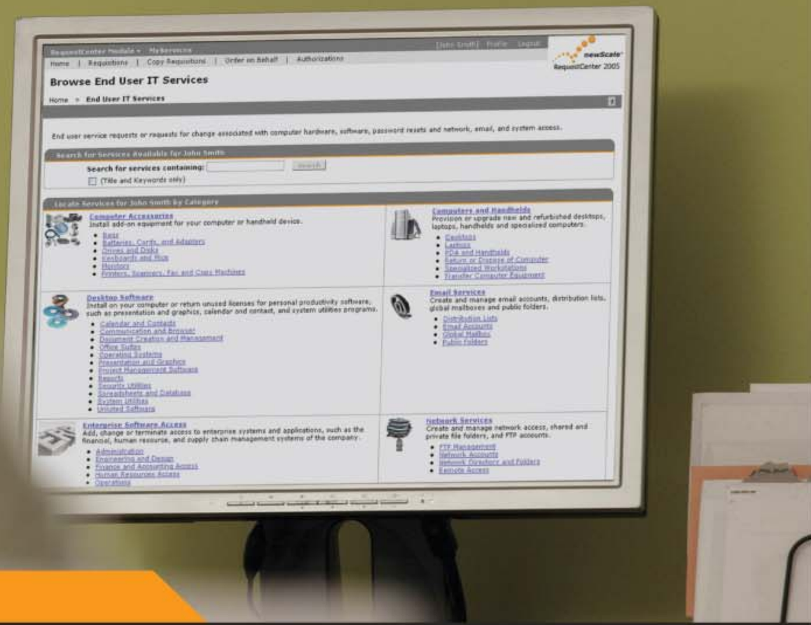
Real-Life Data Mart Design Challenges

Leslie Tierstein


RequestCenter Module ▾ MyServices
[John Smith] Profile Logout
Home | Requisitions | Copy Requisitions | Order on Behalf | Authorizations
newScale
RequestCenter 2005
Search

Browse End User IT Services

Home > End User IT Services

End user service requests or requests for change associated with computer hardware, software, password resets and network, email, and system access.

Search for Services Available for John Smith

Search for services containing:

□ (Title and Keywords only)

Locate Services for John Smith by Category

Computer Accessories
Install add-on equipment for your computer or handheld device.
- Bags
- Batteries, Cards, and Adapters
- Drives and Disks
- Keyboards and Mice
- Monitors
- Printers, Scanners, Fax and Copy Machines

Computers and Handhelds
Provision or upgrade new and refurbished desktops, laptops, handhelds and specialized computers.
- Desktops
- Laptops
- PDA and Handhelds
- Return or Dispose of Computer
- Specialized Workstations
- Transfer Computer Equipment

Desktop Software
Install on your computer or return unused licenses for personal productivity software, such as presentation and graphics, calendar and contact, and system utilities programs.
- Calendar and Contacts
- Communication and Browser
- Document Creation and Management
- Office Suites
- Operating Systems
- Presentation and Graphics
- Project Management Software
- Reports
- Security Utilities
- Spreadsheets and Database
- System Utilities
- Unlisted Software

Email Services
Create and manage email accounts, distribution lists, global mailboxes and public folders.
- Distribution Lists
- Email Accounts
- Global Mailbox
- Public Folders

Enterprise Software Access
Add, change or terminate access to enterprise systems and applications, such as the financial, human resource, and supply chain management systems of the company.
- Administration
- Engineering and Design
- Finance and Accounting Access
- Human Resources Access
- Operations

Network Services
Create and manage network access, shared and private file folders, and FTP accounts.
- FTP Management
- Network Accounts
- Network Directory and Folders
- Remote Access

newScale®
Service Catalog Leader

Overview

- Data mart design challenges addressed in this project:
  - Dynamically defined dimensions
  - Potentially 100’s of dimensions
  - Loading XML data into the data mart
  - An incremental refresh that includes updates
  - Supporting multiple databases with a standard toolset and code base
  - Designing the physical database to map to the BI tools’ business view

Overview

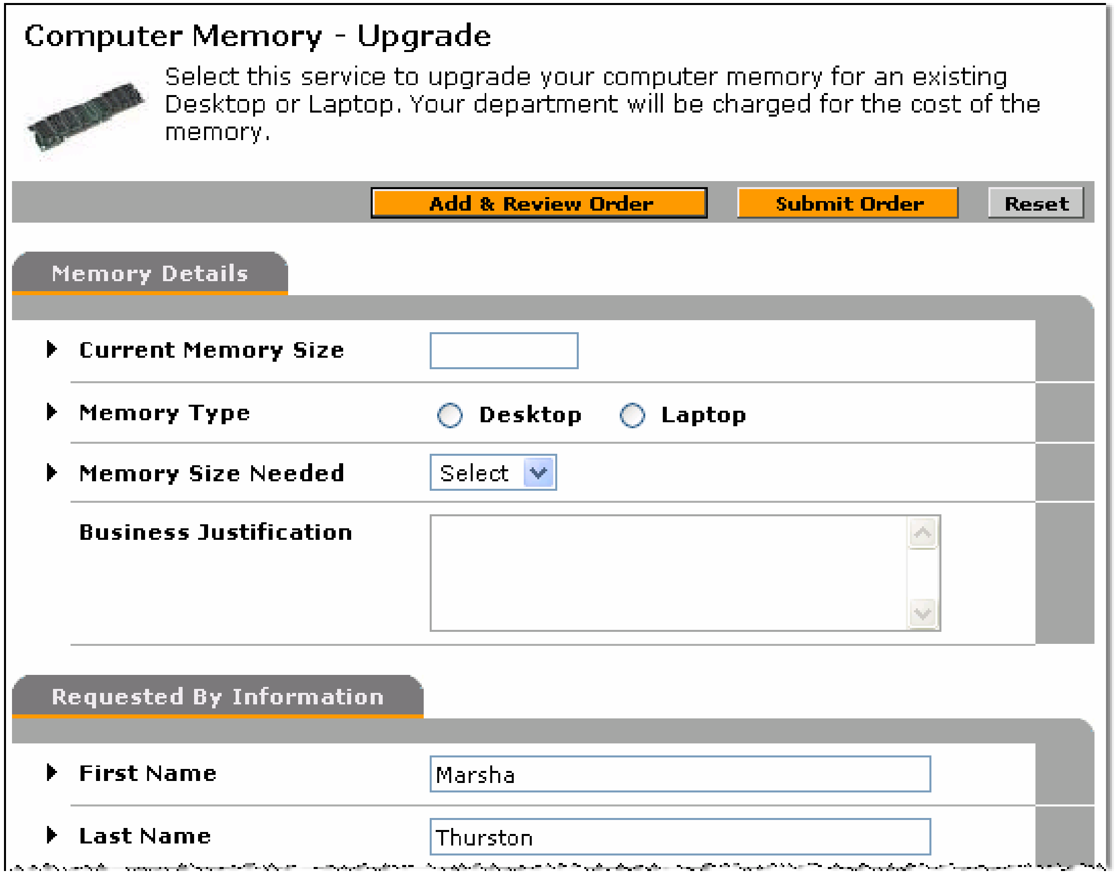
- Data mart design challenges not addressed in this project:
  - Performance tuning for very large databases (VLDB's)
  - Data cleansing and data integrity
  - Integrating data from heterogeneous databases or multiple applications

The OLTP System

- Service Catalog Management
- An IT department defines a catalog of all the services it provides to users
  - End User Computing: Add Visio
  - Advanced Computing: Configure a server
- Users request a service
  - System tracks the service request, all tasks required to complete it

The OLTP System (1)

- Requestor View – Upgrade Memory
  - User fills out and submits an order (requisition)



Computer Memory - Upgrade

Select this service to upgrade your computer memory for an existing Desktop or Laptop. Your department will be charged for the cost of the memory.

