Satyam
What Business Demands.

Sizing Logical Data in a Data Warehouse – A Consistent and Auditable Approach

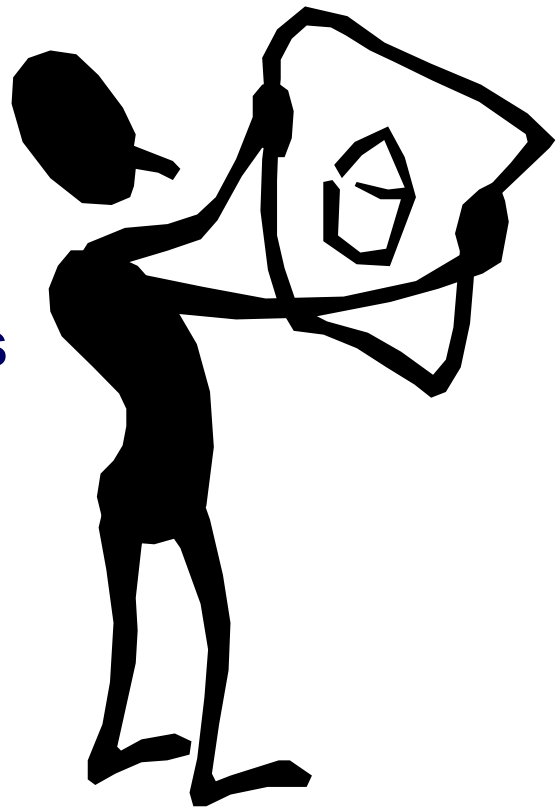
Priya Lobo CFPS
Satyam Computer Services Ltd.
69, Railway Parallel Road, Kumarapark West,
Bangalore – 560020, INDIA
Priya_Lobo@satyam.com
www.satyam.com

14th September 2006

Agenda

Satyam
What Business Demands.

- **Why Domain-specific guidelines for Data Warehouses ?**
- **Unique Data Warehouse Characteristics**
- **Quick Review of Related IFPUG 4.2 rules**
- **Sizing Data in a Data Warehouse, using IFPUG 4.2 rules**
- **Conclusion**
- **Questions**



Why Domain Specific Guidelines for Data Warehouses?

Satyam
What Business Demands.

- **IFPUG's CPM 4.2 provides a generic set of Software Sizing practices intended for sizing every domain or application**
- **Domain-specific counting guidelines are necessary to size functionally unique applications**
- **Data Warehouse (DW) systems are a special context for the application of software functional measurement and hence require domain specific sizing guidelines**

Why Domain Specific Guidelines for Data Warehouses?

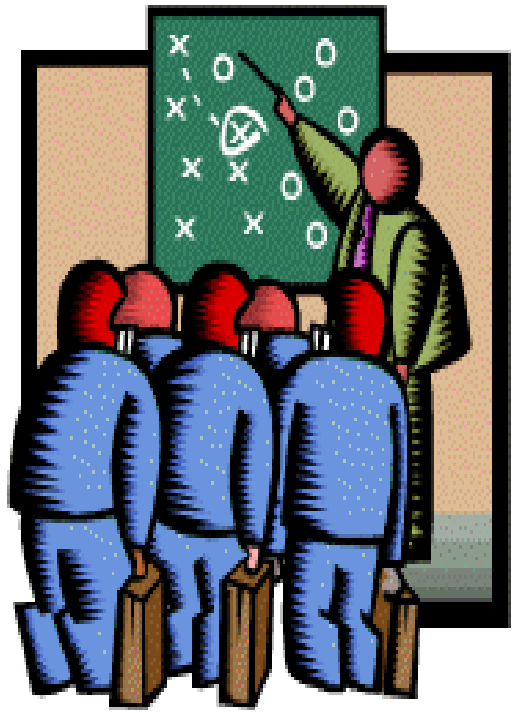
Satyam
What Business Demands.

- The Data Component within a DW is unique (different from traditional, operational systems)
- Existence of different Data Structures, Data Types and Data Redundancy in a DW make it difficult to force a logical view of the Data
- A Consistent and Auditable approach is important to sizing logical data in a DW
- This approach, based on IFPUG 4.2 has been successfully adopted at multiple sites both for development and enhancement DW projects

Why Domain Specific Guidelines for Data Warehouses?

Satyam
What Business Demands.

