Shared File System Use in ETL and Data Warehousing

Dan Pollack

Principal System Administrator

America Online Inc.

dpollack@aol.net

Overview

- ETL / Data warehouse environment
- How data is collected and processed
- Where does file sharing fit in?
- What works and what doesn’t
- What’s next?

ETL / Data warehouse Environment

- Data gathered from many sources
- Transformation occurs on approximately 25 separate systems
- Data is loaded into about a dozen database systems
- About 400TB of raw storage, 500GB of RAM, and 300 CPUs across all transformation and database systems

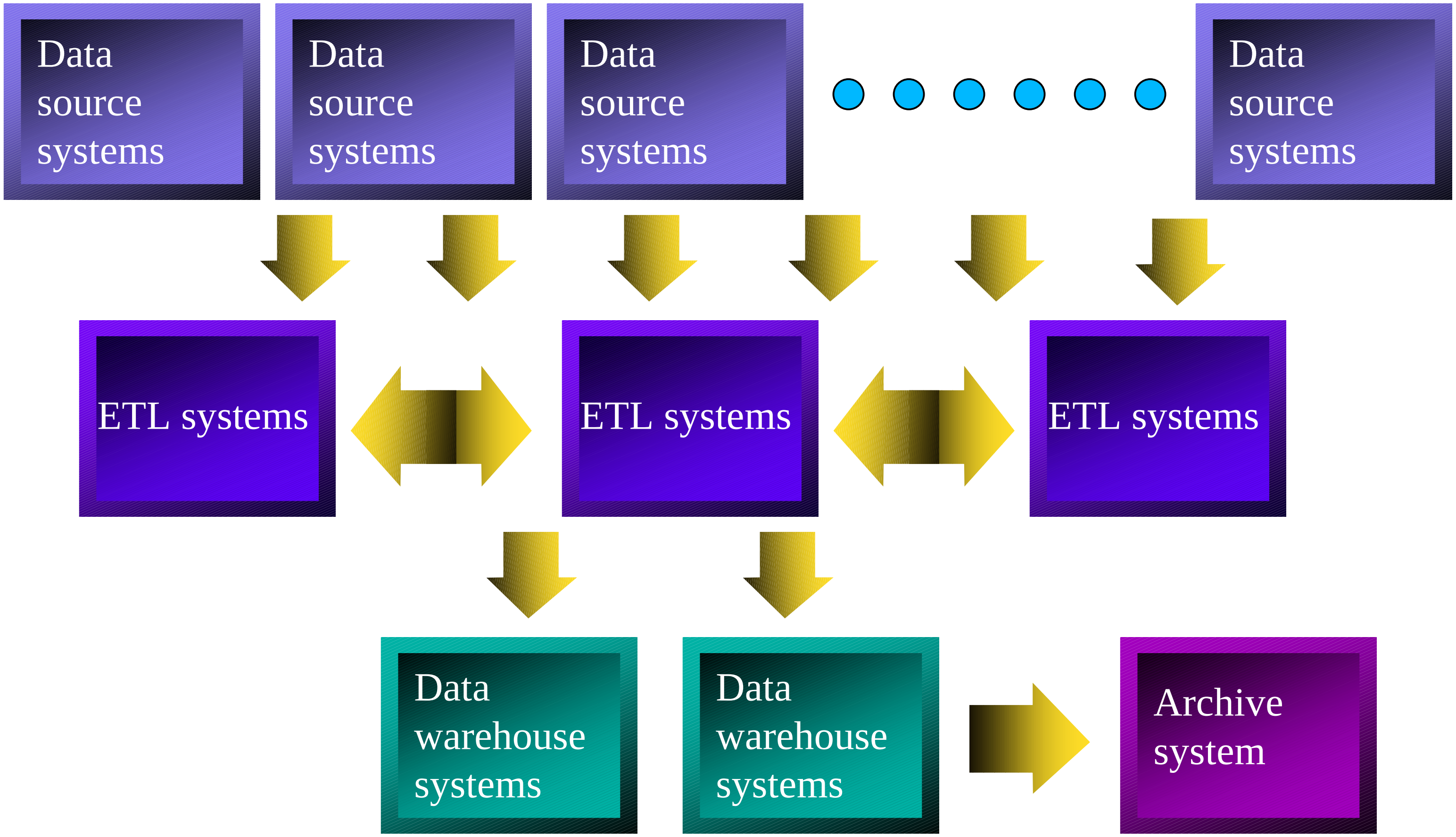
How data is collected

- Log files are transferred to transformation systems via IP data transfer methods (rcp, ftp, scp, etc)
- Central data collection file systems are used for each data set or group of related data sets
- Flag files are used to indicate complete data transfers

How data is processed

- Log files are read into transformation software tools
- Output of transformation process is written to data set output file system
- Data sets are transferred to database systems for load via IP data transfer methods
- Data sets are archived on a central system

ETL / Data warehouse data flow



Where does file sharing fit in?

- Data set collections
- Tools and configuration files
- Temporary work areas
- Output data sets
- Data archive

What works and what doesn’t

- Data set collections – No
  - Input data sets are typically used by only one process
  - Space contention for unrelated projects is a problem
- Tools and configuration files – Yes
  - Tools and configuration files are common across many projects
  - As number of systems increases data duplication gets out of hand

What works and what doesn’t

- Temporary work areas – No
  - Access contention during peak workload windows
  - No reuse of temporary files
- Output data sets – Yes
  - Multiple database systems use the same data set
  - Multiple data transfers are time consuming
- Data archive – Yes
  - Common area for all old data sets is convenient for access
  - Limited availability requirements

What's next?

- Improved availability of shared file system
- Improved access and transfer speed of shared file system
- Simplified shared file system configuration

Questions?