Data Warehousing for Developers

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

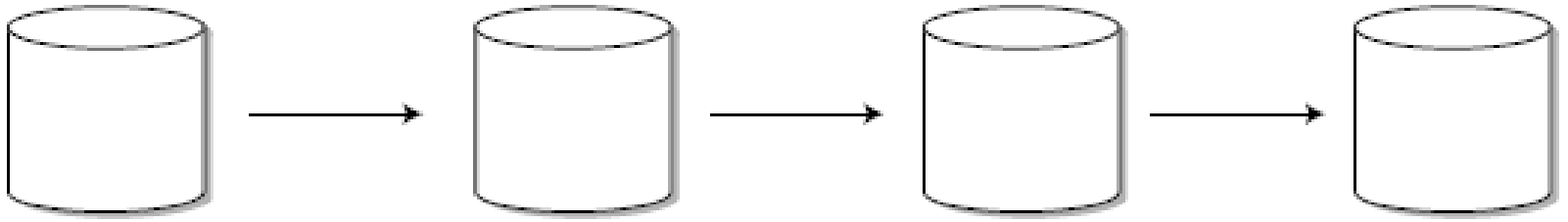
- Data Credibility
  - No time basis
  - Levels of extraction
  - Different reporting criteria
- Productivity
  - Locate information
  - Analyze and custom extract data
- Inability to Transform Data into Information
  - Lack of historical information
  - Absence of relevant data

OLTP vs. Data Warehouse

OLTP	Data Warehouse
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	Atomic/data warehouse	Departmental	Individual

• Detailed
• Day to day
• Current valued
• High probability of access
