Data Warehousing
SS ZG515

BITS Pilani
Pilani Campus

PC Reddy
Guest Faculty – WILP, BITS Pilani

innovate achieve lead

BITS Pilani

Pilani Campus

Data Warehousing – Lecture 3

DW Components , Architecture and Infrastructure

Lecture 3 Outline

innovate achieve lead

- Review Lecture 1 & 2
- Data Warehouses and Data Marts
- Data Warehouse Architecture Types
- Data Warehouse Lifecycle concepts
- Database Normal Forms
- Data Warehouse infrastructure

Review Lecture 1 & 2

innovate achieve lead

Bill Inmon's paradigm:

Data warehouse is one part of the overall business intelligence system. An enterprise has one data warehouse, and data marts source their information from the data warehouse. In the data warehouse, information is stored in 3rd normal form.

Ralph Kimball's paradigm:

Data warehouse is the conglomerate of all data marts within the enterprise. Information is always stored in the dimensional model.

DW and Data Marts

innovate achieve lead

Data Warehouse

- Corporate/Enterprise-wide
- Union of all data marts
- Data received from staging area
- Queries on presentation resource
- Structure for corporate view of data
- Organized on E-R Model.

Data Mart

- Departmental
- A Single business process
- STAR join(facts and Dim)
- Technology optimal for data access and analysis
- Structure to suit the departmental view of data

Top-down vs Bottom-up Approach

innovate achieve lead

Top-down approach

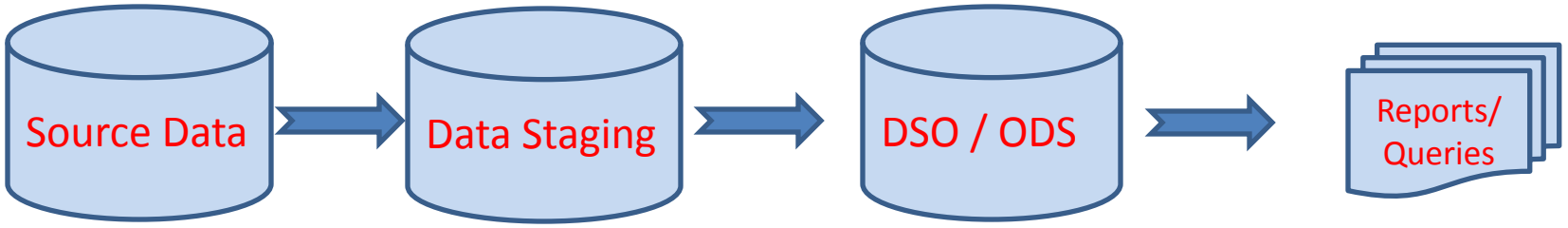
- Bill Inmon
- Normalized data model
- Enterprise view of data
- Single, central storage of data
- Takes longer to build
- High exposure to risk and failure.

Bottom-up approach

- Ralph Kimbal
- De-normalized data model
- Collection of conformed data marts which gives enterprise view
- Inherently incremental
- Less risk of failure and allows project team to learn and grow.

