contract type>
                <attribute>
                         <name> expiration_date </name>
                         <type>  date </type>
                         <nature> mandatory </nature>
                </attribute>
                <attribute>
                         <name> untimely_transportation_
                                  means</name>
                         <type>  transportation </type>
                         <nature> mandatory </nature>
                </attribute>
                    …
</FOM>
```

© 2007 Hewlett-Packard Company

Semantic types

- Enumeration

```
<type>
      <name> transportation </name>
      <kind> enumeration </kind>
      <values> airplane, ship, truck, trailer </values>
</type>
```

- Format

```
<type>
      <name> date </name>
      <kind> format </kind>
      <values> mm/dd/yy, month dd year, mm-dd-yyyy
                                                </values>
</type>
```

© 2007 Hewlett-Packard Company

Regularities

- Each attribute (fact type) can have one or more associated regularities

- Structural regularity
  - Regularities in the structural component (location) of an attribute
  - E.g., untimely_transportation_mean à Shipment and Delivery section

- Phrasal regularity
  - Regularities in the surrounding words
  - E.g., for the start_date attribute of a term clause
    - for the period 01/01/2004 to
    - starting from 01/01/2004 “until

- Grammatical regularity
  - Regularities in the parts of speech (e.g., noun, verb, adjective, etc) of surrounding words, and/or in the syntactic relations between them (subject, etc)
  - Take advantage of clausal structure provided by a syntactic analyzer and PoS tagger

© 2007 Hewlett-Packard Company

Extracting facts

- Text mining
  - Information extraction techniques
    - Learn patterns for regularities
    - Train on annotated set
    - Generate rules
      - Antecedent: pattern for combination of regularities
      - Consequent: attribute name for corresponding fact
    - One or more rules for each attribute
      - Different kinds of contract types
      - Different contract templates for a same contract type
      - Text variations
    - Rules database

© 2007 Hewlett-Packard Company

Rule example

- Rule for attribute expiration_date from an LTA term clause instance

```
<Rule>
     <id> 153 </id>
     <FOM_object> 235 </FOM_object>    //id for LTA FOM object
     <antecedent>
          <structural_component>
               <section> TERM </section>
               <clause> TERM </clause>
          </structural_component>
          <surrounding_component>
                    ‘period’ date ‘to’ (date)
          </surrounding_component>
     </antecedent>
     <consequent>
          <attribute> expiration_date </attribute>
     </consequent>
</Rule>
```

© 2007 Hewlett-Packard Company

Generating rules

- Information extraction technique
  - Each instance-attribute tag pair becomes a “seed” to grow a new rule that covers the seed
  - Top-down algorithm to induce rules
    - First finds the most general rule that covers the seed
      - Anchors the boundaries of the fact
    - Then, extends the rule by adding terms one at a time
      - Metric to select a new term: expected error of the rule
- The technique is made more efficient by the use of structural tags
  - At rule generation time the structural context narrows the validation space
  - At rule application time the structural context narrows the search space

© 2007 Hewlett-Packard Company

Summary

- The intelligent enterprise monitors and optimizes its business processes and interactions with business partners.
- Better business process management is “essential” (and independent of automation)
- Today, this is very difficult to do, although tools are appearing to address pieces of the problem.
- Our approach: a Business Process Intelligence solution (Business Cockpit) that combines process modeling, metrics definition, generic DW schema and ETL generation, and analytics
- But, many research challenges remain.

© 2007 Hewlett-Packard Company

References

- Fabio Casati, Malu Castellanos, Umeshwar Dayal, Norman Salazar,"A Generic Solution to Warehousing Business Process Data." VLDB 2007.
- Fabio Casati, Malu Castellanos, Norman Salazar, Umeshwar Dayal, "Abstract Process Data Warehousing." ICDE 2007.
- Fabio Casati, Malu Castellanos, Umesh Dayal, Ming-Chien Shan, "A Metric Definition, Computation, and Reporting Model for Business Operation Analysis." EDBT 2006.
- Ming C. Hao, Daniel A. Keim, Umeshwar Dayal, Jörn Schneidewind, "Business Process Impact Visualization and Anomaly Detection" Information Visualization Journal 2006.
- Malu Castellanos, Fabio Casati, Mehmet Sayal, Umeshwar Dayal, "Challenges in Business Process Analysis and Optimization." Proc. TES Workshop, Springer-Verlag, 2005.
- Malu Castellanos, Fabio Casati, Umesh Dayal, Ming-Chien Shan, iBOM: A Platform for Business Operation Management." ICDE 2005.
- Fabio Casati, Malu Castellanos, Umesh Dayal, Ming-Chien Shan, "Probabilistic, Context-Sensitive, and Goal-Oriented Service Selection." ICOSOC 2005.
- Malu Castellanos, Norman Salazar, Fabio Casati, Umesh Dayal, Ming-Chien Shan, "Predictive Business Operations Management." DNIS 2005.
- Mehmet Sayal, Ming-Chien Shan, "Analysis of Numeric Data Streams at Different Granularities." IEEE International Conference on Granular Computing, July 2005.
- Malu Castellanos, Fabio Casati, Umeshwar Dayal, Ming-Chien Shan, "A Comprehensive and Automated Approach to Intelligent Business Process Execution Analysis." Distributed and Parallel Databases 16(3): 239-273, 2004
- Malu Castellanos, Norman Salazar, Fabio Casati, Ming-Chien Shan, Umesh Dayal, "Automatic Metric Forecasting for Management Software." OVUA Workshop 2004.
- Daniela Grigori, Fabio Casati, Malu Castellanos, Umesh Dayal, Ming-Chien Shan, Mehmet Sayal. "Business Process Intelligence." Computers in Industry 53 (3), April 2004.
- Ming C. Hao, Daniel A. Keim, Umeshwar Dayal: VisBiz, "A Simplified Visualization of Business Operation." IEEE Visualization 2004
- Malú Castellanos, Fabio Casati, Umeshwar Dayal, Ming-Chien Shan, Intelligent Management of SLAs for Composite Web Services. DNIS 2003.
- Fabio Casati, " Eric Shan, Umeshwar Dayal, Ming-Chien Shan:, "Business-oriented management of Web services." Commun. ACM 46(10)
- Fabio Casati, Umeshwar Dayal, Ming-Chien Shan, " Business Operation Intelligence." DNIS 2002.
- Mehmet Sayal, Fabio Casati, Umeshwar Dayal, Ming-Chien Shan, "Business Process Cockpit." VLDB 2002.
- Angela Bonifati, Fabio Casati, Umesh Dayal, and Ming-Chien Shan, "Warehousing Workflow Data: Challenges and Opportunities." VLDB 2001.

© 2007 Hewlett-Packard Company

Thanks!

© 2007 Hewlett-Packard Company