ETL & OLAP

MODULE 3

By: Manoj Kumar.

AGENDA

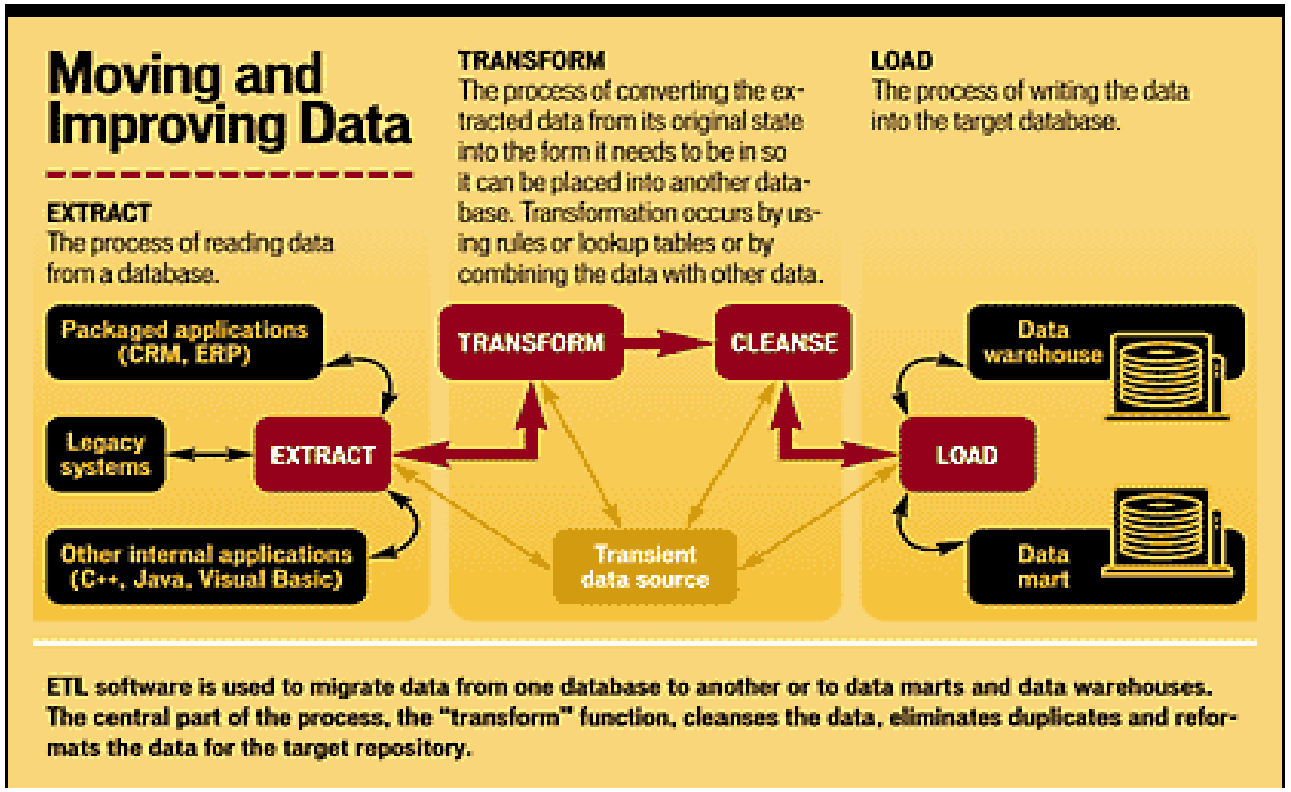
1. ETL Concepts
2. Data Extraction, Transformation & Load
3. OLTP & OLAP Concepts

1.ETL Concept

ETL DEFINITION

- ETL -Extract, Transform and Load

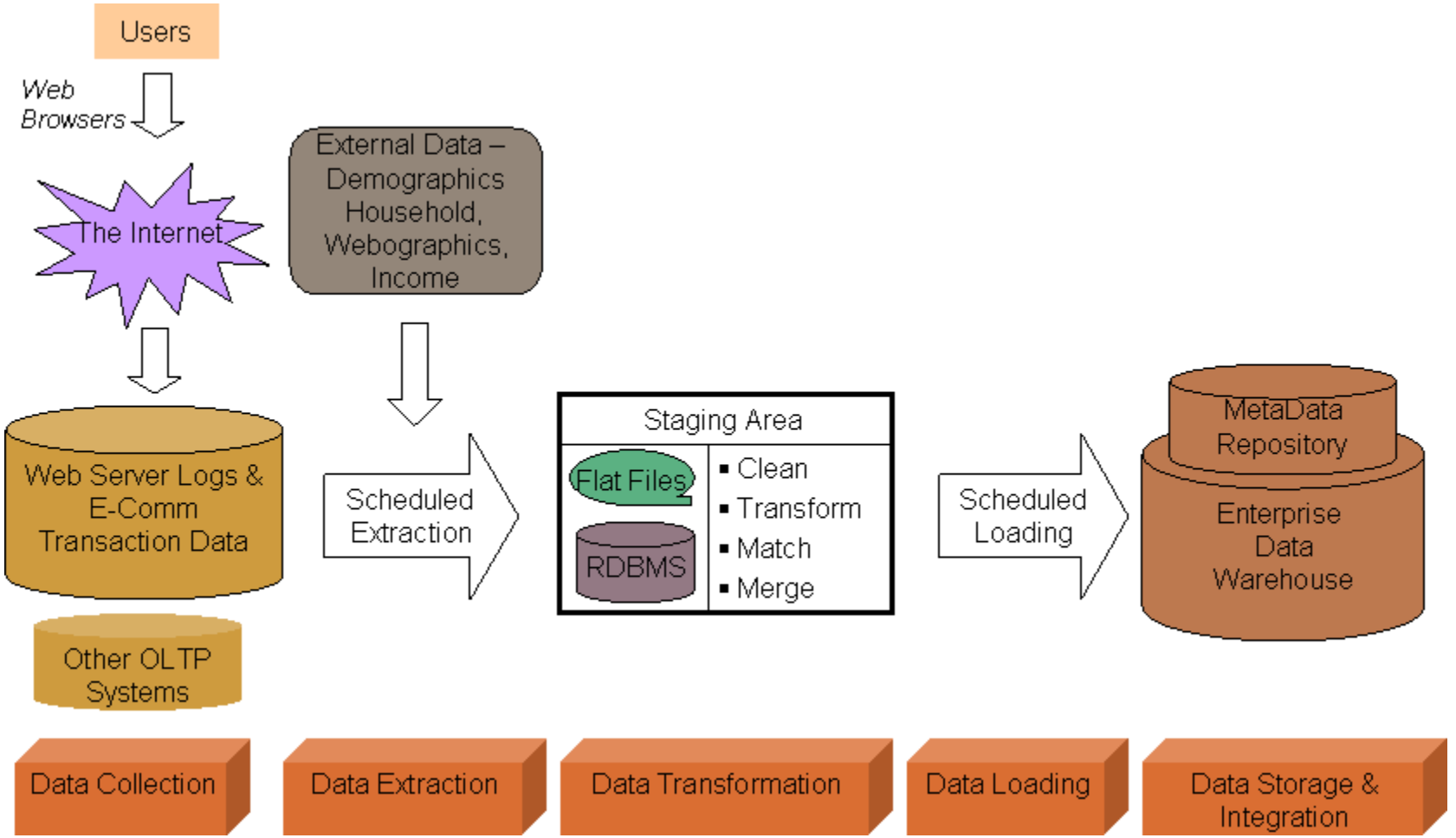
- ETL is the set of processes by which data is extracted from various sources transformed and loaded into target systems



IMPORTANCE OF ETL

- ETL technology is an important component of a complete enterprise data integration solution and the cornerstone of many strategic technology initiatives that benefit greatly from data profiling, data quality and metadata management functionality
- Support data extraction, cleansing, aggregation, reorganization, transformation, and load operations
- Generate and maintain centralized metadata
- Closely integrated with RDBMS's
- Filter data, convert codes, calculate derived values, map many source data fields to one target data field
- Automatic generation of data extract programs
- High speed loading of target data warehouses
- Employs Middle Ware for near Real Time ETL

ETL Framework



ETL ACTIVITIES

Data Extraction:

- Rummages through a file or
- database
- Uses some criteria for selection
- Identifies qualified data and
- Transports the data over onto
- another file or database

Data Extraction – Cleanup

- Restructuring of records or fields
- Removal of Operational-only data
- Supply of missing field values
- Data Integrity checks
- Data Consistency and Range checks, etc...

Data transformation

- Integrating dissimilar data types
- Changing codes
- Adding a time attribute
- Summarizing data
- Calculating derived values
- Renormalizing data

- Data loading
- Initial and incremental loading
- Updation of metadata

2. Data Extraction, Transformation & Load

DATA EXTRACTION OVERVIEW

- Extraction is the operation of extracting data from one or more source systems for further use in a data warehouse environment
- This is the first step of the ETL process
- Data can be extracted to a file or table as required
- After the extraction, this data can be transformed and loaded into the data warehouse.
- Internal Source Data: Extracted directly from the source system. Process can connect to the source system or to an intermediate system that stores the data in a preconfigured manner.
- External Source Data: Not extracted directly from the source system.
    - Sample source systems:
    - Mainframe Sources
    - Flat files: Fixed length, Delimited
    - XML Sources
    - Web log sources
    - ERP System Sources

DATA EXTRACTION METHODS

- Full Extraction
  - Data is extracted completely from the source system
  - No need to keep track of changes to the data source
- Initial Extraction
  - This will be the historical data staging extracts
  - Full Extraction from source
  - Static extracts from source table to staging tables using a tool or hand coding
- Incremental Extraction : Only Changed Data will be extracted
- Real Time Extraction
  - Captures event-driven data by online systems
  - The goal of real-time data extraction is to keep the warehouse refreshed, with minimal delay, as changes happen in the operational systems data
- Batch Extraction
  - A high-performance data extraction solution for extracting large volumes
  - It synchronizes data at a set time, perhaps once a day, or once a week

REAL TIME VS BATCH

Pros Quick and relatively easy to write scripts for doing exports and imports. Does not usually require Additional hardware.
•Almost all applications provide utilities for exporting and importing. Event driven, allowing for unified applications that support a business process that fits how people really want to work Systems can interact in a transactional manner providing true synergy.

