

Data Vault™

“The Next Super Model”

(Patent Pending Architecture)

Presented by

Kent Graziano

Supervisor, Enterprise Data Warehouse

Denver Public Schools

Slides courtesy of

Dan Linstedt

© Core Integration Partners, Inc 2001-2004

455 Sherman St, Suite 207

Denver, CO 80203 USA

www.DanLinstedt.com – Home of the Data Vault

Agenda

- What and Why?
 - Evolution of the Data Model
 - Business and Technical Justification.
 - Modeling Pros and Cons
- Components of a Data Vault
- Constructing a Data Vault
 - Business Case Modeling
 - Dimension to a Data Vault
 - 3NF to a Data Vault
- Conclusions and Q&A



Standard Survey

- Who are you?
 - Logical Data Modelers
 - Physical Data Modelers
 - DBA
 - Data Warehouse Architect
 - Managers
- Experience
 - Data Warehouse & Data Mart Design?
 - Less than 1 yr?
 - 1-5 yrs?
 - Over 5 years?

What are the Problems Faced Today?

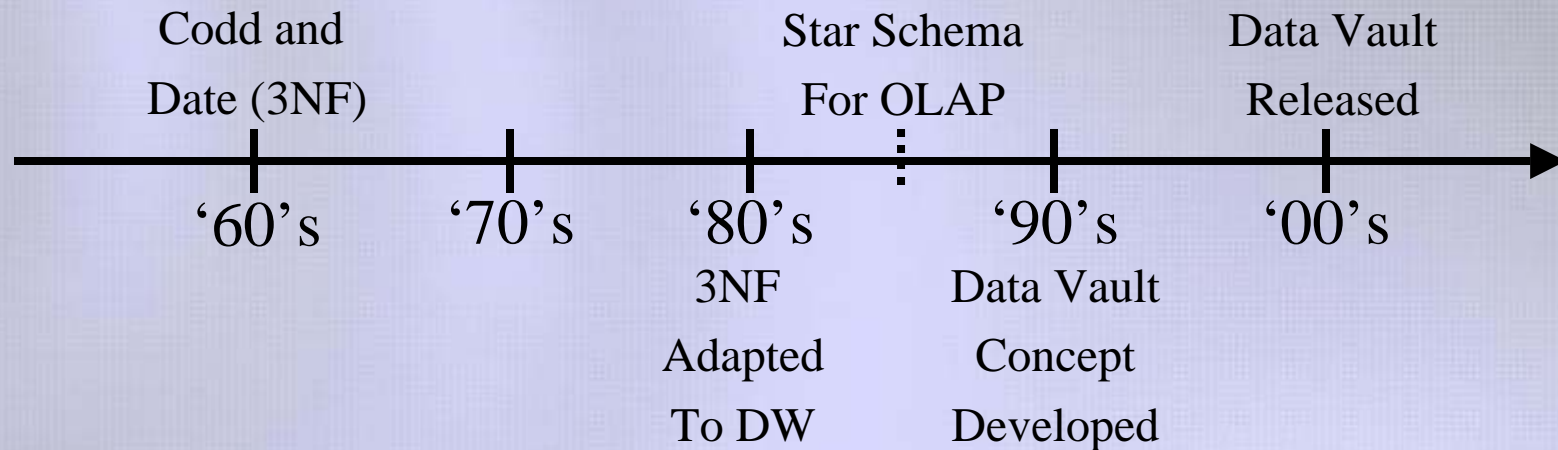
Business

- Lack of a single view of a customer.
- Lack of visibility into ALL information across the enterprise.
- Competition does it better, faster, cheaper.
- Unable to identify and forecast business trends and their impacts.
- WHERE'S THE KNOWLEDGE? OR IS IT JUST ALL DATA?

Technical

- Near-Real-Time (Active)
- Huge Data Volumes
- Massive Data Dis-Integration
- Spread-Mart
- Convergence of Operational and Strategic Questions
- Duplication of data in the ODS, Warehouse, and Data Marts!
- Dimension-itis!!
- ODS vs EDW
- Fact Table Granularity
- JUNK tables, Helper Tables