Memory Details

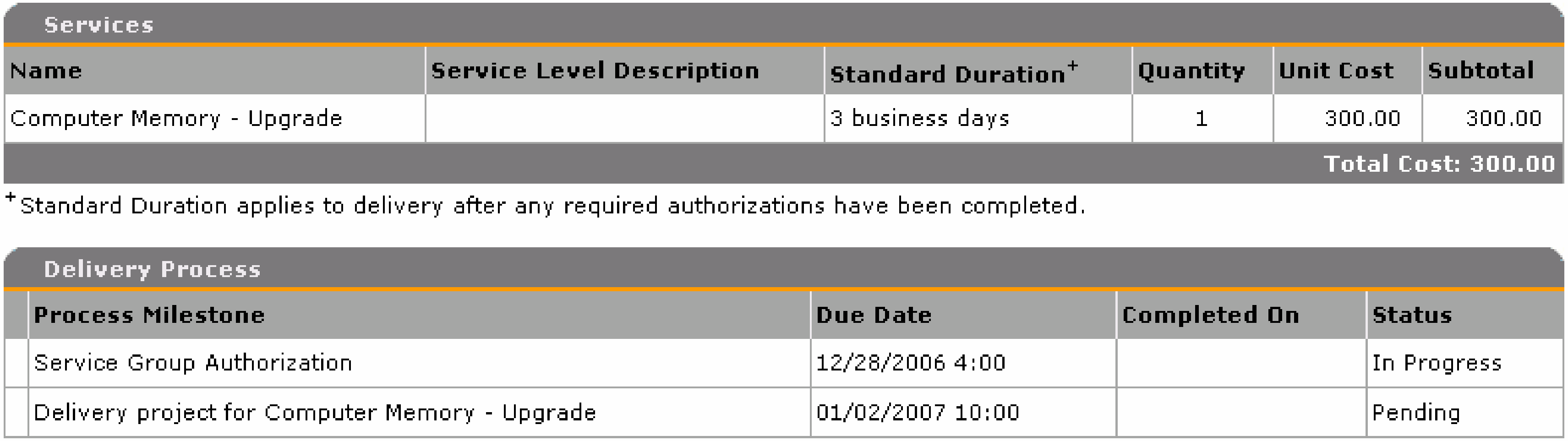
Current Memory Size	
Memory Type	○ Desktop ○ Laptop
Memory Size Needed	Select
Business Justification	

Requested By Information

First Name	Marsha
Last Name	Thurston

The OLTP System (2)

- IT View – Upgrade Memory
  - User request has criteria for completion
  - Request is routed to appropriate people for authorization and task delivery



Services

Name	Service Level Description	Standard Duration[+]	Quantity	Unit Cost	Subtotal
Computer Memory - Upgrade		3 business days	1	300.00	300.00
					Total Cost: 300.00

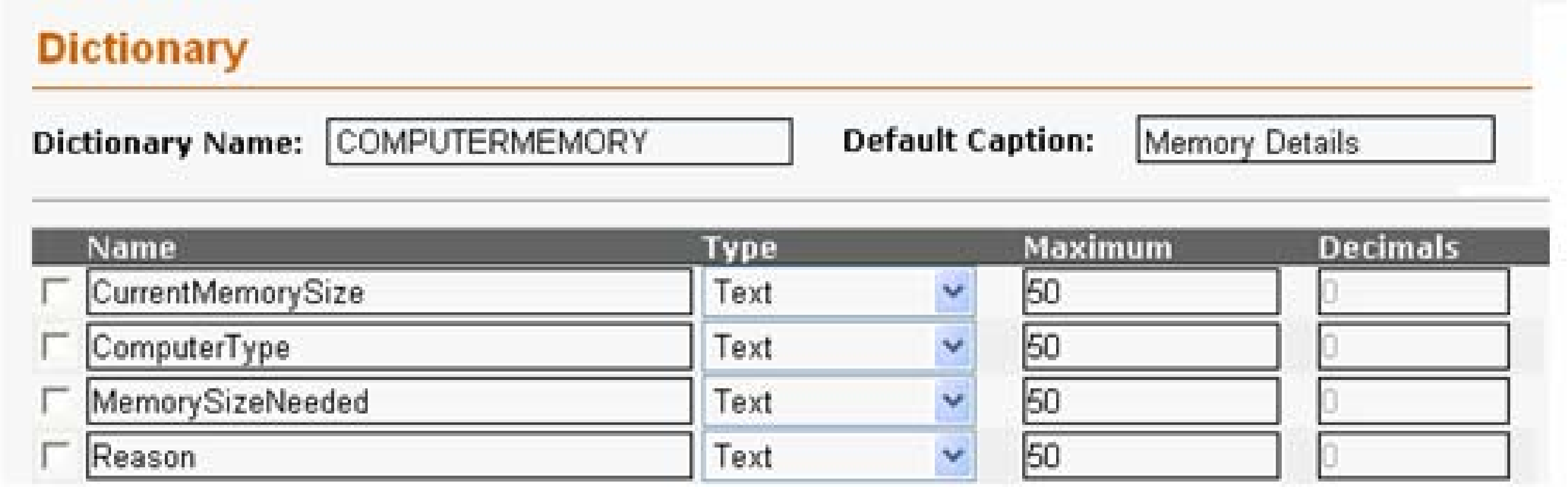
[+] Standard Duration applies to delivery after any required authorizations have been completed.

