

# Understanding Data Warehousing

[by Alex Kriegel]

2008

# Things to Discuss

- Who Needs a Data Warehouse?
- OLTP vs. Data Warehouse
- Business Intelligence
- Industrial Landscape
- Which Data Warehouse: Bill Inmon vs. Ralph Kimball
- Relational vs. Dimensional
- Data Warehouse Life Cycle

# Data Concepts

- Data vs. Information
  - Data – an observable and recordable fact
  - Information – integrated collection of data used for decision making/analysis
- Data = OLTP (On-Line Transaction Processing)
  - Optimized for insert/update/delete speed
  - Contains transient detailed data
  - Often used for analysis
- Information = OLAP (On-Line Analytical Processing)
  - Optimized for querying
  - Contains aggregated data
  - Is NOT (usually) used for transaction processing

# Problems with OLTP Reporting

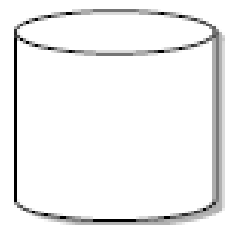
- Data Credibility
  - No time basis
  - Levels of extraction
  - Different reporting criteria
- Productivity
  - Locate information
  - Analyze and custom extract data
- Inability to Transform Data into Information
  - Lack of historical information
  - Absence of relevant data

# OLTP vs. Data Warehouse

<b>OLTP</b>	<b>Data Warehouse</b>
Application/Transaction oriented	Business/Subject oriented Serves managerial community
Thousands of users Run repetitively	Few users (typically under 100) Run heuristically
Current data Accurate, at the moment of access	Historical data Snapshots in time
Normalized data (many tables, few columns per table, no redundancy)	De-normalized data (few tables, many columns per table; redundancy is a fact of life)
Static structure, variable content	Flexible structure
Compatible with SDLC Managed in entirety	Different Life Cycle Managed by subsets
Continuous updates High availability Small amounts of data used Simple to complex queries	Batch updates* Relaxed availability Large amounts of data used Usually very complex queries

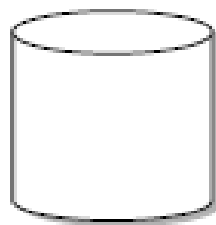
# Data Evolution

## LEVELS OF THE ARCHITECTURE



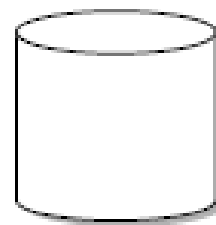
Operational

- Detailed
- Day to day
- Current valued
- High probability of access
- Application-oriented



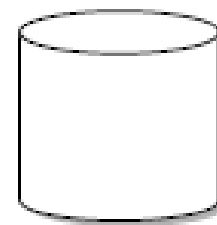
Atomic/data  
warehouse

- Most granular
- Time variant
- integrated
- Subject-oriented
- some summary



Departmental

- Parochial
- Some derived; some primitive
- Typical departments
  - Accounts
  - Marketing
  - Engineering
  - Actuarial
  - Manufacturing



Individual

- Temporary
- Ad hoc
- Heuristic
- Non-repetitive
- PC, work-station based

# Foundation of BI

- Data Warehouse and RDBMS
- Alternative data storage
  - Proprietary data structures
  - XML
  - N-Dimensional CUBE

# Tools of Trade

- Reporting
- Dashboards
- Ad-hoc Query
- Applications Integration (e.g. MS Office)
- Drill-down capability
- Advanced visualization