• Application-oriented
	• Most granular
	• Time variant
	• integrated
	• Subject-oriented
	• some summary
		• Parochial
		• Some derived; some primitive
		• Typical departments
		  • Accounts
		  • Marketing
		  • Engineering
		  • Actuarial
		  • Manufacturing
			• 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

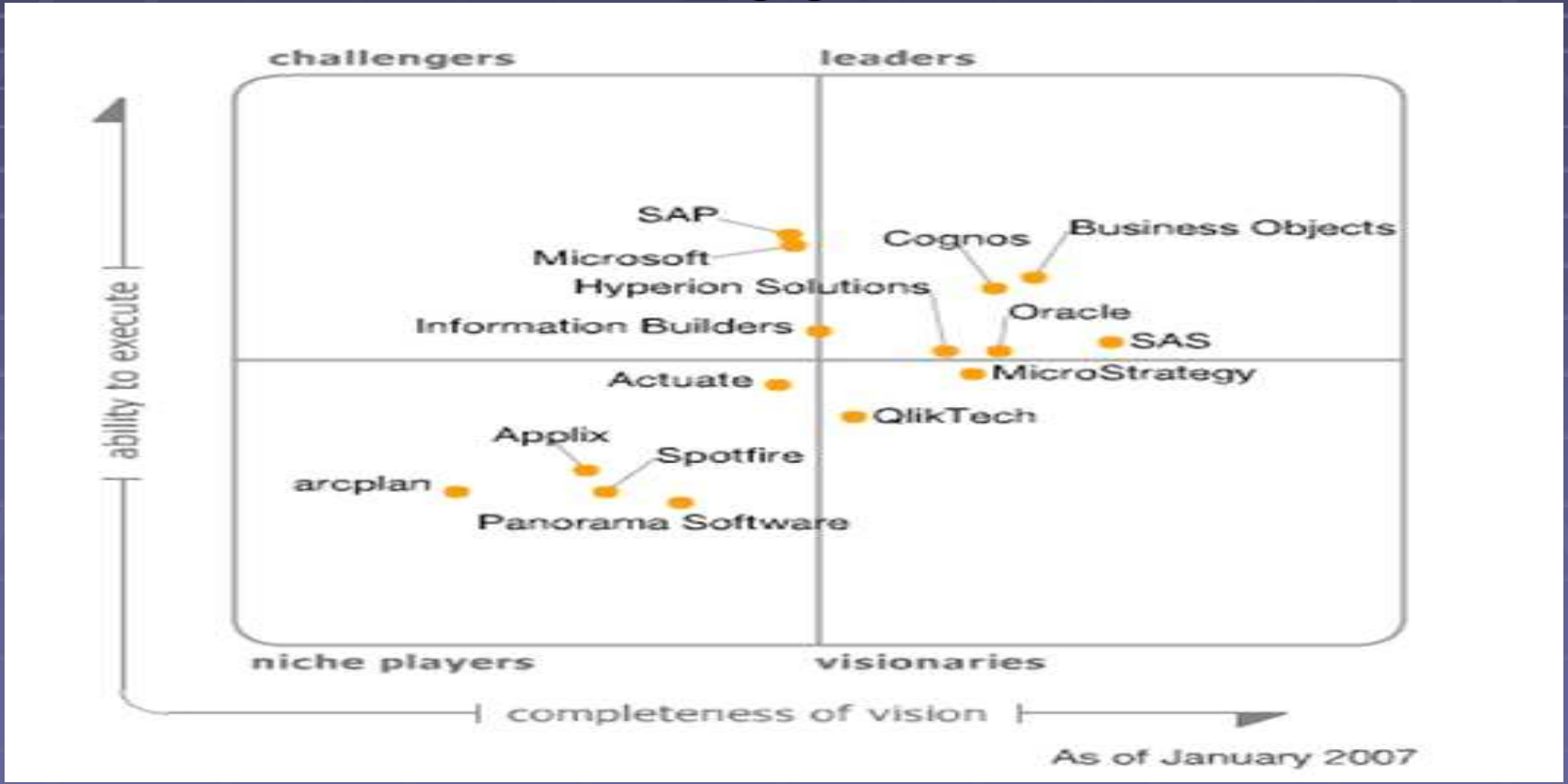
- Pure BI players (hanging there…)
  - Actuate
  - Information Builders
  - Microstrategy
  - SAS
- Mergers and Acquisitions
  - Cognos (IBM)
  - Business Objects/Crystal Decisions (SAP)
  - ProClarity (Microsoft)
  - Hyperion/Brio (Oracle)

