etl-architecture.pdf

ETL ARCHITECTURE

ETL Architecture Options - Overvie

12/08/2004

ETL Architecture Options - Guidelines

be tool dependant, client environment dependant or it can be based on the best p
Explain different plug-inns available with the ETL tool. Explain the command l
Wipro's experience.

- Provide different option of architecture by which the tool can be configured and effectively in the organization, the option may be based on setup of repository s pros and cons for each of the options. Limit the number of option to max 4 and the one appropriately which will suite the client's requirement and environment
- Make sure that you are handling the geographically distributed architecture is r make sure that the architecture suggested is flexible enough to expand as neede
- Provide different option for Disaster recovery solutions and Highly Availability solutions 12/08/2004

ETL Architecture Options - Guidelines

- **Architecture Option 1**

  Explain the architecture with the diagram and provide pros and cons. E this option can be implemented in client environment. Probably this car distributed architecture (explained in ETL Framework session)

- **Architecture Option 2**

  Explain the architecture with the diagram and provide pros and cons. E this option can be implemented in client environment. Probably this car centralized architecture (explained in ETL Framework session)

- **Architecture Option 3**

- **Recommendation for ETL Infrastructure**

  Explain the recommended architecture option for ETL infrastructure an justification. Explain how client can be maximum benefited from the re architecture option in their current environment. Also provide small roa client regarding how they should grow with this option.

12/08/2004

ETL Design Considerations

12/08/2004

- Modularity
- Consistency
- Flexibility
- Speed
- Heterogeneity
- Meta Data Management

12/08/2004

ETL Design Considerations

- **Modularity**

  ETL systems should contain modular elements which encourages reuse them easy to modify when implementing changes

(either commit or rollback).

Flexibility

ETL systems may be appropriate to accomplish some transformations i
and some on the source data system; others may require the developme
applications

ETL Design Considerations

- **Speed**

  ETL systems should be as fast as possible.

- **Heterogeneity**

  ETL systems should be able to work with a wide variety of data in diffe

ETL Architectures - Types

12/08/2004

ETL Architectures - Types

- **Based on the No of ETL Processes:**

  - Traditional Architecture
  - Conformed Architecture

12/08/2004

ETL Architectures - Types

Homogenous Architecture

A homogenous architecture for an ETL system is one that involves only source system and a single target system

- Features

  - **Single data source**

    Data is extracted from a single source system, such as an OLTP

  - **Rapid development**

type.

Light data transformation

No data transformations are required since the incoming data is
format usable in the data warehouse.

12/08/2004

- **Light structural transformation**

  Because the data comes from a single source, the amount of structural c
  as table alteration is also very light

- **Simple research requirements**

  The research efforts to locate data are generally simple: if the data
  is in the source System, it can be used. If it is not, it cannot

12/08/2004

Heterogeneous Architecture

A heterogeneous architecture for an ETL system is one that extracts dat multiple sources

- Features
  - Multiple data sources
  - More complex development

    The development effort required to extract the data is increased there are multiple source data formats for each record type.
  - Significant data transformation

    Data transformations are required as the incoming data is often format usable in the data warehouse.

12/08/2004

Heterogeneous Architecture

- Substantial research requirements to identify and match dat

Heterogeneous Architecture

ETL Architectures - Types

Traditional ETL Architecture

Create individual ETL processes for each source system

A traditional ETL architecture would create one ETL process to perform logic necessary to transform the source data into its target destination.

ETL Architectures - Types
Traditional ETL Architecture

- Advantages
  - There are fewer ETL processes to create and maintain
- Disadvantage
  - Here each individual ETL process redundantly perform many of the s
  - During development, this results in an increase in the amount and co ETL code to create and test.
  - Upon Implementation, modifications require more effort as changes must be made in multiple places

12/08/2004

ETL Architectures - Types
Conformed ETL Architecture

12/08/2004

Conformed table

Advantages

- **Modularization of ETL processes**

  The creation of smaller, less complex ETL processes makes troubleshoo problems and creating future enhancements easier.

