2007 National BDPA Technology Conference
Theme: "Global Strategy for the IT Professional"

Business Intelligence Simplified

Akin S. Walter-Johnson III
Principal Software Architect
PeerLabs, Inc
akin@peerlabs.com

August 14 – 18, 2007

Washington, DC

Outline

- **What is Business Intelligence**
- **Business Intelligence Philosophy**
  - **Corporate Information Factory (Inmon)**
  - **Bus Architecture (Kimball)**
- **Functional Architecture**
  - **Data, Intelligence and Presentation**
- **Business Intelligence Project**

Introduction

- ## Who am I
  Software and Data Architect

- ## What the Presentation is about
  Simplify Business Intelligence
  Decompose into functional areas
  (Data, Intelligence and Presentation)

- ## What the Presentation is not about
  Product Promotion or recommendation
  Platform or Operating system promotion
  You will not leave here an expert, however empowered.

Introduction

- **What are your expectations ??**

What is Business Intelligence

Data → ⚙ *Information* ⚙ → *Knowledge*

- **Business intelligence is the process of transforming data into information and information to knowledge.**

What is Business Intelligence

Data

```
United States,374,PRD-BIKES-26
France,134,PRD-BIKES-16
```

- A stream of **bit and bytes**

- Data is **text and numbers**, when you can identify and **interpret** it becomes **information**, facts and figure.

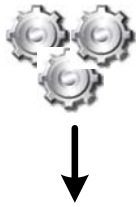# What is Business Intelligence

Data

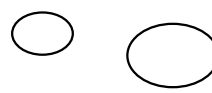| Region	Sales Unit	Product
United States	374,	PRD-BIKES-26
France	134,	PRD-BIKES-16

Information

- Data is **organized, structured** and **aggregated.**
- Information can answer questions such as " How much did we sell last month in France".

What is Business Intelligence

Information

Knowledge

How much will we sell ???

- When information becomes Knowledge then you are looking into the future observing patterns and making predictions.

- Typical question could be "How much will we sell next Year in Chicago before 2nd Quarter".

What is Business Intelligence

- **A BI system is a comprehensive set of processes and functions that can answer questions about the present and attempt to predict the future.**

Business Drivers

- **Online Transaction Processing not suitable**
- **Share information in real time**
- **Empower decision makers at all levels.**
- **Improve visibility and prediction**
- **Integration of business units.**
  - **Sales, Store inventories, and Distribution inventories.**

BI Philosophies

Ralph Kimball

"… The data warehouse is nothing more than the union of all the data marts …"

Ralph Kimball Dec. 29, 1997.

Bill Inmon

"… You can catch all the minnows in the ocean and stack them together and they still do not make a whale."

Bill Inmon Jan. 8, 1998.

(Minnows) A large group of small freshwater fishes. Small or insignificant

BI Philosophies

Ralph Kimball

(BOTTOM-UP)

- **Let everybody build what they want when they want it, we'll integrate it all when and if we need to.**

- **BUS Architecture**

Bill Inmon

(TOP-DOWN)

- **Don't do anything until you've designed everything.**

- **Corporate Information Factory**

BI Philosophies

Ralph Kimball

Pros:

Fast to build,

Quick ROI

Cons:

- Difficult to maintain at enterprise level
- Redundant resources
- Difficult to integrate data marts

Bill Inmon

Pros:

Tightly integrated
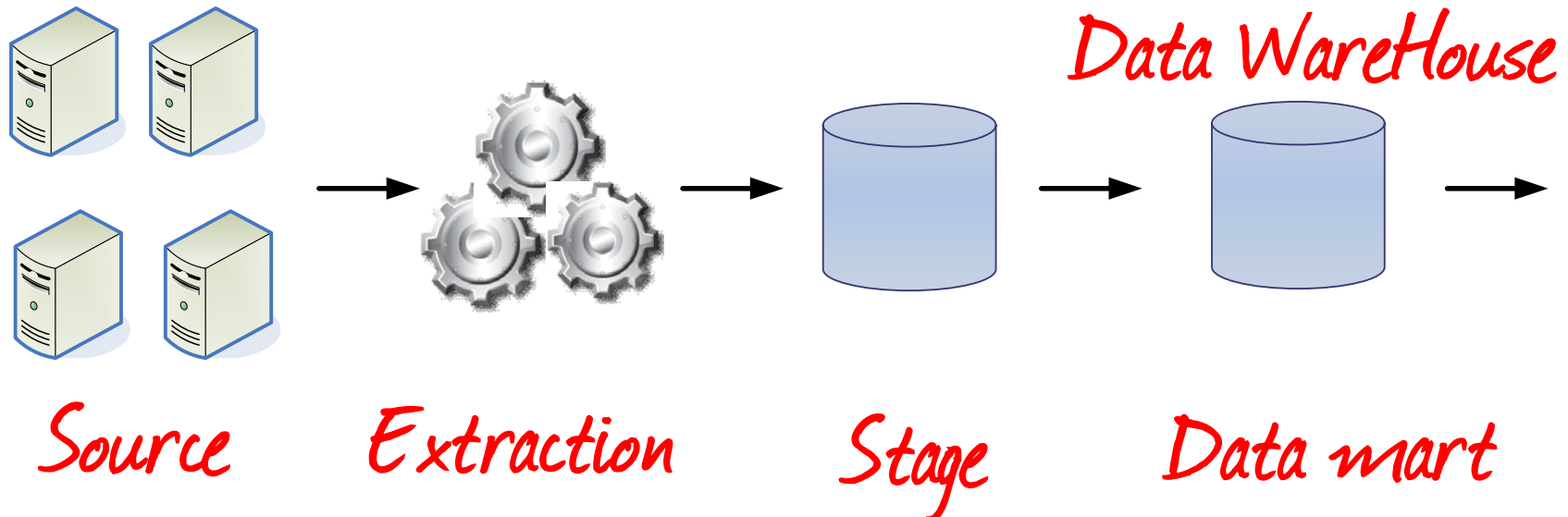
Easy to maintain

Cons:

- Time intensive
- Rigid Structure

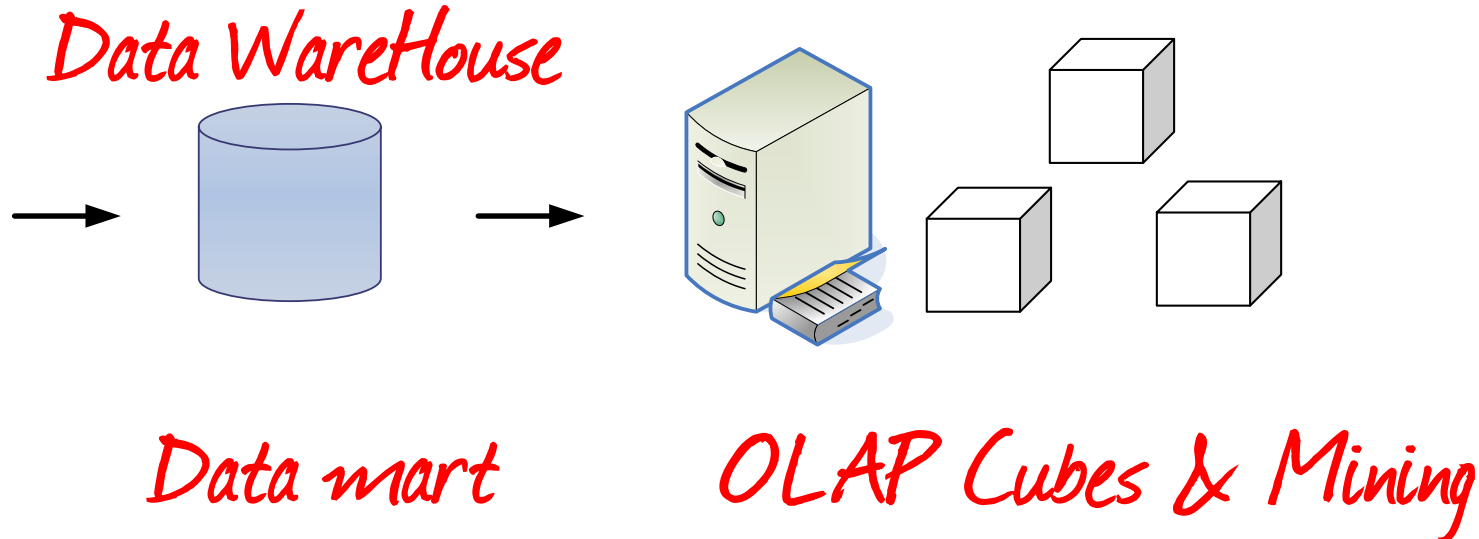# BI Functional Architecture

Any BI system can be decomposed into

- *Data*

- *Intelligence*

- *Presentation*

BI Data layer

Data WareHouse

Source   Extraction   Stage   Data mart

- **Cleaning & integrating** (ETL)
- **Loading** of data into either a data mart or data warehouse

BI **Intelligence layer**

Data WareHouse

Data mart   *OLAP Cubes & Mining*

- **OnLine Analytical Processing & Mining**
- **Answering questions based current data**
- **Observing patterns and making predictions.**