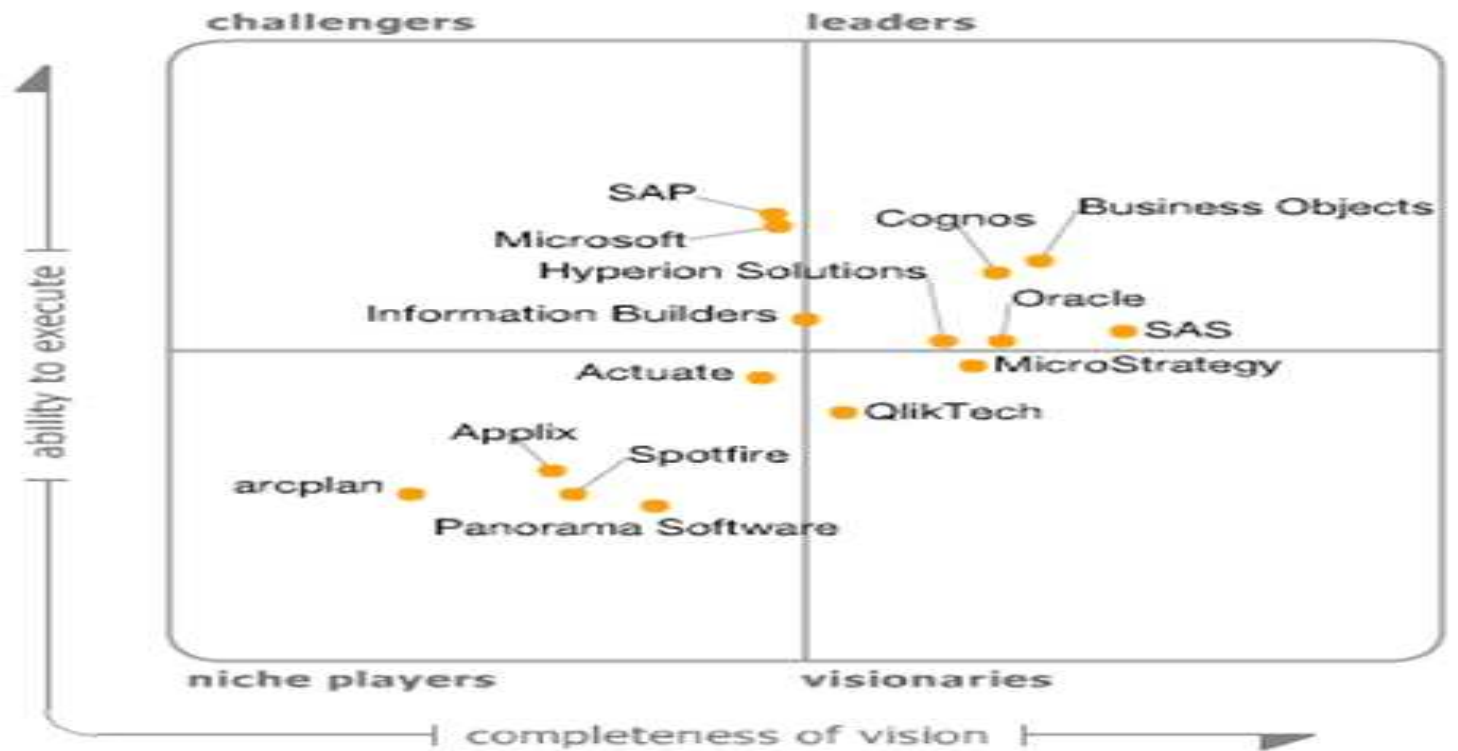
# BI : Industry Dynamics

- Pure BI players (hanging there...)
  - Actuate
  - Information Builders
  - Microstrategy
  - SAS
- Mergers and Acquisitions
  - Cognos (IBM)
  - Business Objects/Crystal Decisions (SAP)
  - ProClarity (Microsoft)
  - Hyperion/Brio (Oracle)

# OLAP Market Share (2006)

■ Microsoft Corporation	31.6%
■ Hyperion (Oracle)	18.9%
■ Cognos (IBM)	12.9%
■ Business Objects (SAP)	7.3%
■ MicroStrategy	7.3%
■ SAP AG	5.8%
■ Cartesis SA (SAP)	3.7%
■ Applix (Hyperion)	3.6%
■ Infor	3.5%
■ Oracle Corporation	2.8%

# Gartner's Magic Quadrant for Business Intelligence Platforms 2007



As of January 2007

# Two Approaches to Data Warehousing\*

- Bill Inmon CIF (Corporate Information Factory)
  - Relational (Normalized) DW
- Ralph Kimball Bus Architecture
  - Dimensional DW