Cons Not event driven--does not facilitate notification or change in another application at the time of a change in first application.
•Business processes that span applications are clumsy and do not behave as one usually requires additional hardware, software, and expertise . Some applications do not provide APIs or interfaces needed for real-time integration. Expensive when compared to batch process.

EXTRACTION TO-DO LIST

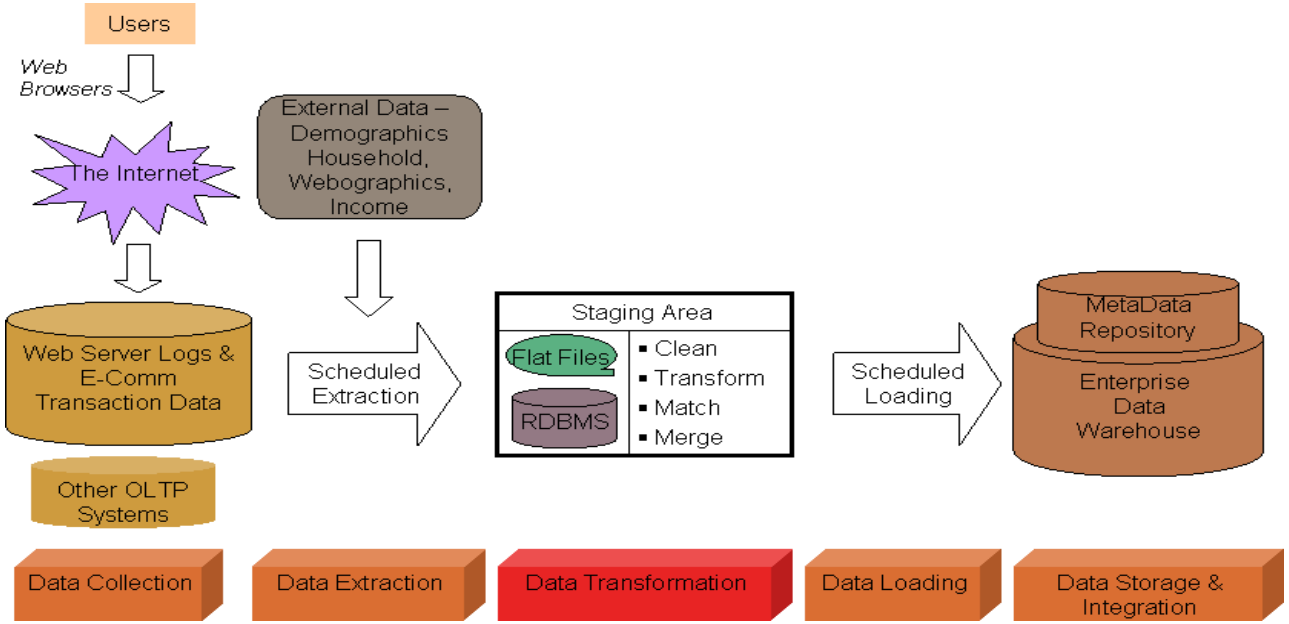
- Following steps should be complete before starting any physical ETL development
  - Logical data map provided by DW architect-describes the process end to end.
  - Identify data source candidates
  - Analyze source systems with a a data-profiling tool
    - Data discovery
    - Anomaly detection
  - Receive walk-through of logical plan and business rules
    - Business rules defined at the data modeling stage are different from those required by the ETL team.
  - Receive walk-through of data warehouse data model
  - Validate calculations and formulas

COMPARISON OF ETL TOOL – DATA EXTRACTION

Feature	Informatica PowerCenter	Ascential Data Stage XE	Ab Initio
Multiple Sources	For Extraction: DB/2, DB/2 /400,Flat Files, IMS, Informix, MS SQL Server, MS Access, Oracle, Sybase, UDB, VSAM, ODBC, Others.	QSAM: Sequential flat files ISAM: VSAM: KSDS, RSDS, ESDS - support GROUPS, multi-level arrays, REDEFINES, and all PICTURE clauses. DB2, Adabas, Oracle OCI (For releases 7 and 8) , Sybase Open Client , Informix CLI , OLE/DB for Microsoft SQL Server 7, ODBC.	Oracle, TeraData, DB2, Informix, SQL Server, XML files, Sybase, Flat Files, VSAM Cobol Files and others. Flat files can be imported as Target using Data set component.
Types of Extracts supported	Incremental loads, Event driven loads, Entire copy of source table (refresh) is supported within PowerCenter .	Following extract methods are possible •Incremental •Transaction Event •Full Ascential also provides Change Data capture for popular RDBMS.	
Code Generation	PowerCenter does not generate code,all the mappings developed will be inform of GUI interface.	Only Datastage XE/390 version automatically generates and optimizes native COBOL code and JCL scripts that run on the OS/390 mainframe.	It has GUI interface and also generates code of scripts using mp, which is shell based development environment for implementing applications using Ab initio
Real-Time Data Integration	Provides Real Time features using PowerCenter RT and EAI tools	Supports real time integration. DataStage can operate in real-time, capturing messages or extracting data at a moment's notice on the same platform that also integrates bulk data.	Has a product called Continuous Flows to implement real-time data access.
Change-Data-Capture Feature	No inbuilt feature to capture the changed records in the source systems. However this can be handled using custom code in the source qualifier. If the source is SAP then using IDOC's feature of the SAP it is possible to capture the changed records.	Ascential provides a number of ways to capture changed data.	Ab Initio supports a wide variety of change data capture strategies.

DATA TRANSFORMATION ACTIVITIES

- Integrating dissimilar data types (Lookups & Joins)
- Changing codes
- Adding a time attribute
- Summarizing data(Aggregation)
- Calculating derived values
- Renormalizing data



AGGREGATION TECHNIQUES

- Aggregation is a process in which information is gathered and expressed in a summary form, for analysis.

- Aggregates are used for two primary reasons
  - To save storage space
  - To improve the performance of business intelligence tools

Transformation:

1. Data cleansing.(INTCAP, UPPER, Standardization)
2. Data Merging. (Joins, Union, CONCAT)
3. Data Scrubbing.(REV:PRI*QUNT)
4. Data Aggregation.(SUM(),AVG(),MIN(), MAX(), COUNT())

Loading:

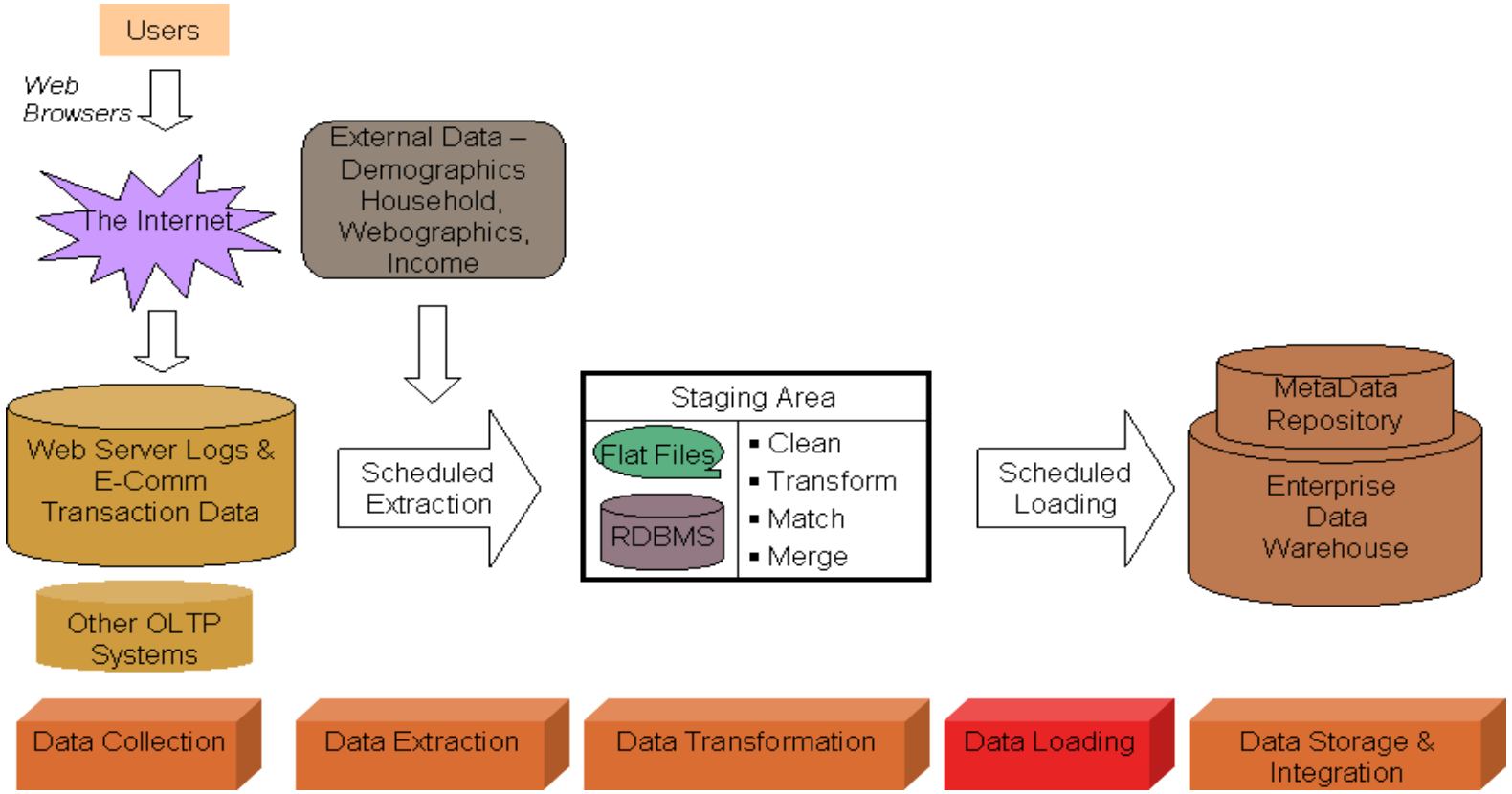
1. Initial Load. BULK LOAD.
2. Incremental Load. REAL TIME LOAD.