Delivery Process

	Process Milestone	Due Date	Completed On	Status
	Service Group Authorization	12/28/2006 4:00		In Progress
	Delivery project for Computer Memory - Upgrade	01/02/2007 10:00		Pending

The OLTP System (3)

- Service Designers’ View
  - Standard information (duration, cost, delivery plan, authorizations) defined via UI
  - Service-specific data entry requirements defined via “dictionaries”



Dictionary

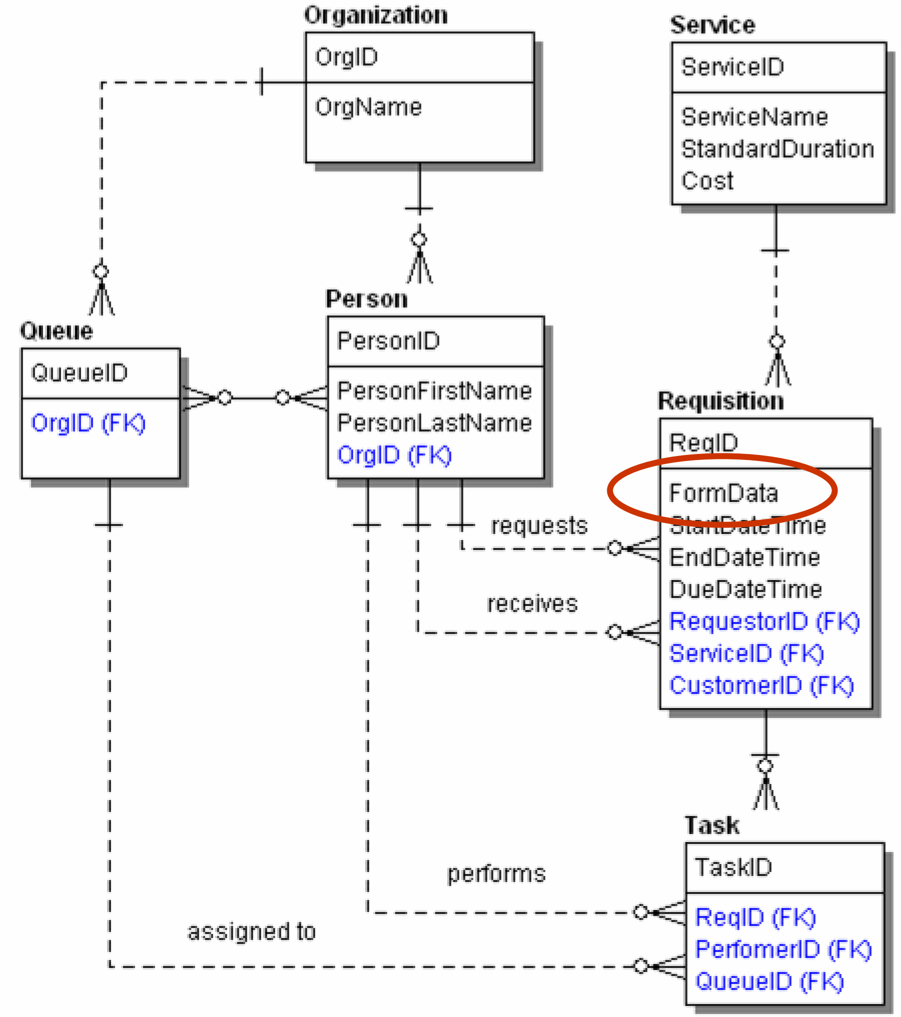
Dictionary Name: COMPUTERMEMORY

Default Caption: Memory Details

	Name	Type	Maximum	Decimals
☐	CurrentMemorySize	Text	50	0
☐	ComputerType	Text	50	0
☐	MemorySizeNeeded	Text	50	0
☐	Reason	Text	50	0

The OLTP System

- Vastly simplified ERD of transactional data
- Metadata on service definitions is also maintained
- Where is form (dictionary-based) data?



Data Mart Requirements

- Data mart must include form data
- Refresh needs to run in nightly maintenance window
  - Minimize impact on global accounts
- Data mart must include data on both open (in progress) and completed requisitions and tasks
  - Refresh must perform "updates"

Extracting Form (XML) Data

- Extract XML data from the requisition BLOB
  - Cognos8 includes Composite, which can extract data from XML
    - Performance is problematic
  - Adding a custom plug-in in the ETL tool provided insufficient power/flexibility
  - Custom Java program was needed
- Would the custom program perform acceptably?

ETL Performance Studies

- 100,000 requisitions per month (an order of magnitude higher than current usage)

Data Points	#	Units
Average # of requisitions started per day	5000	Requisitions
Average # of entries per requisition	1.5	Entries per Requisition
Days from requisition start until it's closed	10	Days
Average # of tasks in a service	6	Tasks per Service
Average # of actions taken per task	1.6	Actions per Task
Average # of dictionary data elements	10	Data Elements per Dictionary
Average # of dictionaries per service	4	Dictionaries per Service
% requisition entries with reportable dictionaries	70%	Requisition Entries
Time to extract one data element to data mart	8	Milliseconds

ETL Performance Studies

Analysis		Units
Requisitions open at any one time	50,000	Requisitions
Requisitions entries open at any one time	75,000	Requisition Entries
Tasks open at any one time	45,000	Tasks
Actions recorded per day	72,000	Actions
Requisitions entries changed in any one day	72,000	Requisition Entries
Requisition entries with reportable dictionaries changed per day	50,400	Requisition Entries
Requisition entry dictionaries changed per day	201,600	Requisition Entry Dictionaries
Requisition entry data elements changed per day	2,016,000	Requisition Entry Data Elements
Time to extract changed data elements	**16,128**	**Seconds**
	269	**Minutes**
	4.48	**Hours**

ETL Performance Studies

- Averages are good, but know your outlyers
  - If only to be able to warn users about potential issues
  - Hotel/Hospitality Company:
    - 30% of services have 100-160 fields, with a maximum of 250. About 60% have 70-100 fields
    - There are dictionaries with about 40 fields, with a maximum of 43.
- Inserts are easy, updates … NOT

Updating Data Mart Contents

- Data Mart contents need to be updated
  - Tasks and requisitions are added to the data mart as soon as the requisition is submitted
  - Status of requisitions may change, as well as completion date and form data
  - Task status and info is updated as the task is performed (completed)
- Logical "update" <> DML UPDATE

Updating Data Mart Contents

- Argh! Updating data mart contents is expensive
  - Updates require indexes
  - Inserts perform much better without indexes
  - Deletes (from large tables) are a disaster
- Avoid updates by placing data to be updated in a partition that is truncated
  - Refreshed data is written to the same partition or to a permanent partition when complete

Updating Data Mart Contents

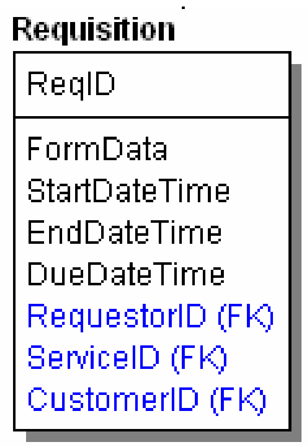
- Still an issue
- Partitioning not implemented because:
  - Small table size and growth rates
  - Lack of support in ETL tool
  - SQLServer (partition views only!)
  - What is partition key?
- Using insert/update model
- Time will tell if this is robust enough

Dimensional Modeling

- Table, View or Metadata?
  - Design the dimensional model in conjunction with the business view of the model
    - Date dimension(s)
    - People dimension(s)

Date Dimensions (1)

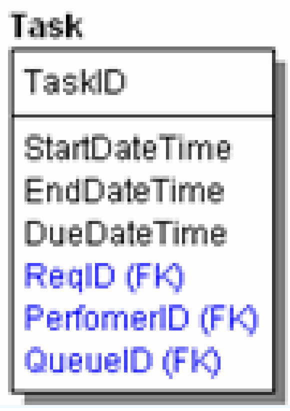
- Date required as a dimensional key
  - Due Date – for task and requisition
  - Closed Date – for task and requisition
  - Start Date – for task and requisition
- Complete date and time may also be needed
  - SLAs and OLAs