# BI Presentation Layer
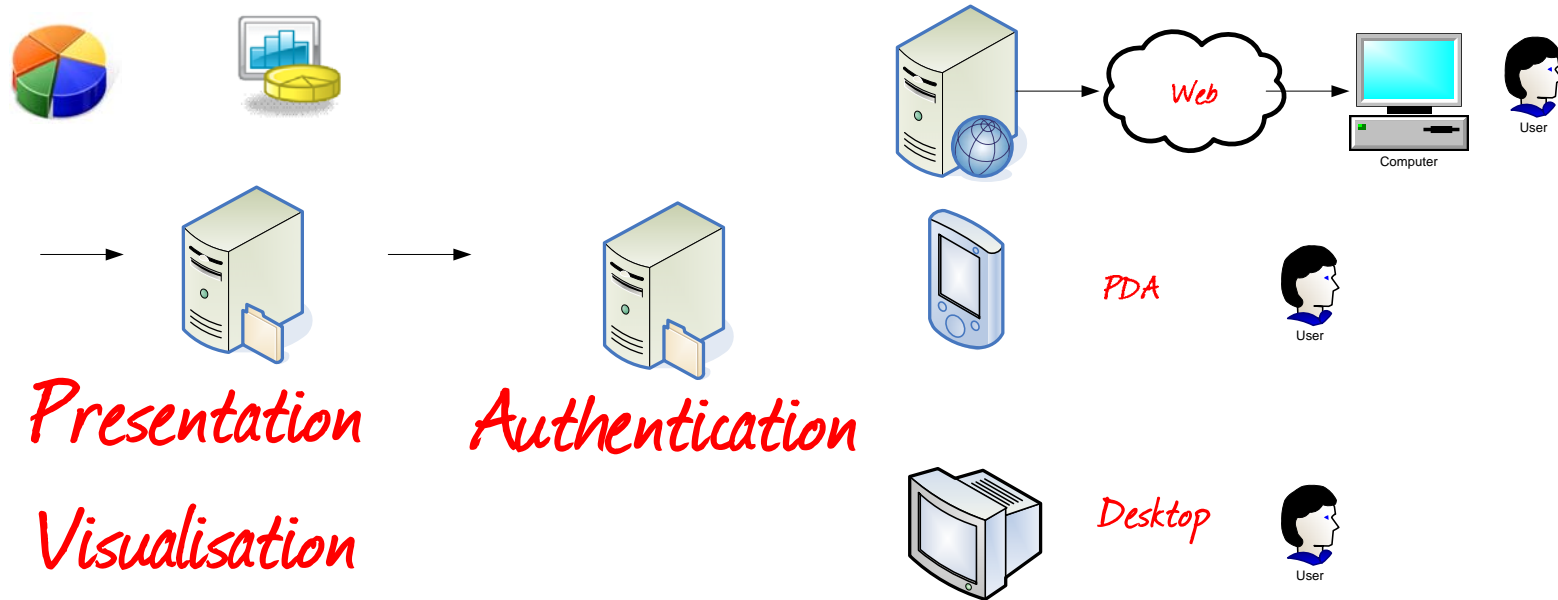
- **Presentation of data**
- **Security and the delivery of data.**
  - Single sign on environment authentication.
  - Authorization Logic in database or Auth module

Data layer

- **ETL (Extract, Transform & Load) ETT transfer**
- **Operational Data Store**
- **Data warehouse**
- **Data mart**
- **Data Quality**
- **Facts / Measures**
- **Dimension**
- **Hierarchy**
- **Surrogate Keys**
- **Slowly changing Dims**
- **Star schema**
- **Snowflake**
- **Metadata**

Data layer ETL

Sales
Manufacturing
Marketing

Extract Transform
Staging
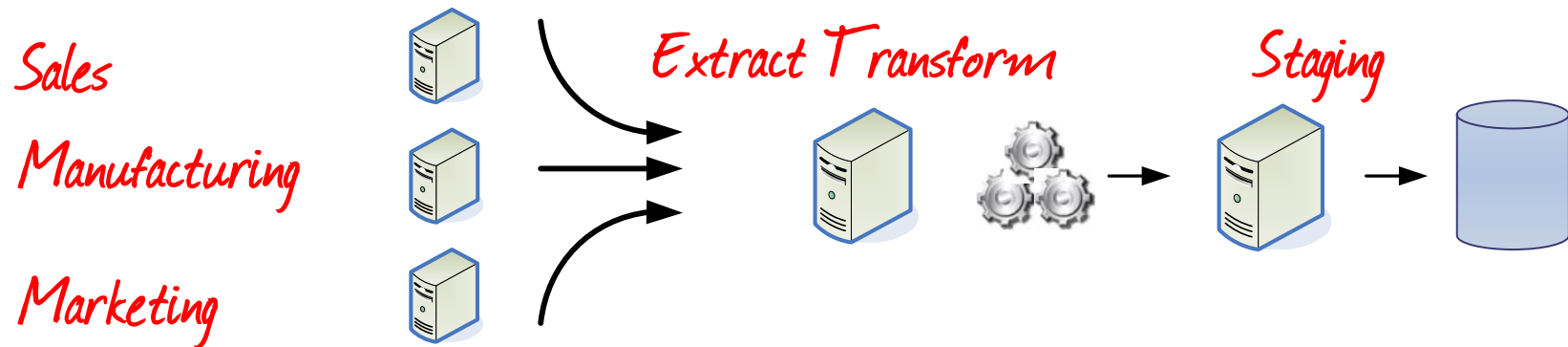
- **Data consolidation** from various systems
- **Constructing** into the desired format suitable for the warehouse

Data layer ETL

- Use **export** utility to **extract** data into a commonly accepted format
- Extraction from the source can be done **fully or incrementally** depending on the amount of time it takes to complete the process.
- **Full extraction** is easier than incremental small cases. **(Pump & Dump)**
- Parallel processing can be used.

Data layer Data Quality

Logically Dirty ➝ **Home Address**:
1600 Pennsylvania Ave Washington DC

Physically Dirty ➝ **SSN**
123a45683

- **You will get dirty data**
- **Clean Data requires validation & consistency**
- **Data is made appropriate for a specific use**
- **A state of completeness, Timeliness and accuracy.**

Data layer Data Quality

- Dirty data should not be allowed into the system it would lead to incorrect results.

- Quality control measures should be in place to manage dirty data.

- A member of the business staff should assist the BI team with the quality control process.

- Physically dirty data can be identified using tools, but logically dirty data can only be identified by a trained eye.

- Someone familiar with the business.

