



Data Warehouse Architecture

James Serra – Data Warehouse/BI/MDM Architect

JamesSerra3@gmail.com

<http://JamesSerra.com/>

About me

- In IT for 28 years
- Worked as desktop/web/database developer, DBA, BI and DW architect, MDM, PDW
- Been perm, contractor, consultant, business owner
- MCSE for SQL Server 2012: Data Platform and BI
- SME for SQL Server 2012 certs
- Currently a consultant working with MDS at Schlumberger as a MDM Technical Lead
- Contributing writer for SQL Server Pro magazine
- Blog at JamesSerra.com

Agenda

- Why use a data warehouse?
- Fast Track Data Warehouse (FTDW)
- Appliances
- Data Warehouse vs Data Mart
- Kimball vs Inmon (Normalized vs Dimensional)
- Populating a Data Warehouse
- ETL vs ELT
- Normalizing and Surrogate Keys
- SSAS Cubes
- SQL Server 2012 Tabular Model
- End-User Microsoft BI Tools

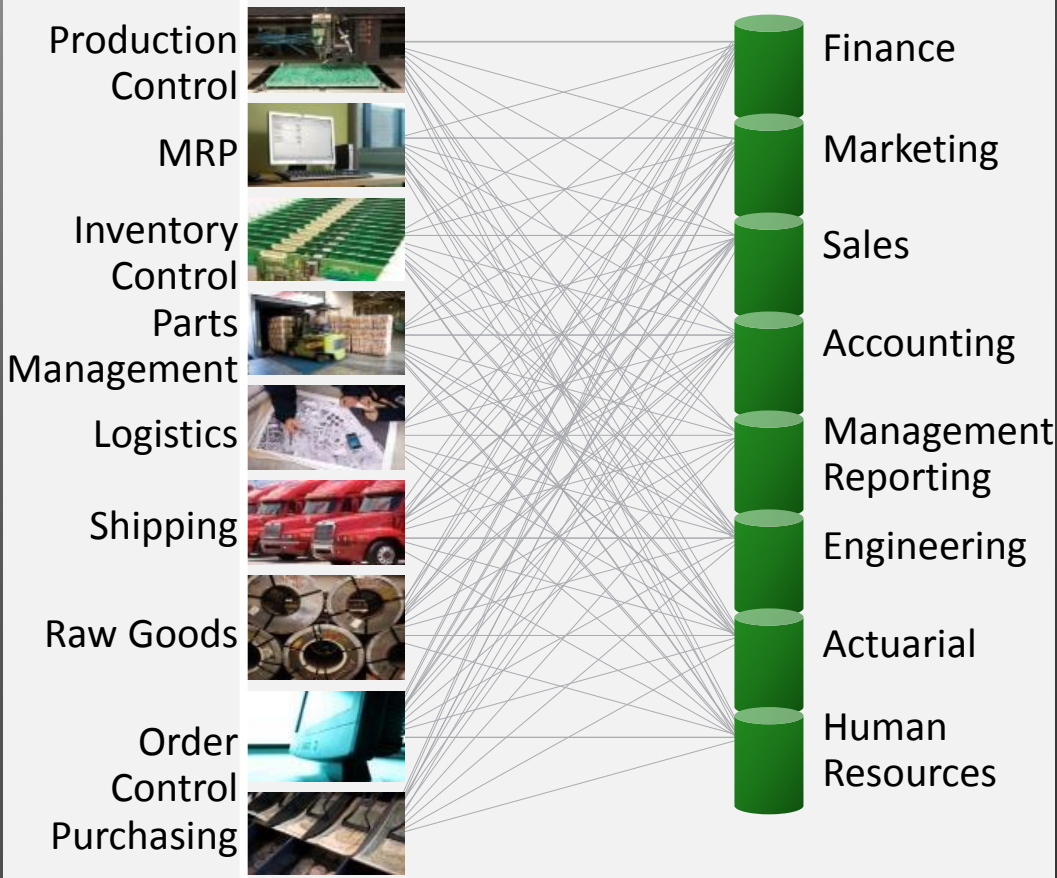
Why use a Data Warehouse?

