Introducing Technical Architecture

Context

- What is Operational or OLTP or Database Systems?

- Limitations of OLTP Systems

Queries not answered by OLTP Systems

- What are the three most popular areas in each city for the renting of property in 2004 and how does this compare with the figures for the previous two years?
- What would be the effect on property sales in the different regions of Britain if legal costs went up by 3.5% and Government taxes went down by 1.5% for properties over £100,000?
- Which type of property sells for prices above the average selling price for properties in the main cities of Great Britain?
- What is the relationship between the total annual revenue generated by each branch office and the total number of sales staff assigned to each branch office?

Business Intelligent Systems

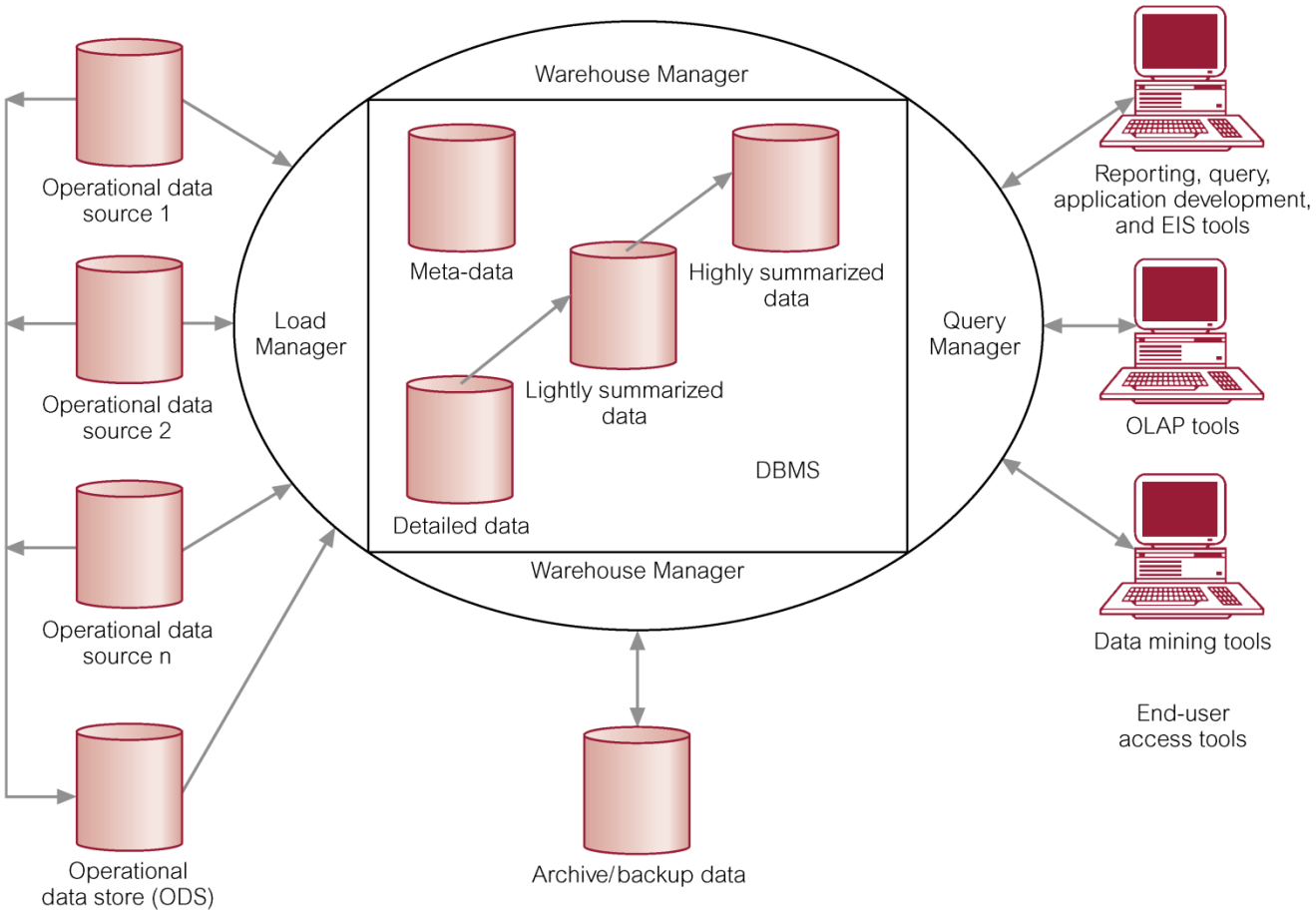
- What is BIS?