Requisition

ReqID

FormData
StartDateTime
EndDateTime
DueDateTime
RequestorID (FK)
ServiceID (FK)
CustomerID (FK)



Task

TaskID

StartDateTime
EndDateTime
DueDateTime
ReqID (FK)
PerfomerID (FK)
QueueID (FK)

Date Dimensions (2)

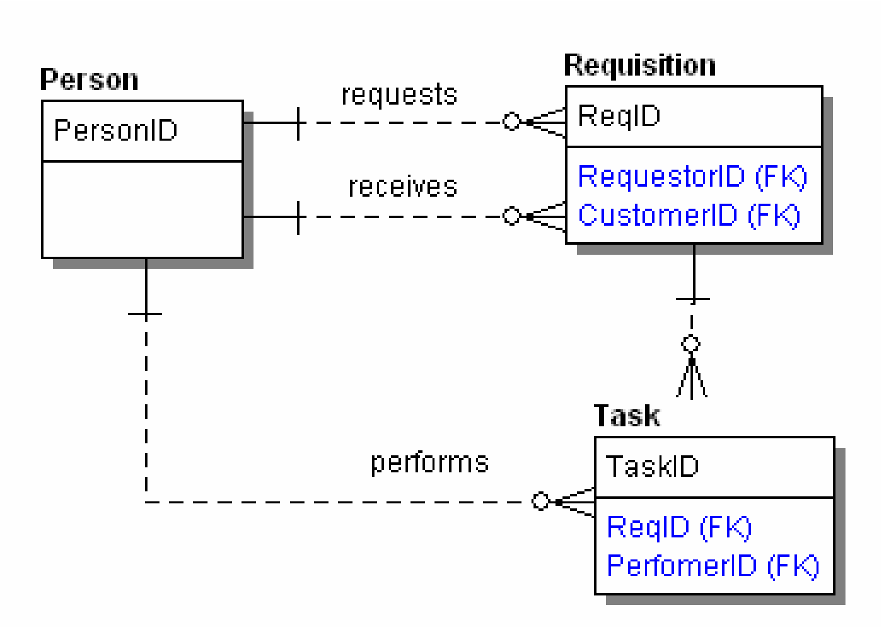
- How many database objects for the date dimensions?
  - 3 tables, 3 subject areas
  - 1 table, 3 views, 3 subject areas
  - 1 table, 3 subject areas
- Pros and cons of each approach
  - ETL, debugging, user interface in ReportNet
  - Where are relationships defined?
    - In the BI tool
    - In the database

Date Dimensions (3)

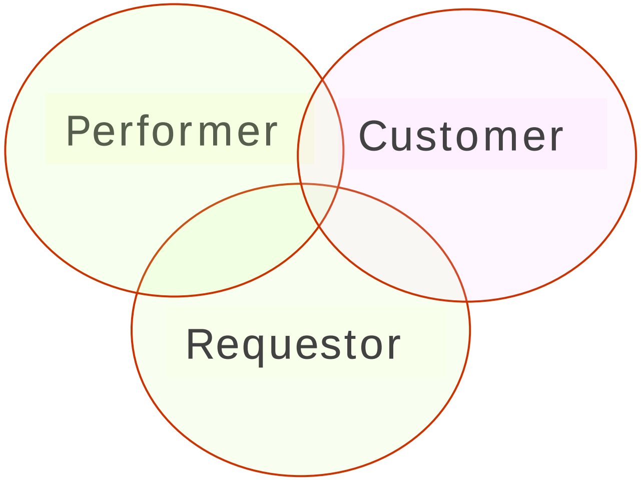
- Date dimensions are densely populated
  - Likely to have an attribute for every weekday date value
- Date dimensions can be statically populated
  - Rather than via fact/dimension build
- Solution: 1 table, 3 views, 3 subject areas
  - In Oracle, constraints can be defined in the view and created in the BI tool

People Dimensions (1)

- Required Dimensions
  - Customer – for requisition
  - Requestor – for requisition
  - Performer – for task



People Dimensions (2)



- How many database objects?
  - Do people fulfilling different roles intersect? And how much?
    - Customers and Requestors may overlap 100%
    - Performers are a much smaller group
  - Implications for prompts and filters
    - Prompted filter should only show people who have filled the corresponding role

Dimensional Modeling

- Dynamic data model to accommodate form data
  - Other companies have done this, but their scenarios are not as complex
    - Oracle Noetix Views and Noetix data warehouse – Flex fields

Dimensional Model

- How to model the dimensions which correspond to dynamically defined dictionaries
  - Each attribute in a dictionary is a dimension – rejected
    - W-A-A-Y too many dimensions
  - “Junk”, “Degenerate” or “Demographic” dimensions
  - “Reference” dimensions