All these solutions are for data warehouses only (not OLTP).

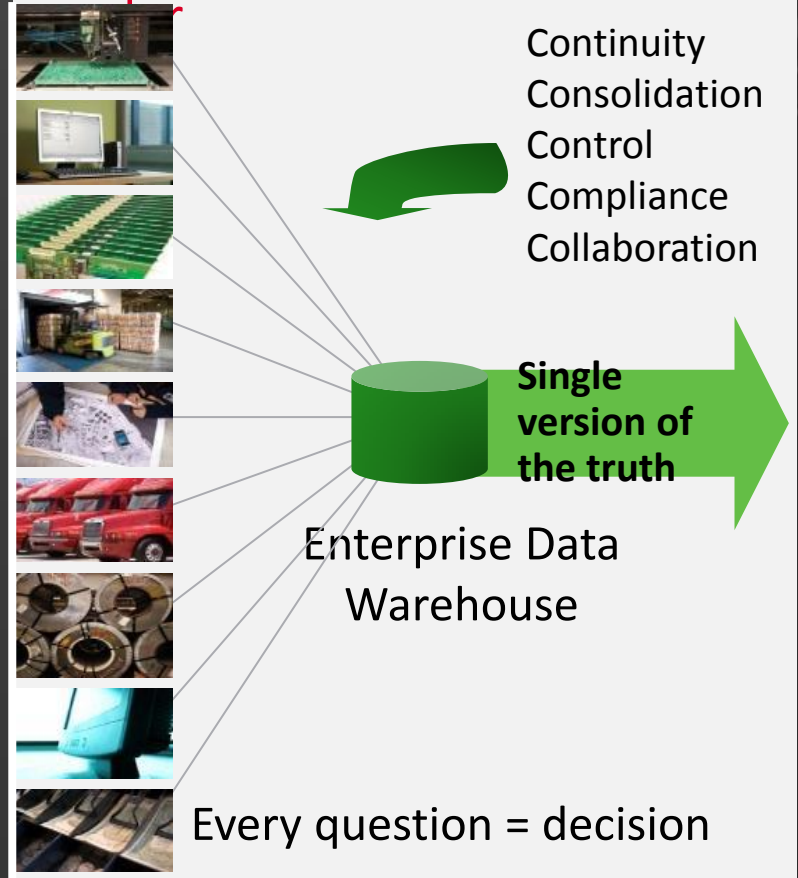
- Reduce stress on production system
- Optimized for read access, sequential disk scans
- Integrate many sources of data
- Keep historical records
- Restructure/rename tables and fields
- Use Master Data Management
- No IT involvement needed for users to create reports
- Improve data quality
- One version of the truth
- Easy to create BI solutions on top of it (SSAS cubes)

Why use a Data Warehouse?