Data Model Evolution



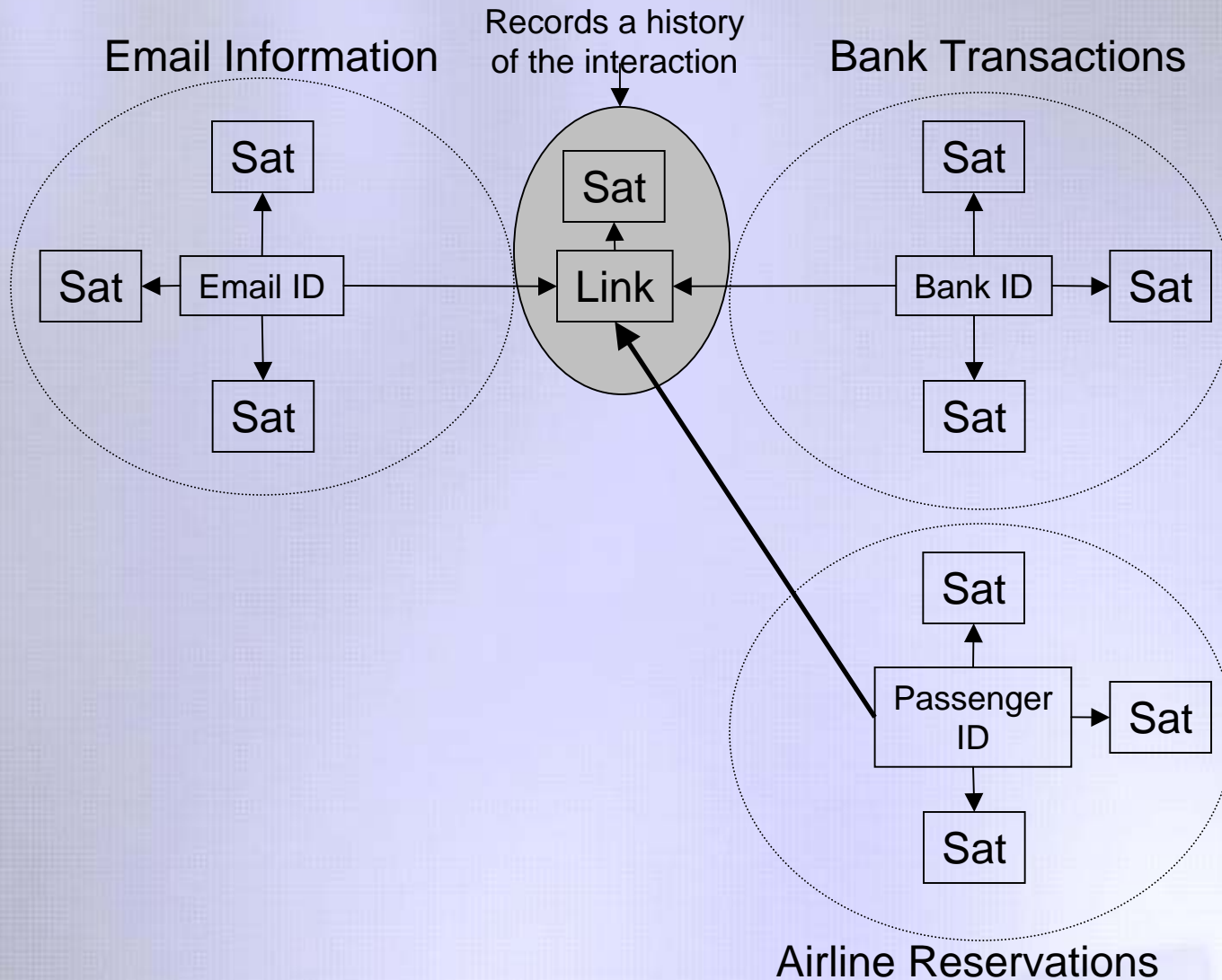
- **3NF was originally built for On-Line Transaction Processing (OLTP) systems. It was *adapted* to meet the needs of data warehousing.**
- **Star Schema was originally architected to solve subject-oriented problems. It was *adapted* to meet the needs of data warehousing.**
- **Data Vault is a hybrid, best of breed solution. The Data Vault is architected and designed to meet the needs of data warehousing. It is **NOT** an adaptation.**

Data Vault Definition

Definition: *The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema. The design is flexible, scalable, consistent and adaptable to the needs of the enterprise. It is a data model that is architected specifically to meet the needs of today's enterprise data warehouses.*

- Extensive possibilities for data attribution.
- Power of historical relationships.
- All data relationships are key driven.
- Relationships can be dropped and created on-the-fly.
- Can be used as a Data Mining source.
- Very easy to extend the model.