- **Reusability of post-conform processes:**

  - Enforcing referential integrity (looking up foreign key assignn performing inserts and/or updates to the final target is made sim

12/08/2004

- Extensibility

  Confirmed architecture allows us to add new source systems.

- Disadvantages

  - There are more objects and processes that must be created and

12/08/2004

ETL Architectures - Types Based on Repository

- 3 types of ETL Architectures based on repository
  - Single Domain Distributed Repository Architecture
  - Single Domain Single Repository Architecture
  - Distributed Domain Architecture

12/08/2004

Single Domain Distributed Repository Architecture

12/08/2004

Single Domain Single Repository Architecture

12/08/2004

Distributed Domain Architecture

12/08/2004

12/08/2004

The ETL process flow Types

- ☐ TRANSFORM THEN LOAD
- ☐ LOAD THEN TRANSFORM
- ☐ TRANSFORM WHILE LOADING

12/08/2004

The ETL process flow Types

TRANSFORM THEN LOAD

The data is manipulated outside the database to cleanse and sort it, with loaded into the database

Transform and load are

- If the data is transformed outside the database, the external tools used t
  data may not scale as effectively as the database does and will become a

- Depending on the architecture, the external mechanism has to control th
  of the ETL process and provide recovery and restart ability

The ETL process flow Types
LOAD THEN TRANSFORM

12/08/2004

The main risks and disadvantages of Load then Transform are

- Here, extra disk storage for the staging tables is required.
- The transformation process is interrupted by storing not only intermedi
  but also the original raw data from the source systems in the Database.

12/08/2004

The ETL process flow Types

TRANSFORM WHILE LOADING

The raw data is selected directly from a stream of data from the produc
or flat files, transformed by applying one or more table functions to it, a
written to the database

12/08/2004

ETL Extraction Methodology

12/08/2004

Important considerations for extraction

- The extraction method to choose is highly dependent on the source syst from the business needs in the target data warehouse environment.
- The estimated amount of the data to be extracted and the stage in the ET also impact the decision of how to extract, from a logical and a physical

12/08/2004

Methods to Extract data

- Logical Extraction Method
  - There are two kinds of logical extraction:
    - Full Extraction
    - Incremental Extraction
- Physical Extraction Method
  - There are two kinds of physical extraction
    - Online Extraction
    - Offline Extraction

12/08/2004

Logical Extraction Method

- Full Extraction
  - The data is extracted completely from the source system
- Incremental Extraction
  - Only the data that has changed since a well-defined event bacl will be extracted.
    - Two ways to accomplish
      - Change data capture technique
      - 12/08/2004
      - Entire tables extraction

- Change data capture technique
  - Extract only the most recently changed data

12/08/2004

- Tables from the source systems are extracted to the data wareh staging area, and these tables are compared with a previous extract from the source system to identify the change

Physical Extraction Methods

- Online Extraction
  - The data is extracted directly from the source system itself.
- Offline Extraction
  - The data is not extracted directly from the source system but is explicitly outside the original source system

12/08/2004

Extract data in two ways

- **Extraction Using Data Files**

  - Most of the database systems provide mechanisms for exportir unloading data from the internal database format into flat files

- **Extraction Via Distributed Operation**

  - Using distributed-query technology, one database (oracle) can query tables located in various different source systems, such as database (oracle)

12/08/2004

12/08/2004

Transportation

- Transportation is the operation of moving data from one system to anot

- The most common requirements for transportation are in moving data f

  - A source system to a staging database or a data warehouse dat
  - A staging database to a data warehouse
  - A data warehouse to a data mart

12/08/2004

Three basic choices for transporting data in warehouses

- Transportation Using Flat Files
- Transportation through Distributed Operations
- Transportation Using Transportable Table Spaces

12/08/2004

- Transportation Using Flat Files

  - The most common method for transporting data is by the trans files, using mechanisms such as FTP or other remote file system protocols.