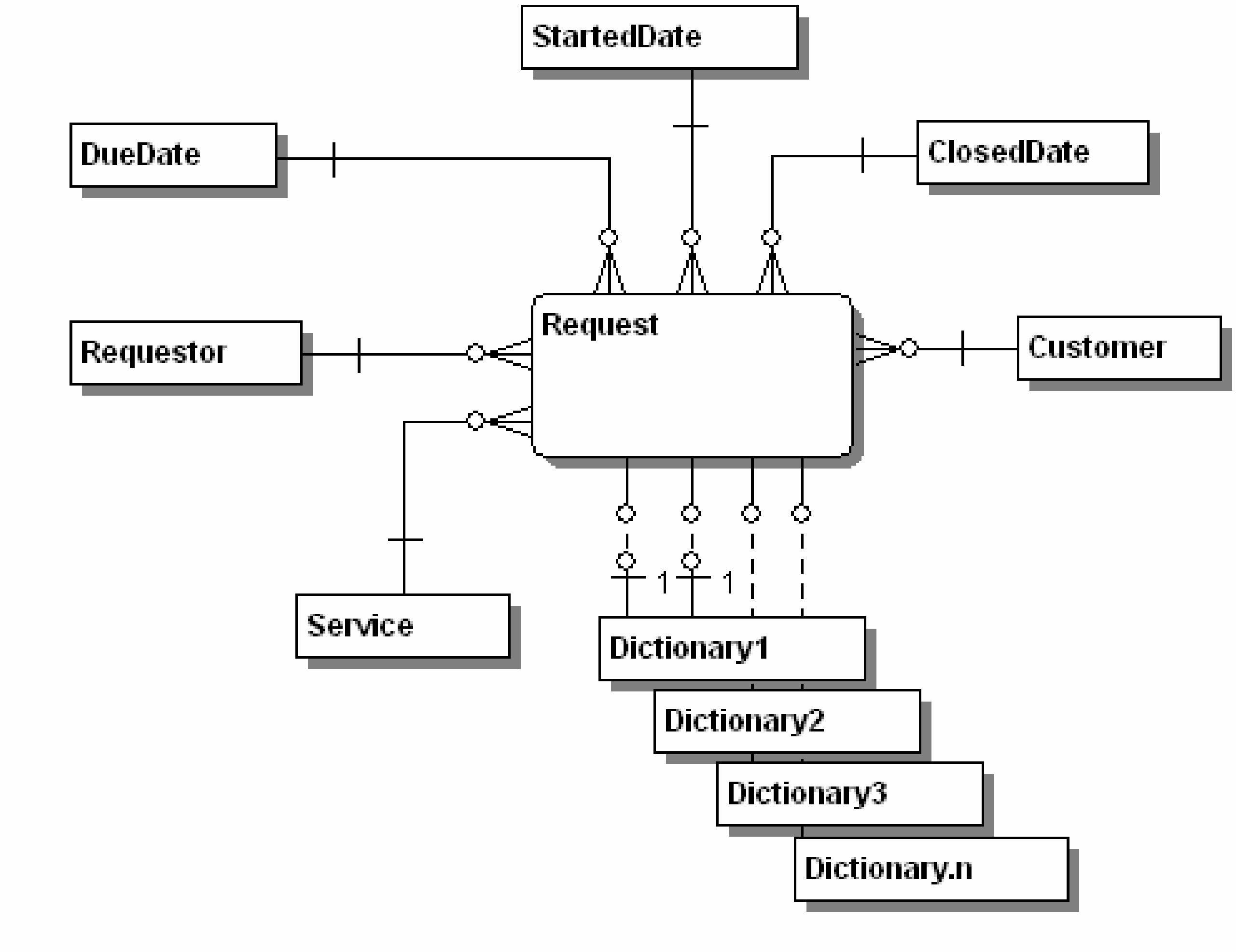
Junk Dimensions

- Populate the dictionary dimension using a unique combination of values for each dictionary across all requisitions
- The fact table has a column for a relationship with each dimension – many will be "N/A"
- A complex fact/dimensional build would be required
- See Kimball, Design Tips 46, 48

Reference Dimensions

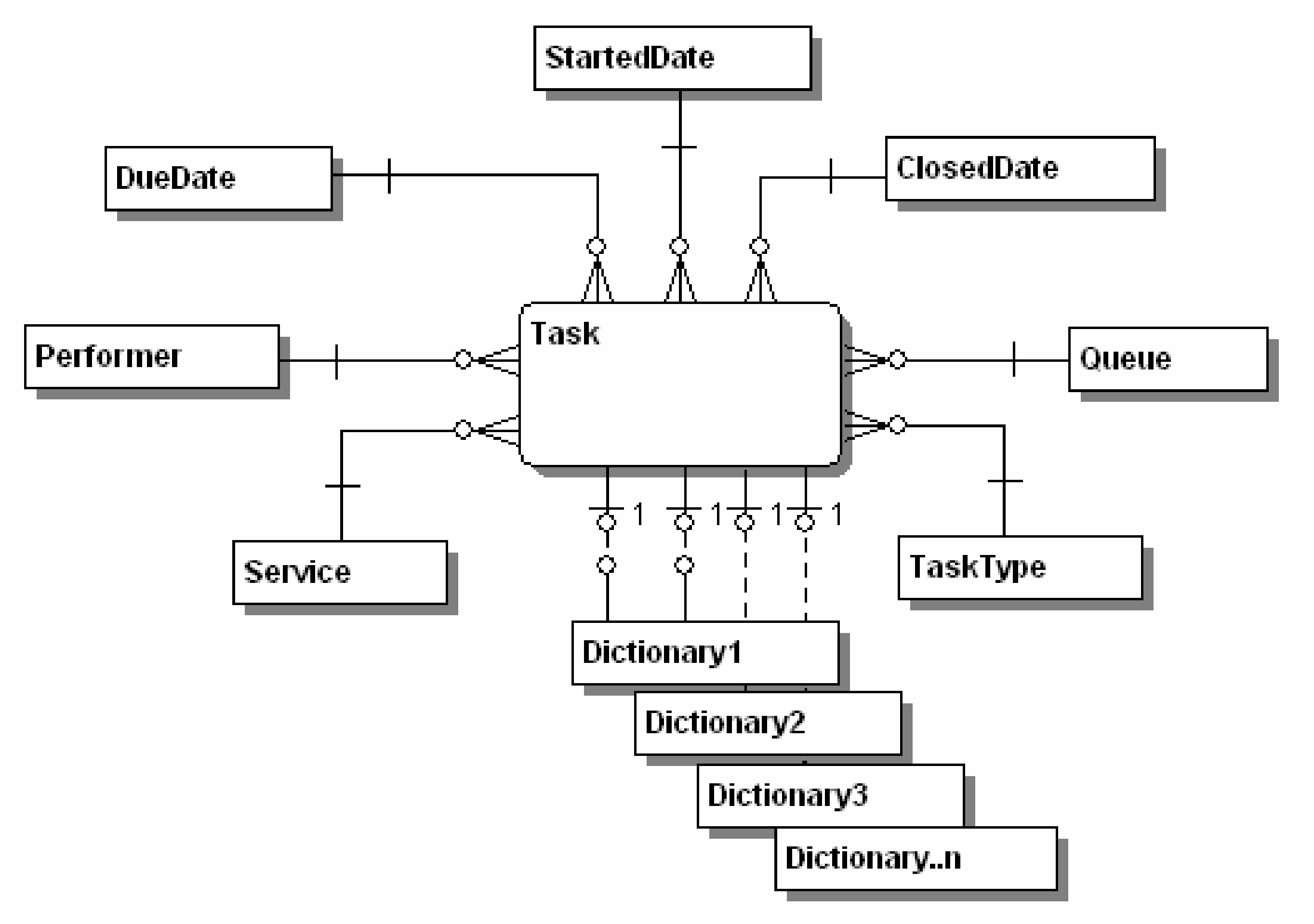
- Each dimension has a 1:1 relationship with the fact table
- Best to not have basic analytical metrics and frequent query topics in the reference dimension
- Greatly reduces the size of the fact table
- See Kimball, Design Tip 86 (http://www.kimballuniversity.com/html/designtips.html)