  BIS=Datawarehouse+Business Intelligence
- Why BIS?

  To take informed decisions at right time to grow business
- Who are the users of BIS?

  Managers

Typical Architecture of a Data Warehouse



Data Warehousing Concepts

- A subject-oriented, integrated, time-variant, and non-volatile collection of data that support s management's decision-making process (Inmon, 1993).

Subject-oriented Data

- The warehouse is organized around the major subjects of the enterprise (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and product sales).

Integrated Data

- The data warehouse integrates corporate application-oriented data from different source systems
- The integrated data must be made consistent to present a unified view of the data to the users.

Time-variant Data

- Data in the warehouse is only accurate and valid at some point in time or over some time interval.

Non-volatile Data

- Data in the warehouse is not updated in real-time but is refreshed from operational systems on a regular basis.

Benefits of Data Warehousing

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

Comparison of OLTP Systems and Data Warehousing

OLTP systems	Data warehousing systems
Holds current data	Holds historical data
Stores detailed data	Stores detailed, lightly, and highly summarized data
Data is dynamic	Data is largely static
Repetitive processing	*Ad hoc*, unstructured, and heuristic processing
High level of transaction throughput	Medium to low level of transaction throughput
Predictable pattern of usage	Unpredictable pattern of usage
Transaction-driven	Analysis driven
Application-oriented	Subject-oriented
Supports day-to-day decisions	Supports strategic decisions
Serves large number of clerical/operational users	Serves relatively low number of managerial users

Data Warehouse Queries

- **The types of queries that a data warehouse is expected to answer ranges from the relatively simple to the highly complex and is dependent on the type of end-user access tools used.**

- **End-user access tools include:**
  - **Reporting, query, and application development tools**
  - **Executive information systems (EIS)**
  - **OLAP tools**
  - **Data mining tools**

Problems of Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands

Problems of Data Warehousing

- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration

Components of BI Systems

- Backroom
- Presentation server
- Front room

Some BI Tools

- MS SQL Server 2012 SSIS,SSAS,SSRS
- Oracle BI Standard Edition One
- Pentaho
- SAP BUSINESS INTELLIGENCE:
- MS Power BI (Analysis and Reports) through couresra,EdX
- Talend (ETL)
- MicroStrategy (Analysis and Reports)
- IBM Cognos Analytics (Analysis and Reports)
- Google Data Cloud (Analysis and Reports)
- Tableau (Analysis and Reports)