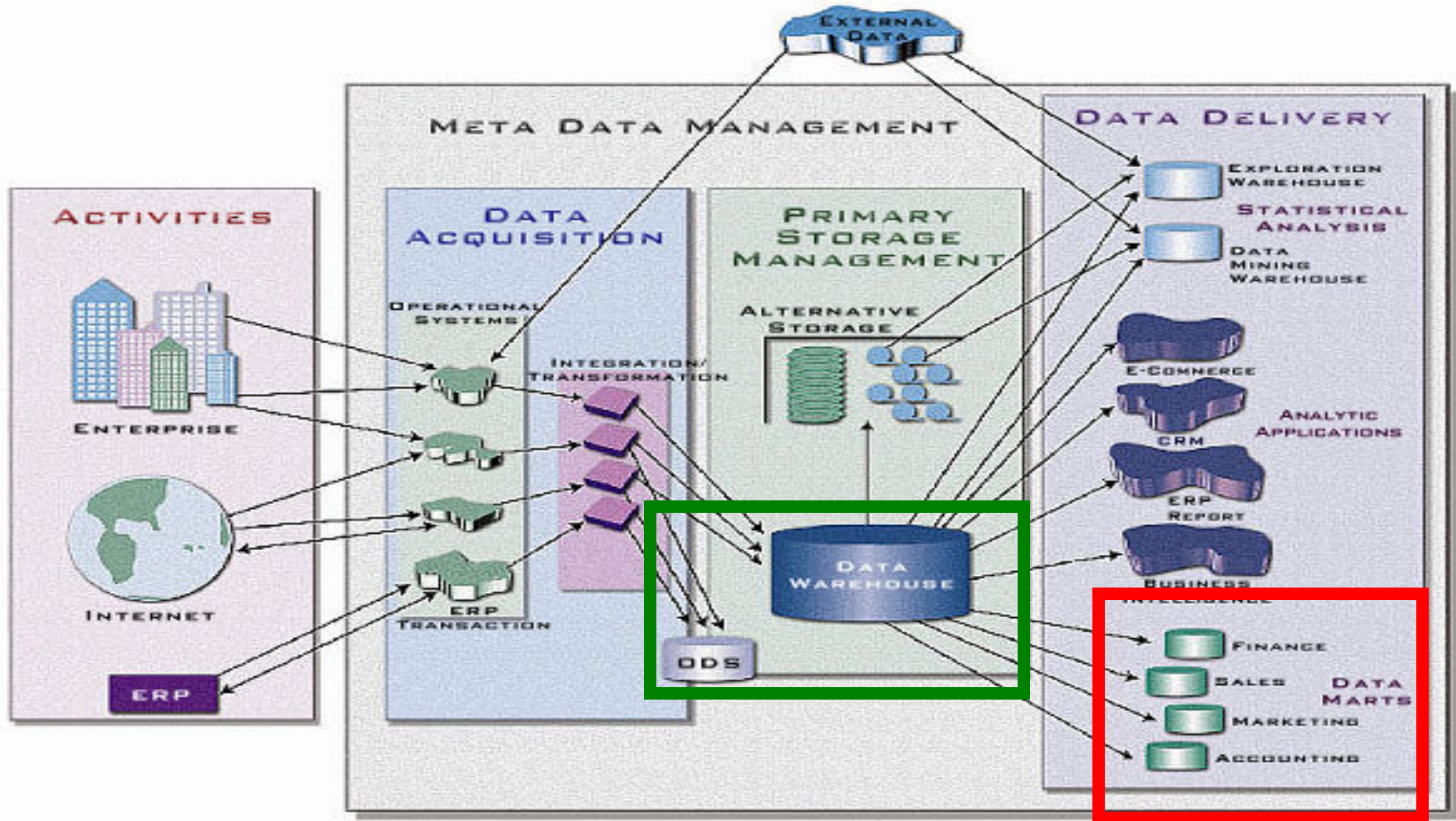
Preview: What it Looks Like



Why?

- Why do we need it?
 - We finally have a Data Model that will work for small, medium, or large business
 - Anyone building a Data Warehouse can use these techniques.
 - We've got issues in constructing the data warehouse from 3rd normal form, or Star Schema Form.
 - There are inherent road blocks to each method that we must solve technically through our Data Model.

