Preview of TPC-ETL: A Benchmark Under Development

TPC Technology Conference at VLDB

August 24, 2009

Presented by:

Jerrold Buggert (Unisys, Former TPC Chairman)

for

Len Wyatt (Microsoft), Brian Caufield (IBM), Daniel Pol (Hewlett-Packard)

Agenda

- Current state of ETL performance comparison
- Benchmark goals
- Business scenario
- Scoping the source, transformations and destination
- Open issues
- Summary

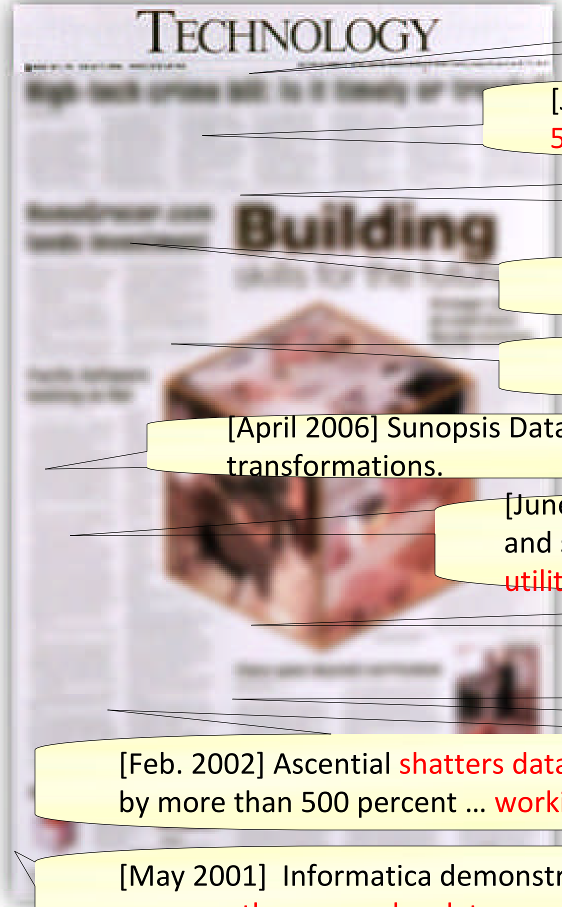
Background

- The TPC is a non-profit corporation founded in 1988 to define transaction processing benchmarks and to disseminate objective, verifiable performance data to the industry
- Current benchmarks include
  - TPC-C: Online transaction processing (distribution centers)
  - TPC-E: Online transaction processing (brokerage)
  - TPC-H: Decision support for ad hoc queries
- ETL Benchmark subcommittee formed in Nov. 2008
  - The following slides illustrate preliminary work and direction
  - The benchmark will evolve during development
- Related material is in the paper “Principles for an ETL Benchmark” to be presented at the TPC Technical Conference, in conjunction with VLDB 2009
- In some usages, “ETL” is different from “ELT”
  - That distinction is not relevant for purposes of this benchmark

Current state of ETL performance comparison

- There is no standard ETL benchmark
- Vendors publish one-off benchmark results
  - Many claim “World Record” performance
  - No way to compare the results
  - Some use TPC-H data as a common data set
- Some vendors and consultants argue for a realistic standard of comparison

Claims are out there…



[Dec. 2008] Syncsort and Vertica Shatter Database ETL World Record

[April 2007] SAS smashes ETL world record while establishing new, real-world benchmarks.

[Jan. 2007] Jaspersoft launches Jasperetl…Performance tests indicate performance up to 50% faster than other leading commercial ETL tools.

[Aug. 2006] Informatica sets world record data integration performance.

[May 2006] SAS, Sun Microsystems establish new data integration performance world record.

[Feb. 2008] ETL World Record … SSIS … Over 1 TB of TPC-H data was loaded in under 30 minutes.

[April 2006] Sunopsis Data Conductor demonstrates indisputable superiority for high volumes, complex transformations.

[June 2005] Informatica and Sun achieve record-setting results In data integration performance and scalability test … Data sets for the tests were generated by the industry-standard TPC-H utility dbgen.

[April 2005] Unisys and SAS deliver record-breaking ETL benchmark result.