Prior to Sizing Logical Data in a DW, using IFPUG 4.2 rules, lets look at some Unique Data Warehouse Characteristics and Related IFPUG 4.2 rules …



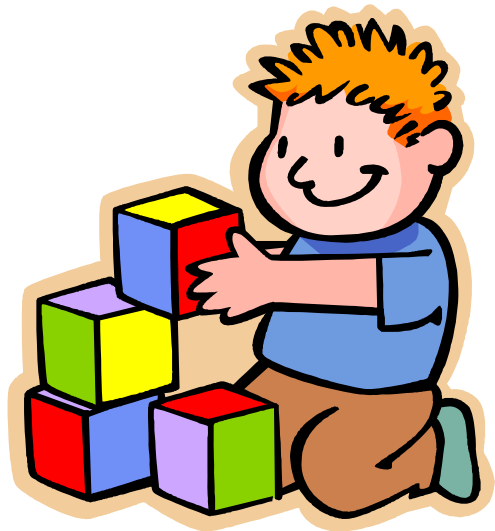
Unique Data Warehouse Characteristics

Satyam
What Business Demands.

- A Data Warehouse is a logical collection of information gathered from many different operational databases, to create a business intelligent application

- Contains cleansed and organized data to support business analysis activities and decision-making tasks

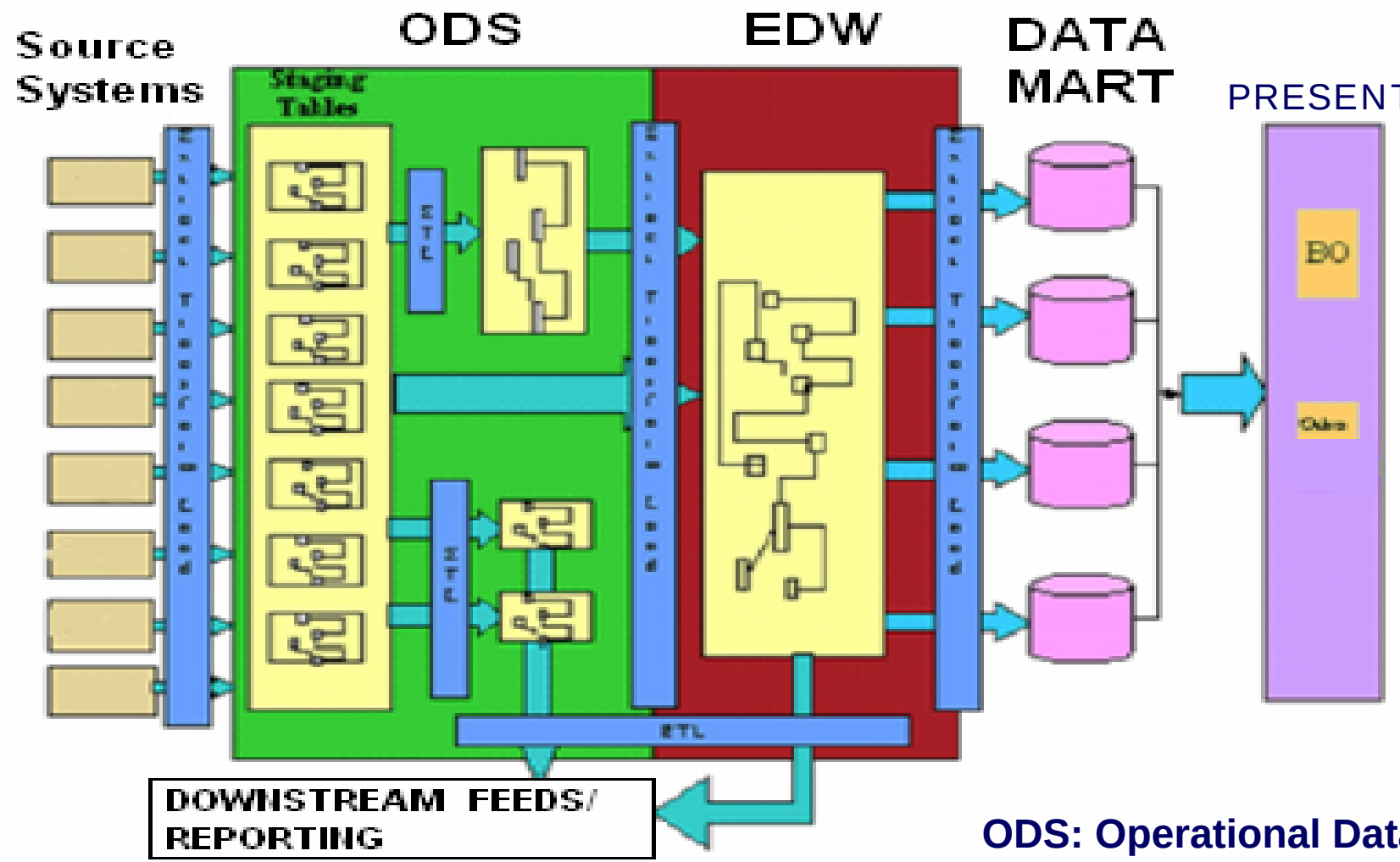
- Periodically, one imports data from enterprise resource planning (ERP) systems, mainframe systems and other related business software systems into the data warehouse for further processing