COMPARISON OF ETL TOOL – DATA TRANSFORMATION

Feature	Informatica PowerCenter	Ascential Data Stage XE	Ab Initio
Data Profiling Capability	PowerCenter 7's native data profiling can run in batch or interactive mode, and uses either predefined or customized sets of rules to analyze the data without having to extract it first.	ProfileStage provides data profiling capabilities.	NA
Code Re-usability	Good features to make the code re-useable. Supported through re-useable transformations, Mapplets and also some extent short cuts.	Code Reusability is supported. Ascential's Quality Manager provides a framework for developing a self-contained and reusable Project which consists of business rules, analysis results, measurements, history and reports about a particular source or target environment.	Transformations can be shared
Handling duplicate records	This needs to be handled by using Aggregator Transform in the Mapping.	It can match and deduplicate records character by character exactly or match and deduplicate even non-exact record matches (and provide a probable likelihood that two records match), in the absence of common keys.	Ab Initio provides fundamental building blocks for implementing de-duplication

DATA LOADING – INITIAL LOAD

- First load into the Data warehouse
- Very similar to a system conversion process
- Load the Live Operational Data



HISTORICAL LOAD

- Extension of the initial load process
- Loading the historical data (static data)
- Format of Archived data different from operational data
- Different set of conversion scripts needed
- Lengthy and Complex process

BULK LOAD

- Inserts a large amount of data to the target database

- It improves the performance of a session.

- If there is no changes occur in records then for better performance use bulk load instead of record by record insertion.

- It limits your ability to recover because no database logging occurs

- Normally happens for Historical load and Initial load

ETL PROCESS FLOW

A typical ETL process has the following job flow:

- Extract dimensions and write out metadata
- Extract facts and write out metadata
- Process dimensions
  - Surrogate key/slowly changing process/key lookup etc
  - Data quality checks – write out metadata
- Process facts
- Process aggregates
- Load dimensions before the fact table to enforce referential integrity
- Load facts
- Load aggregates

DIFFERENT SCHEDULERS AVAILABLE

- Tool Specific
  - Very much specific to tools
  - Difficult to manage many tool specific schedulers in an enterprise
  - Not efficient to handle complex job dependencies
- Autosys
  - Scheduling solution for any enterprise
  - One scheduler can manage all the different tool specific schedules
  - Handles complex job dependencies
  - Superior Job Recovery
- CRON jobs
  - Traditional CRON jobs are not GUI based
  - Works efficiently in Unix
  - Excellent Error Tracking/Management

COMPARISON OF ETL TOOL- DATA LOAD

Feature	Informatica PowerCenter	Ascential Data Stage XE	Abinitio GDE
Loading Capabilities	Better	Good	API based utilities and Direct loading capabilities available (e.g: SQL loader in Oracle can be initiated from Abinitio, MLoad utilities can be initiated from Abinitio)
Loading to various platforms	Supports: Sun Solaris, AIX, HP-UNIX, Windows 2000, Compaq Tru64	Windows NT (Intel and Alpha Platforms), UNIX AIX, HP-UX, Sun Solaris, COMPAQ Tru64, Linux. Data Stage XE 390 works on OS/390 platform.	Ability to load data into all OS.
Loading many to	Supports	Supports	Supports

COMPARISON OF ETL TOOL- DATA LOAD (CONTD.).

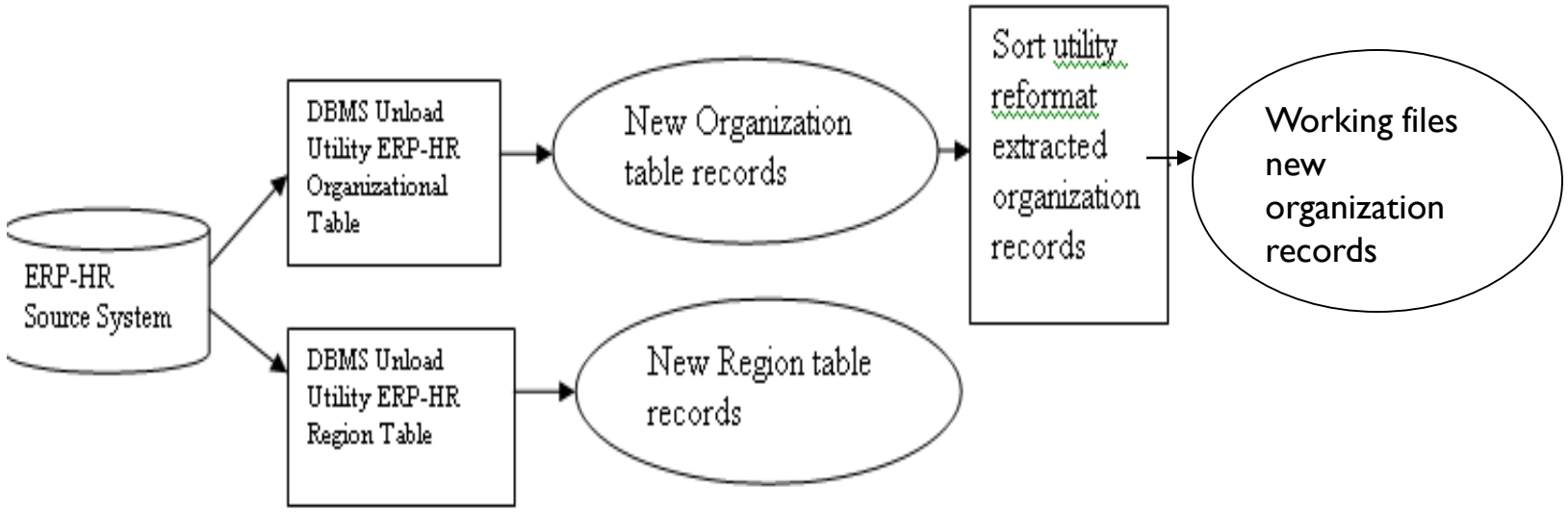
Feature	Informatica PowerCenter	Ascential Data Stage XE	Abinitio GDE
Loading one to many table	Supports	Supports	Supports
Automation of loading	It supports some extent, but not all the features. So third-party tools needed for other requirements.	Supports	Supports
Scheduling feature	•Mappings are run using the Workflow manager of PowerCenter. The Mappings are wrapped with Sessions.	•The included Job Sequencer handles complex job relationships including parallel and sequential execution, event occurrences and multiple conditions	•Inbuilt Scheduler for Abinitio to be purchased exclusively.

ETL PROCESS EXAMPLE

- ETL Process involves 5 stages which provide a modular and adjustable transformation process for the target table that can adapt easily to changes in the source systems or the warehouse model designer
  - Stage 1. Source verification: performs the access and extraction of data from the source system and builds a temporal view of the data at the time of extraction
  - Stage 2. Source alteration: perform a variety of transformations unique to the source, depending on business requirements
  - Stage 3. Common interchange: applies business rules and/or transformation logic that is frequent across multiple target tables
  - Stage 4. Target load determination: performs final formatting of data to produce load-ready files for the target table; identifies and segregates rows to be inserted vs. updated (if applicable); applies remaining technical meta data tagging; and processes data into the RDBMS
  - Stage 5. Aggregation final stage, uses the load- ready files from Stage 4 to build aggregation tables needed to improve query performance against the warehouse

ETL PROCESS EXAMPLE

- Stage 1. Source Verification
  source system is a human resources (HR) ERP system
  target is an organization dimension table that happens to use type 2 slowly changing dimensions