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%

Source: OlapReport.com

Gartner's Magic Quadrant for Business Intelligence Platforms 2007



Two Approaches to Data Warehousing*

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

* Not mutually exclusive

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)

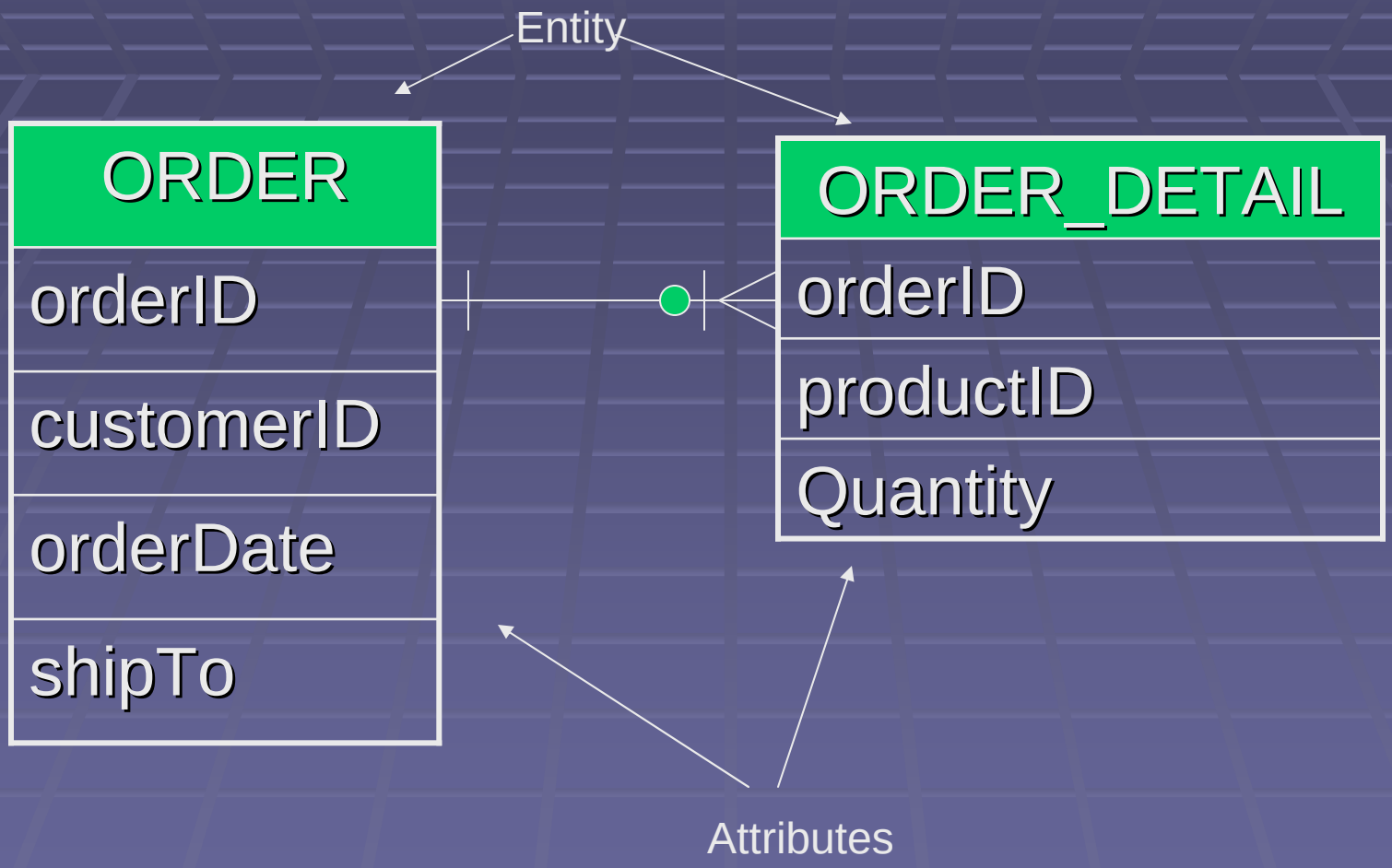
DW Definitions IV

- 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

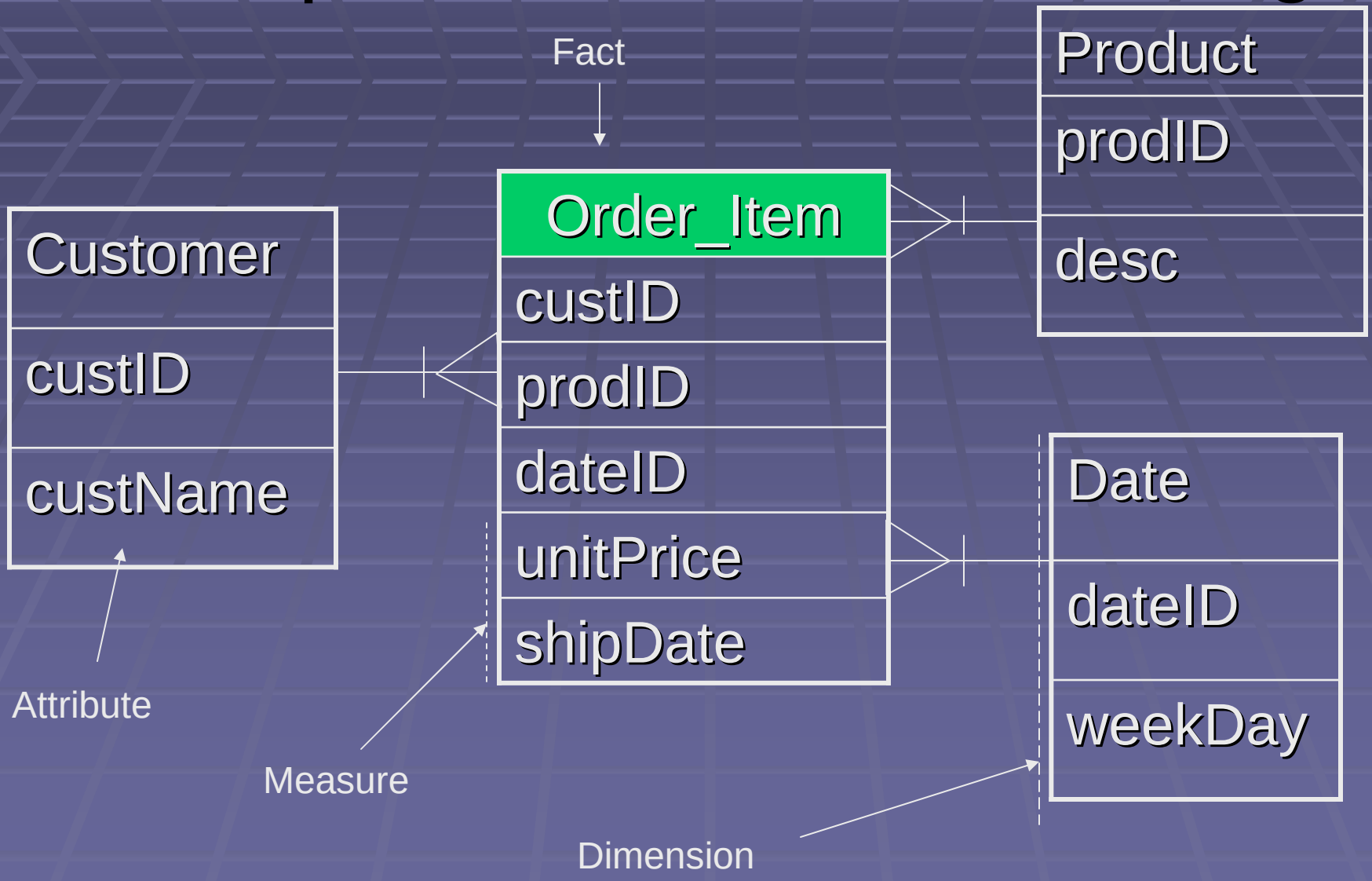
Basic Dimensional Concepts