\* Not mutually exclusive

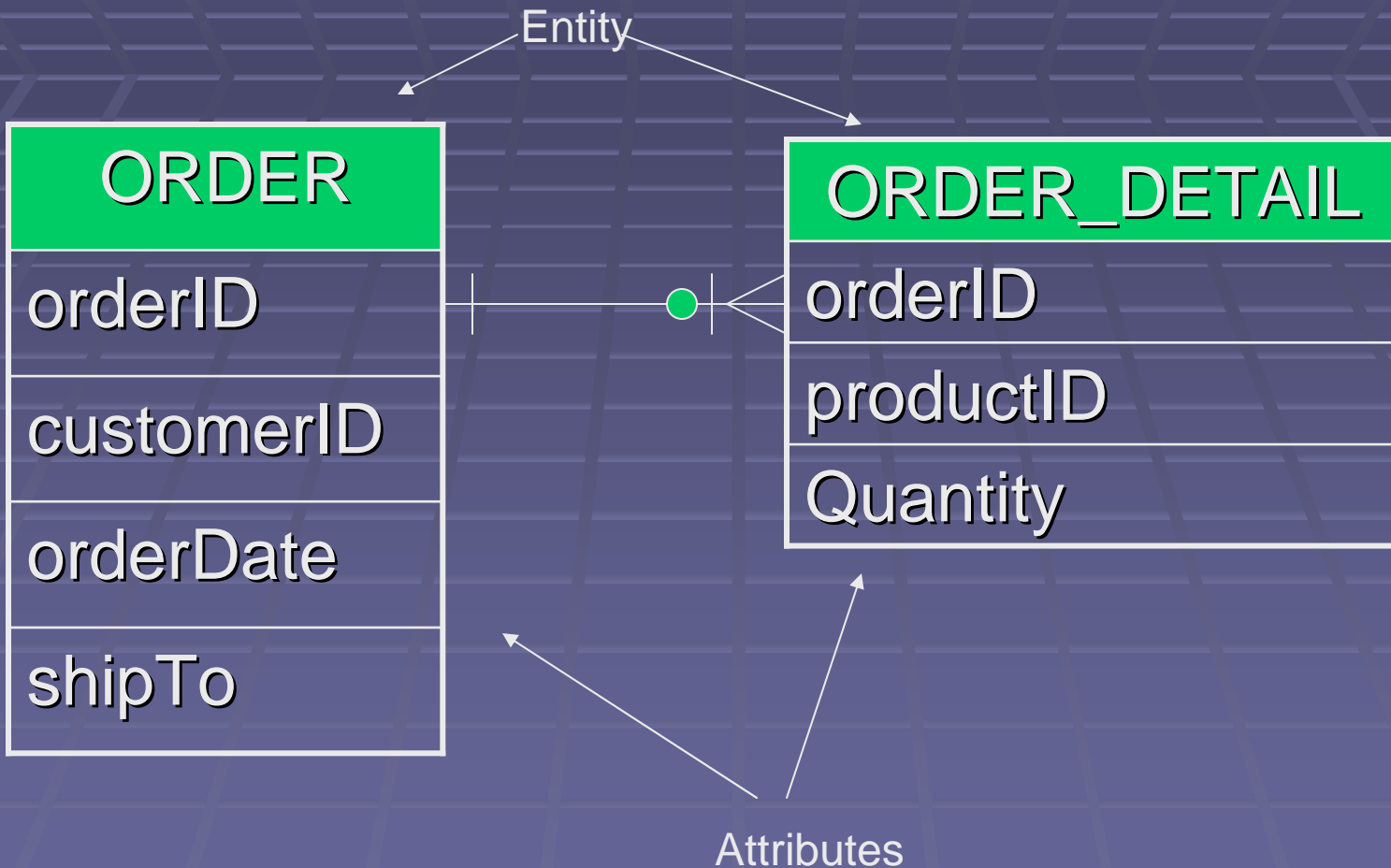
# Relational Concepts I

- Normalization
  - Granularity vs. Performance
- Entity/Relationship Diagrams
  - Entity
  - Attribute
  - Domain
  - Relation
  - Key