Unique Data Warehouse Characteristics

Satyam
What Business Demands.

DATA WAREHOUSE STRUCTURE



ODS: Operational Data Store

EDW: Enterprise DW

Unique Data Warehouse Characteristics

Satyam
What Business Demands.

Enterprise Data Warehouse (EDW)

- Contains data captured from operational systems, cleaned, transformed, integrated and loaded into a separate Subject Oriented database
- Data is accumulated to show a historical record

Data Mart (DM)

- Subset of corporate data (historical, summarized) that is of value to a specific business unit or set of users
- Contains multi dimensional data (can be hybrid as well)
- Dependent Data Mart - data loaded from EDW
- Independent Data Mart - data loaded directly from ODS

Unique Data Warehouse Characteristics

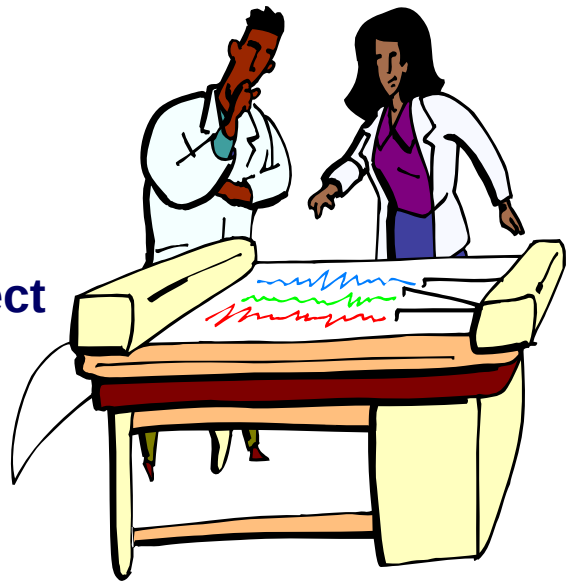
Satyam
What Business Demands.

Source Systems

- Core operational systems that feed data into the DW

Operational Data Store (ODS)

- Stores data received from source systems
- Is an extract of the operational data



Presentation Area

- Area where the organized data is available for direct querying by users & data access tools
- Data is dimensional and atomic