Data layer Data Quality

- **Handling dirty data**
  - **Discard rows**
  - **Replace with null** values and load
  - If **primary key** is dirty then is can't be loaded
  - Generate exception reports for dry cleaning
  - (Regular Expressions) for parsing
  - 10 lines of perl code = 1 line of awk code

Data layer Operational Data Store

Savings

Mortgage

Credit Card
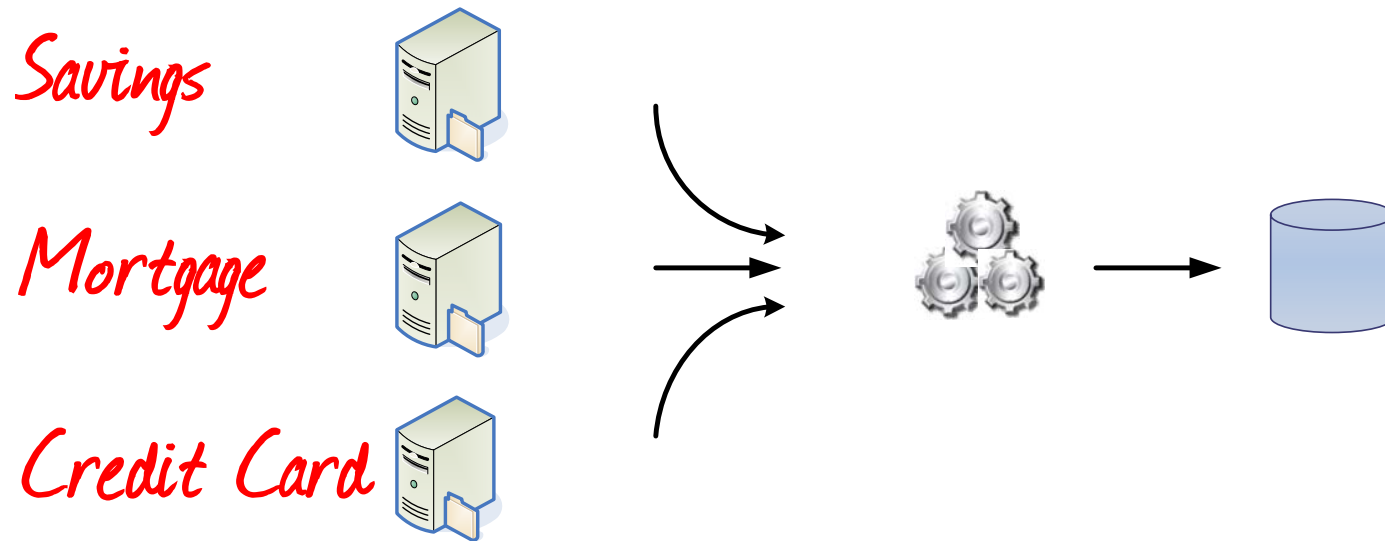
- ODS is not the Data warehouse
- ODS integrates dispersed legacy systems.
- A bank which has different systems for loans, checks, mortgages etc, will use an ODS

Data layer Operational Data Store

- ODS is a **current and volatile** collection of data used to support the tactical decision-making process for the enterprise
- ODS is volatile so **no history**
- ODS is **updated through** the course of **business operations**
- ODS operates at the **transactional level**.
- ODS can feed the Data Warehouse

Data layer Data Warehouse

Savings

Mortgage

Credit Card

ODS   Warehouse
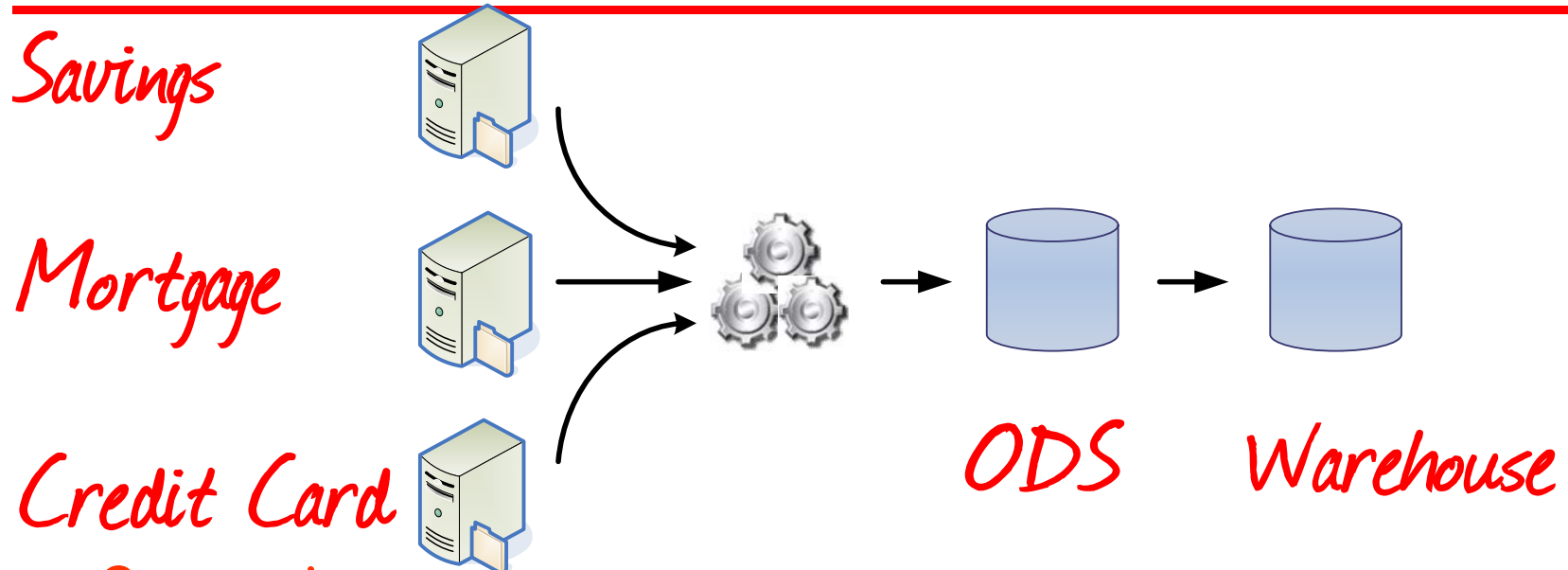
- **Corporate memory**
- Common view of enterprise data
- **Separate system** from the operational database
- **Non-volatile (Historical)** collection of data used to support the strategic decision-making process for the enterprise