# Basic Relational Concepts II

- Relationships
  - Primary and Foreign Keys
  - One-to-One
  - One-to-Many
  - Many-to-Many
- Data Integrity
  - Domain Integrity
  - Entity Integrity
  - Referential Integrity

# Example of Relational Design



# DW Definitions I

- Data Warehouse
- Data Mart
- Staging Area
- Operational Data Store
- Star Schema
- Snowflake Schema
- ETL (Extract-Transform-Load)

# DW Definitions II

- Data Warehouse
  - Typically contains the full range of business intelligence available from all sources
  - Data is cross-divisional
- Data Mart
  - Typically contains specialized subset of data
  - Data is division specific

# DW/DM Common Characteristics

- Subject oriented (the data is organized around subjects)
- Nonvolatile (once placed in the warehouse, is not *usually* subject to change)
- Integrated (the data is consistent)
- Time variant (historical data is recorded)

# DW Definitions III

- Staging Area
  - temporary area used for data cleansing and transformations; usually structurally compatible with target schema
- Operational Data Store (ODS)
  - is a database designed to integrate data from multiple sources to facilitate operations, analysis and reporting
  - the integration often involves cleaning, redundancy resolution and business rule enforcement
  - Is updateable and contains highly granular information (as opposed to DW proper)

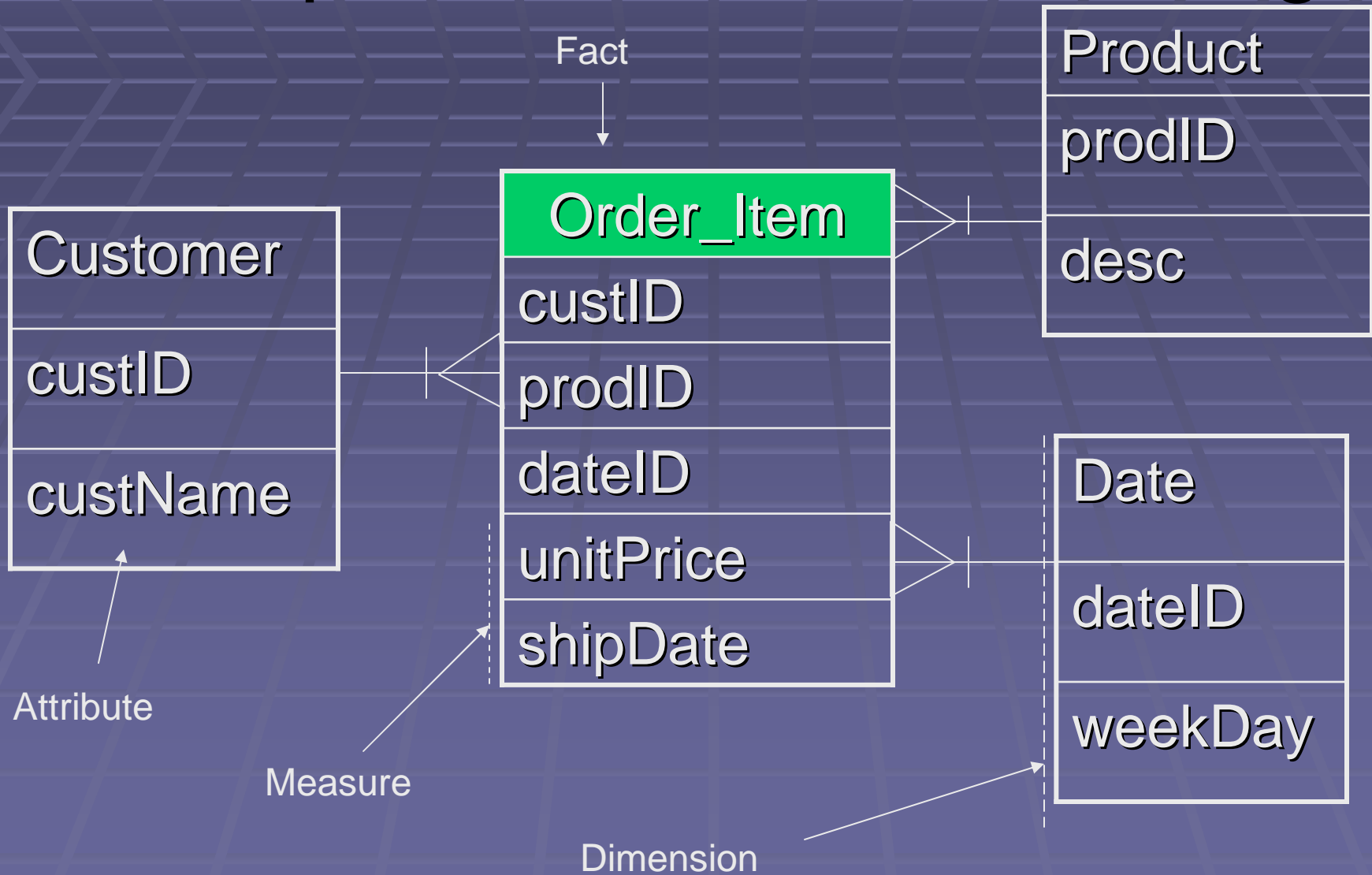
# DW Definitions IV

- Star Schema
  - Single FACT table surrounded by Dimension Tables
  - There could be many star schemas within database
- Snowflake Schema
  - Extension of star schema where one or more dimensions could be split into dimensions of their own