Technical Architecture Overview

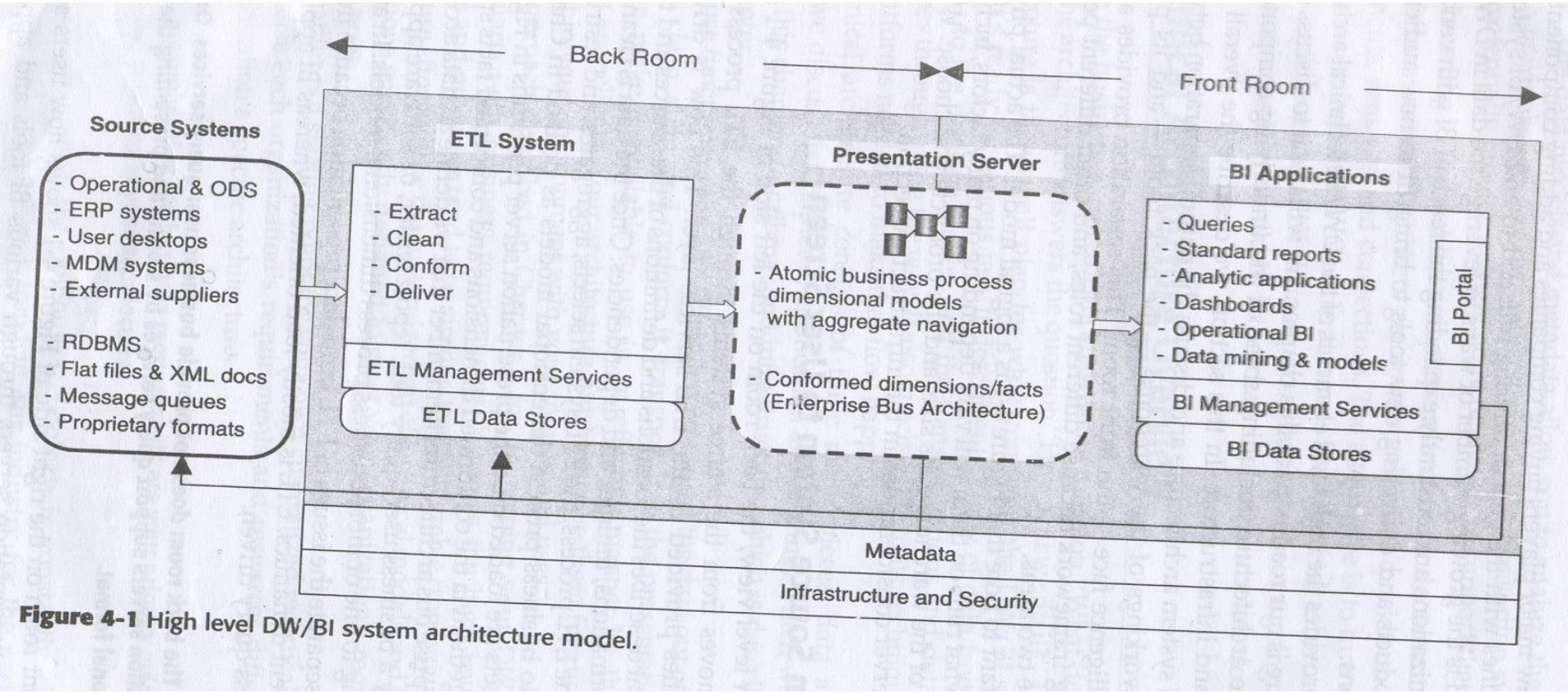


Figure 4-1 High level DW/BI system architecture model.

Flow from Source System to user screen

- Moves through ETL service layer
- ETL selects, aggregates and restructures data into business process dimensional models
- Loaded onto presentation server and tied via conformed dimensions and conformed facts

Common Architecture Features

- Metadata driven
- Technical metadata

Defines objects and processes

Tables, fields,datatypes,indexes

IN ETL sources, targets,transformations and frequency

Business Metadata

- What data you have?
- Where it comes from?
- Display name and content description fields are basic Examples Business metadata

Process Metadata

- Results of various operation in warehouse
- In ETL process start time ,end time, disk reads, disk writes
- Used to monitor user access and popularity of Data Warehouse

Flexible service layers

- SOA
- Service interface definations
- Quality of service characteristics
- Repositories for business policies
- Governance of services
- Lifecycle management

Back Room Architecture

- General ETL Requirements

Productivity Support

-needs to provide basic development environmental capabilities like code library management checkin/out,version control production and development system builds

Usability

-must be usable.

-employs GUI to define and ETL task

-system documentation

Metadata Driven

-ETL process must be meta data driven.