  Advantages

  - Source systems and data warehouses use different operating and database systems, using flat files is the simplest way to data between heterogeneous systems with minimal transfor

  - When transporting data between homogeneous systems, fla often the most efficient and most easy-to-manage mechanis transfer

12/08/2004

Transportation through Distributed Operations

- Distributed queries, either with or without gateways, can be an effective for extracting data.

12/08/2004

Transportations Using Transportable Table Spaces

- Using transportable table spaces, Oracle data files (containing table dat and almost every other Oracle database object) can be directly transpor database to another.

Disadvantage

- Source and target systems must be running Oracle8*i* (or hig be running the same operating system, must use the same c

12/08/2004

Loading Methodology

12/08/2004

Loading Mechanisms

- SQL*LOADER
- External Tables
- OCI and Direct Path API's
- Export /Import

12/08/2004

Staging Area

STAGING AREA

12/08/2004

Definition

- A place where raw data is brought in, cleaned, combined, archived, and one or more data marts.
- It is also used to get data ready for loading into a presentation server.

12/08/2004

Area.

12/08/2004

Data Staging Area Roles

- Integrates data from many application source systems so there is one co
  System wide enterprise view of the data.

12/08/2004

Characteristics of Staging Area

- Facilitates moving data from different sources on different schedules.
- Provides a place to check data cleanliness and correctness.
- Is the enterprise-wide integration of data.
- Data is stored at the lowest level of detail available.

Need For Staging Area

- Data used in the data warehouse is extracted from the data sources, clea transformed into the data warehouse schema.
- The data is checked for consistency and referential integrity.
- Promotes effective data warehouse management.
- Data transformation in the data source systems can interfere with OLTP performance.
- Allows DW professionals to assess the data quality problems before the to the warehouse.

12/08/2004

Creating Staging Area

- Create tables and other database objects to support
  - the data extraction
  - cleansing
  - transformation operations required to prepare the data for load data warehouse.
- We can create a separate database for the data staging area, or can creat

12/08/2004

Creating Staging Area

- It should include tables to contain the
  - incoming data
  - tables to aid in implementing surrogate keys
  - tables to hold transformed data.
- Design will depend on
  - the diversity of data sources
  - the degree of transformation necessary to organize the data fo loading
  - the consistency of the incoming data.

12/08/2004

Data Staging Techniques

- Surrogate key creation and maintenance
- Processing Slowly Changing Dimensions
- Combining from Separate Sources
- Data Cleaning
- Processing Names and Addresses
- Validating One-to-One and One-to-Many Relationships
- Fact Processing
- Aggregate Processing

12/08/2004

Data Staging Techniques

- Surrogate key creation and maintenance
  - create a surrogate key .
  - every DW key should be a surrogate key

12/06/2004

Type 1: Overwrite the Value

"Rewriting History" - no history .

Type 2: Add A Dimension Row

"

Data Staging Techniques

- Combining from Separate Sources

  - Dimensions are derived from several sources. the merge opera on the some criteria.

- Data Cleaning

  - Data cleaning may involve checking the spelling of an attribut checking the membership of an attribute in a list

- Names and addresses have been cleaned and put into standard

Data Staging Techniques

- Fact Processing
  - The incoming fact records will have production keys, not data keys. The current correct correspondence between a production data warehouse key must be looked up at load time
- Aggregate Processing
  - Each load of new fact records requires that aggregates be calcu augmented. It is very important to keep the aggregates synchron base data at every instant

12/08/2004

Data Staging Techniques

- [ ] A many-to-one relationship , (eg.zip code-to state) ,can be va sorting on the "many" attribute and verifying that each value ha value on the "one" attribute.

12/08/2004

Data Staging Storage Types

- Flat Files
- Relational tables

12/08/2004

Is Data Staging Relational ???

- whether the data staging area is relational or has more to do with seque processing of flat files. Ralph Kimball [1] concludes that

" Most data staging activities are not relational, but rather they are sequer processing. If your incoming data is in flat-file format you should finish staging processes as flat files before loading it into a relational databas states that if both the source and target databases are relational it may appropriate to retain this format and not convert to flat files "

12/08/2004

Data Staging Components

- Data Staging Application Server
- Data Staging Repository
- Metadata and Meta model Repository

12/08/2004

Data Staging Components

- **Data Staging Application Server**
  - temporarily stores and transforms data extracted from OLTP da
  .