# Basic Dimensional Concepts

- Fact table contains
  - Dimension Keys (foreign key to Dimension tables)
  - Facts (actual data being measured)
  - Facts are preferably numeric
  - Grain
    - All facts in the FACT table must be of the same level of detail (table's grain)
- Dimension table contains
  - Actual data pertaining to the Fact table

# Example of Dimensional Design



# Basic Data Warehouse Components



# Data Warehouse Lifecycle Management (DWLM) I

- Project Planning
  - Readiness and risk assessment
  - Scope, Time, Resource
  - Roles and Responsibilities
- Gathering and Analyzing Requirements
- UI Analytics/Reporting Interface (technology)
  - Presentation Architecture: OLAP, ROLAP, MOLAP...

# Data Warehouse Lifecycle Management (DWLM) II

- DW Schema Design
  - Dimensional/Normalized Modeling
- Technical Architecture Design
  - Architectural Components and Services
  - Storage/staging consideration
  - Metadata Repository
  - Integrating Data Sources
    - Logical Mapping/Transformation Rules/Data Quality Requirements
    - ETL Process Design (logical/technology/workflow)
    - OLTP data profiling

# Data Warehouse Lifecycle Management (DWLM) III

- Implementation
  - Develop DW schemas
  - Develop ETL/Staging Schema
  - Develop ETL routines
  - Build and Populate Cubes
  - Build Dashboard/Reports (COTS/MDX Queries)
  - Testing (data/throughput/requirements)

# Data Warehouse Lifecycle Management (DWLM) IV

- Deployment
  - Alpha, beta, rollout release process
  - Set up/configuration
  - Training and support
  - Documentation
  - Sizing and Adjustments
  - Performance tuning