Where It Fits - CIF

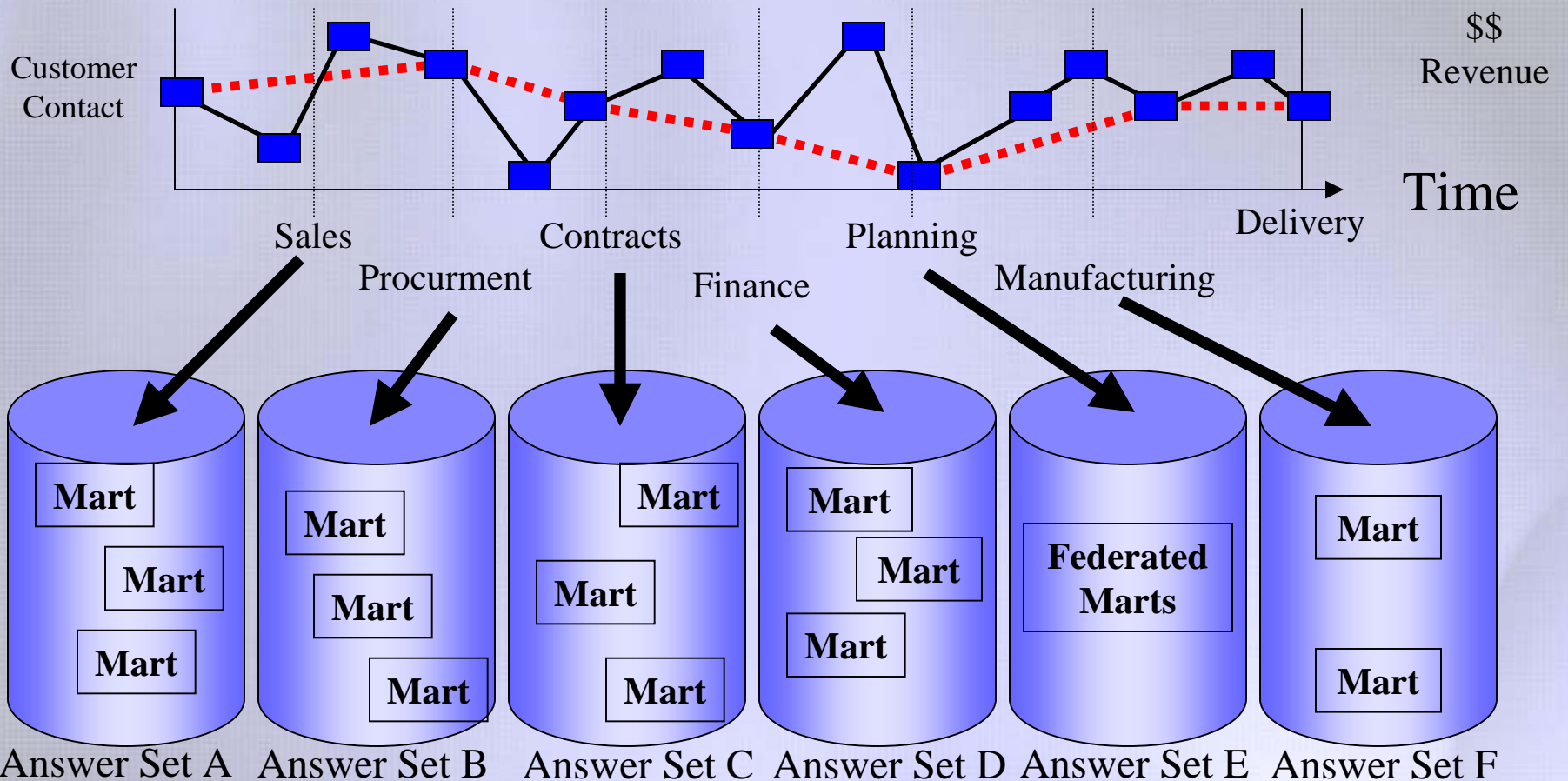


* Source: Bill Inmon and Claudia Imhoff

Technical Justification

- Provides for Multi-Terabyte Information
- Easier Detection of “Dead Data”
- Delta Driven Information
- Generation of Audit Trails
- Standard Implementation Architecture
- Restartable, Consistent Loading Patterns.
- Rapid Build of Data Marts

Business Process Chain Issues



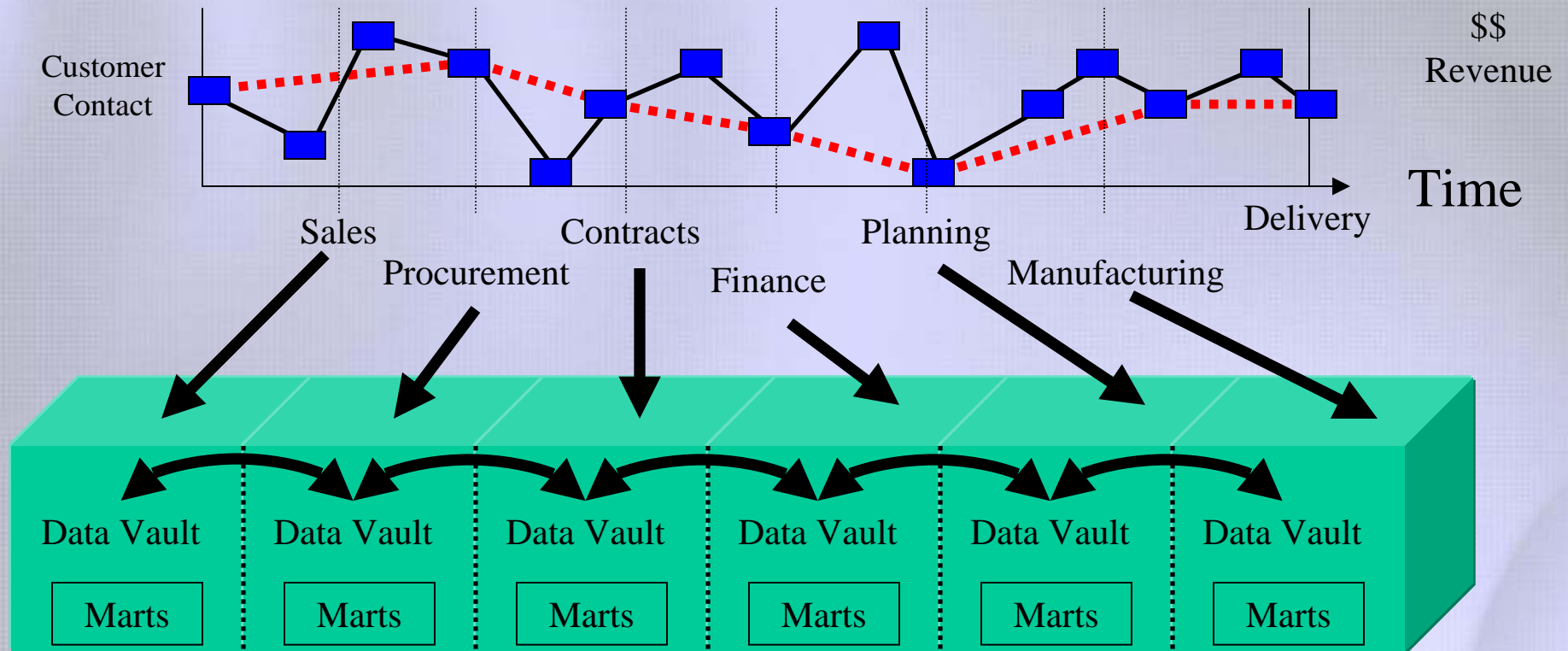
Which Answer set is right?