Legacy applications + data marts = chaos



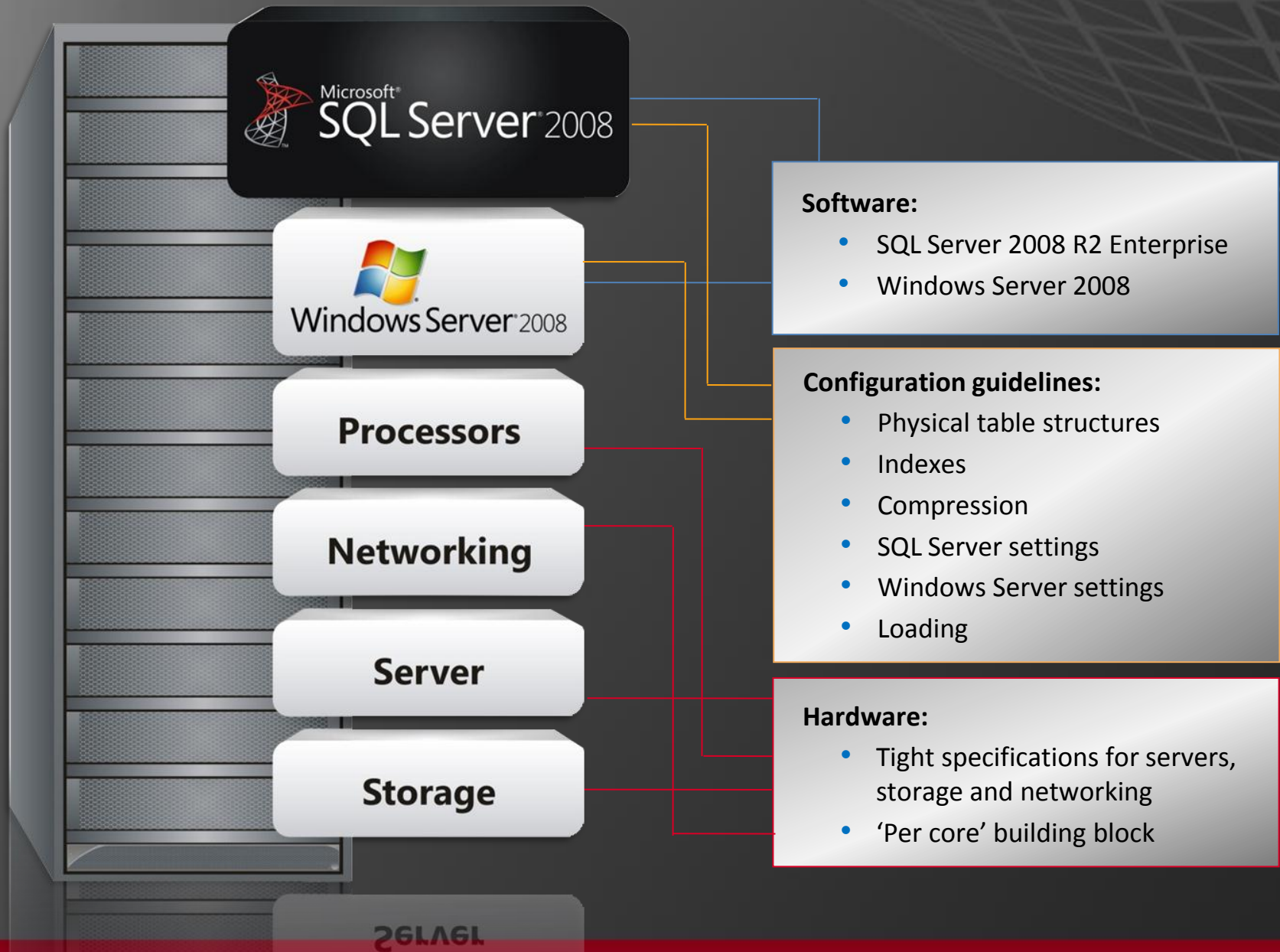
Enterprise data warehouse =



Hardware Solutions

- Fast Track Data Warehouse - A reference configuration optimized for data warehousing. This saves an organization from having to commit resources to configure and build the server hardware. Fast Track Data Warehouse hardware is tested for data warehousing which eliminates guesswork and is designed to save you months of configuration, setup, testing and tuning. You just need to install the OS and SQL Server
- Appliances - Microsoft has made available SQL Server appliances that allow customers to deploy data warehouse (DW), business intelligence (BI) and database consolidation solutions in a very short time, with all the components pre-configured and pre-optimized. These appliances include all the hardware, software and services for a complete, ready-to-run, out-of-the-box, high performance, energy-efficient solutions

Fast Track Data Warehouse



Appliances

- HP Business Data Warehouse Appliance
- HP Business Decision Appliance
- HP Database Consolidation Appliance
- HP Enterprise Data Warehouse Appliance
- Dell Quickstart Data Warehouse Appliance 1000
- Dell Quickstart Data Warehouse Appliance 2000
- Dell Parallel Data Warehouse Appliance