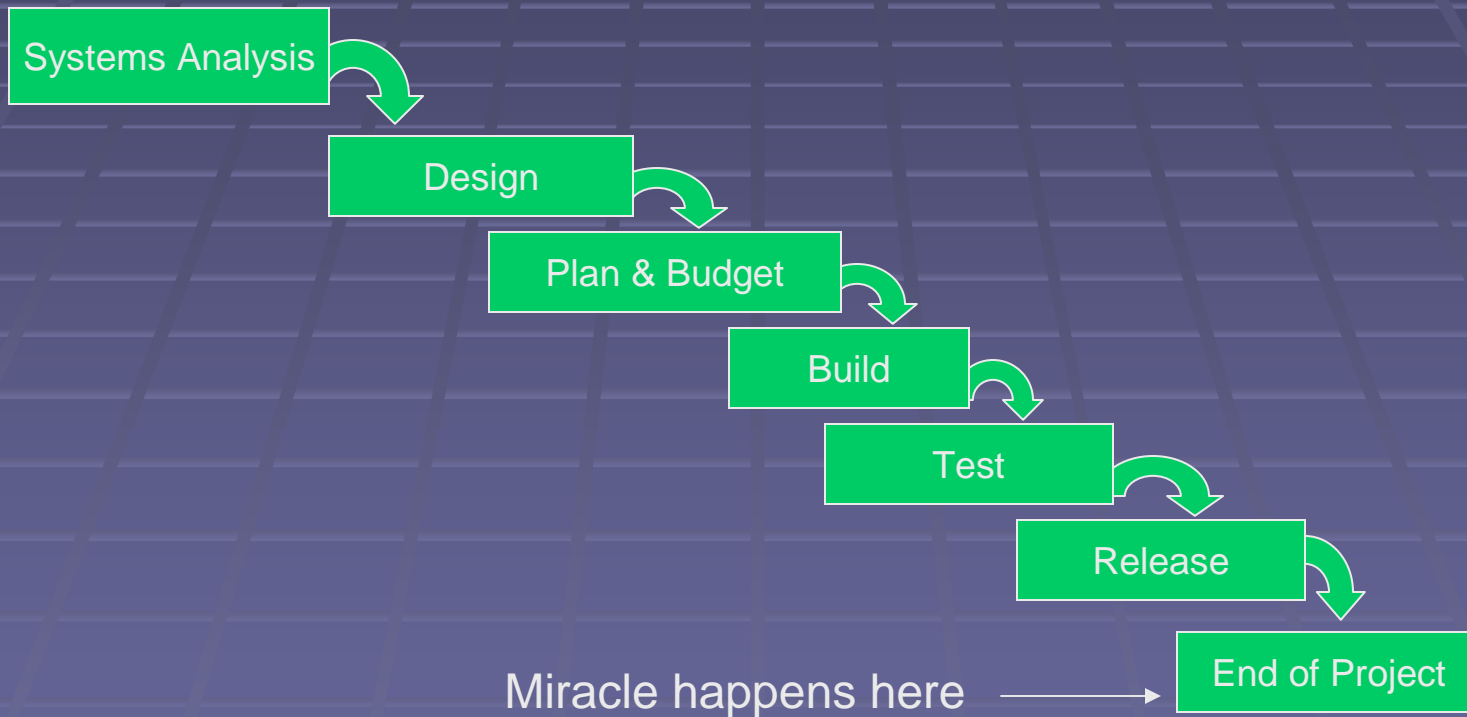
# Data Warehouse Lifecycle Management (DWLM) V

- Managing and Maintenance
  - Support/training
  - Tuning and Optimization
  - Planning for Growth/Enhancements