Requisition Star Schema



Task Star Schema

newScale

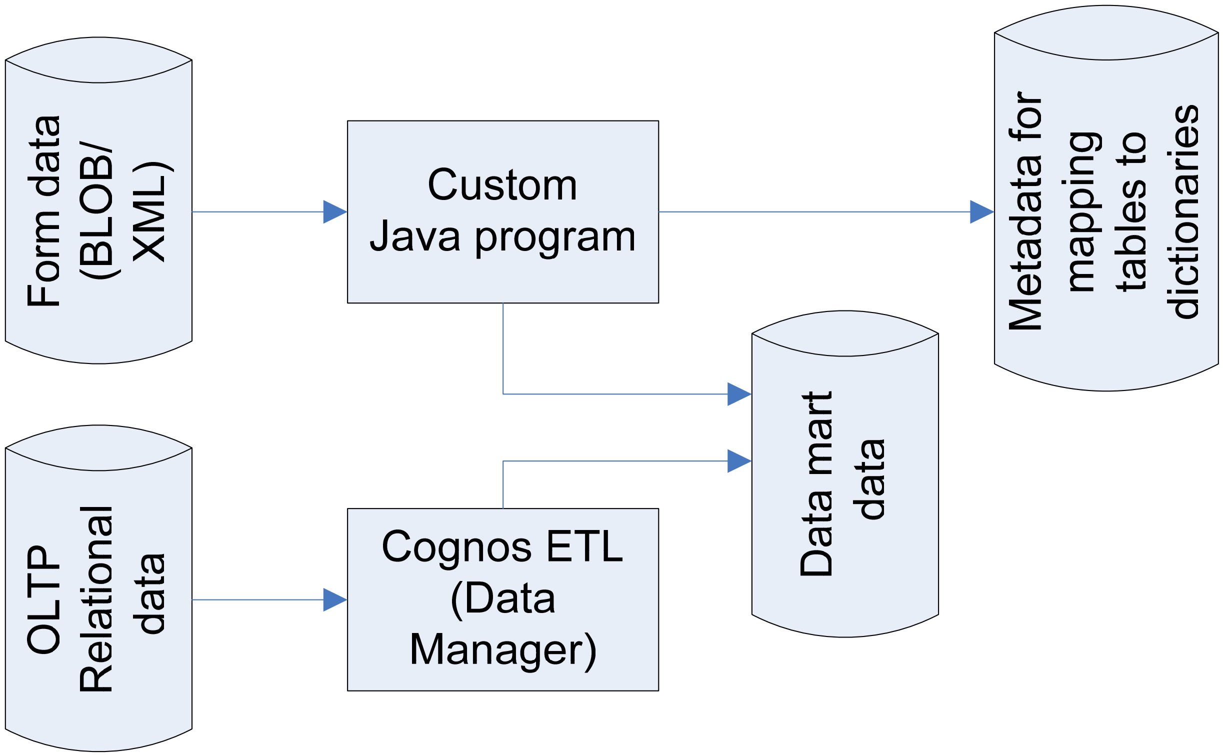


Tool Suite

Tool Type	Cognos Product
ETL tool (maps OLTP data to data mart)	Data Manager
BI Framework (maps database to end-user/business view)	Framework Manager
BI Power User Tool (allows end users to create reports/queries)	ReportStudio QueryStudio
BI End User Tool (provides portal for running reports/queries)	ReportNet

Data Mart Refresh

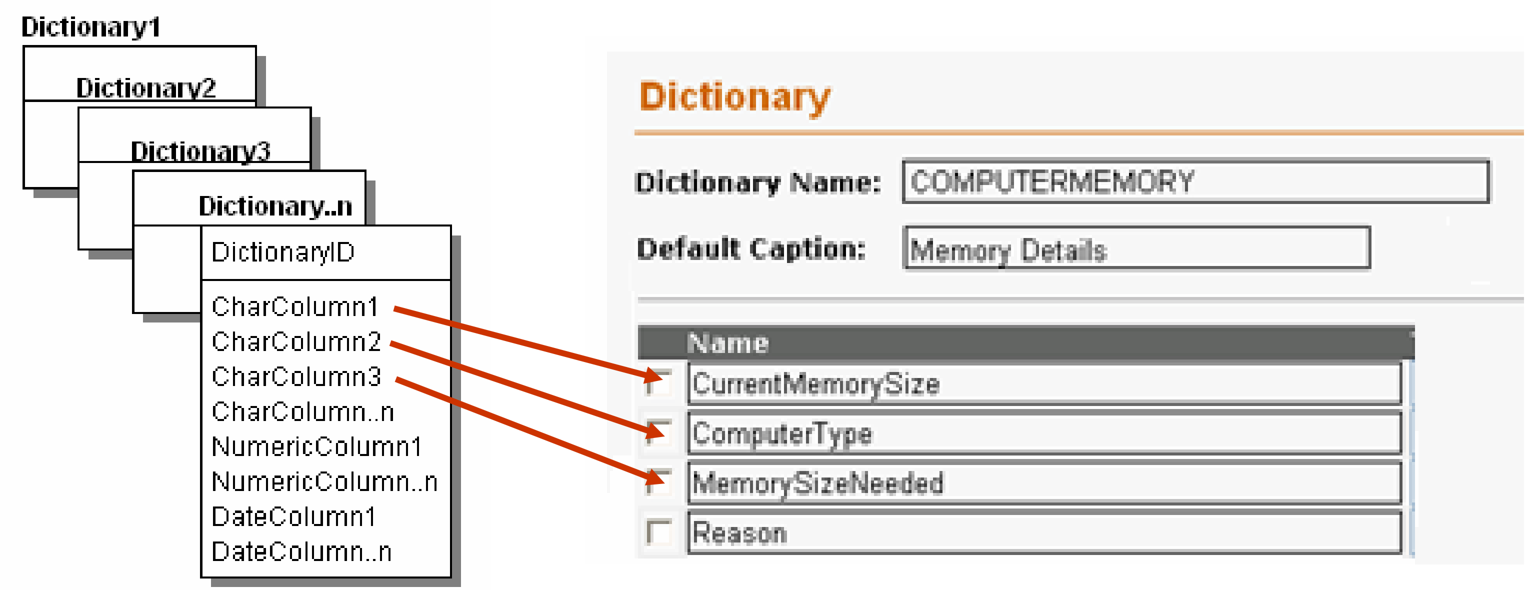
- Custom Java program is used in conjunction with Data Manager ETL



Data Mart Refresh ...

- ... and metadata creation
  - Dictionaries and their attributes must be mapped to tables/columns in the data mart
  - This mapping is used in the BI business view