- 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

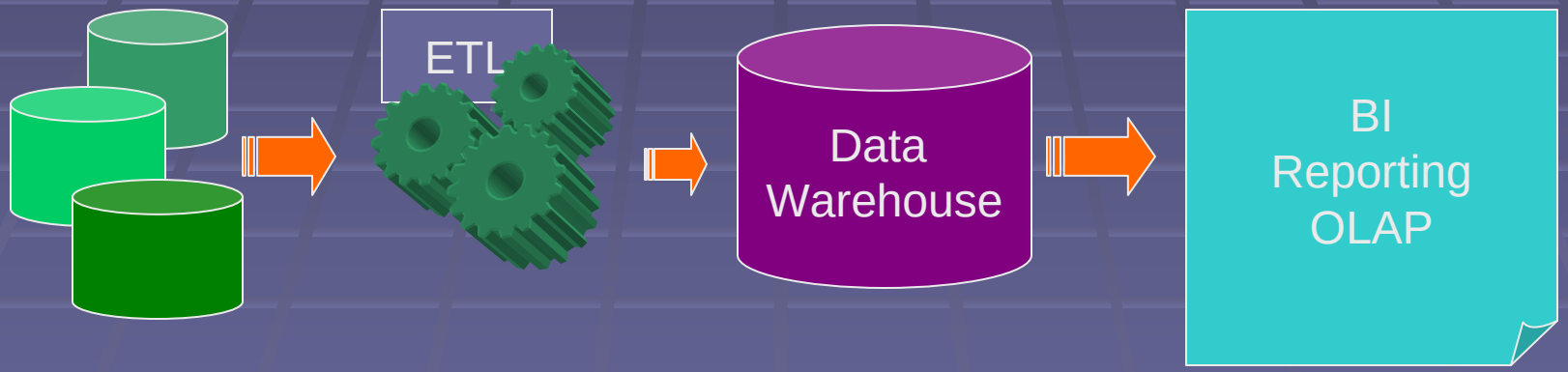
Example of Relational Design



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

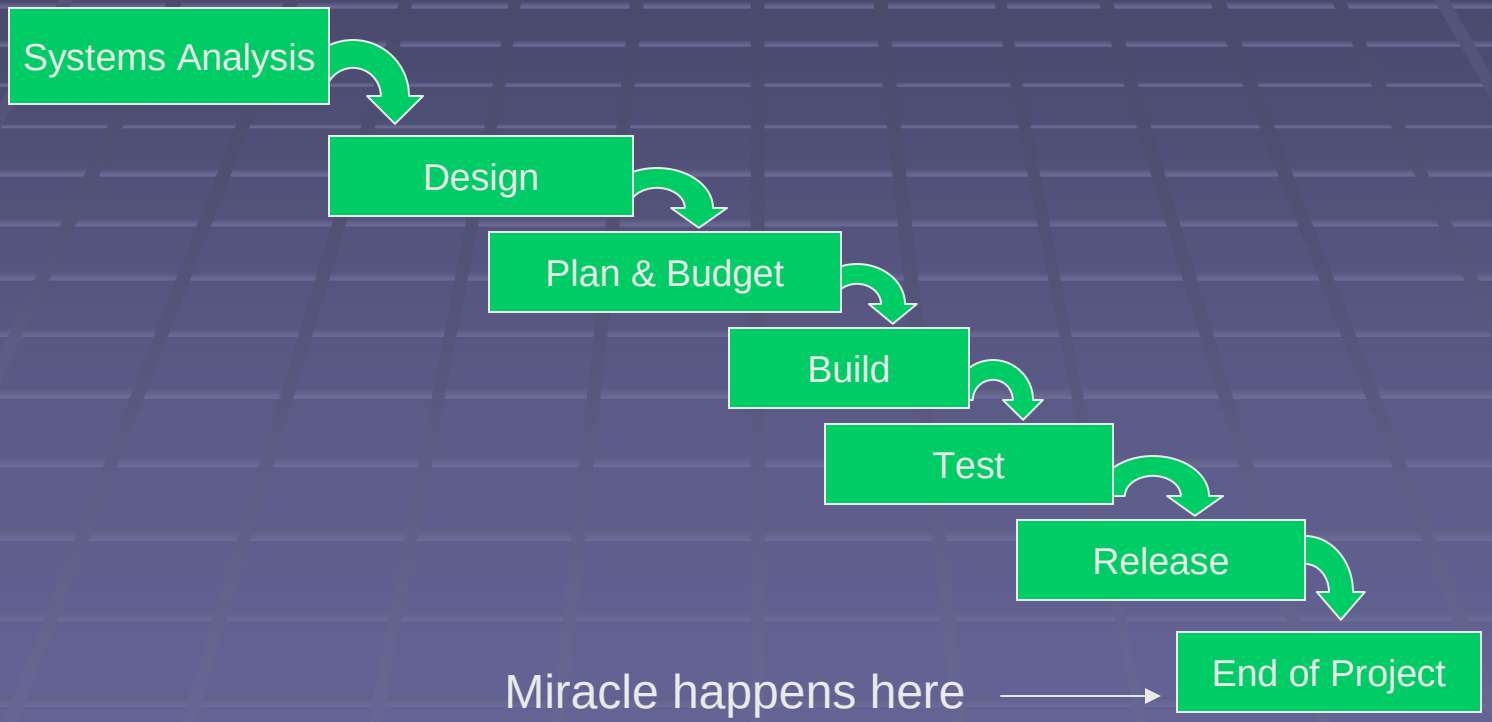
Data Warehouse Lifecycle Management (DWLM) V