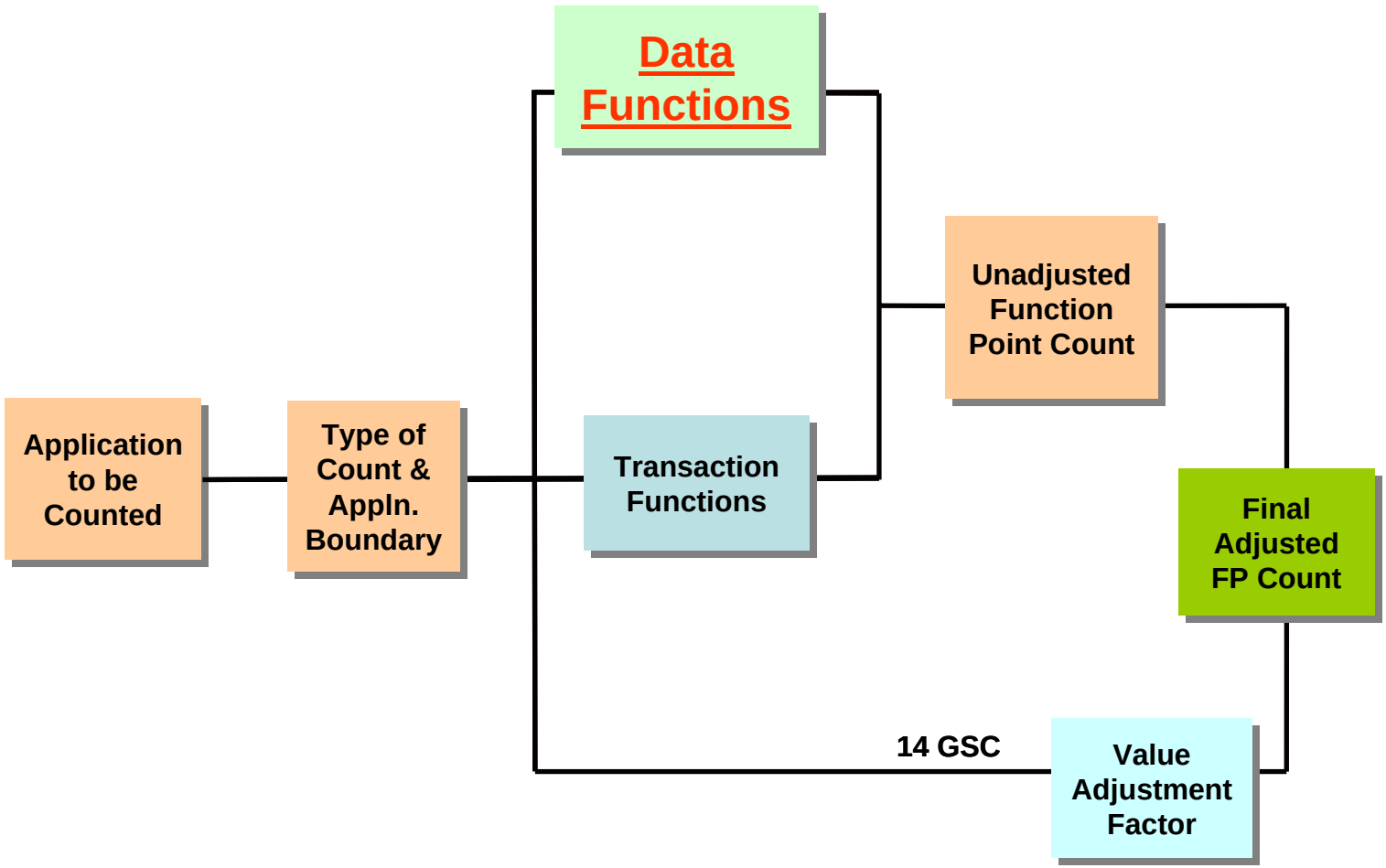
Extraction, Transformation and Loading (ETL)

- Set of processes by which the operational source data is prepared for the DW

Quick Review of Related IFPUG 4.2 rules

Satyam
What Business Demands.

FP Counting Process

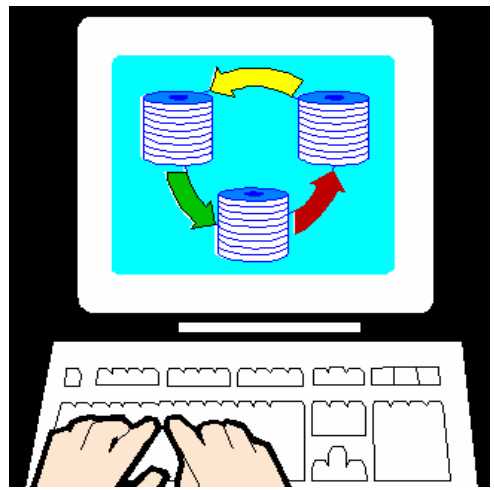


Quick Review of Related IFPUG 4.2 rules

Satyam
What Business Demands.

Data Functions

- Represents the functionality provided to the user to meet internal and external data requirements
- Defined as Internal Logical Files (ILF) and External Interface Files (EIF)
- ILF – User identifiable group of logically related data or control information maintained within the boundary of the application
- EIF – User identifiable group of logically related data or control information referenced by the application, but maintained within the boundary of another application



Quick Review of Related IFPUG 4.2 rules

Satyam
What Business Demands.

Logical Data Types

Business Data

- Core User data; Business Objects
- Satisfies functional user requirements
- Eg: Customer file, Job File, Invoice File

Reference Data

- Supports business rules for the maintenance of business data
- Satisfies functional user requirements
- Eg: Tax Rates, Threshold Settings

Code Data (List Data)