Data layer Data Warehouse

- **Central point of data integration for business intelligence and is the source of data for the data marts**

- **Subject-oriented, integrated and has time-variant**

Data layer Data mart

Source   ETL   ODS   Warehouse   Sales Data Mart / Manufacturing Data Mart / Inventory Data Mart
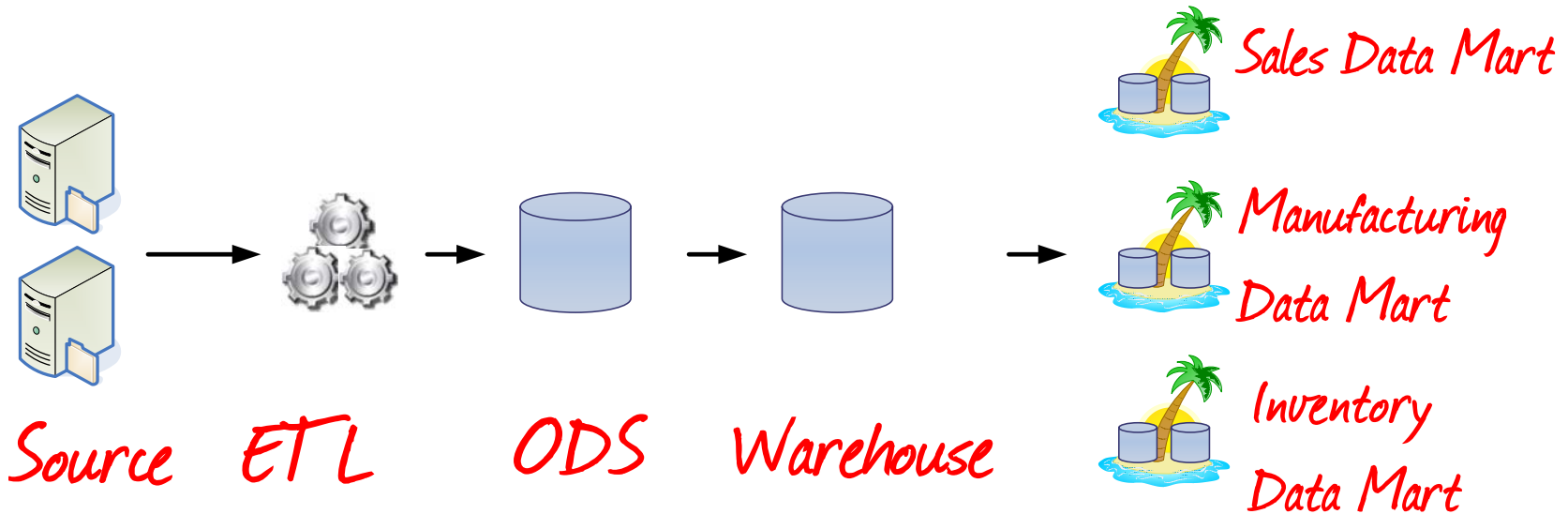
- **Data derived from the data warehouse**
- **Support a specific analytical requirements of a business unit or function**

Data layer Data mart

- **Not required to be co located with Data warehouse**
- **Easy to create**
- **Lower cost than implementing a full Data warehouse**
- **Potential users are more clearly defined than in a full Data warehouse**

Data Layer Facts / Measures

Time_key

Customer_key

Product_key

Sales_region_key

SalePerson_key

Unit price

A Sales transaction

This is the GRAIN

- **Lowest level** of detail (Atomic level)
- **Facts or measures are at the center of the business Intelligence model.**
- **BI system can support different levels of detail (fact)**

Data Layer Facts / Measures

- **They contain numeric values expressing an aspect of the organization transaction and foreign keys to dimension.**

- **The numeric values can be added across a time period (Rollup).**

- **Some numeric values cannot be added an example would be unit price or weight**

- **The Grain of a fact table is the level of information in the table.**

Data Layer Facts / Measures

- **If the grain of the fact table is an individual line item on a customer bill, then all of the numerical facts must refer to this particular line item.**