Data warehouse Architecture types

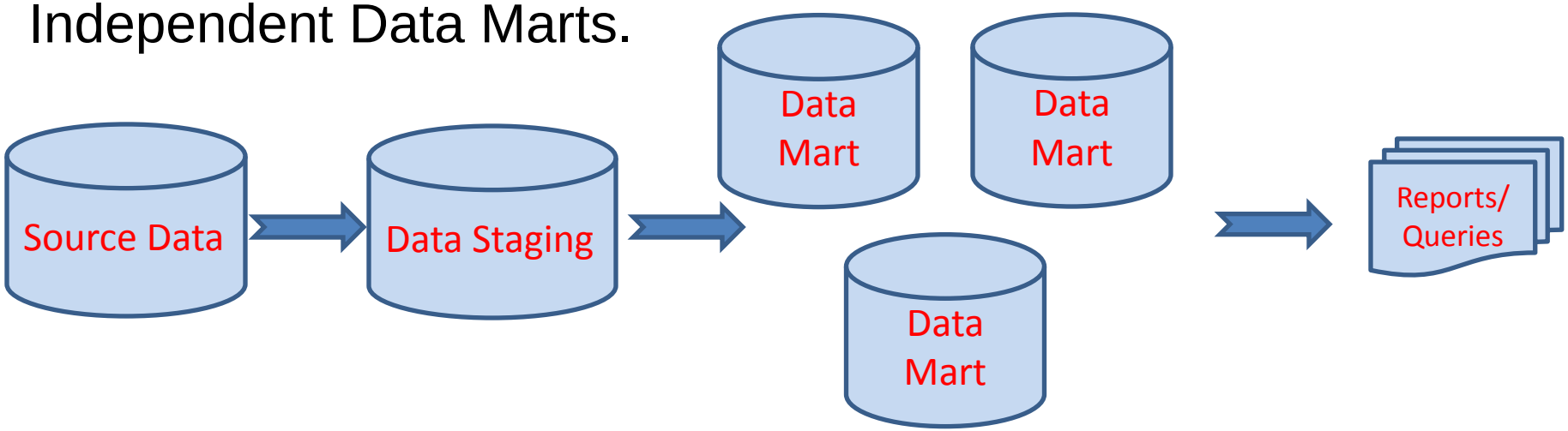
Centralized Data Warehouse



Normalized data in third normal form.
Summarized data at DSO/ODS level
Queries/Reports access central DW.
There are no Separate data marts.

Data warehouse Architecture types

Independent Data Marts.

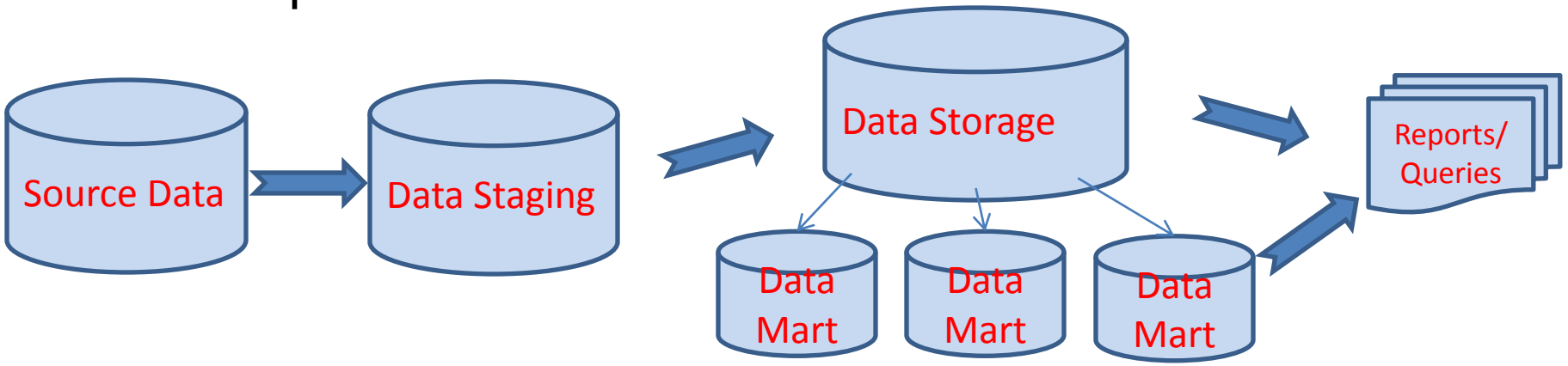


Each data mart in this model serves a particular organizational unit.
Each data mart is independent of one another.
Variances between data marts affect data analysis across data marts.

For example: Sales and Shipments are two independent data marts.
Even though sales and shipments are related, in this model, it is difficult to analyze sales and shipment data together.

Data warehouse Architecture types

Hub and Spoke model.