Metadata for mapping tables to dictionaries



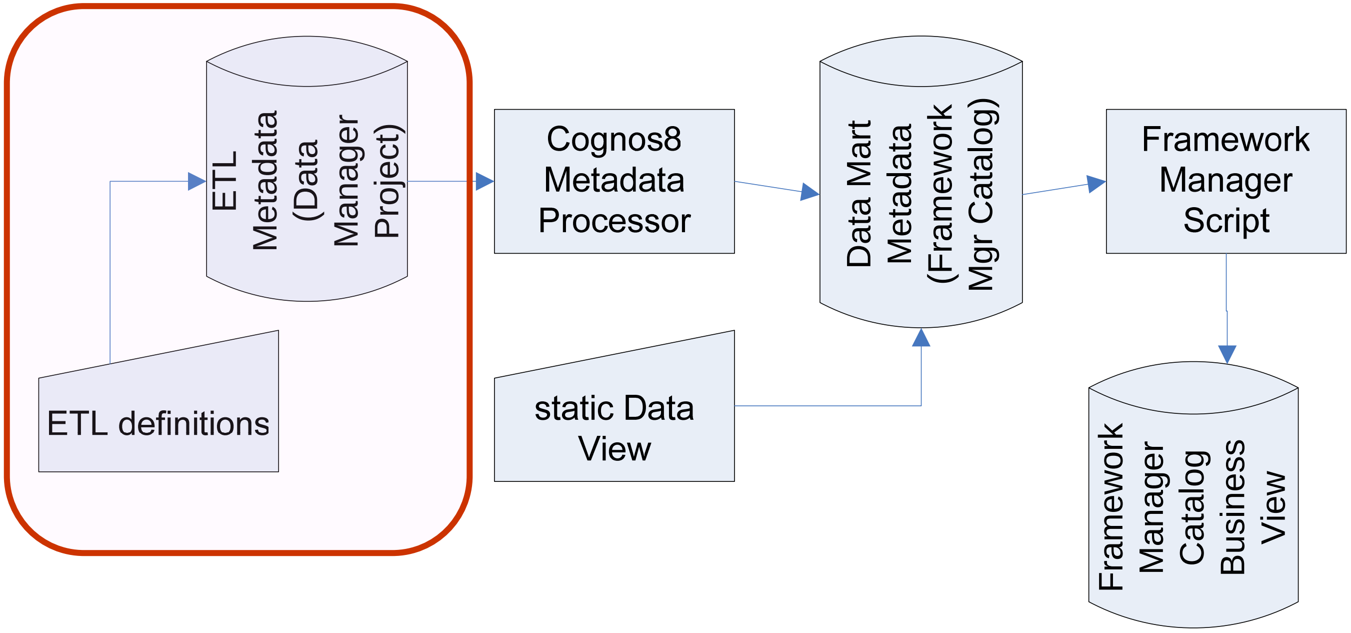
Dictionary

Dictionary Name:	COMPUTERMEMORY
Default Caption:	Memory Details

	Name
□	CurrentMemorySize
□	ComputerType
□	MemorySizeNeeded
□	Reason

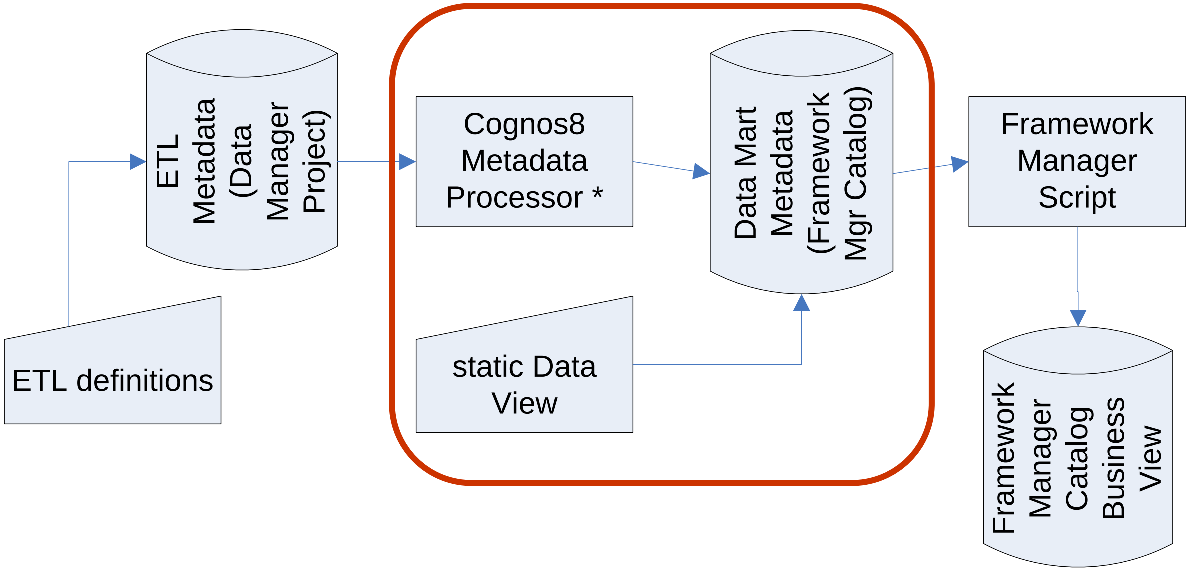
ETL / Metadata Process Flow

- Specify ETL for facts and traditional dimensions
  - Uses Cognos8 ETL tool, Data Manager


ETL Metadata (Data Manager Project)
ETL definitions
Cognos8 Metadata Processor
static Data View
Data Mart Metadata (Framework Mgr Catalog)
Framework Manager Script
Framework Manager Catalog Business View

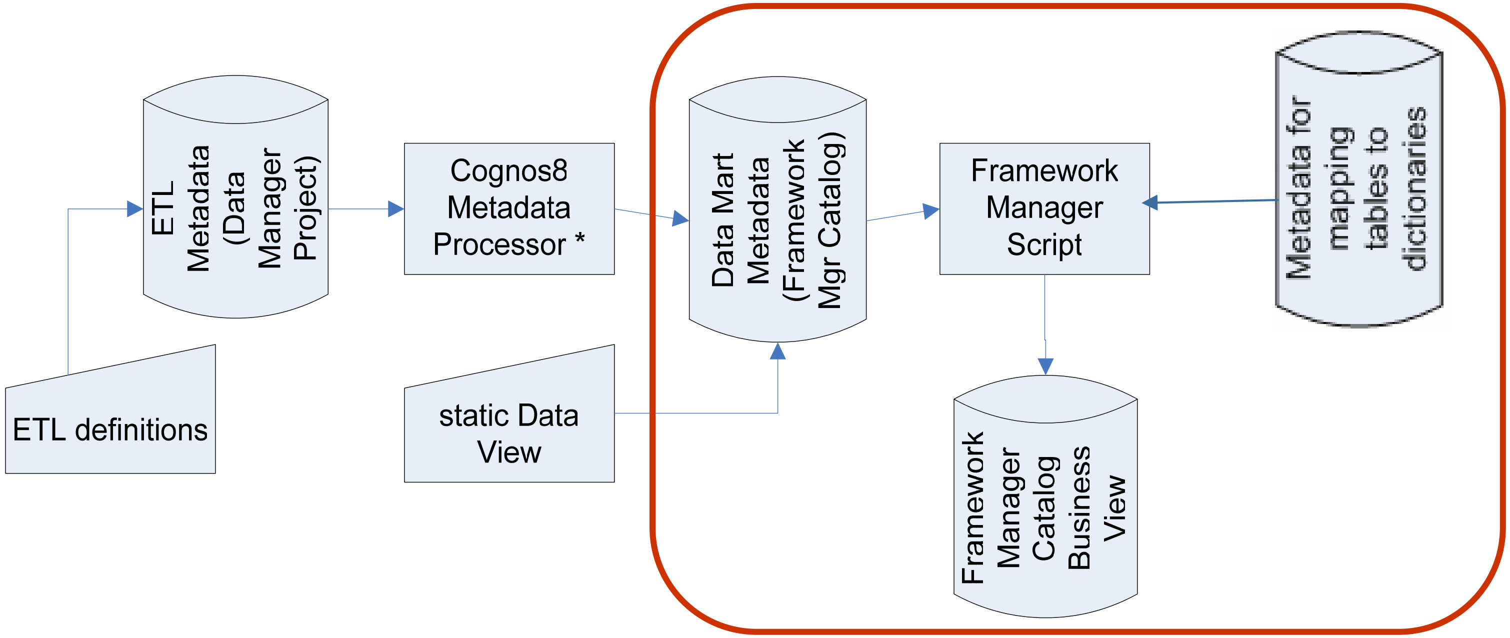
ETL / Metadata Process Flow

- Transform the Data Manager project into a Framework Manager catalog
  - (Sigh) two tools, two different sets of metadata
  - Data Manager’s ETL destination tables become the basis of the BI tool’s metadata