Build vs Buy

Back Room ETL Flow

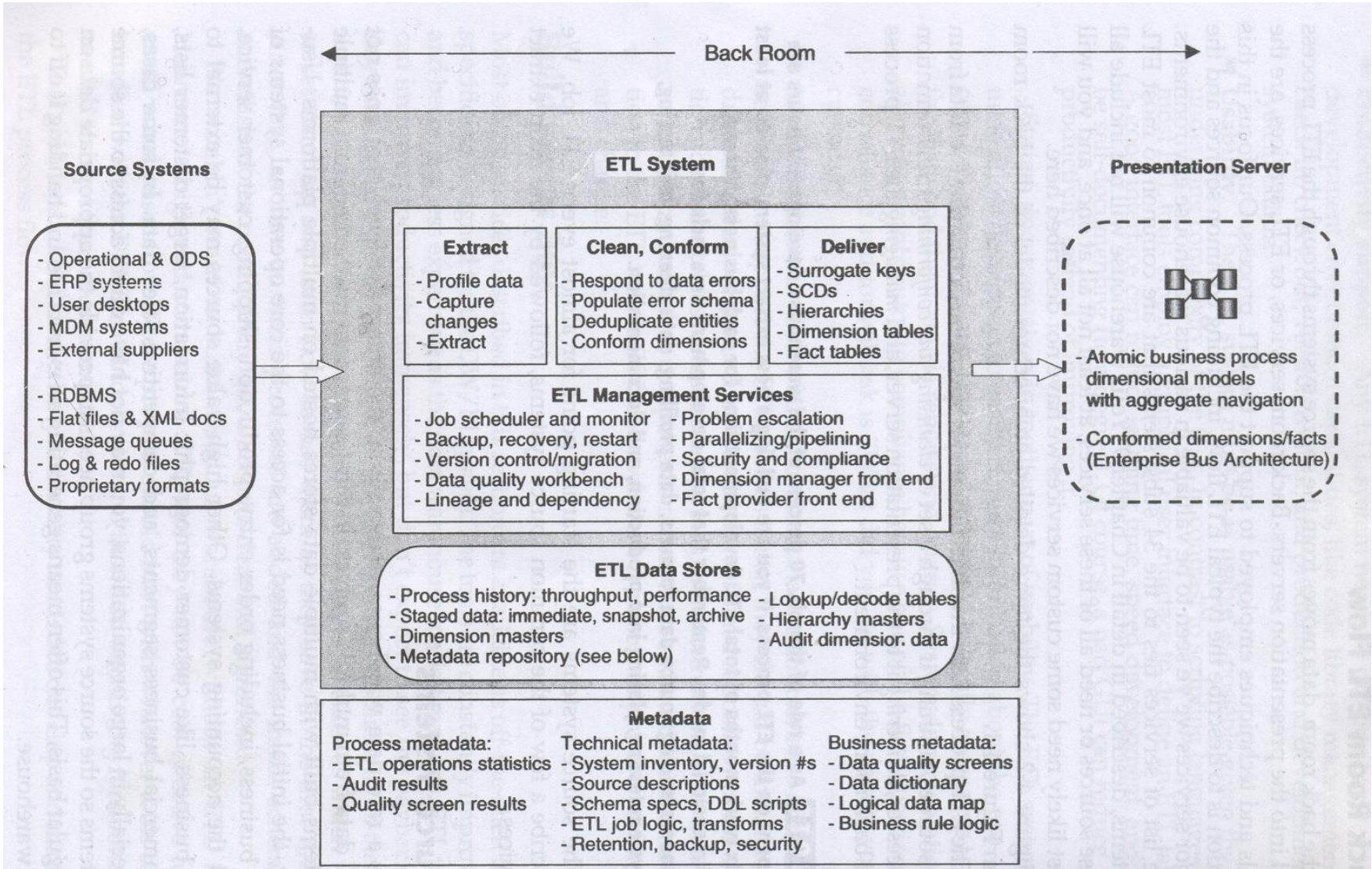


Figure 4-2 Back room system architecture model.

- Source systems

-ERP Systems

-operational data stores

--Xml files

-message queues, log files

-proprietary formats

Extract

- Data profiling
- Change data capture
- Extract system

Clean and conform

- Data cleansing
- Error event handling
- Audit dimension creation
- Deduplicating
- Conforming

ETL Management Services

- Job scheduler
- Backup system
- Recovery and restart
- Version control
- Workflow migration
- Workflow monitor
- Sorting
- Lineage and dependency

Additional Backroom Services and Trends

- Data service providers
- Functional service providers
- Data Delivery Services
- ETL data stores
- ETL System Data Stores
- Data Quality Data stores

ETL Metadata

- Process Metadata

-ETL operations statistics

-Audit results

-Quality screen results

- Technical Metadata

-System inventory including version numbers

-source descriptions

-source access methods

-ETL data store specifications and DDL scripts

- ETL data store policies and procedures
- ETL job logic, extract and transforms
- Exception handling logic