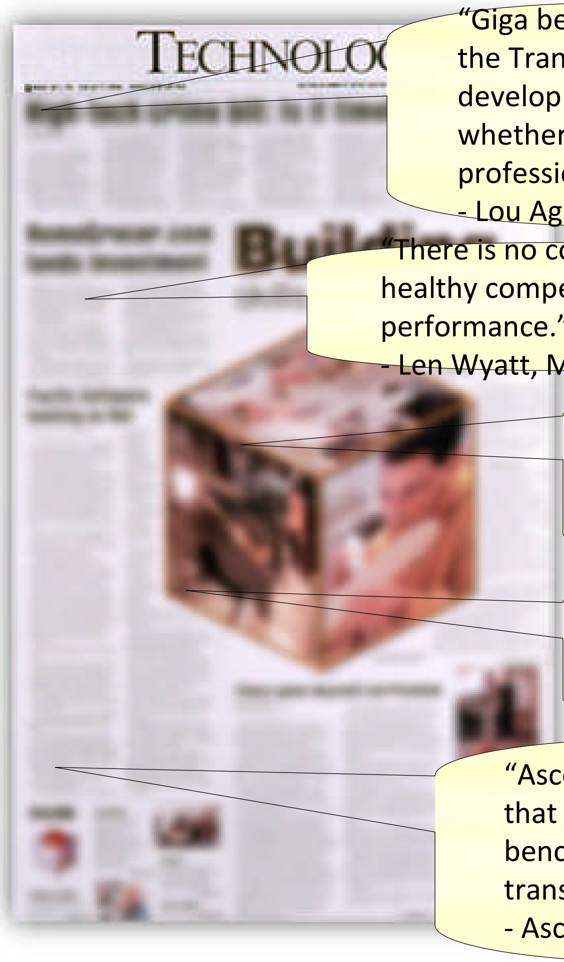
[March 2005] New release of SAS® Enterprise ETL Server sets performance world record.

[Feb. 2002] Ascential shatters data integration performance record; outperforms competitors' published benchmark results by more than 500 percent … working under the same parameters of prior benchmarks announced by competitors.

[May 2001] Informatica demonstrates massive scalability, powerful performance of its data integration platform on HP servers… three complex data mappings representing typical business scenarios.

- Market pressures demand the best claim that can be made
- Lack of standards allow almost any claim to be made

…and so are calls for a standard



"Giga believes industry-standard, independently audited benchmarks, such as those sponsored by the Transaction Processing Performance Council have value. Benchmarks, used properly, can drive development of new technology features that benefit real-world customers … The issue is whether the dynamics of unenlightened self-interest will be able to be contained by professionalism and objectivity."
- Lou Agosta, Forrester Research, March 2002

"There is no commonly accepted benchmark for ETL tools… Industry standard benchmarks can lead to healthy competition, better products, and better publication of the techniques used to get high performance."
- Len Wyatt, Microsoft, Feb. 2008

"Performance is one of the top concerns of people evaluating data integration technology, yet there is a surprising lack of information available. Most ETL benchmarks oversimplify the problems faced by users in order to show impressive numbers."
- Mark Madsen, Third Nature, April 2007

"As ETL performance becomes a higher priority for more enterprises, we applaud efforts to establish benchmarking methods."
- Doug Laney, vice president, Application Delivery Strategies, META Group, Feb. 2002

"Ascential has taken several important steps to create a new industry standard benchmark that better represents real-world customer data integration challenges. This real-world benchmark … will measure how fast data can be extracted from multiple disparate sources and transformed into a common format before being loaded into end user applications."
- Ascential P.R., Feb. 2002

ETL Benchmark: TPC's Unique Position*

OLTP Systems

TPC-C

TPC-E

ETL

TPC-ETL

Decision Support Systems

TPC-H

*Slide illustrates that the TPC is uniquely positioned to create a standardized ETL benchmark; not meant to imply that the benchmark will translate data from TPC-C/E to TPC-H