How do I get a corporate view of my BUSINESS?

Don't Re-Create STOVE-PIPED SOLUTIONS!



Business Process Chain Answers



- Data Vaults are built incrementally. The architecture is top-down, the implementation (build out) is bottom up.
- Data Vaults when linked together become cross-functional views of the business cycle processes, and can play a HUGE role in cycle time reduction and competitive advantage.
- Marts now give consistent answers
- Metadata and Business Rules are implemented at different levels

3rd Normal Form Pros/Cons

(With regard to EDW)

PROS

- Many to Many Linkages
- Handle lots of information
- Tightly integrated information
- Highly structured
- Conducive to near-real time loads
- Relatively easy to extend

CONS

- Time Driven PK issues
- Parent-Child Complexities
- Cascading Change Impacts
- Difficult to load
- Not conducive to BI tools
- Not conducive to Drill-down
- Difficult to architect for an Enterprise
- Not conducive to Spiral/scope controlled implementation.



Kim Loeb Collection

Star Schema Pros/Cons

(With regard to EDW)

PROS

- Good for Multi-Dimensional Analysis
- Subject Oriented Answers
- Excellent for Aggregation Points
- Rapid Development / Deployment
- Great for Some Historical Storage

CONS

- Not cross-business functional.
- Not conducive to data mining.
- Begins to fail under very large loads.
- Unable to provide integrated enterprise information.
- Not conducive to real-time loading.
- Can't handle ODS or Exploration Warehouse Requirements



Data Vault Pros/Cons

(With regard to EDW)

PROS

- Supports Near-Real Time and Batch Feeds
- Supports functional business linking
- Extensible, Flexible
- Provides rapid build/delivery of Star Schema's
- Supports VLDB/VLDW
- Designed for EDW
- Supports Data Mining and A.I.
- Provides granular detail.

CONS

- Requires Business Analysis to be firm.
- Not conducive to today's BI tools.
- Not conducive to OLAP processing.
- Not always friendly to pre-generated aggregate levels.



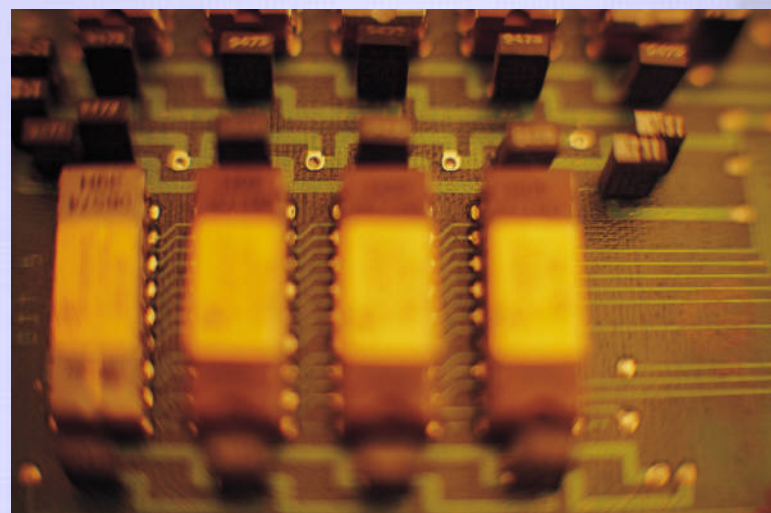
Analogy – The Porche and the Big Rig



- Which would you use to win a race?
- Which would you use to move a house?
- Would you adapt the truck and enter a race with Porches and expect to win?

Agenda

- What and Why?
 - Evolution of the Data Model
 - Business and Technical Justification.
 - Modeling Pros and Cons
- **Components of a Data Vault**
- **Constructing a Data Vault**
 - Business Case Modeling
 - Dimension to a Data Vault
 - 3NF to a Data Vault
- **Conclusions and Q&A**



Data Vault Structural Components

- Hubs = List of Business Keys
- Satellites = Descriptive Information
- Links = Describes Relationship Between Business Keys

- Point In Time = Time Picture of all satellites surrounding a single hub. (Satellite Derivative)
- Optional: User Grouping Tables (Hubs, Links, Satellites)

Common Attributes

- Primary key – PK
- Load Date Time Stamp – LD_DTS
- Load End Date Time Stamp – END_DTS
- Record Source – REC_SRC
- Load Sequence ID (optional) – LDSEQ_ID
- Update User – (optional) UPDT_USER
- Update DTS – (optional) UPDT_DTS

Some of these attributes may not be necessary if a Meta Data Warehouse is in place.