Data Warehouse vs Data Mart

- Data Warehouse: A single organizational repository of enterprise wide data across many or all subject areas
 - Holds multiple subject areas
 - Holds very detailed information
 - Works to integrate all data sources
 - Feeds dimensional model
- Data Mart: Subset of the data warehouse that is usually oriented to specific subject
 - The logical combination of all the data marts is a data warehouse

In short, a data warehouse as contains many subject areas, and a data mart contains just one of those subject areas

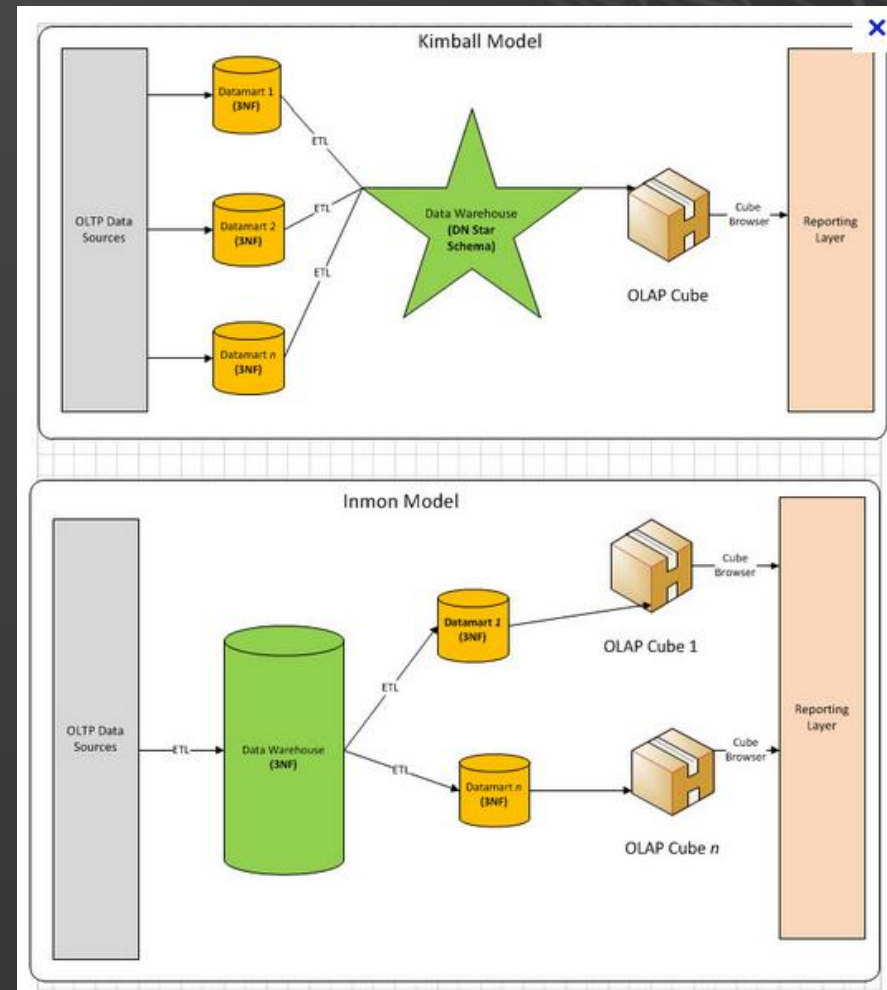
Kimball vs Inmon

- Normalized (Inmon) vs Dimensional (Kimball)
- Normalized:
 - Normalization rules
 - Many tables using joins
- Dimensional:
 - Facts and dimensions
 - Less tables having duplicate data (de-normalized)
 - Easier for user to understand

Kimball vs Inmon

- Top-Down (Inmon) vs Bottom-Up (Kimball)
- Bottom-Up:
 - Data marts
 - Logical data warehouse
 - Decentralized
 - Quick results, iterative approach
- Top-Down:
 - Enterprise data model
 - Centralized
 - Later create data marts
 - More upfront work but less redo