- Usually represents technical requirements
- Eg: State Code, Payment Type code

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

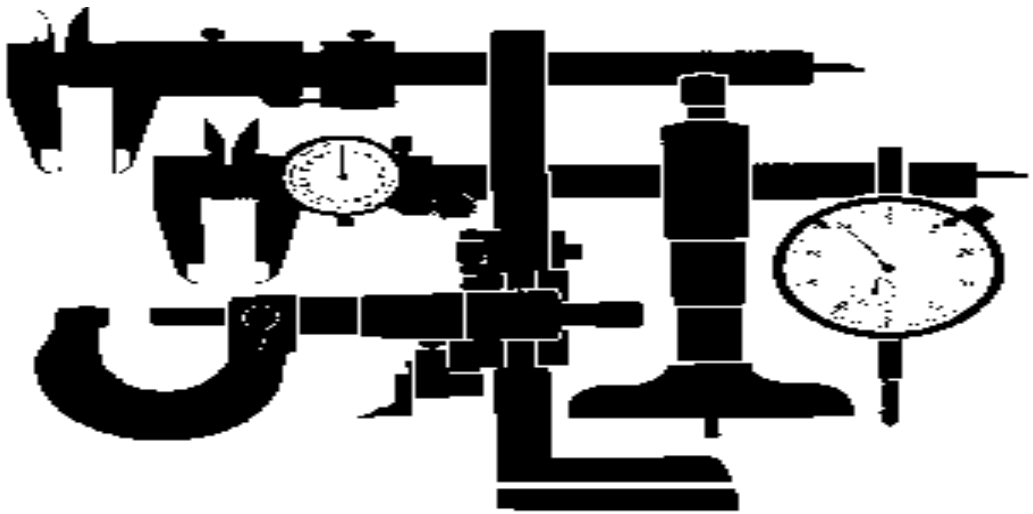
SUGGESTED SIZING APPROACH

[1] Basic Philosophy

[2] Table (Data) Types found in a Data Warehouse

[3] Multidimensional Data Structures

[4] Lessons Learnt



Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

[1] Basic Philosophy

- **For a DW table to be included as a Logical Data File it must display Persistence, have at least one load function which processes the input data or be referenced by at least one output function**
- **The Count should be “Inclusive” rather than “Exclusive”**
- **Philosophy of sizing data is not whether a ‘table’ is to be counted but whether or not the data in the table, when interpreted logically can be viewed as business data, and if so has not been included elsewhere in the count**

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Data in a DW is (generally) defined as belonging to :

- **Operational Data Store (ODS)**
- **Enterprise Data Warehouse (EDW)**
- **Data Mart (DM)**
- **Presentation Layer (Cubes may be present here)**

If the data in any of these areas show persistence and have at least 1 Process loading it, then count as logical files, using the respective approach for the table type or multidimensional schema as suggested in the following slides…

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

[2] Table (Data) Types found in a Data Warehouse

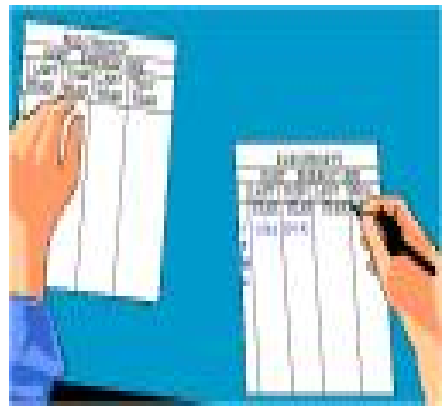
- Metadata
- Translation tables
- Staging tables
- Temporary tables
- Aggregates
- Logs/Event Files
- Release or Publish tables
- View tables
- Existence tables
- Facts
- Dimensions

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Metadata

- Data about Data
- Keeps track of what is where in a DW
- Examples - Data Dictionary Files, Event Files, User Profiles Files
- Reference Data as defined in CPM 4.2



Count metadata that is recognizable to the user (including administrator) as Logical files.

If maintained within the application count as ILF

If referenced (not maintained) by the application count as EIF

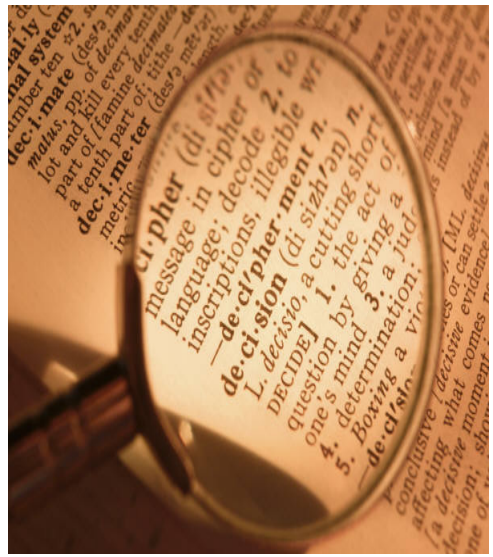
Do not count technical metadata like 'update frequency', 'physical-logical' file mapping