Hub Structure

A Hub is a list of unique business keys.

Sample Data Set "CUSTOMER"

ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT
2	ABC925_24FN	10-2-2000	CONTRACTS
3	DKEF	1-25-2000	CONTRACTS
4	KKO92854_dd	3-7-2000	CONTRACTS
5	LLOA_82J5J	6-4-2001	SALES
6	HUJI_BFIOQ	8-3-2001	SALES
7	PPRU_3259	2-2-2000	FINANCE
8	PAFJG2895	2-2-2000	CONTRACTS
9	929ABC2985	2-2-2000	CONTRACTS
10	93KFLLA	2-2-2000	CONTRACTS

Primary Key

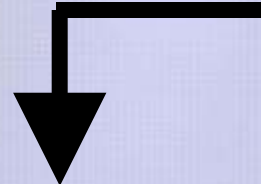
<Business Key>

Load DTS

Record Source

Satellite Structure

A Satellite is a time-dimensional table housing detailed information about the hub's business keys.



ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT
2	ABC925_24FN	10-2-2000	CONTRACTS

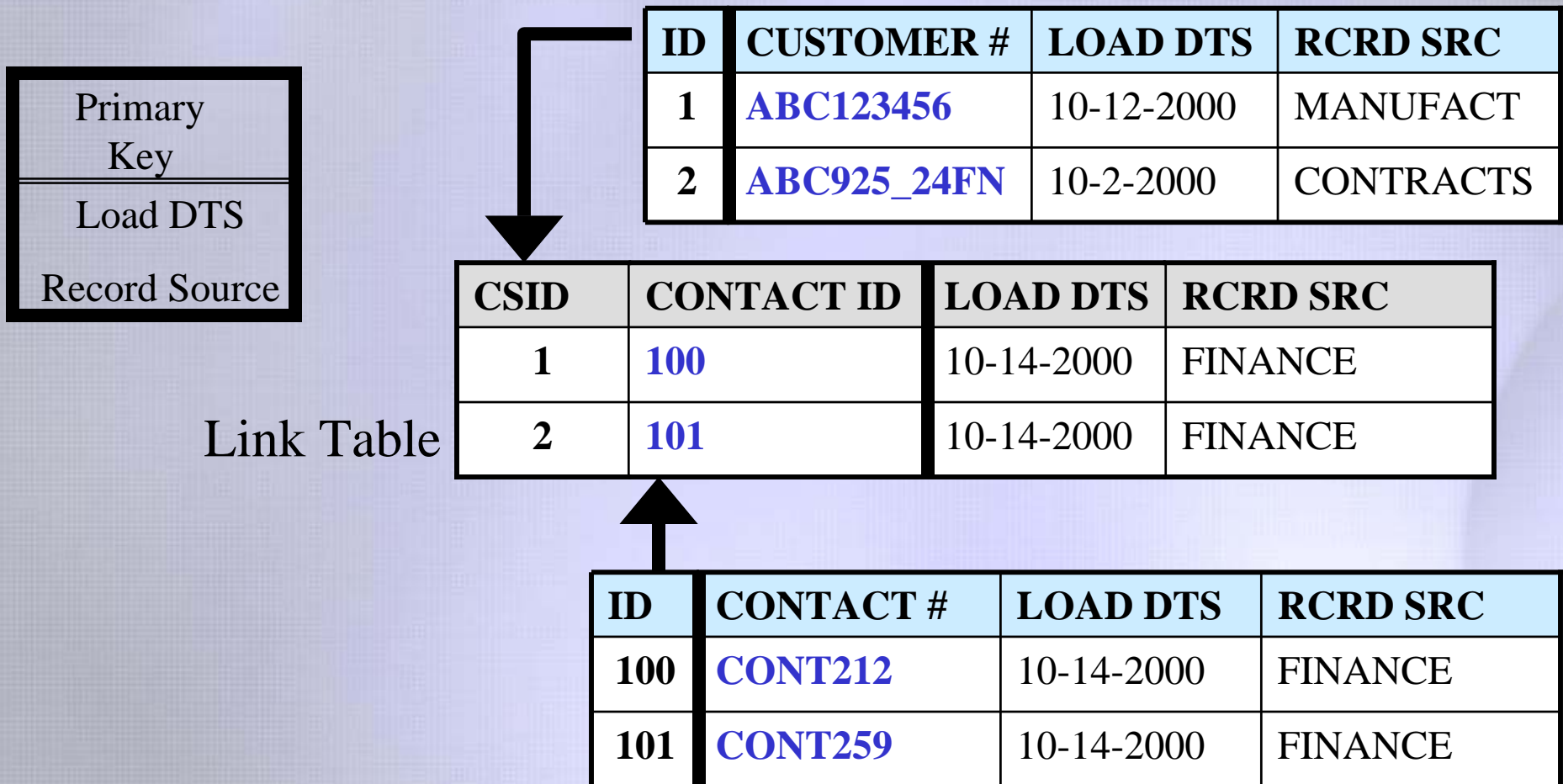
Primary Key Load DTS
Detail Business Data
Aggregation Data
{Update User} {Update DTS} Record Source