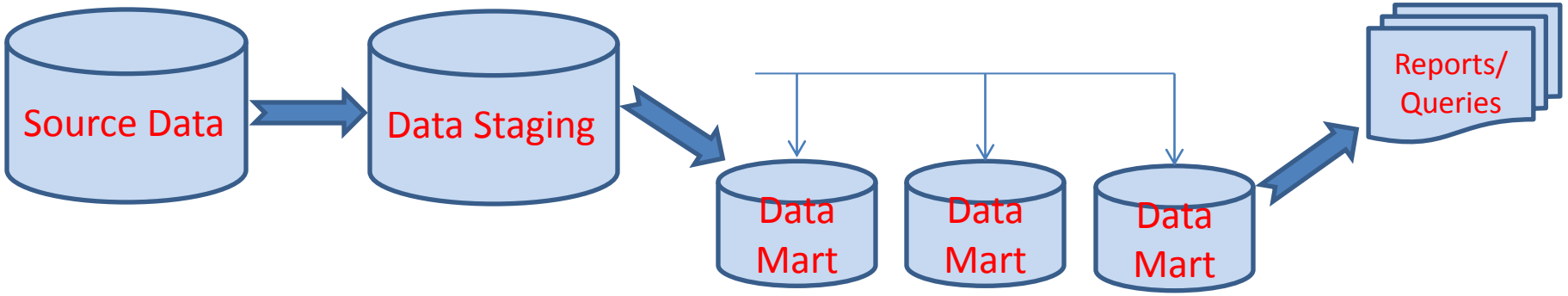


Inmon CIF (Corporate Information Factory) Approach.
Centralized DW in third normal form.
Dependent Data marts obtain data from Centralized DW.

Each dependent Data mart may have
- Normalized
- Denormalized
- summarized/dimensional data structures.

Data warehouse Architecture types

Hub and Spoke model.



Kimbal's conformed approach
Business dimensions from first data mart is shared among other data marts,
Conformed dimensions will give logical integrated DW with enterprise view,