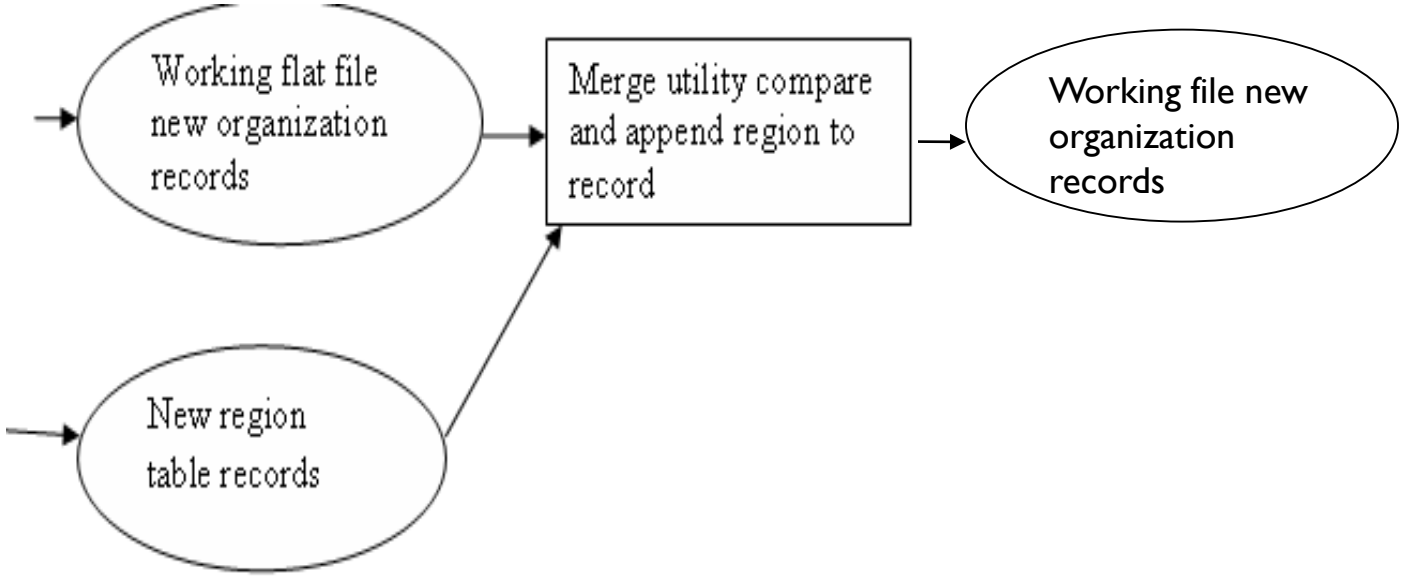
Source verification, alteration, and common interchange stages.

ETL PROCESS EXAMPLE (CONTD.).

- Stage 2. Source Alteration

  We append data from secondary sources.

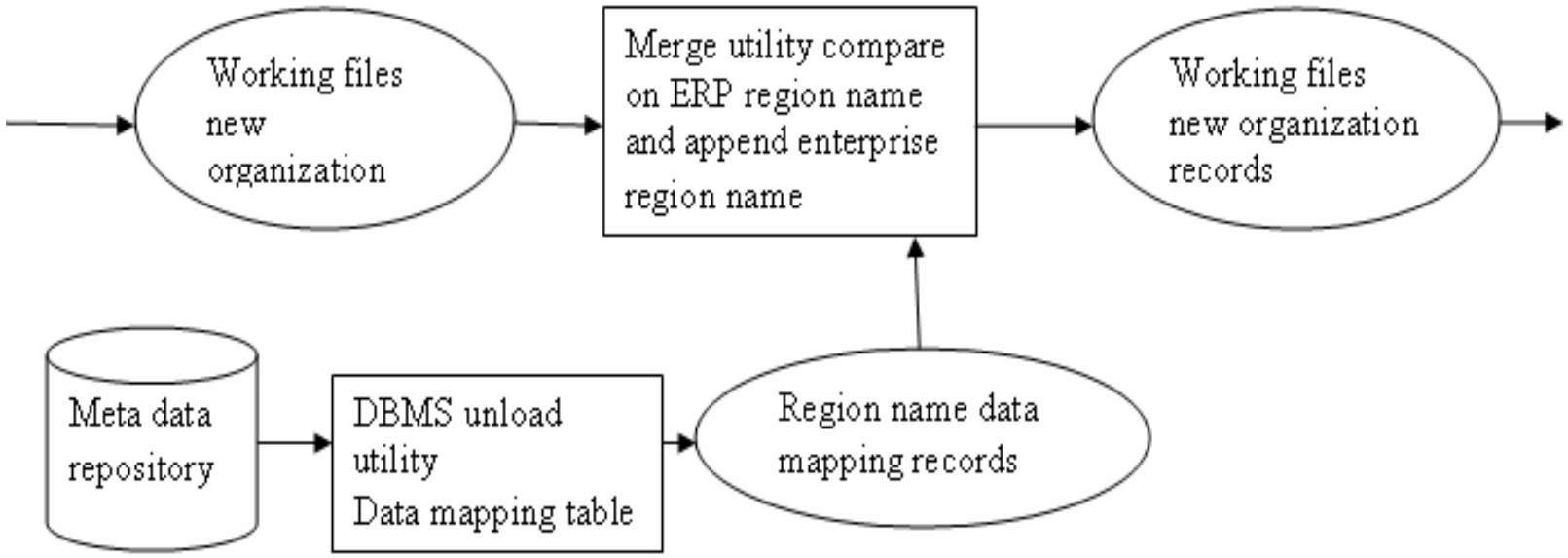
  In this case the HR ERP Region table — to the primary organizational extract file



ETL PROCESS EXAMPLE (CONTD.).

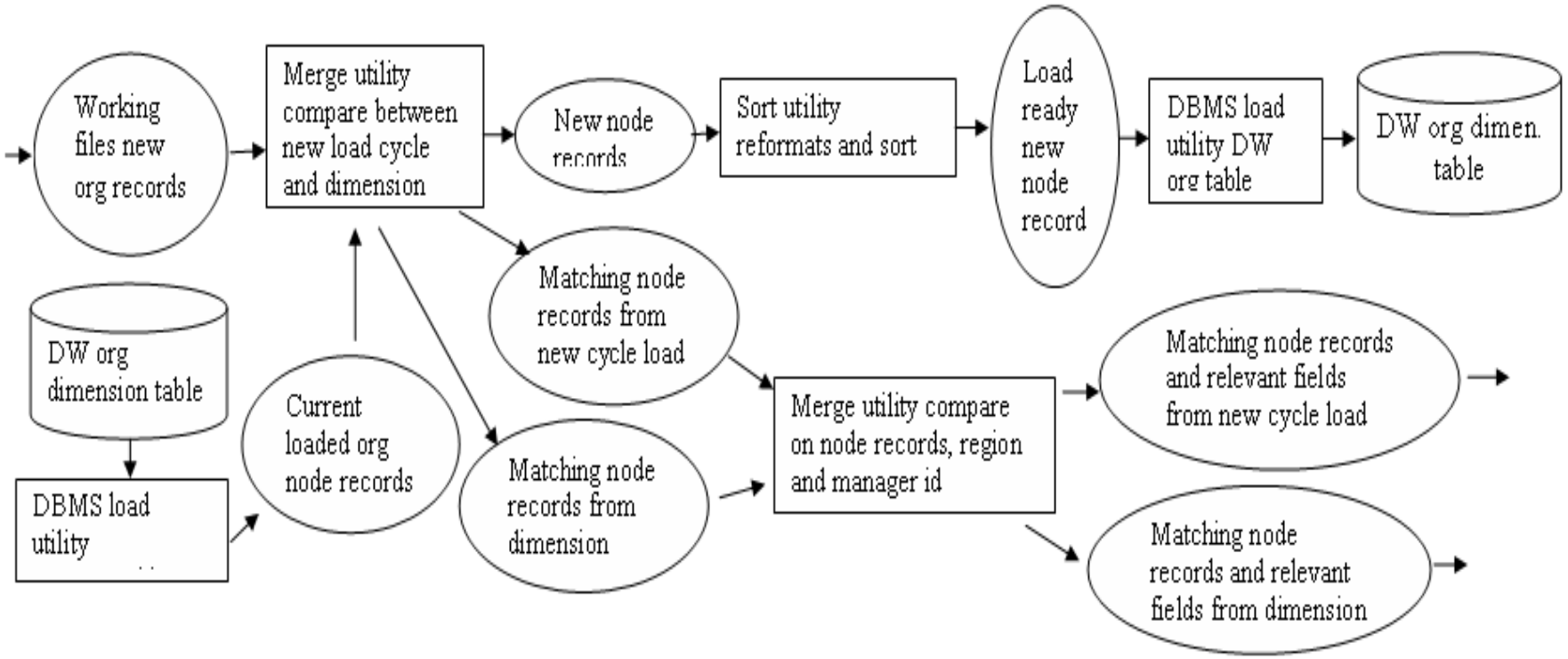
- Stage 3. Common Interchange

we find that the region name values stored in the HR ERP system do not conform to the established enterprise definitions thus we need to use the merge infrastructure utility to update the organization record region names to reflect the enterprise versions …….



ETL PROCESS EXAMPLE (CONTD.).