CSID	LOAD DTS	NAME	RCRD SRC
1	10-12-2000	ABC Suppliers	MANUFACT
1	10-14-2000	ABC Suppliers, Inc	MANUFACT
1	10-31-2000	ABC Worldwide Suppliers, Inc	MANUFACT
1	12-2-2000	ABC DEF Incorporated	CONTRACTS
2	10-2-2000	WorldPart	CONTRACTS
2	10-14-2000	Worldwide Suppliers Inc	CONTRACTS

CUSTOMER NAME SATELLITE

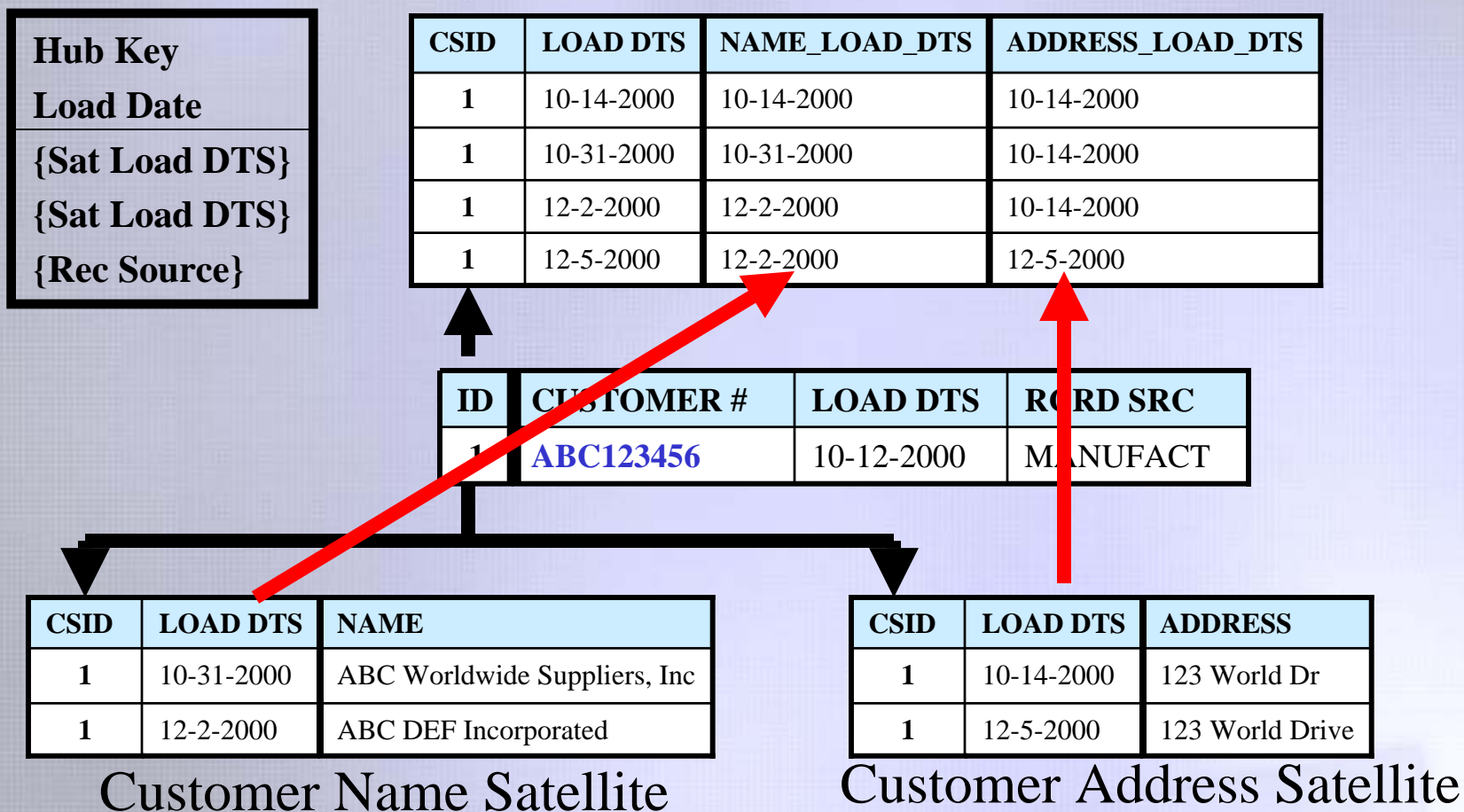
Link Structure

A Link is a many to many, representing the connection between information between business elements.



Point In Time Structure

A structure which sustains integrity of joins across time to all the satellites that are connected to the hub.



Hybrid PIT and Satellites

Only current picture / most recent is stored in the PIT Table.