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

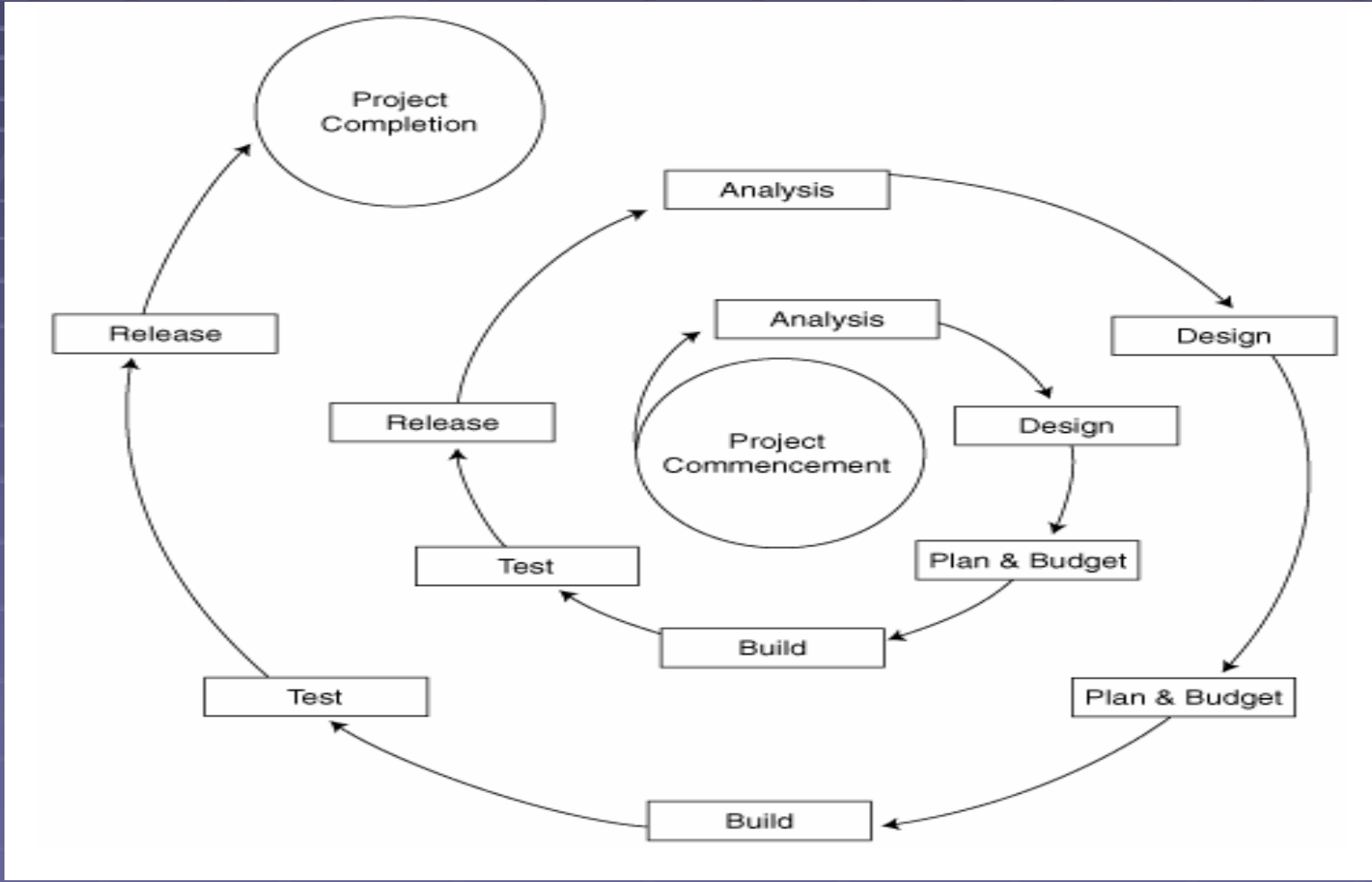
DW Life Cycle Models

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

Waterfall Model



Spiral Model



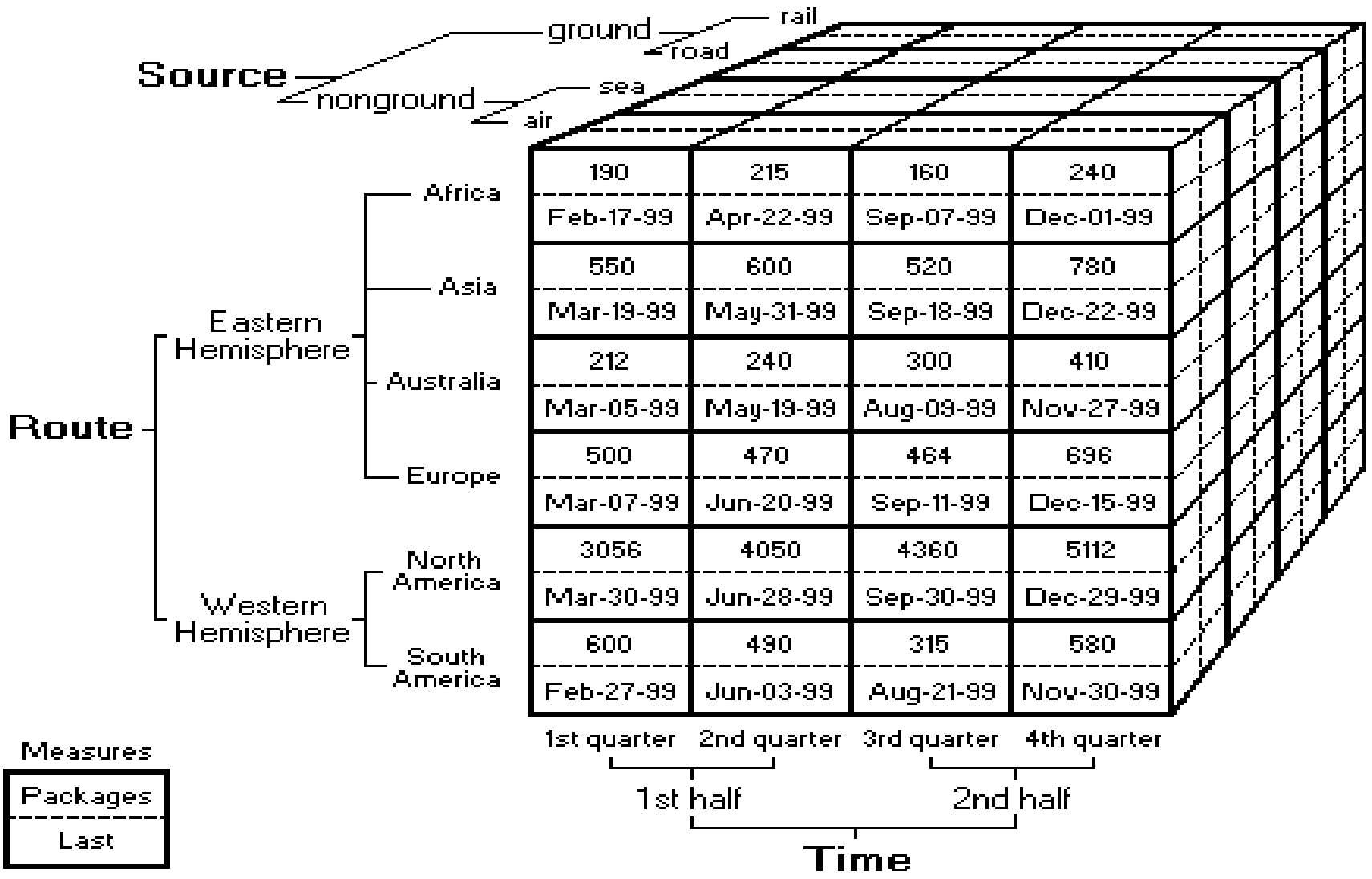
Ralph Kimball's Four Principles

- **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.

OLAP CUBE

- Multidimensional structure containing dimensions and measures
  - Measure is based on fact table
  - Dimension is based on dimension table
- Measure values are contained at intersection between dimensions (cells)

OLAP CUBE


Source
ground
rail
road
nonground
sea
air
Route
Eastern Hemisphere
Western Hemisphere
Africa
Asia
Australia
Europe
North America
South America
190
Feb-17-99
215
Apr-22-99
160
Sep-07-99
240
Dec-01-99
550
Mar-19-99
600
May-31-99
520
Sep-18-99
780
Dec-22-99
212
Mar-05-99
240
May-19-99
300
Aug-09-99
410
Nov-27-99
500
Mar-07-99
470
Jun-20-99
464
Sep-11-99
696
Dec-15-99
3056
Mar-30-99
4050
Jun-28-99
4360
Sep-30-99
5112
Dec-29-99
600
Feb-27-99
490
Jun-03-99
315
Aug-21-99
580
Nov-30-99
1st quarter
2nd quarter
3rd quarter
4th quarter
1st half
2nd half
Time
Measures
Packages
Last

Types of CUBE

- Regular
- Linked
  - Allow for reuse of cubes across servers
  - Local caching helps reduce query loads
- Distributed
  - Cubes can be broken down into partitions
- Virtual
  - Similar to concept of views in a relational database
  - Could be used to combine cubes
- Local
  - Used by PivotTable Service to provide off-line access to parts of a cube

OLAP CUBE Characteristics

- Storage mode
  - MOLAP
    - Data and aggregations compressed and stored in a file
  - ROLAP
    - Data and aggregations stored in RDBMS
  - HOLAP
    - Aggregations stored in a file, data remains relational
- Aggregation level

Populating OLAP CUBE

- Full process
  - Invalidates cube and recreates structure
  - Retrieves all measure data and dimensional keys from underlying fact table
- Refresh data
  - Retrieves all measure data and dimensional keys from underlying fact table
  - Allow uninterrupted end-user access (shadowing)
- Incremental update
  - Can be used to add new data to a cube
  - Need a consistent way to recognize new and modified data within the underlying fact table