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Translation Tables

- Holds translation information used during Data Load process
- Usually exists only for implementation purposes
- Code Data as defined in CPM 4.2



Do not count, as it does not contribute to functional size

As with Code data, at the maximum only one translation table may be counted per application

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Staging Tables, Temporary tables

- Contains Temporary Data
- Just another instance of same data counted elsewhere in the application
- Not User recognizable

Do not count Staging and Temporary tables as Logical Files

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Aggregates

- Contains Aggregated data created for use by other processes or for performance reasons
- Usually exists as Aggregate Fact table

Count each Aggregated fact group as 1 Internal Logical file, if used by other processes

Do not count Aggregate tables, if created for performance reasons

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Views

- Query statements that create logical copies of tables

Do not count Views as Logical Files

Existence (Exception) Tables

- Defines business rules (Eg.- which products will be offered in which regions during which promotions)
- Reference Data as defined in CPM 4.2

Count as Logical files.

If maintained within the application count as ILF

If referenced (not maintained) by the application count as EIF

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Release/Publish tables

- Data is assembled for further use
- Usually data stored here is temporary (not persistent)

Do not Count Release/Publish tables as Logical files (unless it contains persistent data)

Logs/Event Files

- Metrics data about the DW may be logged
- Maintains information about the currency of the data within the DW.

Count as Internal Logical Files

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Fact tables (Measures)

- Factual or quantitative data
- Contain detailed or summarized data values
- Links to dimensional data
- Is usually meaningful to the user only when linked to dimensional data
- Examples - sales, profits, counts, claims
- Business Data as defined in CPM 4.2

Count each fact table as an RET in a "logical" star schema/cube

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Dimension tables (Attributes)

- Describes business data in a Fact table
- Usually organized into hierarchies (to roll up and drill down)
- Is usually meaningful to the user only when linked with fact data
- Examples - products, regions, time period (months roll up to quarters & quarters to years)

Count each dimension table as an RET in a "logical" star schema/cube (which is an ILF)

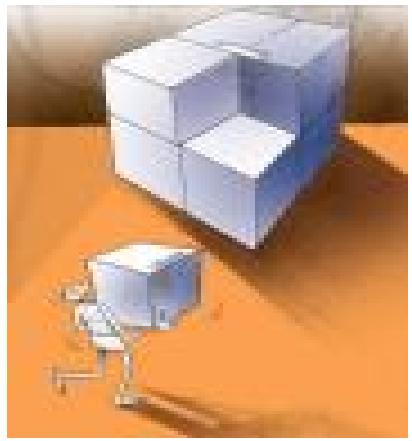
If it serves additional business purpose, Count as a separate ILF

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

[3] Multi-dimensional Data Structures

- **Categorizes data in order to enable business analysis**
- **Has two components - Fact and Dimensions**
- **Common Design Schemas**
  - **Star**
  - **Snowflake**
  - **Cube**



Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

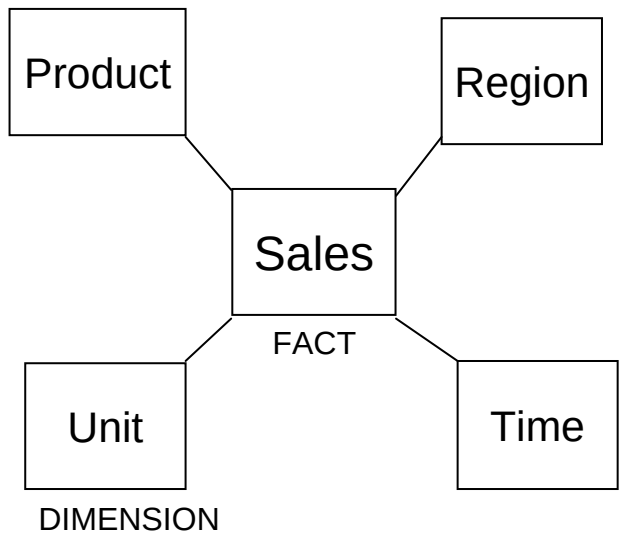
Star Schema

- **Generic representation of a dimensional model in a relational database**
- **Multiple Dimensions provide descriptive information on data in the central Fact table**

Count each star schema as an ILF.