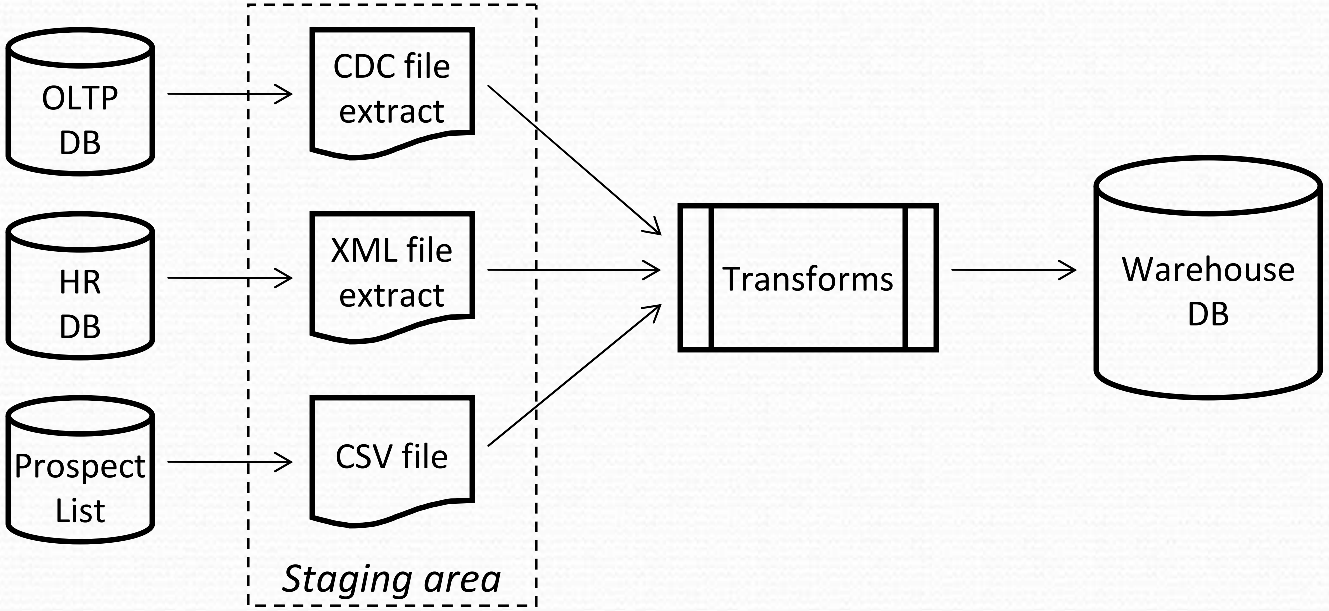
Note: This is a compelling enough idea that two new companies are joining the TPC explicitly to work on the ETL benchmark

TPC’s ETL benchmark goals

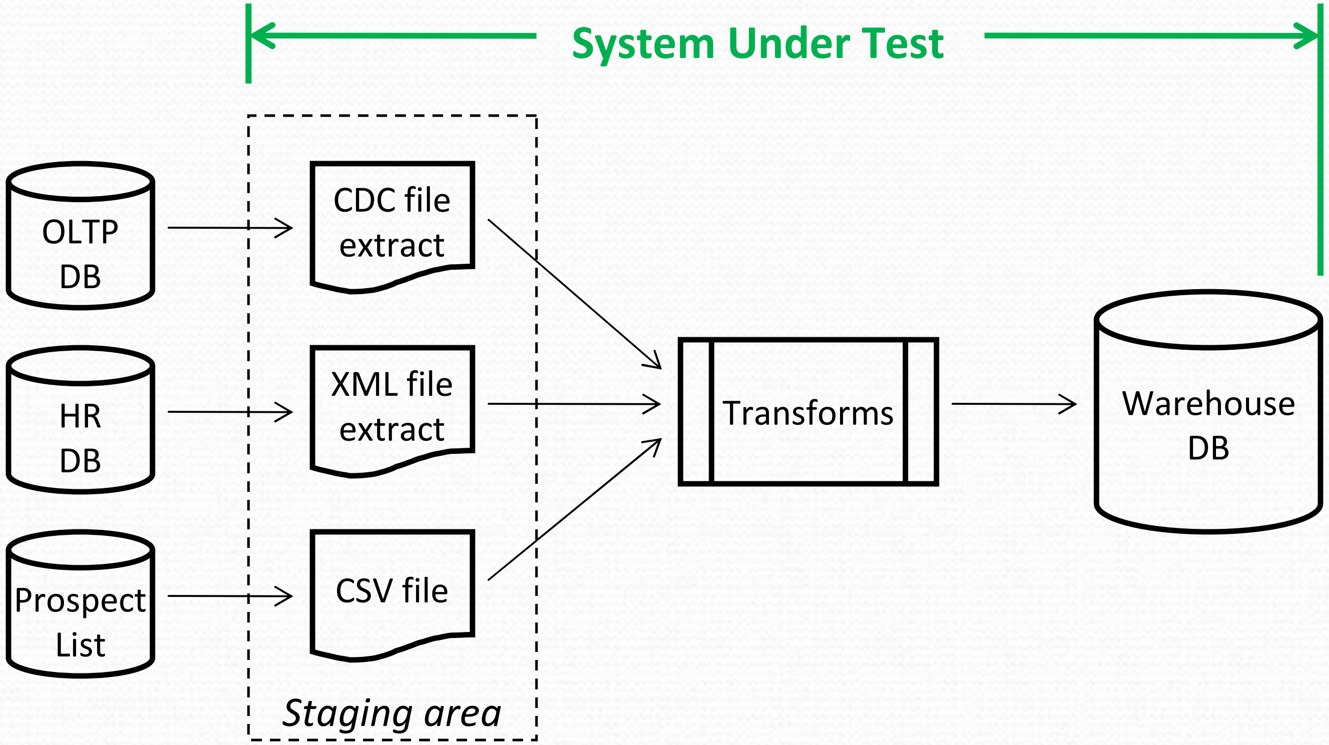
- Provide customers with reliable, comparable performance data based on a meaningful scenario
- Encourage vendors to create better performing systems
- Encourage disclosure of methods used to obtain performance
  - Customers should be able to replicate vendor results
- Expand the market through increased credibility of ETL tools due to:
  - An industry organization setting ground rules
  - A recognized system of measurements