Hybrid: Data Vault



Populating a Data Warehouse

- Frequency of data pull
- Full Extraction – All data
- Incremental Extraction – Only data changed from last run
- Determine data that has changed
 - Timestamp - Last Updated
 - CDC
 - Partitioning
 - Triggers
 - MERGE
- Online Extraction – Data from source
 - Replication
 - Database Snapshot
 - Availability Groups
- Offline Extraction – Data from flat file

ETL vs ELT

- Extract, Transform, and Load (ETL)
 - Transform while hitting source system
 - No staging tables
 - Processing done by ETL tools (SSIS)
- Extract, Load, Transform (ELT)
 - Uses staging tables
 - Processing done by target database engine (SSIS: Execute T-SQL Statement task instead of Data Flow Transform tasks)
 - Use for big volumes of data
 - Use when source and target databases are the same
 - Use with PDW

ELT is better since database engine is more efficient than SSIS

Database engine: Transformations

SSIS: Data pipeline and workflow management

Normalizing and Surrogate Keys

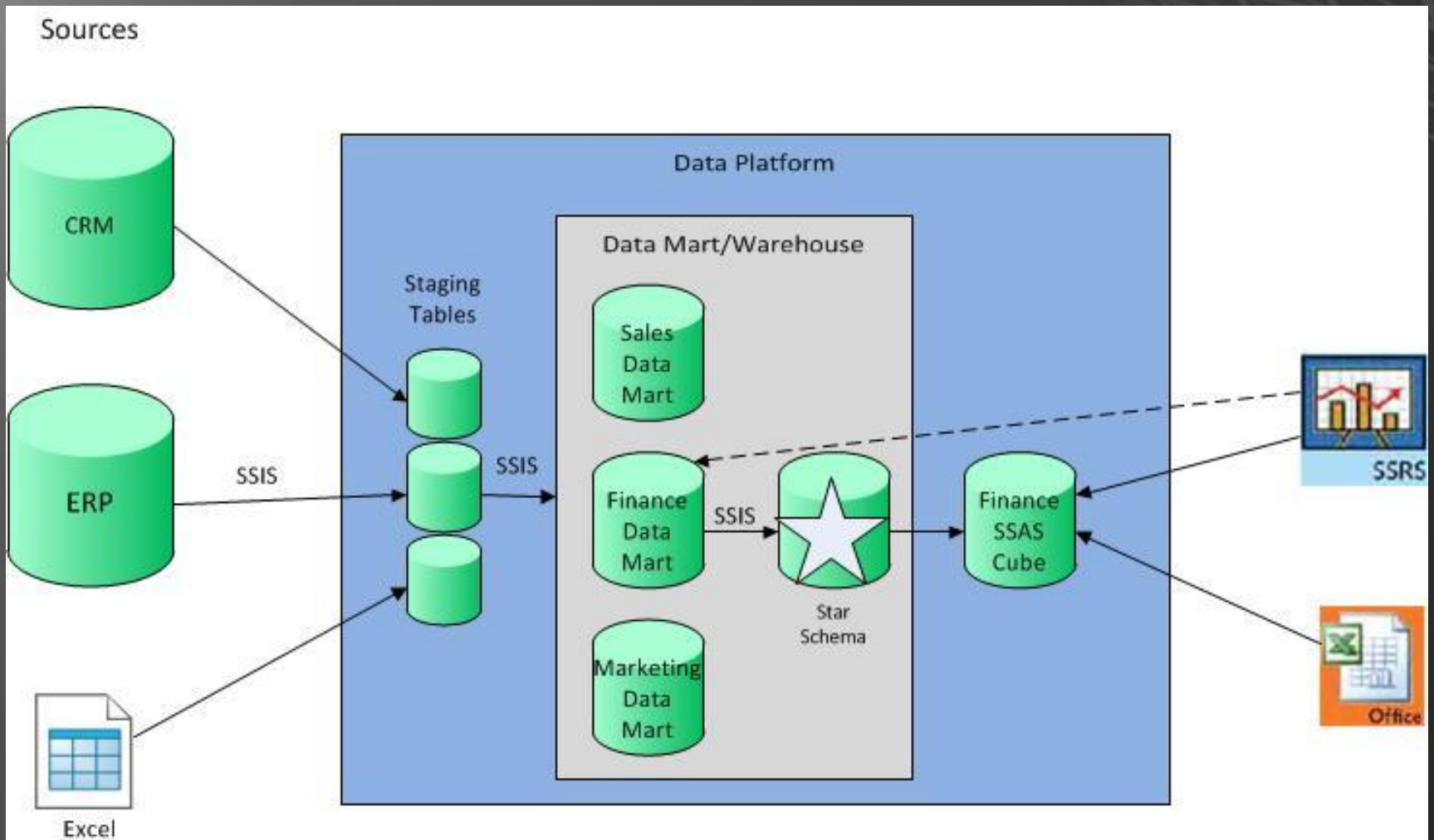
- Normalize to eliminate redundant data and setup table relationships
- Surrogate Keys – Unique identifier not derived from source system
 - Embedded in fact tables as foreign keys to dimension tables
 - Allows integrating data from multiple source systems
 - Protect from changes in the source system
 - Allows for slowly changing dimensions
 - Allows you to create rows in the dimension that don't exist in the source (-1 in fact table for unassigned)
 - Improves performance (joins) and database size by using integer type instead of text