Bottom up approach

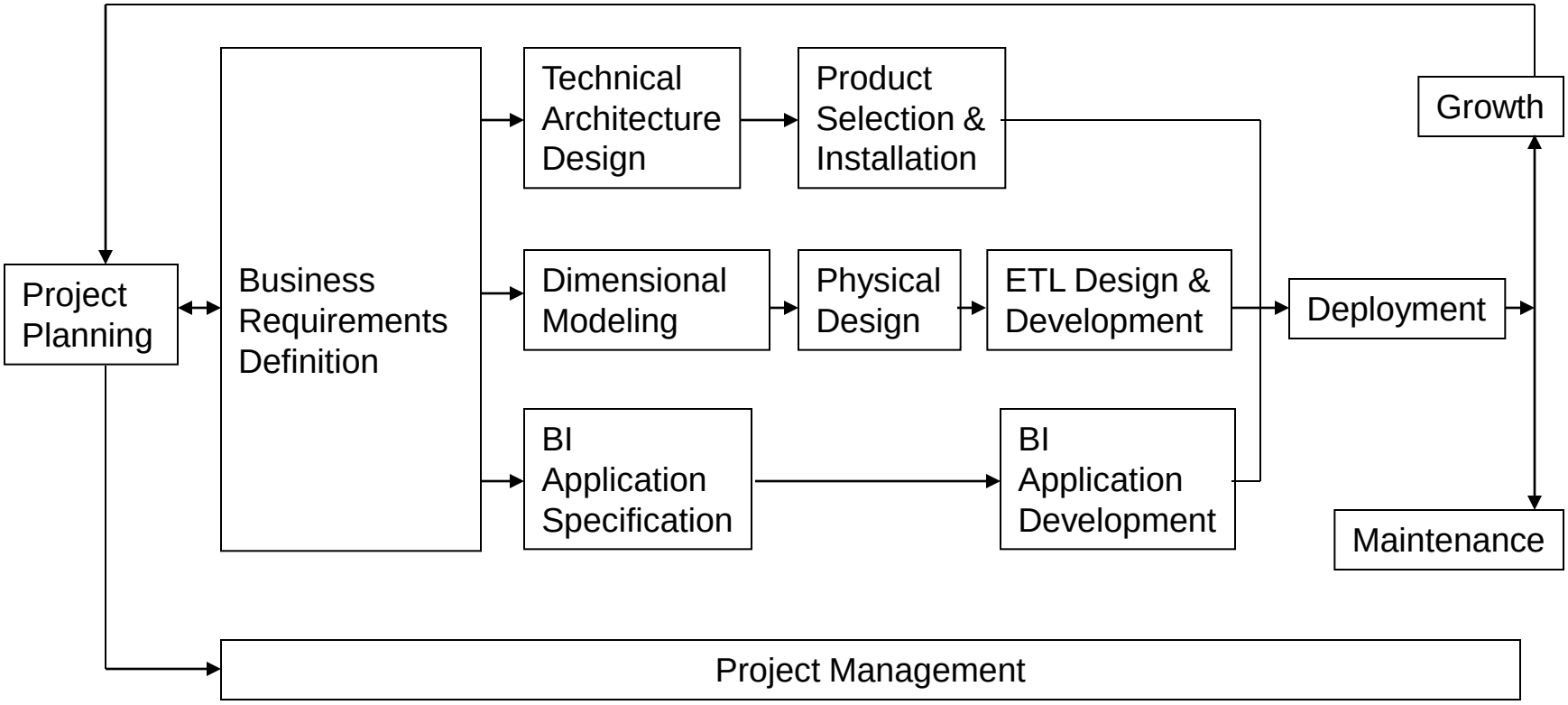
Characteristics of DW/BI

- High profile and high impact
- High risk
- Highly political
- Requires sophisticated and complex data gathering
- Requires intensive user access, training and support
- *high maintenance*

DW Lifecycle Principles

- Focus on the business
- Build an information infrastructure
- Deliver in meaningful increments: six to twelve month timeframes
- Deliver the entire solution: query and display tools in addition to the database

DW Lifecycle



Project Roles

- Business Sponsor* – *approves and pays for the project*
- DW/BI manager – *organizational DW sponsor*
- Project manager – *day to day leader*
- Business project lead – *business community interface*
- Business systems analyst – *business requirements*
- Data modeler – *detailed data analysis*
- Systems Architect – *system components*

Specialized Roles

- Data warehouse DBA
- OLAP designer
- ETL system developer
- DW/BI management tools/Application developer

DW Infrastructure

innovate achieve lead

Issues:

- Performance drain on the operating environment
- Technical skills of the data warehouse implementers
- Operational issues such as funding requirements
- Shop standards

DW Infrastructure

innovate achieve lead

Platforms:

- Source system
- Staging area
- Application server
- Desktop tools

DW Infrastructure
Database server

innovate achieve lead

Size: 500 GB to 250 TB
- < 500 GB is small
- 500 to 5 TB is medium
- over 5 TB is large

Volatility: what is the nature and frequency of the update process

Users: the number of users as well as their level of knowledge

Number of business processes (marts.)
- In some cases there may be a separate platform for each one.
- There may be an additional central server for management roll ups

DW Infrastructure
Database server

Nature of use

- ad hoc queries from power users
- standardized queries
- data mining

Technical support: the hardware, the operating system and the database engine may all require specialized support

Software requirements may dictate platform.

DW Infrastructure
Operating systems

innovate achieve lead

- Mainframes
  - transaction oriented
  - complex administration
  - not parallel
- Open system (UNIX) servers
  - specialized environment
- NT servers
  - relatively small capacities (limited numbers of processors and less efficient performance)

DW Infrastructure Hardware

Single processor at a time system

Symmetric Multiprocessing (SMP)

- Multiple processors
- Shared memory
- Common bus

Massively Parallel Processing (MPP)

- Multiple processors
- Distributed memory
- Distributed bus

DW Infrastructure Performance

- Indexing
- Physical organization
- Caching and blocking
- Data distribution
- Memory
- Chip architecture

DW Infrastructure
Database Engine

Relational

- Well understood
- Include DW support for star joins and fast access
- Flexible

Multidimensional (MOLAP)

- Extremely fast
- Pre-calculated combination facts

DW Infrastructure Front Room

Configuration of desk-top

Client/Server

Web

Supplemental tools

DW Infrastructure
Operational management

- Load window (operational scheduling)
- Backup
- User support
- Change management