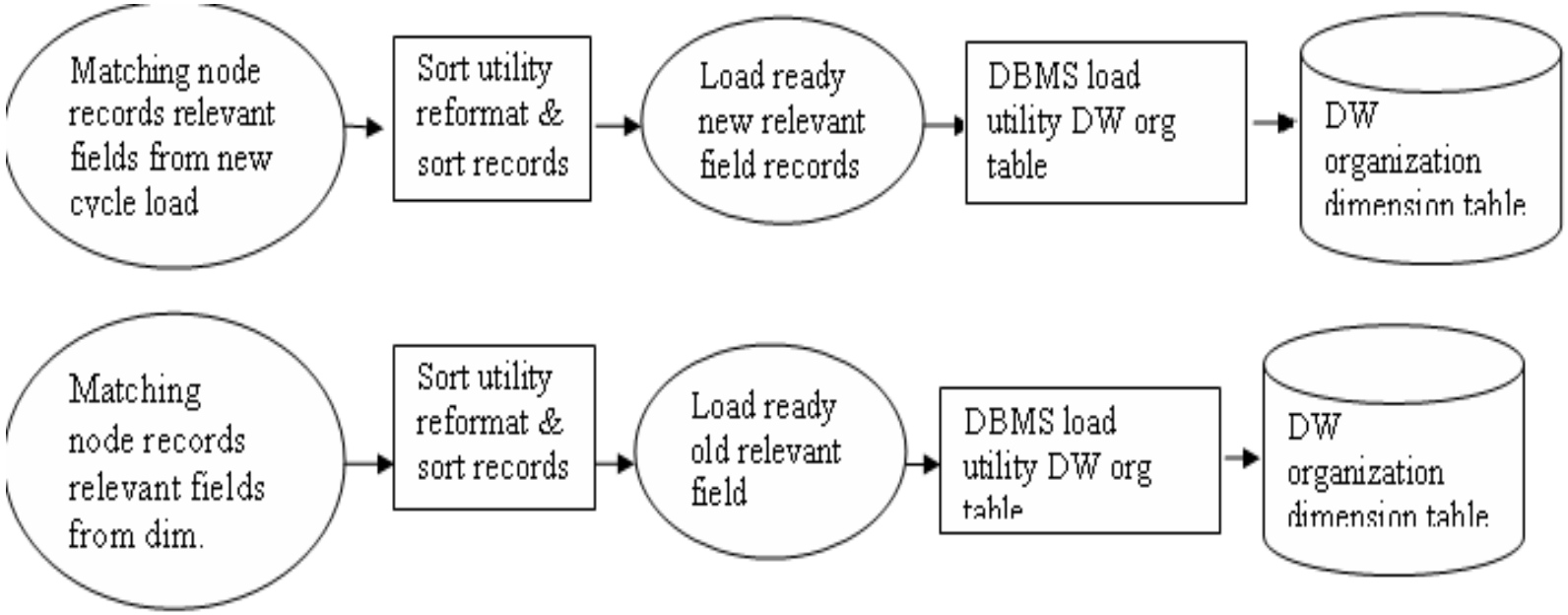
- Stage 4. Target Load Determination
  we compare the current load of organization records against those previously loaded in earlier batch cycles



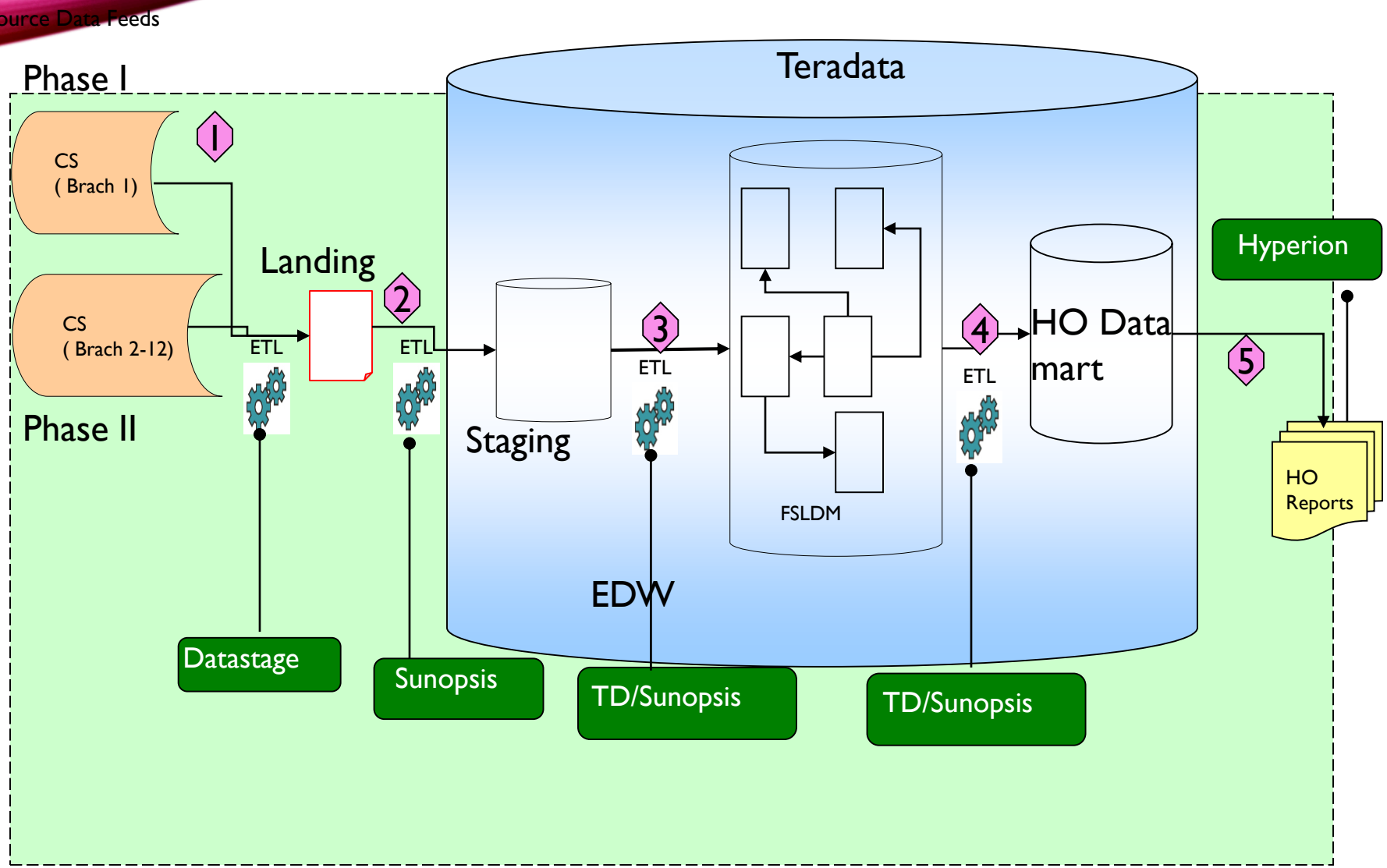
ETL PROCESS EXAMPLE (CONTD.).

- Stage 5. Aggregation

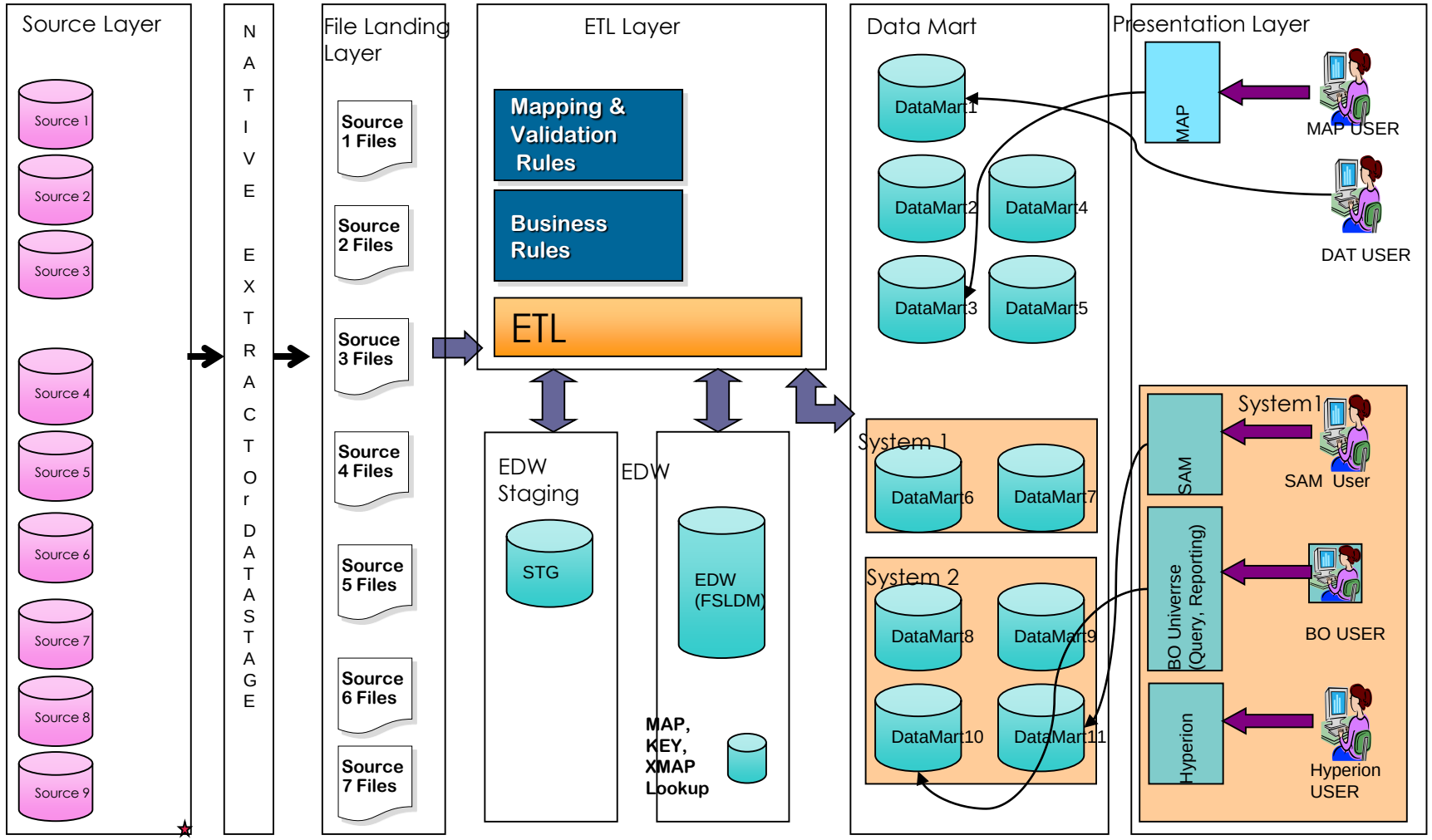
  we flag new rows for insertion: current load-cycle records that have relevant columns that do not match their corresponding organization dimension table rows, new region names, or manager IDs