SSAS Cubes

Reasons to use instead of data warehouse:

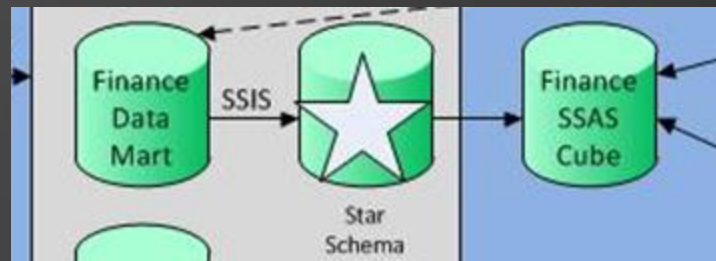
- Aggregating (Summarizing) the data for performance
- Multidimensional analysis – slice, dice, drilldown
- Hierarchies
- Advanced time-calculations – i.e. 12-month rolling average
- Easily use Excel to view data
- Slowly Changing Dimensions (SCD)

Data Warehouse Architecture



SQL Server 2012 Tabular Model

- New xVelocity in-memory database in SSAS
- Build model in Power Pivot or SSDT
- Uses existing relational model
- No star schema, no extra SSIS
- Uses DAX
- Faster and easier to use than multidimensional model



End-User Microsoft BI Tools

- Excel PivotTables
- SQL Server Reporting Services (SSRS)
- Report Builder
- PowerPivot
- PerformancePoint Services (PPS)
- Power View

Resources:

- Data Warehouse Architecture – Kimball and Inmon methodologies: <http://bit.ly/SrzNHy>
- SQL Server 2012: Multidimensional vs tabular: <http://bit.ly/SrzX1x>
- Data Warehouse vs Data Mart: <http://bit.ly/SrAi4p>
- Fast Track Data Warehouse Reference Guide for SQL Server 2012: <http://bit.ly/SrAwsj>
- Complex reporting off a SSAS cube: <http://bit.ly/SrAEYw>
- Surrogate Keys: <http://bit.ly/SrAIRp>
- Normalizing Your Database: <http://bit.ly/SrAHnc>
- Difference between ETL and ELT: <http://bit.ly/SrAKQa>
- Microsoft's Data Warehouse offerings: <http://bit.ly/xAZy9h>
- Microsoft SQL Server Reference Architecture and Appliances: <http://bit.ly/y7bXY5>
- Methods for populating a data warehouse: <http://bit.ly/SrARuZ>
- Great white paper: Microsoft EDW Architecture, Guidance and Deployment Best Practices: <http://bit.ly/SrAZug>
- End-User Microsoft BI Tools – Clearing up the confusion: <http://bit.ly/SrBMLT>
- Microsoft Appliances: <http://bit.ly/YQIXzM>