ETL Metadata (Data Manager Project)
Cognos8 Metadata Processor *
Data Mart Metadata (Framework Mgr Catalog)
Framework Manager Script
ETL definitions
static Data View
Framework Manager Catalog Business View

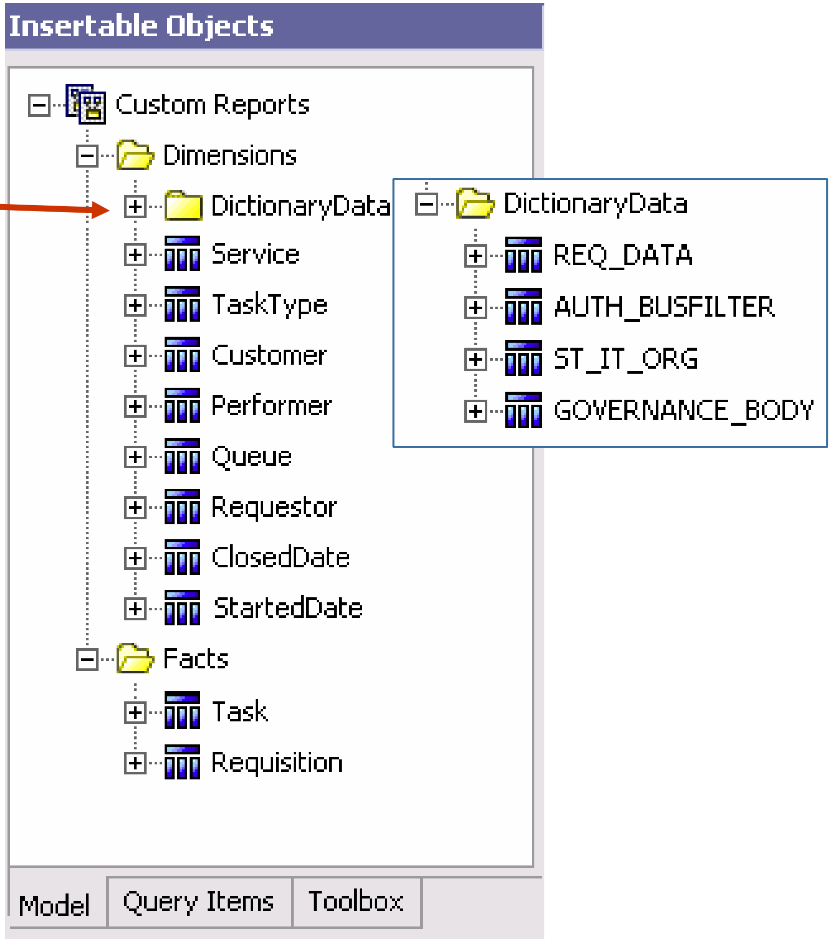
ETL / Metadata Process Flow

- For dictionaries (“junk” dimensions) use the Framework Manager API to create a business view
  - Appropriate names for dictionary dimensions and data elements


ETL Metadata (Data Manager Project)
Cognos8 Metadata Processor *
Data Mart Metadata (Framework Mgr Catalog)
Framework Manager Script
Metadata for mapping tables to dictionaries
ETL definitions
static Data View
Framework Manager Catalog Business View

End User Business View

- Dimensions
  - Dynamic dimensions
    - Based on dictionaries designated as reportable
    - Different for each site
  - Universal/static dimensions
    - Found at all sites

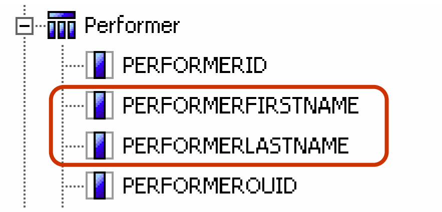


Insertable Objects

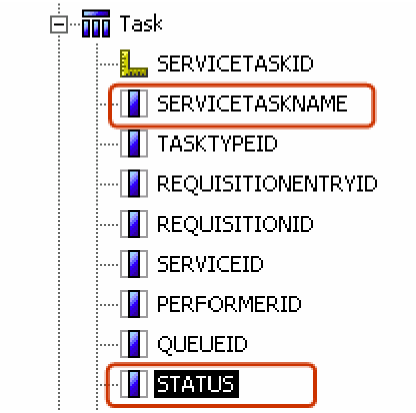
- Custom Reports
  - Dimensions
    - DictionaryData
      - REQ_DATA
      - AUTH_BUSFILTER
      - ST_IT_ORG
      - GOVERNANCE_BODY
    - Service
    - TaskType
    - Customer
    - Performer
    - Queue
    - Requestor
    - ClosedDate
    - StartedDate
  - Facts
    - Task
    - Requisition

ReportStudio

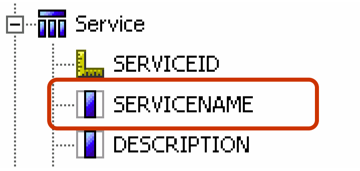
- Create a report listing tasks performed, the performer, the service in which they were performed, and the current status



Performer
- PERFORMERID
- PERFORMERFIRSTNAME
- PERFORMERLASTNAME
- PERFORMEROUID



Task
- SERVICETASKID
- SERVICETASKNAME
- TASKTYPEID
- REQUISITIONENTRYID
- REQUISITIONID
- SERVICEID
- PERFORMERID
- QUEUEID
- STATUS



Service
- SERVICEID
- SERVICENAME
- DESCRIPTION

Conclusions (1)

- XML extraction
  - Java programming was much more efficient than ETL tools available to us
- Dimensional modeling
  - Junk dimensions, reference dimensions, degenerate dimensions – who knew?
  - Subscribe to Kimball
  - Hire short-term consultants

Conclusions (2)

- ETL Performance
  - Satisfactory
  - Need sophisticated scheduler to track all jobs
- Ad-hoc Report Performance
  - Insufficient testing on the largest user data volumes (IMHO)
  - The software went live on January 31, 2007
  - Ask me in a few months

Update – Six Months In

- Performance issues
  - Reporting behavior is acceptable except for building filter selection lists on large fact table
  - Rework was required for customer load at client with largest customer base
- Database issues
  - Incompatibilities between Oracle SQL, SQLServer SQL, DB2 SQL, and Cognos SQL
  - Implement in database vs implement in tools

Update – Six Months In

- Functional/design issues
  - Should have gone with two people-based tables
    - Requestor/Initiator and Performer
  - Treatment of dimensions with all blank attributes needs more attention
  - Better use of framework to incorporate display definitions – better metadata required

About the Author

- Leslie Tierstein is a Senior Technical Architect at newScale, Inc, in Foster City California.
- She can be reached at ltierstein@earthlink.net or leslie.tierstein@newscale.com

- This paper is available on line at: http://home.earthlink.net/~ltierstein