Multi-Dimensional Extensions (MDX) to SQL

- MDX is the query language for OLAP CUBES specification,
- Created (1998) and maintained by Microsoft
- Enjoys industry-wide support (but NOT a standard)
  - IBM DB2® Alphablox cubes support a subset of the MDX syntax and functions
  - IBM/Cognos (with acquisition of Applix)
  - Oracle (through Hyperion/Essbase)
  - Teradata (through Microsoft)
  - SAP BW (SAP NetWeaver BI)
  - SAS
  - Open Source: PostgreSQL and MySQL w/third party tools

MDX Elements

- Complex Syntax
- Similar to SQL (DDL, DML)
  - CREATE (cube, member, mining model etc)
  - DROP, REFRESH, USE LIBRARY (COM)
  - INSERT INTO, SELECT, UPDATE CUBE
- Extensive use of wizards recommended
- OLAP and Data Mining functions

Alternatives and alternative MDX implementations

- Sybase IQ
- Oracle SQL for OLAP
- Open Source
  - Mondrian OLAP Server (MDX)
    - JasperSoft
    - BIOLAP
    - CompiereBI
    - Pentaho
    - SpagoBI
  - OpenI (by Loyalty Matrix)
  - XMLA (XML for Analysis)
  - mdXML
  - Eclipse BIRT (Business Intelligence and Reporting Tool)

XML for Analysis

- API for OLAP
  - Set of XML interfaces using SOAP (as opposed to OLE DB for OLAP – ODBO)
- Supported by Hyperion, Microsoft, SAP, SAS
- Abstained: IBM, Oracle
- Not much development since Hyperion was acquired by Oracle

Microsoft SQL Server DW

- Microsoft SQL Server Analysis Services (SSAS)
  - First as Microsoft SQL Server OLAP Services with version SQL Server 7.0
  - Integrated in SQL Server 2000
  - Expanded in SQL Server 2005 and 2008
- Complete implementation of MDX specification
- Integrated with 2007 Office System and ProClarity

IBM DW

- DB2 Data Warehouse EE (v.9.1) components directly related to BI
  - DWE Design Studio (subset of Data Architect)
  - SQL warehousing tool
  - DWE Cube Views
  - DWE Intelligent Miner
  - DB2 Alphablox (v.8.4)

Oracle DW

- Oracle Warehouse Builder
  - ETL, data cleansing
- Oracle Data Mining
  - Intelligence Discovery
- Oracle OLAP
  - Analytic Queries
  - OLAP cube materialized views
  - No MDX

Word of Caution

Most data warehouse project plans focus entirely on technical aspects of initial implementation and change is thereafter seen as a production issue. This does not reflect reality. Enterprises need to adopt a methodology where both business and IT align themselves during the entire life cycle of the data warehouse…

Frank Buytendijk, Vice President,

Gartner, Inc.

Data Mining Applications

- In ***retail chains***, data mining has been used to identify the purchasing patterns of customers and associating this data with demographic characteristics such as the age and class profile of customers. Such patterns are useful in making decisions such as what products to sell in which retail stores and when
- In the ***insurance industry***, data mining has been used to analyze the claims made against insurance policies and hence feed into actuarial decisions such asthe pricing of particular policies
- In the ***finance industry***, banks have used data mining techniques to identify fraudulent credit card use among its transaction data
- In the ***medical domain***, data mining may be applied to identify successful medical treatments for particular medical complaints

Data Mining Models

- Groupings and predictive analysis based on relational or OLAP data
- Interprets data based on statistical information (referred to as cases)

Data Mining II

- Predictive modeling
- Database segmentation (profiling)
- Link analysis
- Deviation detection

- Palgrave

Basic Relational Concepts I

- Normalization – Lossless Decomposition
  - 1NF
    - Eliminating Redundancy [on row level] (each attribute (column) of a tuple (unique row) must contain a single value; scalar)
  - 2NF
    - Is in 1NF, and all its attributes (columns) are dependent on the entire candidate key
  - 3NF
    - Is in 2NF, all non-key attributes (columns) are mutually independent
  - 4NF (Boyce/Codd Normal Form)
- Considered a variation of third normal form
    - Special case of relations with multiple candidate keys
  - 5NF
    - Addresses rare case of join dependencies
    - Is rarely used