Business Metadata

- Data quality screen specifications
- Data dictionary
- Business rule logic

Presentation Server Architecture

- Business requirements for information

-Access to data from all major business processes

-Access to both summary and atomic data

-Single source for analytical data

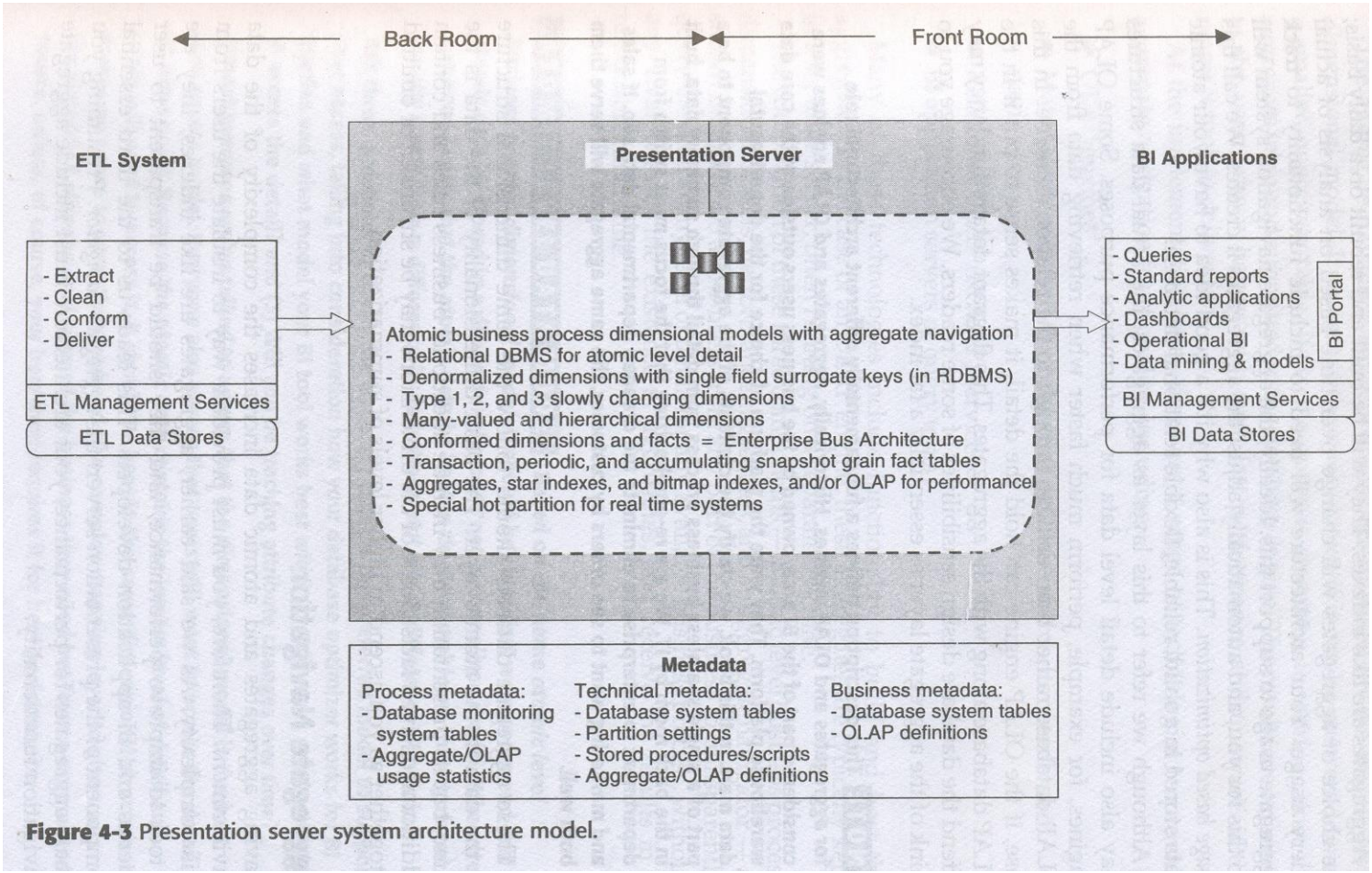


Figure 4-3 Presentation server system architecture model.