12/08/2004

transformation and loading activity.

- The archival repository stores cleaned, transformed records and attributes for later loading into data marts and data wareho

Data Staging Components

- Metadata and Meta Model Repository
  - The data staging process is driven in an essential way by meta including business rules.
  - Metadata is used along with administrative tools to guide data transformations, archiving, and loading to target data mart and d warehouse schemas.

12/08/2004

Staging scenarios

2 staging scenarios

Scenario 1- a data staging tool is available .

The data is already in a database. The data flow is set up so that it come
source system, moves through the transformation engine, and into a sta
database.

12/08/2004

Staging scenarios

12/08/2004

General Data Staging Requirements

- **Productivity support**

  The data staging services needs to provide basic development environm capabilities like code library management check in/check out, version c production and development system builds.

- **Usability**

  The data staging system must be as usable as possible, it should have g

General Data Staging Requirements

- **System Documentation**

  The data staging system need to provide a way for developers to easily information about the processes that they are creating.

- **Metadata driven**

  Metadata is used along with administrative tools to guide data transformations, archiving, and loading to target data mart and dat schemas.

Staging Metadata

12/08/2004

Data Staging Metadata

The metadata needed to get the data into a staging area and prepare it fo
into one or more data marts are:

- Data acquisition information

12/08/2004

- Transformation and Aggregation

  Audit, Job Logs and documentation

Data Staging Metadata

- Data acquisition information

  - [ ] Data transmission scheduling
  - [ ] File usage in the data staging area including duration, volatilit ownership.

12/08/2004

Data Staging Metadata

- Dimension table Management
  - Definitions of conformed dimensions and conformed facts.
  - Job specifications for joining sources, stripping out fields, and attributes.
  - Slowly changing dimension policies for each incoming descrij attribute.
  - Current surrogate key assignments for each production key, in fast lookup table to perform this mapping in a memory.
  - Yesterday's copy of production dimensions to use as the basis f compare.

12/08/2004

Data Staging Metadata

- [ ] DBMS load scripts.
- [ ] Aggregate definitions.
- [ ] Aggregate usage statistics ,base table usage. statistics and pote aggregates.
- [ ] Aggregate modification logs.

12/08/2004

Data Staging Metadata

- **Audit, Job Logs and documentation**
  - [ ] Data lineage and audit records (where exactly did this record and when ?)
  - [ ] Data transformation run time logs, success summaries and tim
  - [ ] Data transformation software version numbers.
  - [ ] Security settings for extract files, extract software , and extract
  - [ ] Security settings for data transmission (e.g. passwords ,certific
  - [ ] Data staging area archive logs and recovery procedures.

Data Staging Metadata

- DBMS metadata
  - DBMS system table contents.
  - Partition settings.
  - Indexes.
  - Disk stripping specifications.
  - Processing hints.
  - DBMS-level security privileges and grants.
  - View definition.
  - Stored procedures and sql administrative scripts.
  - DBMS backup ,backup procedures and backup security.

Data Staging Metadata

- Front room metadata
  - Business names and descriptions for columns,tables and group
  - query and report definitions.
  - Join specification tool settings.
  - Network security user privilege profiles.
  - Usage and access maps for data elements,tables,views and rep

12/08/2004

College Sidekick

Terms

Academic Integrity

Privacy Policy

Cookie Policy

Do not Sell or Share My Personal Info

SUBJECTS

Accounting

Aerospace Engineering

Anatomy

Anthropology

Arts & Humanities

Astronomy

Biology

Business

Chemistry

Civil Engineering

Computer Science

Communications

Economics

Electrical Engineering

English

Finance

Geography

Geology

Health Science

History

Industrial Engineering

Information Systems

Law

Linguistics

Management

Marketing

Material Science

Mathematics

Mechanical Engineering

Medicine

Nursing

Philosophy

Physics

Political Science

Psychology

College Sidekick

SOCIAL