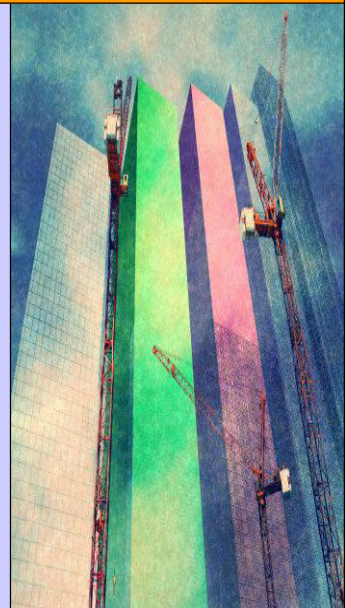# Active Warehouse

Proof of Concept in a W2K / WinXP Environment

DI Robert M. Bruckner
Vienna University of Technology
Institute of Software Technology
Information & Software Engineering
http://www.ifs.tuwien.ac.at
bruckner@ifs.tuwien.ac.at

May 2002

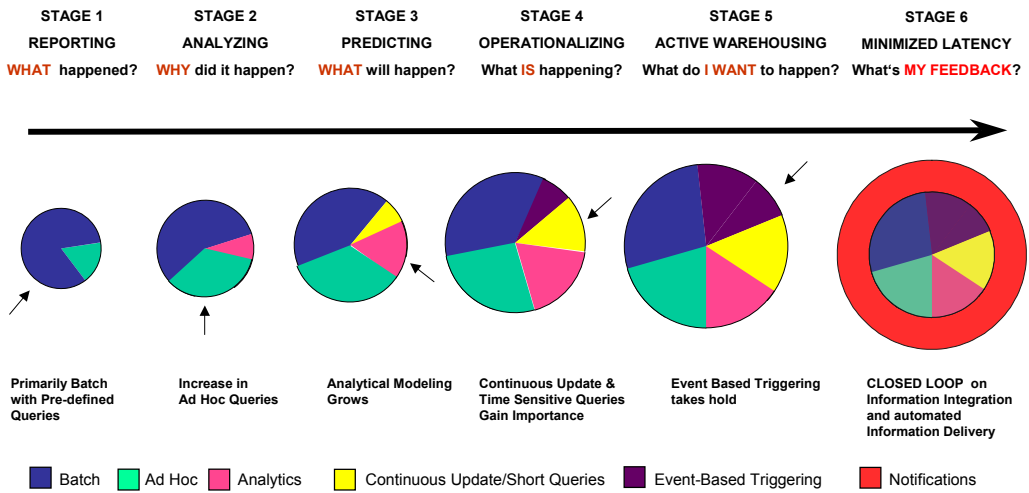# Agenda

- Active Warehouse
- Scientific Viewpoint: Minimized Latency in Data Warehouses
- Need for Messaging Infrastructure

- Active Warehouse: Architecture Overview
- Setting up MSMQ
- MSMQ in a W2K / WinXP Environment
- High-Performance Data Feeds for MSMQ
- TPump Job-Scheduling
- Demo

- Conclusion

IfS TU
technische universität wien

STAGE 1	STAGE 2	STAGE 3	STAGE 4	STAGE 5	STAGE 6
REPORTING	ANALYZING	PREDICTING	OPERATIONALIZING	ACTIVE WAREHOUSING	MINIMIZED LATENCY
WHAT happened?	WHY did it happen?	WHAT will happen?	What IS happening?	What do I WANT to happen?	What's MY FEEDBACK?

Primarily Batch
with Pre-defined
Queries

Increase in
Ad Hoc Queries

Analytical Modeling
Grows

Continuous Update &
Time Sensitive Queries
Gain Importance

Event Based Triggering
takes hold

CLOSED LOOP on
Information Integration
and automated
Information Delivery

■ Batch ■ Ad Hoc ■ Analytics ■ Continuous Update/Short Queries ■ Event-Based Triggering ■ Notifications

---

IfS TU
technische universität wien

▶ **Traditional Data Warehousing**
   – Focus on "ivory tower" decision makers
   – Long-term decision making
   – Strategic focus

▶ **Active Data Warehousing**
   – Expand scope to include "in the field" decision makers
   – Day-to-day (minute-to-minute) decision making
   – Tactical focus with strategic implications

**Business needs both strategic and
tactical decision support capabilities.**

- ► **Decrease the time** it takes to make the business decisions
- ► Minimize latency between the **cause and effect** of a business decision
- ► Notify the business of actionable **recommendations**
- ► Effectively **close the gap** between business intelligence systems and business processes

Analytical decisions integrated into operational processes combined with closed loop analytics.