SOURCE TO EDW EXAMPLE



SOURCE TO EDW EXAMPLE30



OLTP & OLAP Concepts

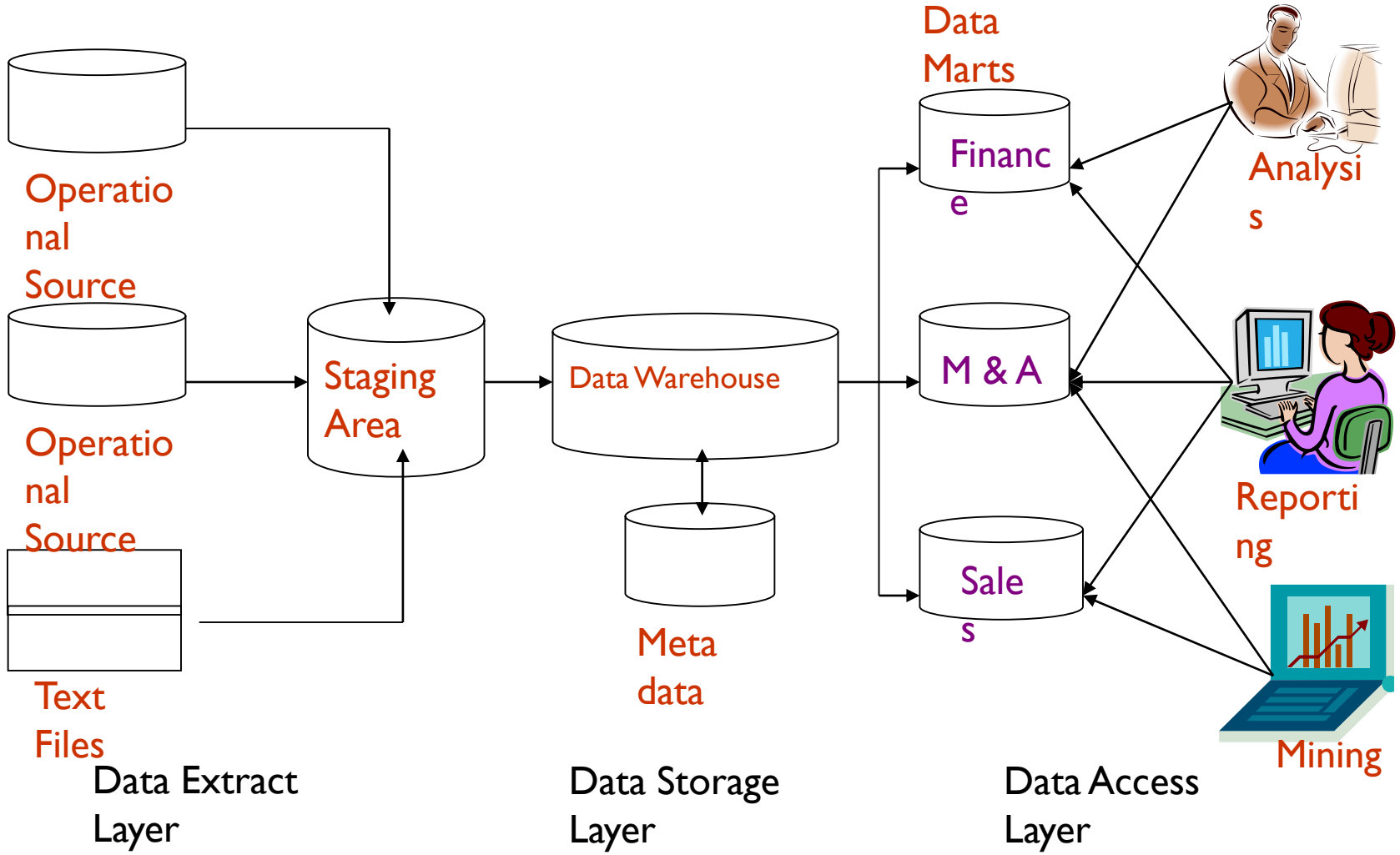
INTRODUCTION TO BIDW ARCHITECTURE

A typical data warehouse consists of …

- Source data
- Staging area
- Warehouse
- Data Marts
- Reports
- Analytical environment

A TYPICAL DATA WAREHOUSE ARCHITECTURE

A typical data warehouse architecture looks like ..



TRADITIONAL DECISION MAKING & ITS LIMITATION

- Spreadsheets and SQL are traditionally used as tool for analysis and decision making

- Limitations of Traditional Techniques
  - It is very difficult to define the aggregation levels, views in spreadsheets
  - SQL does not have a natural way of providing flexible view reorganizations that will transpose the data
  - Common analytic functions such as cumulative average and total are not supported in SQL
  - Extensive programming
  - Redundant reporting

DATA ACCESS OVERVIEW AND ANALYSIS

- It is the process of timely access and analysis of data
- It is the means by which the End Users ‘see’ the data warehouse or the ODS or the Operational Systems

DATA ACCESS TOOLS

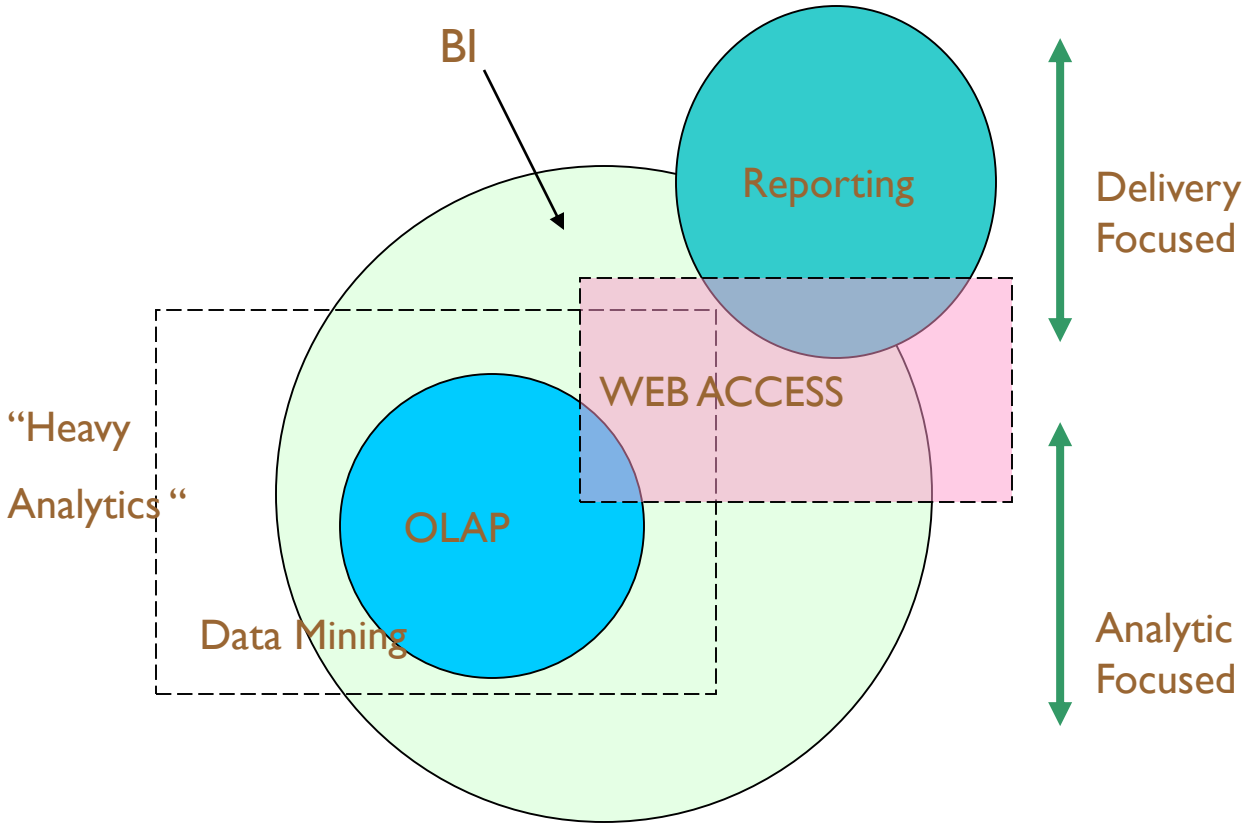
- A Data Access Tool
  - Encapsulates technical complexity like physical structure of the database from the users
  - Facilitates easy and controlled access to authorized users
  - Provides helpful metadata to the users
  - Provides a variety of features for analyzing the data

REQUIREMENTS OF DATA ACCESS TOOLS

- The requirements of a data access tool include
  - High performance and scalability to many users
  - Ease of use - intuitive, consistent interface to all functions of tool
  - Short learning curve
  - Semantic layer that maps business terms to database schema
  - Rapid slicing across dimensions and attributes
  - Tight integration with central metadata repository
  - Scheduling of reports should be possible

DATA ACCESS AND ANALYSIS - CATEGORIES

- Different categories of Data Access ..
  - Reporting
  - OLAP
  - Data Mining
  - Web Access



DATA ACCESS AND ANALYSIS – CATEGORIES (CONTD.).