Complexity :

- **Count each Fact subgroup as 1 RET**
- **Count each Dimension subgroup as 1 RET**
- **If a Dimension is used in multiple star schemas, count each as 1 RET**
- **Do not count redundant DETs that occur in the same star**



Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

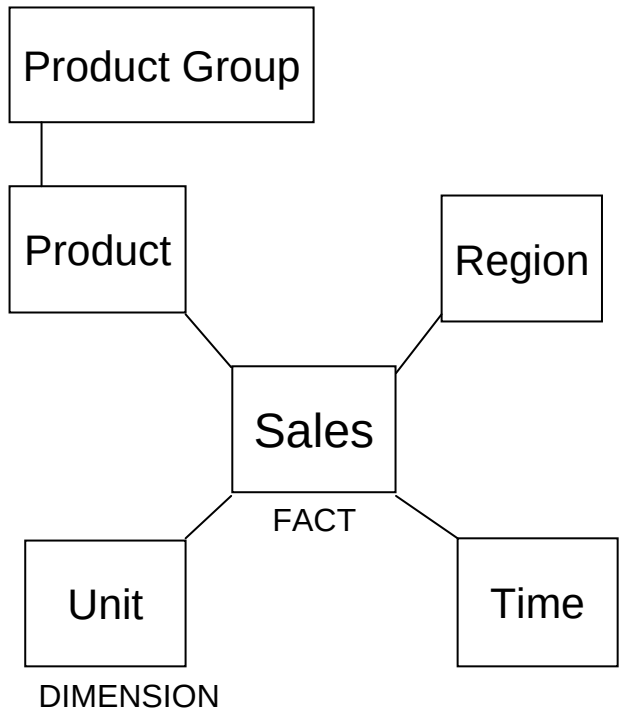
Snowflake Schema

- **Expanded version of (Adds Hierarchy to) the Star Schema**

Count each snowflake schema as an ILF.

Complexity :

- **Count each Fact subgroup as 1 RET**
- **Count each Dimension subgroup as 1 RET**
- **Where the hierarchical dimensions are exploded to their levels, the second order tables do not represent other RETs**
- **Do not count redundant DETs that occur in the same snowflake**

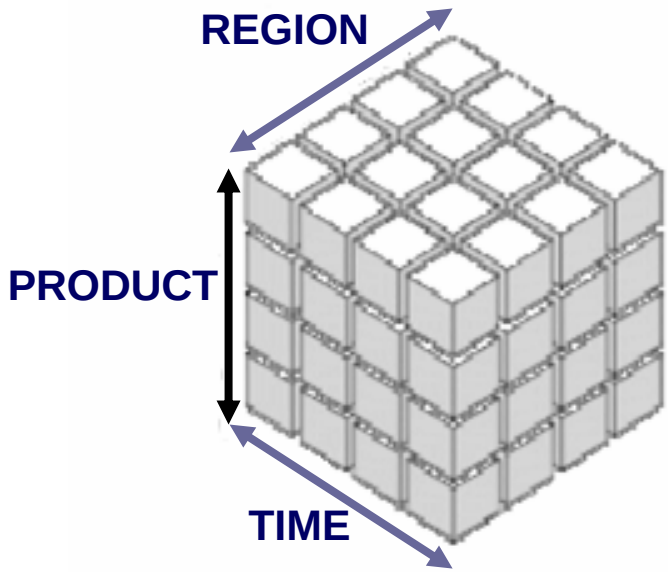


Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

Cube

- A multidimensional cube represents a similar structure with the cube axes representing Dimensions and the cube the Facts



Count each cube as an ILF.

Complexity :

- Count each Fact subgroup as 1 RET
- Count each Dimension subgroup as 1 RET
- Do not count redundant DETs that occur in the same cube

Sizing Data in a DW using IFPUG 4.2

Satyam
What Business Demands.