- Detail Atomic Data

-analytic queries require detail or summary data

-atomic data are built with conformed dimensions as per enterprise bus matrix

-stored in rdbms than in OLAP

Aggregates

-organizations have large data sets

-summary query takes long time

-pre-aggregating data during load process

-choice of aggregates changes based on analysis of actual query usage

-usage based optimization

Aggregate Navigation

-aggregate like indexes improves performance

-include aggregate navigation facility

-input from user query

-see if query can be answered using smaller aggregate table

-so query is rewritten to work against aggregate table and submitted to database engine

Technologies providing aggregate navigation facilities

- OLAP engines
- Materialized views
- ROLAP services
- BI application servers or query tools

Design Disciplines within presentation server

- Contents of PS includes

-Denormalized Dimension tables with single field surrogate keys

-Type 1,2,3 slowly changing dimensions

-many valued and hierarchical dimensions

-conformed dimensions and facts based on enterprise bus architecture

-transaction, periodic snapshots and accumulating snapshots fact tables

Adjusting PS architecture

- Includes partitioning warehouse onto multiple servers either vertically or horizontally
- VP is component based partitioning
- Server for atomic level, aggregates and server for aggregate management and navigation
- Separate servers for background ETL processing

Horizontal Partitioning

- Load distribution based on data sets
- Separate presentation servers for specific business process dimensional models