- **A common mistake is to mix facts at different levels.**

Data Layer Facts / Measures

- **Actual quantity shipped** (Fact 1)
- Forecast (budget) quantity shipped (Fact2)
- Problem is budget is at a higher level of granularity such as planned monthly sales.
- Never mix facts at different levels create a separate fact table.
- Build your fact table with lowest Grain
- Facts take up about 95 % of BI system

Data Layer Dimensions

CustomerID	Name	Address
123	Bill	20 West Ave
234	Jane	10 East street

- **Dimensions describe Lookup objects**
- **They give meaning to the fact values**
- **Examples Customer, Product, Employee & Time**
- **Each row describes an instance**
- **" Show sales by customer ". Customer is the dimension**

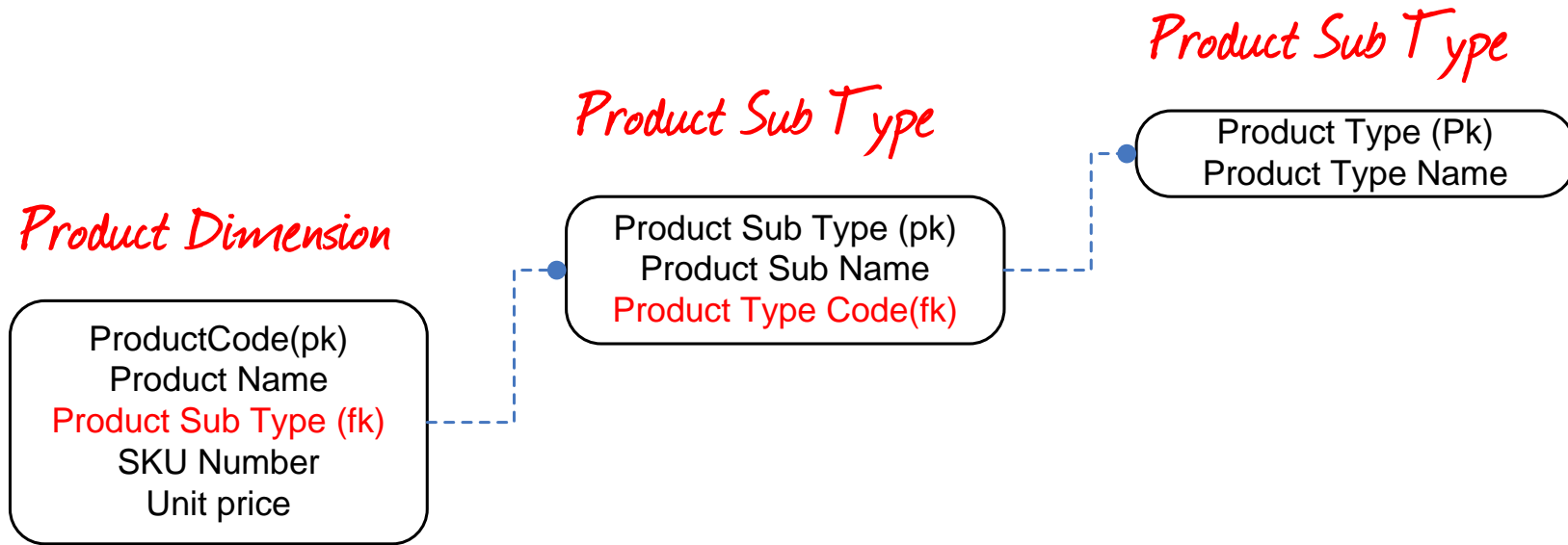# Data Layer Dimensions

- **Dimensions allow you to group your facts/measures because they occur in the facts.**

- **With dimensions you can slice and dice data.**

- **Time dimension is very important because it enables us to present facts in a way we can easily relate to.**

- **Time dimension should be created first and it should be able to support the earliest fact or measure.**

Data Layer Hierarchy

Product Sub Type

Product Sub Type

Product Type (Pk)
Product Type Name

Product Dimension

Product Sub Type (pk)
Product Sub Name
Product Type Code(fk)

ProductCode(pk)
Product Name
Product Sub Type (fk)
SKU Number
Unit price

- **Dimensions can be related via a parent child relationship (Hierarchy)**
- **Allow grouping data within a dimension**

Data Layer Hierarchy

- **One to many between child and parent**
- **Permits the ability to drill down**
- **Product Type to Sub Types**

Data Surrogate Keys

- **Auto number or sequences**
- **New keys** for records in the warehouse
- Separate from Source system key
- Dimensions as lookup tables have a **natural primary key** also know as **business key**.
- If **source system key changes** you are protected
- Allows integration from multiple sources

Data Slowly changing Dims

- **Dimension can change**
- **A customer changes has last name**
- **A sales person is relocated to another region**
- **Company merger**
- **Slowly changing dimensions occurs when the attribute of a record varies over time.**

Data Slowly changing Dims

- **TIME dimension** should be the last to change.

- Dates will not change, especially in the past

- July 4th 1776 would always be July 4th 1776. (Holiday)

- October 1 1960 would always be the same

- A company's quarter can change if the fiscal year changes.