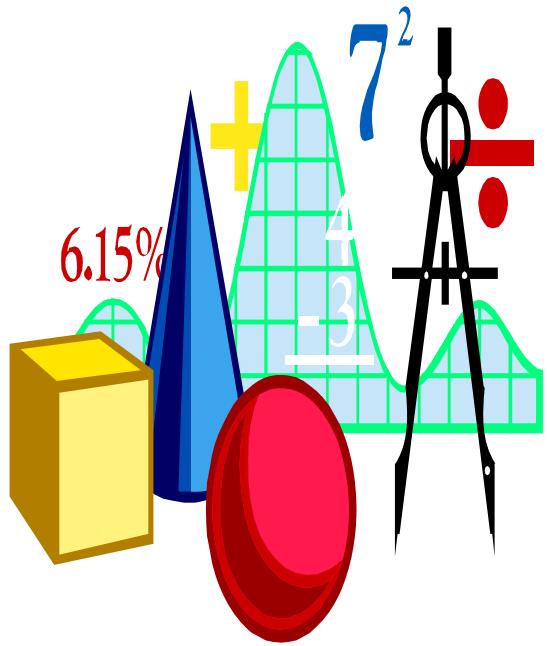
[4] Lessons Learnt

- **Document a local DW Sizing guideline (as each DW has its own uniqueness) to ensure consistency across the organization**
- **Clearly segregate "Code Data" from Business and Reference data as it may often be confused with Reference data in the DW**
- **Use an "inclusive" approach – Consider all data groups (unless no justification may be found in terms of rules for doing so)**
- **Maintain a list of Logical-to-Physical file (temporary and persistent) Mapping as part of the count documentation. This helps identify the logical data groups impacted by an enhancement to the DW**

Conclusion

Satyam
What Business Demands.

- **Suggested domain-specific approach is consistent with IFPUG's 4.2 counting guidelines**
- **Suitable for both development and enhancement projects – this approach has been successfully adopted at multiple sites in both scenarios**
- **This "inclusive" approach ensures all logical data is addressed**
- **The Size obtained through this approach is in line with Industry data - this is indicated by the productivity indicator.**

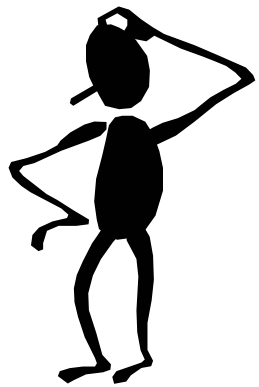


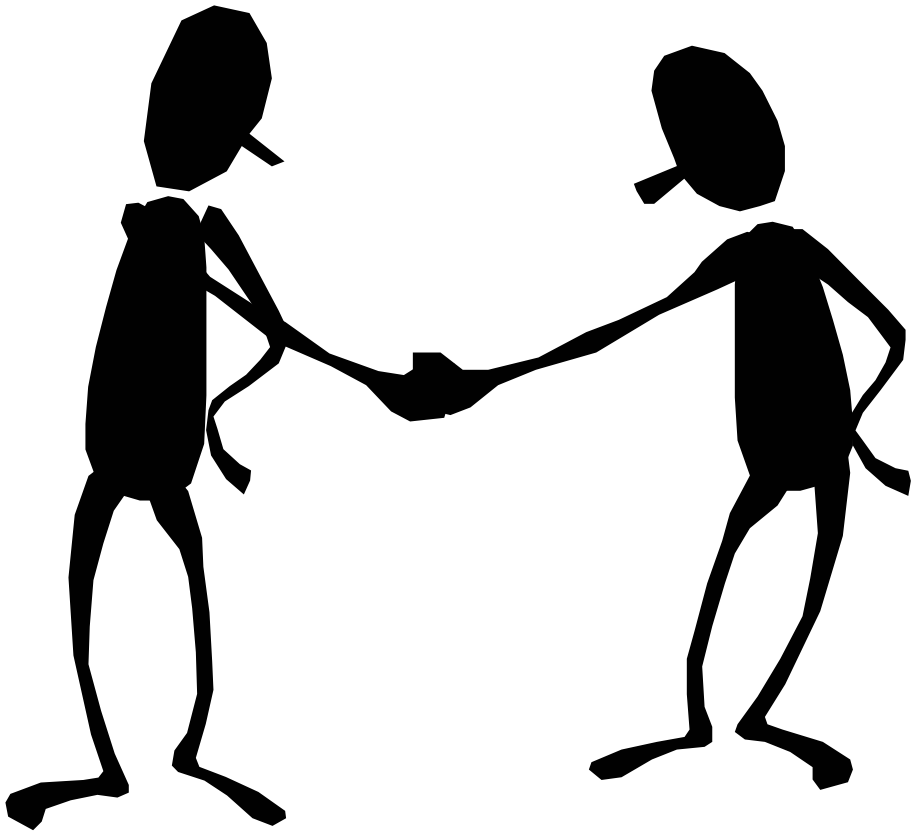
References

Satyam
What Business Demands.

- **IFPUG, *Function Point Counting Practices Manual Release 4.2***
- **W. H. Inmon, *Building the Data Warehouse***
- **Ralph Kimball, *The Data Warehouse Toolkit (Second Edition)***
- **Luca Santillo, *Size & Estimation of Data Warehouse Systems***

Questions???





Thank you