- Reporting
  - A category of data access solution in which the information is presented in the form of reports
  - Reporting tools are also referred as Query and reporting tools

- OLAP (On-Line Analytic Processing)
  - Defined as "Fast Analysis of Multidimensional Information" by the OLAP council
  - Used interchangeably with 'BI'
  - OLAP tools are synonymous with Multidimensional tools or applications

DATA ACCESS AND ANALYSIS - CATEGORIES

- Data Mining
  - A process that uses a variety of statistical and artificial intelligence frameworks to discover patterns and relationships in data
  - Used to make valid predictions in data analysis problems where the exact sequence and nature of queries/questions to be written/asked against the data to make the prediction is not known and the number of variables involved in the analysis is too large to be intuitively handled by structured querying or OLAP tools

- Web Access
  - A category of data access solutions in which information is viewed through a web browser

DATA ACCESS - USERS

Reporting Solution	OLAP Solution	Data Mining Solution	Web Services Solution
Operating personnel, Standard fixed format reports for power users	Analysts, Decision makers	Decision makers	Operating personnel, Analysts, Decision makers

IMPORTANCE OF DATA ACCESS

- Businesses today face challenges like
  - Large volume of data
  - User demands of flexible and timely access to information
  - Extracting value from key business data
- Data Access is the 'last mile' that enables decision makers to
  - Reach the database infrastructure
- Prompt, reliable data access
  - Lowers operating costs
  - Reduces error
  - Increases productivity

REPORTING SOLUTIONS

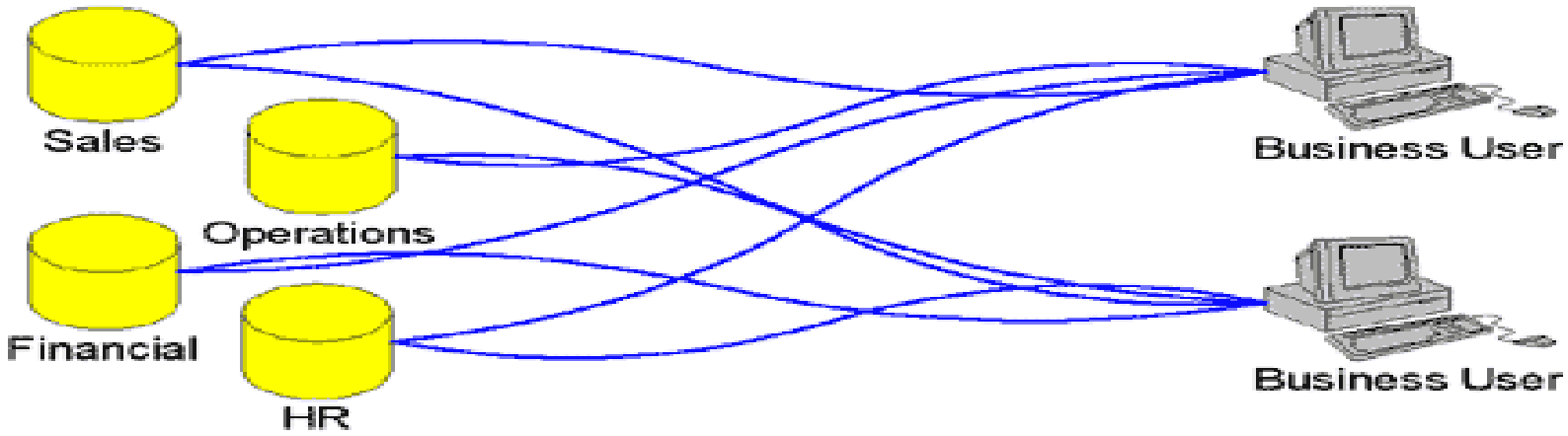
- Reporting solutions provide information in the form of reports
- Companies are data rich but information poor
- Transaction systems generate huge volume of data
- Need to turn this data into actionable information

TYPE OF REPORTING

- Transaction Systems Reporting
- Replicated OLTP Reporting
- Data Mart Reporting
- Enterprise Data Warehouse Reporting

TRANSACTION SYSTEMS REPORTING

- Reports are created from the transaction systems
- Reporting Tool has a native connectivity to
  - The OLTP database or
  - A generic data access protocol such as ODBC, JDBC or OLE DB is made use of



Source : http://www.dmreview.com/portal_ros.cfm?NavID=91&EdID=4903&Topic=14

TRANSACTION SYSTEMS REPORTING (CONTD.).

- This type of reporting is needed for time-sensitive reporting
  - E.g. those reports which need to be up to date with the transactions occurring within the last 24 hours

TRANSACTION SYSTEMS REPORTING - PROS AND CONS

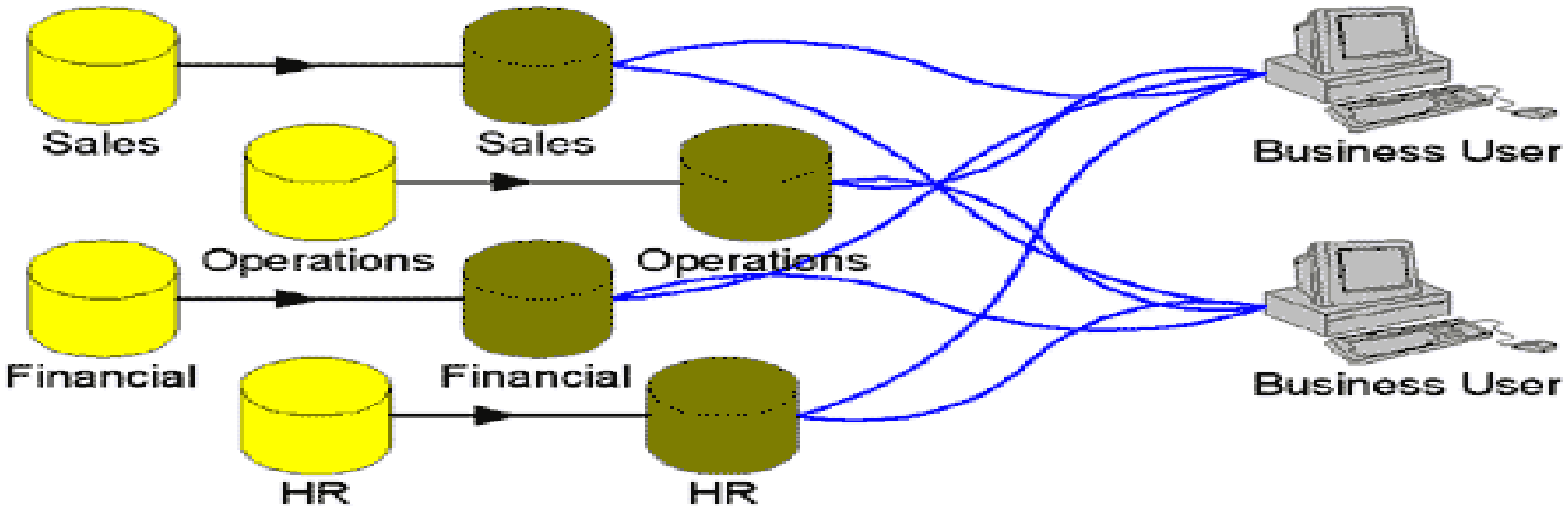
- The primary advantage to reporting from OLTP is speed to information(latest data)
- The data accessed in the OLTP system is the ‘raw’ transaction data and is often ‘dirty’
- Organization of the data is not intuitive to a business user
- Historical data is not available
- When used for Reporting, the transaction system performance is affected

REPLICATED OLTP REPORTING

- An alternative to Transaction System Reporting
- Create an offline data store that is a replica of the production data store
- This data store services the reporting needs of the end user

REPLICATED OLTP REPORTING

- Replication to the offline data store is done by
  - Transaction logs
  - Database replication
  - Batch files



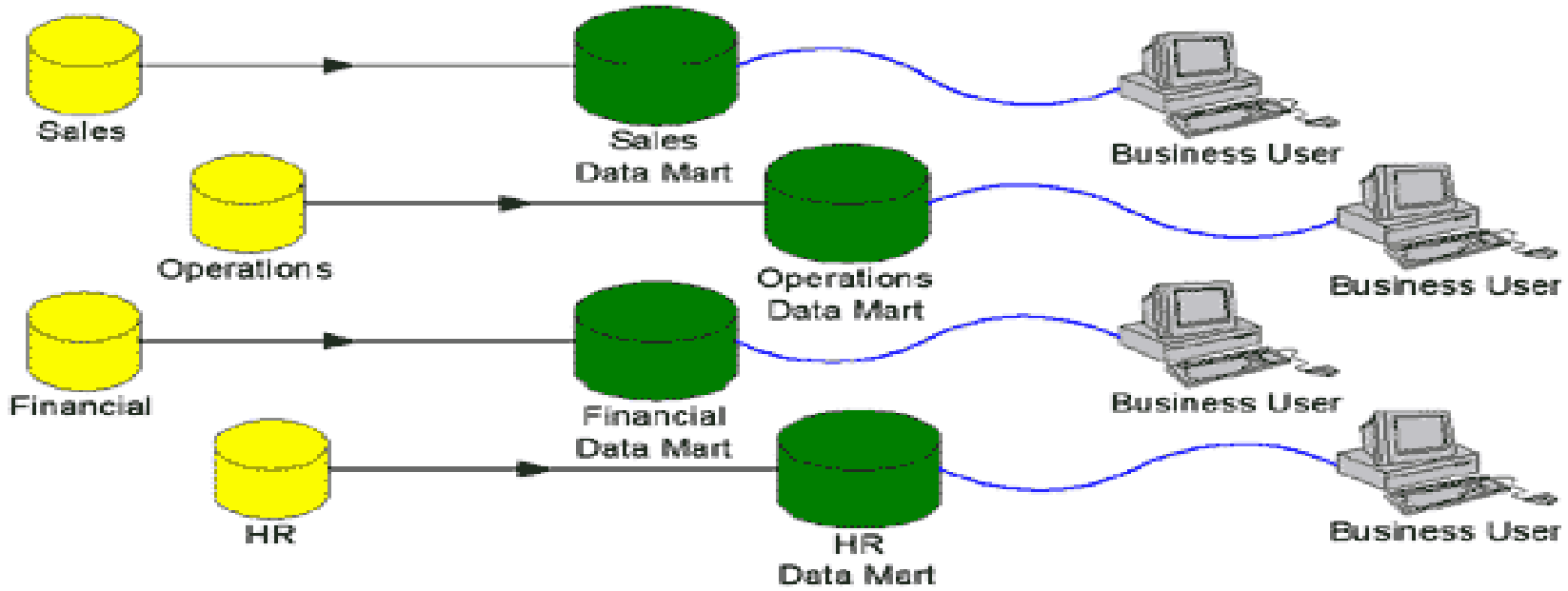
Source : http://www.dmreview.com/portal_ros.cfm?NavID=91&EdID=4903&Topic=14