Data Slowly changing Dims (i)

Before

Customer ID	Name	Address
123	Bill	20 West Ave

After

Customer ID	Name	Address
123	Bill	10 Main st

- **Replace the information, no tracking of history (Overwrite)`**

Data Slowly changing Dims (ii)

Customer ID	Name	Address	Start date	End Date	Status
123	Bill	20 West Ave	1/1/90	5/1/95	Closed
123	Bill	10 Main st	5/1/95	******	Open

Current Record

- **Add attributes to track history (History)**

- **Dates are added to track and manage history.**

- **During design identify dimensions that will change and create attributes to support the change.**
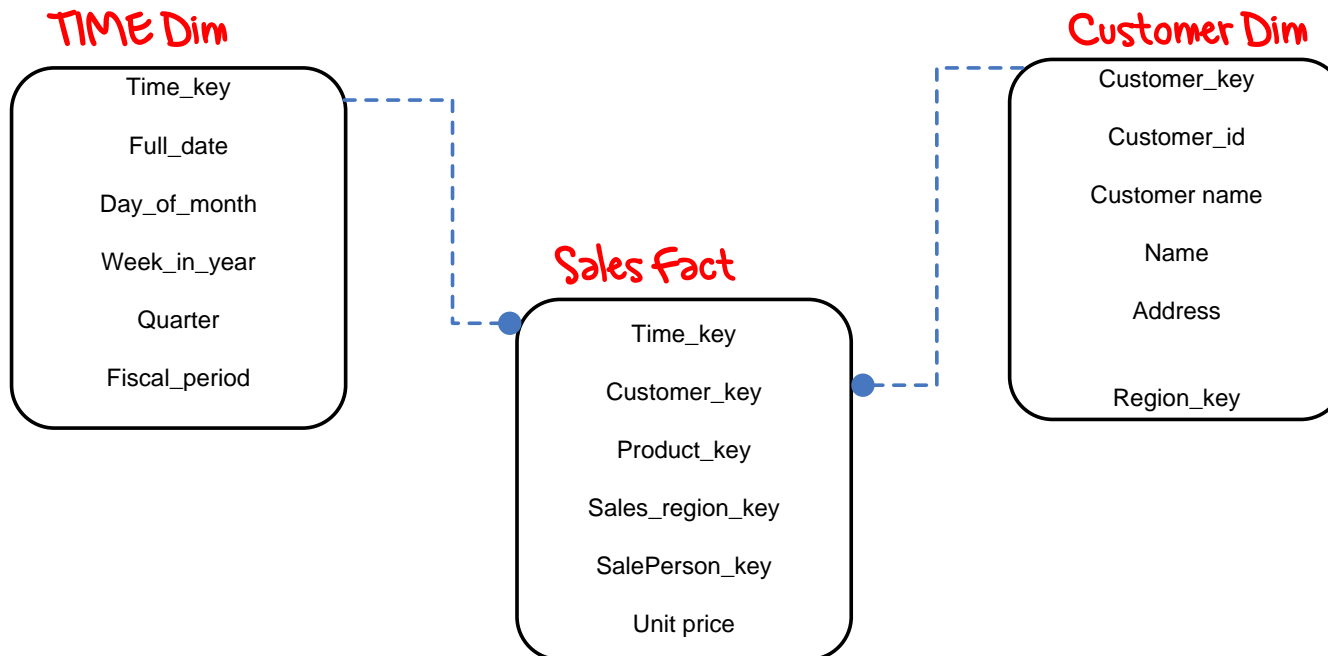
Data Slowly changing Dims (iii)

Customer ID	Name	Address	Old address	End Date
123	Bill	10 Main st	20 West Ave	5/1/95

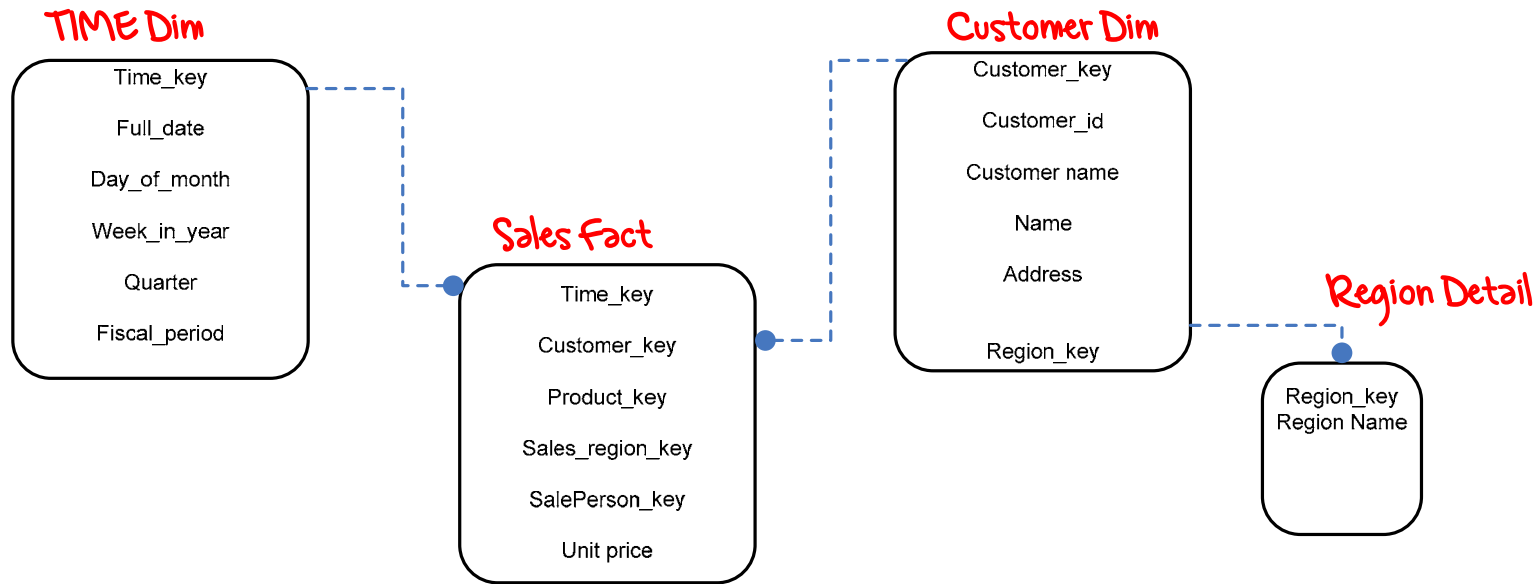- **Update the record but include the previous value, this will only track only the most recent history. (Most recent History)**