Business scenario

- Provides a foundation for the benchmark requirements
- Loading the data warehouse of a brokerage company
  - Multiple data sources, various formats, various rules
  - Populating a dimensional model
  - Historical load and incremental updates



Scope



- ETL process reads from staging area
- Everything that directly influences the running of the ETL process is in scope of the benchmark

Scoping the source side

- Varied data sources
  - OLTP system with Changed Data Capture (CDC)
  - HR system extracts in XML format
  - Marketing data from external provider, as text files
- **The ETL job is modeled as starting from a staging area that holds file extracts from source systems**
  - Many real sites extract to a staging area first
    - For asynchrony, backup and/or auditing
  - Simplifies the benchmark environment
    - Reduces costs and outside influences

Scoping the destination

- There are interactions between the ETL process and the data warehouse. Examples include:
  - Lookups into dimension tables
  - Creation of summary tables
- Some ETL tools send data directly from the transformation process into the database, without an intermediate “stop” that would allow transformations to be measured separately from loading
- Some ETL tools (especially the “ELT” variety) run in the destination server and perform transformations using database functionality
- **Therefore the data warehouse is part of the benchmark SUT**

Scoping the transformations

- Transformations are defined by the source and destination schemas and business rules
- Transformation types expected:
  - Aggregations, lookups and joins, type conversions, complex data types, general data manipulations, data integrity checks & error handling, conditional processing, string operations, handling changing dimensions
- Major ETL scenarios
  - Historical load: The DW is initially created or re-created from historical records (e.g. when the DW schema is restructured)
  - Incremental update: Periodic load of new data into the existing DW
    - Considering modeling daily updates in the benchmark

Open issues

- What reliability characteristics should be required?
  - How should it be measured or validated?
- What should the benchmark metric(s) be?
  - Time-based? Workload-based?
  - Separate or combined metrics for historical load and incremental updates?
- How to scale the benchmark?

- These are all topics of discussion in the ETL subcommittee…

Summary

- The TPC is actively developing a benchmark to represent ETL technology in a meaningful way
- Initial scoping work has been done
- The ETL benchmark subcommittee is carrying out the work
  - Much work remains to deliver a final benchmark of the quality demanded by TPC and our customers
  - The TPC welcomes participation from those with an interest in the ETL space
- Two new member companies are joining the TPC explicitly to participate in the development of this benchmark. Others are encouraged to join as well.

Contacts and Information

- TPC ETL Chairman Len Wyatt
  - LenWy@Microsoft.com
- TPC Web Site
  - http://www.tpc.org
- Joining the TPC
  - Contact admin@tpc.org
  - Academic special offer
  - http://www.tpc.org/information/specialinvitation.asp