REPLICATED OLTP REPORTING - PROS AND CONS

- This approach allows the OLTP to continue to process transactions
- The data is still dirty
- Data organization is not intuitive for the business user
- Historical transactions are not available

DATA MART REPORTING

- Another method is to feed transaction data to a system optimized for reporting(data mart)



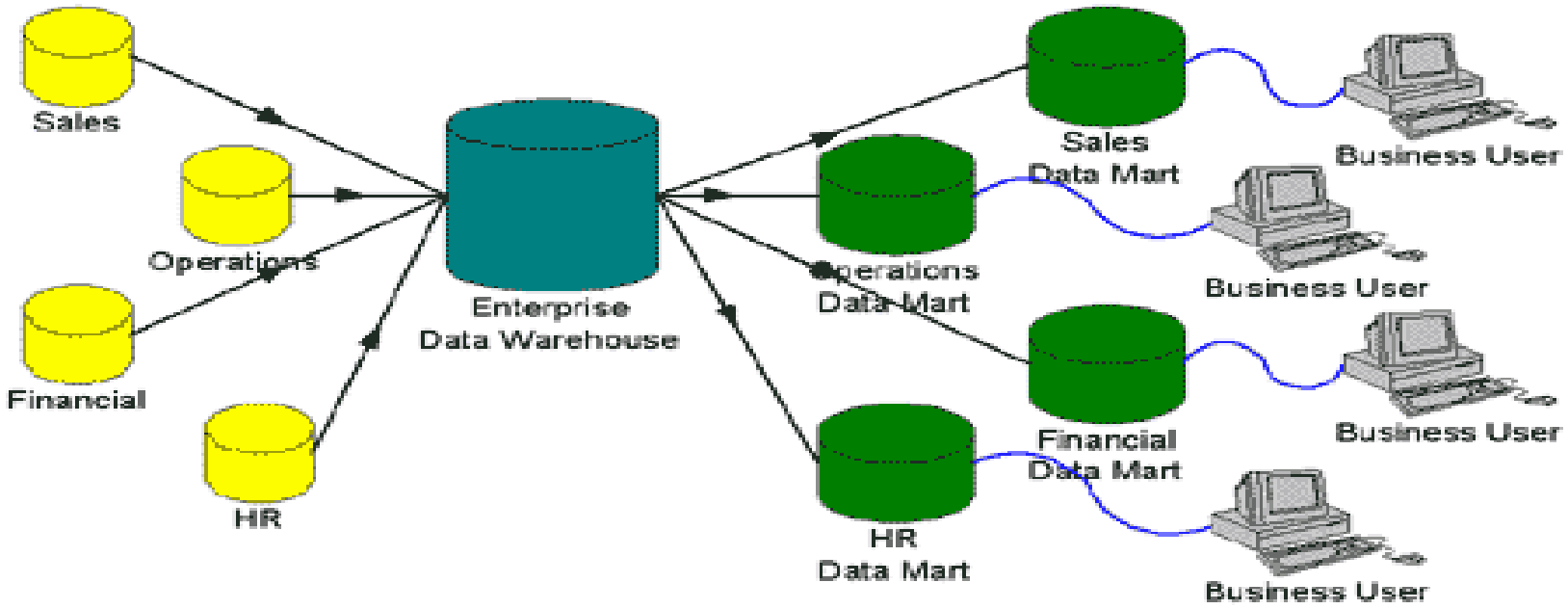
Source : http://www.dmreview.com/portal_ros.cfm?NavID=91&EdID=4903&Topic=14

DATA MART REPORTING – PROS AND CONS

- Data marts are optimized for efficient reporting
- The data organization is intuitive so the business user will be able to browse the data with little or no training
- Historical information will be stored
- Use of data cleansing tools can ensure data quality
- The primary limitation of a data mart approach is scalability
    - Data marts focus on a single department such as sales, inventory, financial analysis or production

ENTERPRISE DATA WAREHOUSE REPORTING

- An enterprise data warehouse (EDW) is designed
  - To combine data from multiple OLTP systems
  - To provide consolidated and cleansed data to an array of data marts



Source : http://www.dmreview.com/portal_ros.cfm?NavID=91&EdID=4903&Topic=14

ENTERPRISE DATA WAREHOUSE REPORTING (CONTD.).

- Historical data is stored
- Data cleansing tools can ensure data quality
- The data organization is intuitive to the business user
- Scalable
- High cost of implementation

QUERY AND REPORTING TOOLS

- Query and reporting tools are used for providing Reporting Solutions
- One type of classification
  - Ad hoc Query Tool
  - Managed Query Tool

QUERY AND REPORTING TOOLS (CONTD.).

- Ad hoc Query Tool
  - Lets the user create a SQL query in a graphical environment
  - Supports limited complex analysis
  - Connects to external data sources through ODBC
  - Examples :
    - Microsoft Access
    - MS Query

QUERY AND REPORTING TOOLS (CONTD.).

- Managed Query Tool
  - Has a semantic layer to hide database complexity from the user
  - Presents the user with a semantic layer in the business terms
  - Has security management, usage limits, report sharing
  - Examples :
    - Oracle Discoverer
    - Business Objects
    - Cognos Impromptu

QUERY AND REPORTING TOOLS (CONTD.).

Transaction/OLTP Reporting	Replicated OLTP Reporting	Data Mart Reporting	Enterprise Data Warehouse Reporting
Proprietary reporting tools e.g., Oracle Reports, Ad hoc Query tools	Proprietary reporting tools e.g., Oracle Reports, Ad hoc Query tools	Ad hoc Query tools, Managed Query tools	Ad hoc Query tools, Managed Query tools

OLAP OVERVIEW

- OLAP – An Introduction
- Distinction between OLTP and OLAP
- Evolution of OLAP
- Functional Requirements of OLAP
- Codd rules for OLAP

OLAP OVERVIEW (CONTD.).

- OLAP – On Line Analytical Processing
  - On Line – Emphasizes live access to data, not static reporting
  - Analytical Processing – Ad-hoc quires, drill-down, roll-up, reporting across various dimension
  - category of technology that enables users to gain insight into their data in a fast, interactive and easy to use manner

- OLAP provides the following 3 features
  - Multidimensional viewing Capabilities -Browsing and Navigation (Slice and dice)
  - Calculation Intensive Capabilities
  - Time Intelligence - Time Series analysis

DISTINCTION BETWEEN OLTP AND OLAP

	OLTP System	**OLAP System**
Source of data	Operational data; OLTPs are the original source of the data	Consolidation data; OLAP data comes from the various OLTP databases
Purpose of data	To control and run fundamental business tasks	Decision support
What the data reveals	A snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries returning relatively few records	Often complex queries involving aggregations

DISTINCTION BETWEEN OLTP AND OLAP (CONTD.).

	OLTP System	**OLAP System**
Processing speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

FUNCTIONAL REQUIREMENTS OF OLAP

- Rich dimensional structuring with hierarchical referencing
- Efficient specification of dimensions and dimensional calculations
- Separation of structures and representation
- Flexibility
- Sufficient speed to support ad hoc analysis
- Multi-user support

KEY OBJECTS WITHIN OLAP

- Key Objects within OLAP
  - Measures – represents factual data
    Example: sales, cost, profit
    - Types of measures
      - Stored measures (Example: revenue, expense)
      - Calculated measures (Example: ratio, averages, profit)

  - Dimensions – identify and categorize data
    Example: product, time, geography, customer
    - Key Components
      - Hierarchies – a logical grouping of data
      - Levels – position in hierarchy
      - Attributes – descriptive information about dimension

TYPE OF OLAP

- MOLAP
- ROLAP
- HOLAP
- DOLAP

TYPE OF OLAP (CONTD.).

- MOLAP
  - Data is stored in multidimensional cubes
  - MDDB technology is proprietary
  - Compilation intensive architecture
  - Load involves series of aggregations across orthogonal dimensions
  - Good to access pre-aggregated data

Example: Cognos, Oracle OLAP, Microsoft Analysis Services, Essbase

- ROLAP
  - Support for large databases with good performance
  - Excellent platform portability
  - Good exploitation of hardware advances such as parallel processing

Example: Microsoft Analysis Services, Microstrategy, Oracle BI

TYPE OF OLAP (CONTD.).

- HOLAP
  - Hybrid approach seeks to provide best of both worlds - MOLAP and ROLAP
  - Create MDB structures for fast analytical needs
  - Create mappings to RDBMS for larger data volumes
  - The engine hides these mappings/ structures from the end user

Example: Microsoft Analysis Services

- DOLAP
  - Desktop OLAP tools work by extracting RDBMS data into local (or server) MD cubes
  - User queries are limited to the predefined dimensions in the hypercube
  - Do not provide shared environment
  - Do not provide read/write capabilities
  - Low cost tools

Example: Business Objects, Brio

Thank You