Data Star schema

TIME Dim

Time_key
Full_date
Day_of_month
Week_in_year
Quarter
Fiscal_period

Sales Fact

Time_key
Customer_key
Product_key
Sales_region_key
SalePerson_key
Unit price

Customer Dim

Customer_key
Customer_id
Customer name
Name
Address
Region_key

Data Snowflake

- **Snow flaking is removing low cardinality attributes from dimension and placing them in secondary tables**
- **Snow flaking is appealing because the data is normalized.**

Data Metadata

- **Greek word after, beyond**
- **Manage data about data**
- **Meta data would help in locating and managing information in the warehouse**
- **How old is the data**
- **Describe tables, attributes & relationship**
- **What subject area is available**
- **Formulas for calculating derived values**

Types of **Metadata**

- **Source**
- **ETL**
- **User**

Data Source Metadata

- **Source systems, where the data comes from**
- **Repositories**
- **Files and Spreadsheet**

Data ETL Metadata

- **Information related to the transformation of data from source to target.**

- **Data cleaning rules**

- **Mapping between source and target**

- **Rules for transformation**

- **Formula for derived values**

Data**User Metadata**

- **Gives the users the meaning of row and columns, the data they are working with.**

- **Application level dictionary**

Intelligence

- **Cubes**
- **OLAP**
- **Data mining**

Intelligence Cubes

- **Cubes contain data from one or more fact tables with the associated dimensions.**

- **The data structure will support more than just rows and columns.**

- **Cubes in geometry are 3 dimensional**

- **In Business Intelligence a cube can have more than 3 dimensions.**

Intelligence Cubes

- **Cubes can exist at detail single transaction level**

- **Cubes can exist at the aggregate level using aggregate values formed by the dimensional hierarchy.**

Intelligence OLAP

- **Online analytical processing was coined by Dr E.F. Codd in mid 90's.**

- **Technology is to make enterprise data management accessible to user and decision makers .**

- **The OLAP environment uses complex analysis (Statistical, forecasting and regression analysis) for effective decision making.**

Intelligence OLAP

- **OLAP is not relational it is based on multidimensional cube of data.**
- **There are 3 flavors of OLAP**
- **(ROLAP , MOLAP & HOLAP)**

Intelligence OLAP

- ## ROLAP
- ## Data is stored in relational format. SQL is used for analysis
- ## Very common and works with may reporting tools
- ## Lack of complex analytical tools

Intelligence OLAP

- **MOLAP**
- **Data is stored in multi dimensional model non relational**
- **Handles large amount of data and support complex analysis**
- **Data is in proprietary format not portable.**
- **Difficult to change without re aggregating.**

Intelligence OLAP

- **HOLAP Hybrid**
- **Combination of ROLAP and MOLAP**

Intelligence Data mining

- **Data mining discovers the hidden knowledge which is not possible with relational and OLAP models**
- **The selection and grouping criteria are not known in advance but derived**
- **"Customers who are buying Pc's are buying digital cameras"**
- **Data mining does not require the data to be in multi dimension or aggregates**

Presentation

- **Email**
- **Mobile**
- **Portal desktop**
- **Desktop**
- **Security Authentication**
- **Security Authorization**

How to make your BI project fail

- **Don't bother with executive sponsorship**
- **Don't define a clear mission of the project**
- **Don't involve the business users or customers during design**
- **Continue to make changes without change management**
- **Use the most recent technology with most features in next release**
- **Ignore all the risks and just forge ahead**

How to make your project succeed

- **Go back to previous slide**

Project Design Pitfalls

- **Mixing Grains in same fact**
- **Identifying and Conforming the Dimensions**
- **Choosing the Facts**
- **Storing Pre calculations in the Fact Table**

Project Role of Consultants

- **The use of consultants in your project should be controlled and limited.**

- **You should retain ownership of the project and don't let them build it for you without any technical knowledge**

- **Know the role and capacity of the consultant**

- **Are you paying for knowledge or augmenting staff ?.**

The End

- **If this ends up on E-Bay, I want a taste**

- **Thank You**

Contact Information

- **Akin Walter-Johnson III**

  http://www.peerlabs.com

  akin@peerlabs.com

  773-443-5266