- ► Support for various data freshness requirements
  - – Minimize propagation delays for particular data

- ► Continuous data integration
  - – Near real-time capturing and loading of data

- ► Timeliness of data / late-arriving data
  - – Represent the history as accurately as possible
  - – Deal with late-arriving data in tactical decision support

- ► Active decision engine
  - – Notifications, event / subscriber paradigm

- ► High availability, scalability, and performance

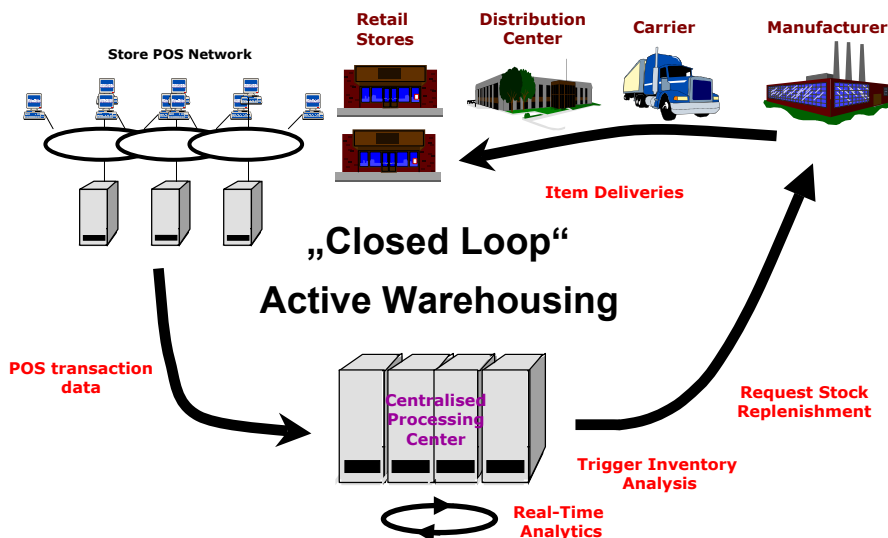# Need for a Messaging Infrastructure

- ▶ A DWH is an enterprise system
  - requiring data from transactional systems
  - delivering data (decisions) to enterprise applications
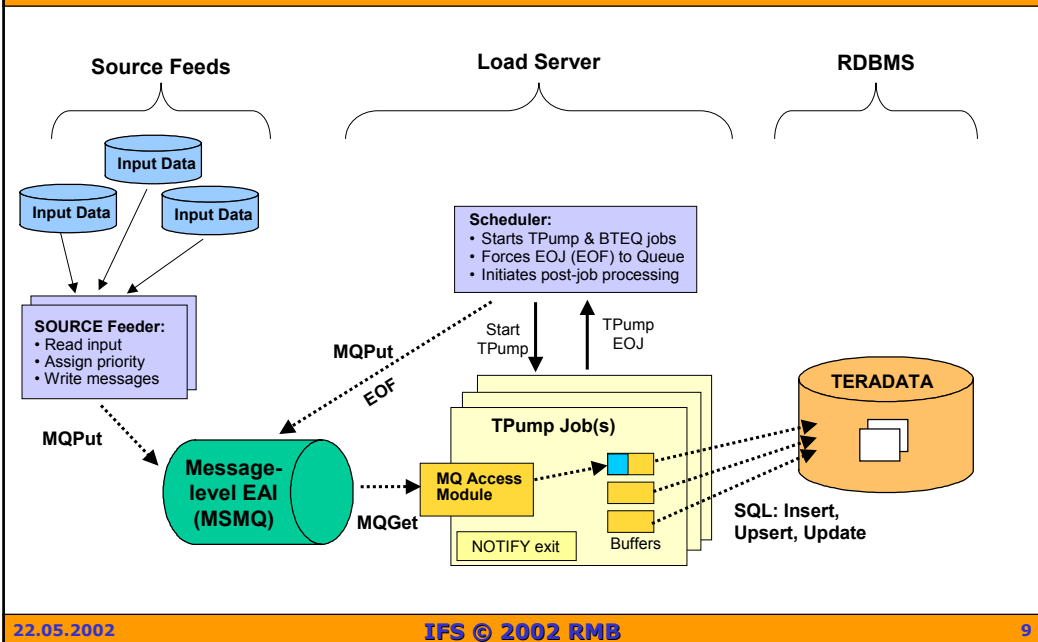- ▶ Need for a various levels of Enterprise Application Integration (EAI)

Continuous data integration ➔ message-level EAI:

- ▶ Asynchronous, connectionless service
  (vs. RPC or socket-based communication)
- ▶ Reliable delivery, transaction support if required
- ▶ Priority-based messaging
- ▶ Decoupling from transactional systems
  ➔ supporting dynamic data rates

Common options: IBM MQSeries®, Microsoft MQ

Active Warehouse: Application Szenario

Store POS Network

Retail Stores

Distribution Center

Carrier

Manufacturer

Item Deliveries

„Closed Loop"

Active Warehousing

POS transaction data

Centralised Processing Center

Request Stock Replenishment

Trigger Inventory Analysis

Real-Time Analytics

Source Feeds   **Load Server**   **RDBMS**

Input Data
Input Data   Input Data

SOURCE Feeder:
• Read input
• Assign priority
• Write messages

Scheduler:
• Starts TPump & BTEQ jobs
• Forces EOJ (EOF) to Queue
• Initiates post-job processing

MQPut

Start TPump

TPump EOJ

MQPut
EOF

MQPut

Message-level EAI (MSMQ)

MQGet

TPump Job(s)

MQ Access Module

NOTIFY exit

Buffers

TERADATA

SQL: Insert, Upsert, Update

MSMQ Features

- Integrated with
  – WinNT4: MSMQ 1.0
  – Win2K: MSMQ 2.0
  – WinXP: MSMQ 3.0
- Security, message persistence, transaction support

- **Public Queues** published through directory service
  – Win2K / MSMQ 2.0: Active Directory at domain controller
  – Win NT 4 / MSMQ 1.0: SQL Server 6.5
- **Private Queues** are not published
  – no directory service overhead

- More Details: http://www.microsoft.com/msmq

MSMQ Setup

▶ Re-install MS DTC (Distributed Transaction Controller)
  → run: `\winnt\system32\dtcsetup.exe`

▶ Install MSMQ
  → Settings → Control Panel → Add/Remove Programs
    → Add/Remove Windows Components
    → Install Message Queuing Services

▶ Optional: Configure MSMQ for **Workgroup Mode**

▶ Create message queues using the management console snap-in
  (run: `compmgmt.msc /s`)

MSMQ – Workgroup Mode

Workgroup Mode:
▶ Running MSMQ on Win2K/WinXP within a Windows domain
  but **not** on the domain controller
▶ Enables only private queues on the local machine
▶ No directory service overhead

Enable Workgroup Mode:
▶ Modify Windows registry
  – HKLM \ Software \ Microsoft \ MSMQ \ Parameters \ setup
  – Add DWord "AlwaysWithoutDS" = 1

MSMQ – Troubleshooting

- ► Event ID 2124: Message Queuing was unable to join the local Windows 2000 domain.
- ► Event ID 2121: Unable to complete Message Queuing Setup.
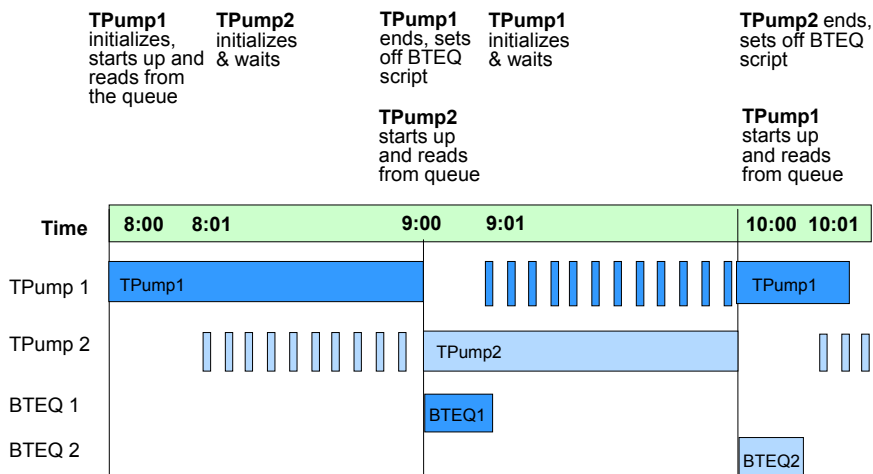- ► Hresult: c00e0075h

→ Re-install MS DTC

→ Configure MSMQ in Workgroup Mode

Role of Source Feeder

- ► Provide application data feeds
- ► Read transaction messages from a file
- ► Add a timestamp to the message
- ► Put message into the queue
- ► Arrival rate is adjustable

- Continuous feed
  - New or changed data
  - MPP Teradata loading
  - Many client platforms and sources (AXS-modules)
- Row level locking
  - Several TPump jobs can run against the same table at the same time from sources varied as Windows, MVS, or Unix
- Concurrent query access possible

- TPump is suitable, when some of the data needs to be updated closer to the time the event or the transaction took place

---

TPump Job-Scheduling

TPump1 initializes, starts up and reads from the queue

TPump2 initializes & waits

TPump1 ends, sets off BTEQ script

TPump2 starts up and reads from queue

TPump1 initializes & waits

TPump2 ends, sets off BTEQ script

TPump1 starts up and reads from queue

Time	8:00	8:01		9:00	9:01		10:00	10:01

TPump 1 — TPump1 ... TPump1

TPump 2 — TPump2 ...