Presentation Server Metadata

- Process metadata

-Database monitoring system tables

information about use of tables throughout presentation server

-aggregate usage statistics including OLAP usage

Technical Metadata

-Database system tables

-partition settings

-stored procedures and SQL scripts

-aggregate definitions such as Mviews

-OLAP system definitions specific to OLAP databases

-Target data policies and procedures

retention,backup,archive,recovery,ownership

Business metadata

- Provided by BI applications
- OLAP definitions

Front Room Architecture

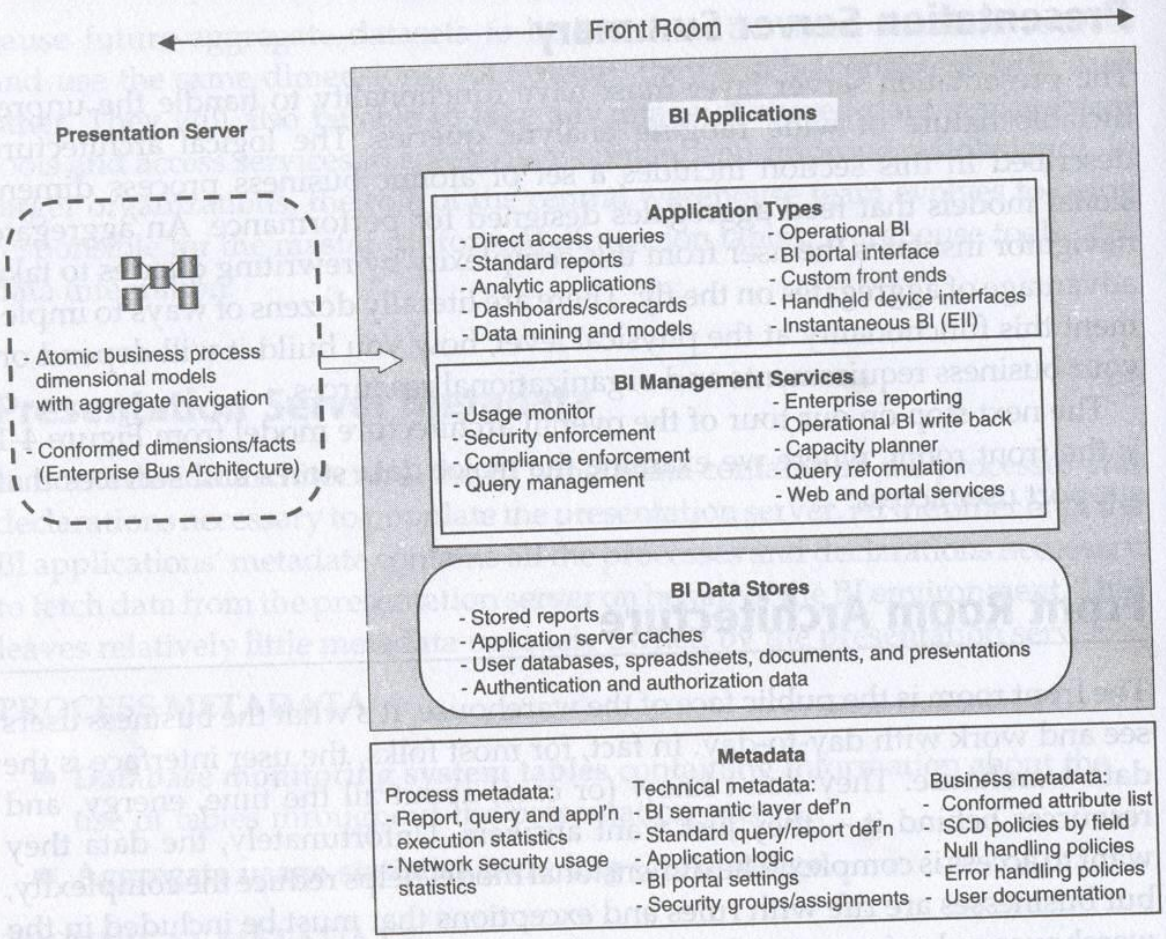


Figure 4-4 Front room technical architecture model.

BI Application types

- Direct access queries

-desktop query tools applications

Standard reports

-regularly scheduled reports typically delivered via BI portal

Analytic applications

-budgeting, forecasting and business activity monitoring

- Dashboards and scorecards
- Datamining and models
- Operational BI-Realtime queries are handled

Application Interfaces

- BI portals
- Handheld device interface
- Instantaneous BI-Realtime integrated DW from source to target

BI Management services

- Shared services

-metadata, security, usage monitoring, query management, enterprise reporting

- Metadata services
- Security services
- Usage monitoring

-performance-usage information to tune performance

-user support-poorly formed queries and constructing more efficient queries

- Marketing-publish usage statistics to inform management
- Planning-Analyze usage , av.query time etc
- Compliance-audit trail of all user activity

Query Management

- Query reformulation
- Query retargeting and drill across
- Aggregate navigation-query can be satisfied by an available aggregate table than summing up detail records
- Query governing and prioritization

  -limits on num of queries/sec

  -time required to execute query in sec based on query plan

Enterprise Reporting Services

- Report development environment
- Flexible report definitions

-compound document layout dashboards with graphs and tables on same page. Formatting control for display and print

- Report Execution Server

-central resource for running and staging for delivery

- Parameter or variable driven capabilities

-change region name parameter and get entire set of reports

- Time & Event based scheduling of report execution

-run at time or event such as completion of ETL process

- Iterative execution or separation of results

-create a copy of report for each region

-report file mailed to region manager

- Flexible report delivery

-via multiple delivery methods like email, web, network directory. In multiple result types like file,html,pdf,excel

- User accessible publish and subscribe capabilities

-make report available to departments

-Subscribe to reports others have made

- Report linking

-click and display department wise report

- Report library with browsing capability

-user interface allows the user to search reprt library using different criteria

- Mass distribution

-mass distribution and viewing by users across organization and customers via internet

- Report Env admin tools
- User access security

-authenticate users and check access rights of users

Web Access

- Web based data access services
- Direct access to specific information
- Monthly summaries of purchases, sliced in various ways

BI Data Stores

- Stored reports

-reports stored at client side cache to improve performance

-faster response to request of previously retrieved result set

- Application server caches

-data oriented services for operational BI, analytic applications and enterprise reporting have their own data stores

- Local user databases
- Disposable analytic data stores
- Results from analytic applications

- Downstream systems

-other systems using facilities of DW/BI

-Budgeting systems using data from warehouse

-Like monthly phone bills for last three years

-Forecasting systems using historical data

-CRMs using DW/BI data

-Need to support such downstream systems

- Data Store Security

-authentication and authorization of data stores

BI Meta Data

- Process Metadata

-report and query execution statistics

-Network security usage statistics

- Technical metadata

-BI semantic layer definitions-business names for tables, columns mapped to presentation server objects

-standard query and report definitions

-application logic

-BI portal settings

- Business Meta data

-Conformed attribute and fact definitions and business rules

-User docs and training materials

Infrastructure

- Infrastructure Drivers-bus req, expertise, policy, growth rates, ETL splits
- Backroom and presentation server infrastructure factors
- -Data size
- -Volatility

  measures dynamic nature of DB.