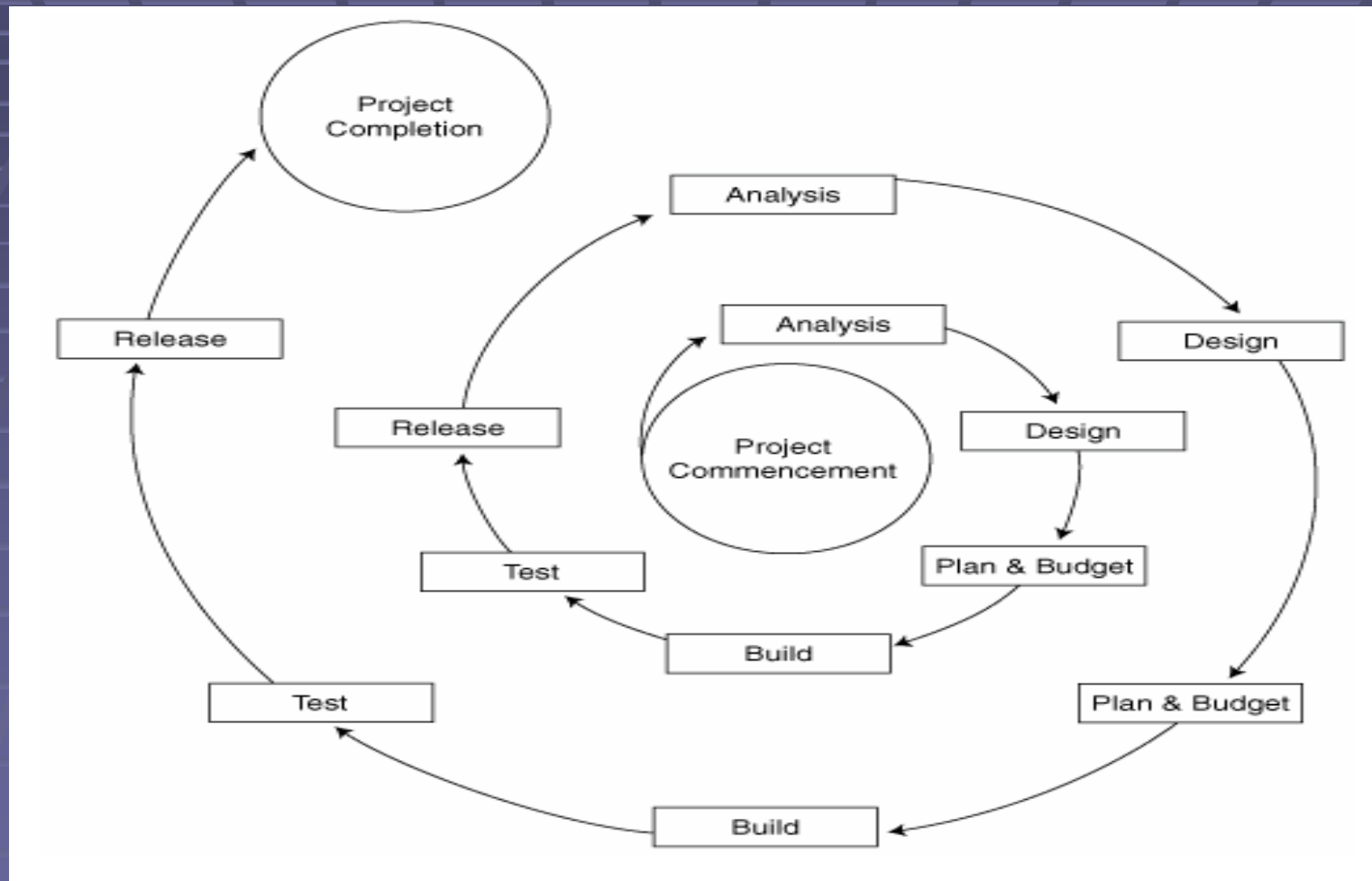
# DW Life Cycle Models

- Waterfall Model
- Spiral Model
- COTS Solutions
  - BusinessObjects
  - Cognos
  - Informatica
  - SAS
  - Kalido

# Waterfall Model



# Spiral Model



# Ralph Kimball's Four Principles

- **Focus on the business:** Concentrate on identifying business requirements and their associated value. Use these efforts to develop solid relationships with the business side and sharpen your business sense and consultative skills.
- **Build an information infrastructure:** Design a single, integrated, easy-to-use, high-performing information foundation that will meet the broad range of business requirements you've identified across the enterprise.
- **Deliver in meaningful increments:** Build the data warehouse in increments that can be delivered in 6 to 12 month timeframes. Use clearly identified business value to determine the implementation order of the increments.
- **Deliver the entire solution:** Provide all the elements necessary to deliver value to the business users. This means a solid, well-designed, quality-tested, accessible data warehouse database is only the start. You must also deliver ad hoc query tools, reporting applications and advanced analytics, training, support, web site, and documentation.