BTEQ 1 — BTEQ1

BTEQ 2 — BTEQ2

## Loading additional POS data into a fact table

- ► Environment: Win2K / WinXP
- ► Teradata V2R4.x and Teradata Demo 4.x

- ► Source Feeder
  - – Source Feeder by Teradata CoE
  - – Java Source Feeder and Queue Administration Tool for MSMQ by IFS
- ► MSMQ Management
- ► TPump Job Scheduler

- ► Feeding into MSMQ or MQSeries queues from files
- ► Further information: vincent.hager@austria.ncr.com
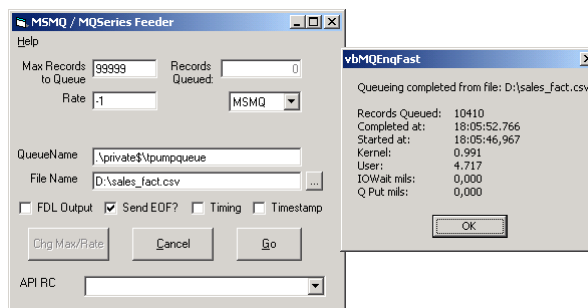
- ▶ **QTool 2.0:** Source Feeder, MSMQ Admin Tool, Scheduler
- ▶ About **3 times faster** than Source Feeder of Teradata CoE
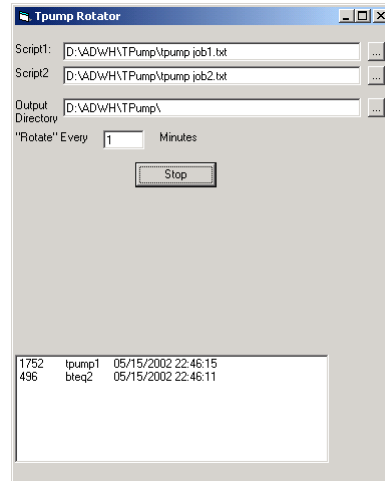- ▶ Further Information: http://www.ifs.tuwien.ac.at/~bruckner/teradata
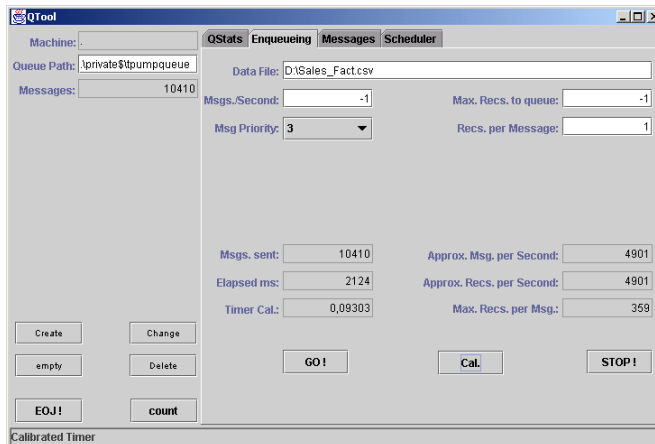
- ▶ Management Console Snap-In, run: `compmgmt.msc /s`
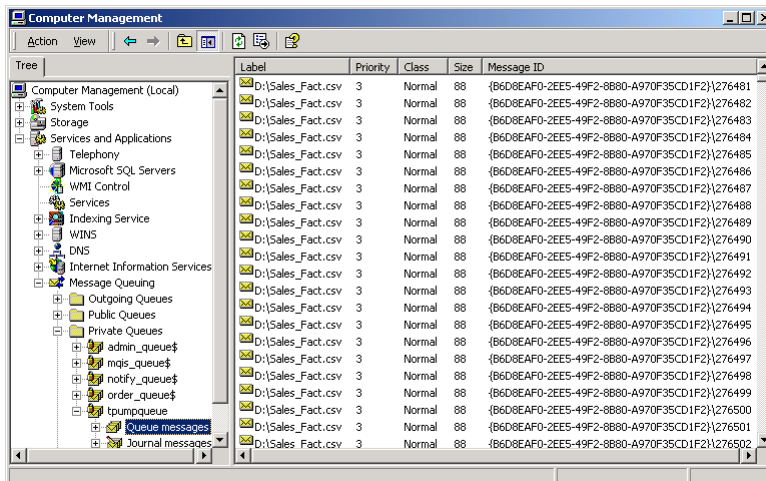
Basic scheduler:

▶ Start TPump job

▶ End TPump job by placing
an EOF message to the queue

▶ Launch post-job processing:

  – BTEQ script to consolidate
   error rows inside Teradata

▶ Monitor load process status/results

Tpump Rotator

Script1: D:\ADWH\TPump\tpump job1.txt

Script2: D:\ADWH\TPump\tpump job2.txt

Output Directory: D:\ADWH\TPump\

"Rotate" Every 1 Minutes

Stop

1752   tpump1   05/15/2002 22:46:15
496   bteq2   05/15/2002 22:46:11

▶ Analyze best practices for various data integration scenarios

▶ Sample data generation (BeDaWa) engine integration with
JMS data feeder to simulate various scenarios for concurrent
data update and user load (pattern analysis based on sample data)

▶ Vision: minimized latency data warehousing
  – service levels for particular source data
  – near real-time data integration
  – systematic temporal support
  – active information delivery based on notifications

Near real-time data integration works!

▶ Carefully analyze your requirements

▶ Teradata provides powerful integrated load utilities
(fastload, multiload, tpump)

▶ The NCR Active Warehouse is an important step
towards minimized latency in data warehouses

Thank You!
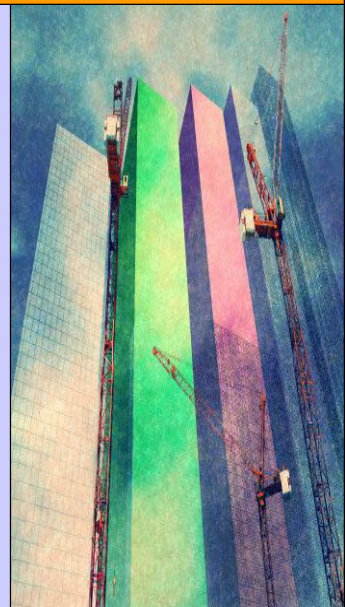
► Roland Braito (TU Vienna)

  QTool Development

► Teradata Division – Austria

  Cooperation and knowledge exchange

  Valuable discussion on various topics (e.g. EAI & DWH)

► Teradata CoE

Further information on:

► Testing the Active Warehouse using Teradata Demo 4.x

► MSMQ 2.0 Features

► MSMQ 3.0 Features

► Working with the ADW-sample provided by
  the Teradata CoE

- Teradata Demo 4.0: Win2K, Win NT4
- Teradata Demo 4.1: Win2K, WinXP (Patch required!)
  Download WinXP Patch:
  http://www.teradata.com/solutions/Files4XP.zip

- TPump Parameter: **sessions = 1**
- MSMQ must be configured in **Workgroup mode**
  on a standalone machine

MSMQ 2.0

IʃS TU
technische universität wien

- Win2K Professional, Win2K Servers
- Win2K security integration (Kerberos)
- Encryption: 40 bit, 128 bit
- Active Directory integration
  (Workgroup mode is possible but tricky to setup)
- Windows Cluster (active/active) support
- 2GB storage limit per machine

- MSMQ - MQSeries bridge is available
- Cross plattform support:
  MQC (Message Queuing Connectors) for Unix, CICS/MVS,
  VMS, AS/400

MSMQ 3.0

- ▶ WinXP Professional, WinXP Servers

- ▶ New Features
  - – Messaging based on HTTP / HTTPS
  - – SOAP extensions for reliable messaging (based on HTTP)
  - – Network load balancing / web-farm support
  - – Multicast messaging
  - – Message trigger concept (based on ECA rules)

- ▶ 1TB storage limit per machine
- ▶ Easier administration & deployment

Working with the ADW – Sample

- ▶ ADW – Sample is provided by Teradata CoE
  - – Source Feeder for MSMQ / MQSeries (VB 6.0)
  - – AXS-module for MSMQ (C++)
  - – TPump Scheduler (VB 6.0)

- ▶ Some minor issues:
  - – Hard-coded paths in source code
  - – No out-of-the-box demo included
  - – No sample TPump Job scripts included
  - – No AXS-module for MQSeries included (due to licensing)

- ▶ Further information: vincent.hager@austria.ncr.com