  Business and technology changes must be considered

  impact size and speed of H/W platform
- -Number of users-active concurrently, peak load
- -Number of business processes
- -Nature of use-front end tools, report scheduling, data mining load

- Service level agreements

  -Performance and availability requirements affect size and quantity of hardware needed.

  -24x7 user support

  -centralized operational systems and warehouse

  -if decentralized operational systems then decentralized presentation server

- Technical readiness

-technical expert manpower in H/W ,OS, Database, Datawarehouse, BI level at setup, management, security, backup and recovery level

- Software availability
- -requirements analysis indicates need for certain capability
- Financial resources
- -amount of money spent on project is usually a function of projects expected value.
- -better understanding of business requirements and target DW/BI is required

Parallel Processing Hardware Architectures

- Symmetric multiprocessing(smp)
- Massively parallel processing(mpp)
- Non-Uniform memory Architecture(NUMA)
- Clusters
- -scaleout
- -scaleup
- -Shared cache
- -Federated data
- -Replicated data

- Partitioning Hardware

- -16 proc, 4 ETL, 8 rdbms,4 reporting

- Considerations common to all parallel architectures

  -software availability, system admin complexities, type and version of os

  -utilities and drivers available for os

  -run on current version of rdbms, ETL tool, development environment, datawarehouse utilities and application servers

Hardware Performance Boosters

- Disk Issues
- -disk drives
- -DW require high bandwidth of IO subsystem than OLTP
- -high end drives require disk subsystem called SAN
- -RAID
- -Fault tolerance
- -Back up
- -Expensive but good value over time

- Memory
- -more is better
- -BI queries are much larger and require several passes through large tables
- -If table in memory performance improve
- -Go for 64 bit os, databases than 32 bit

CPUS

-more and faster is better

-Dual core and quad core

Secondary storage

-configuration includes resources to support backup and archiving

-disk to disk backup

Database Platform Factors

- Major database vendors provides support for DW/BI
- Rest through high end data warehouse appliance
- Aggregate management and navigation

- Characteristics of Relational Engines
- -add capabilities like dimensional model support, star join optimization, bit mapped indexes, data compression and improved cost based optimizations,
- Characteristics of OLAP engines
- -designed to support analysis
- Serve two purposes 1) pre-aggregating 2)complex analysis through more analytically oriented query language
- All database vendors support some OLAP queries

Front Room Infrastructure factors

- Application server
- -memory, Disk, platform sharing,bottlenecks
- Desktop considerations
- -cross platform support-mac,sun
- -desktop os and software-os version, browser version
- -memory

Connectivity and Networking factors

- Band width-large block transfers from source to presentation servers
- Remote access
- File transfer-ssl
- Database connectivity-multiple drivers support
- Directory services-x500,LDAP,user directories, PCs in network,printers or any lookup

MetaData

- Value of Metadata integration
- Impact analysis
- Audit and documentation
- -lineage analysis
- -understanding contents and source of table, column
- -lineage analysis use audit metadata to determine origin of fact or dimension row
- Metadata quality and management
- -multiple copies on diff systems get out of sync
- -DW/BI developer should know this ahead of time
- -less problematic in business meta data
- -provide common extra field for description

- Option for metadata integration
- -DW/BI involves products from different vendors
- -need for standards
- -common warehouse metamodel(CWM) with XMI
- -XMI is xml based metadata interchange standard
- -MetaObject facility (MOF)is underlying storage facility for CWM
- -single source DW/BI system vendors-many vendors provide all tools with common metadata management and their tools share central metadata repository
- -core vendor support-offer metadata management systems