CSID	NAME_LOAD_DTS	ADDRESS_LOAD_DTS
1	12-2-2000	12-5-2000

ID	CUSTOMER #	LOAD DTS	RCRD SRC
1	ABC123456	10-12-2000	MANUFACT

CSID	LOAD DTS	LOAD_END_DTS	NAME
1	10-31-2000	12-2-2000	ABC Worldwide Suppliers, Inc
1	12-2-2000	NULL	ABC DEF Incorporated

Customer Name Satellite

CSID	LOAD DTS	LOAD_END_DTS	ADDRESS
1	10-14-2000	12-5-2000	123 World Dr
1	12-5-2000	NULL	123 World Drive

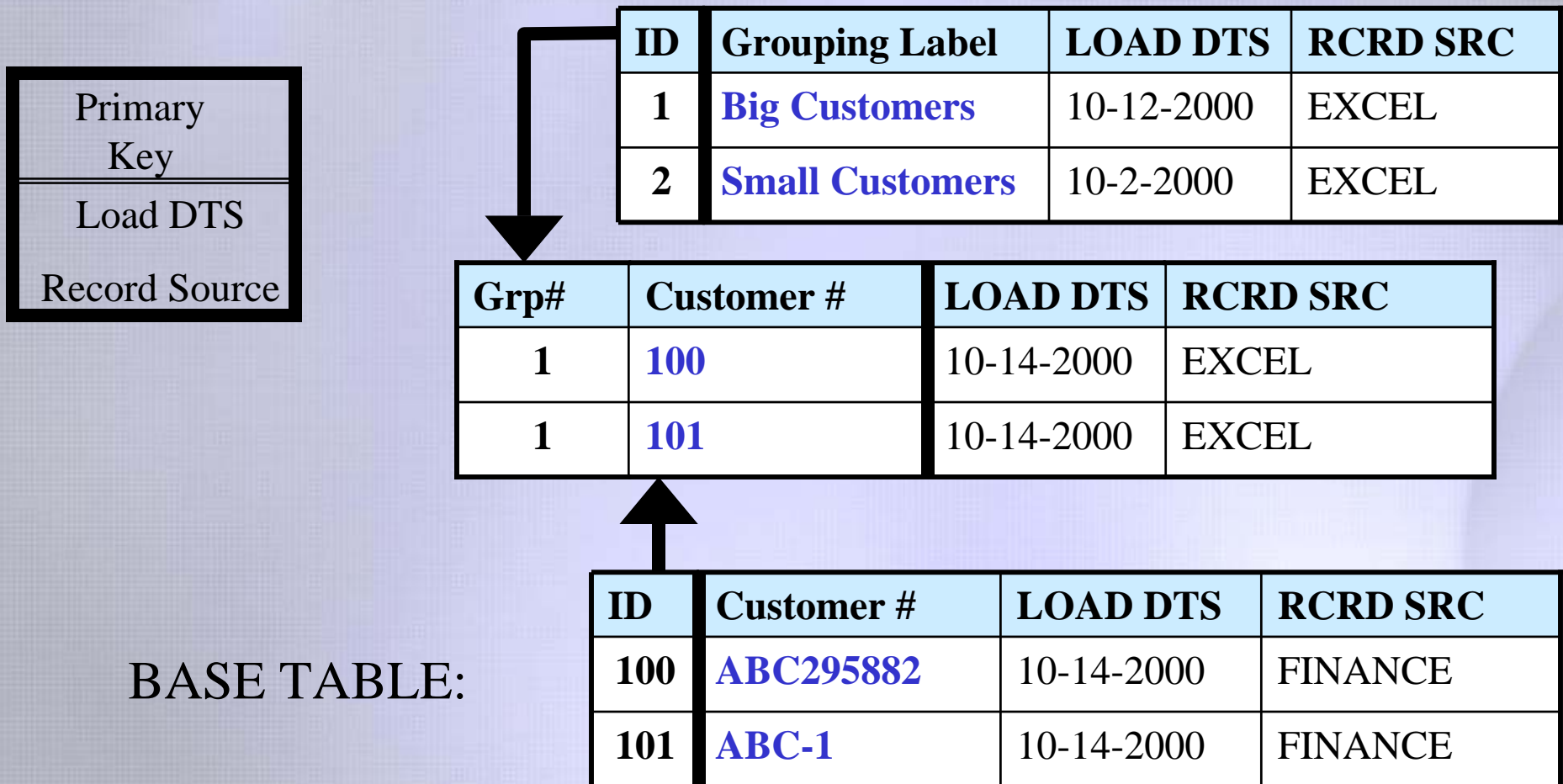
Customer Address Satellite

What is a User Grouping Set?

- A User grouping Set is a set of tables, that **Links** a user defined information hub to a Source System Feed Hub.
- The information is typically housed on the users desktop, and is not available through the source system(s).
- The information provides the user with a customized view from a reporting standpoint and does not affect the underlying information.
- User Grouping tables can define keys (parent to child) information to roll up under specific print labels.

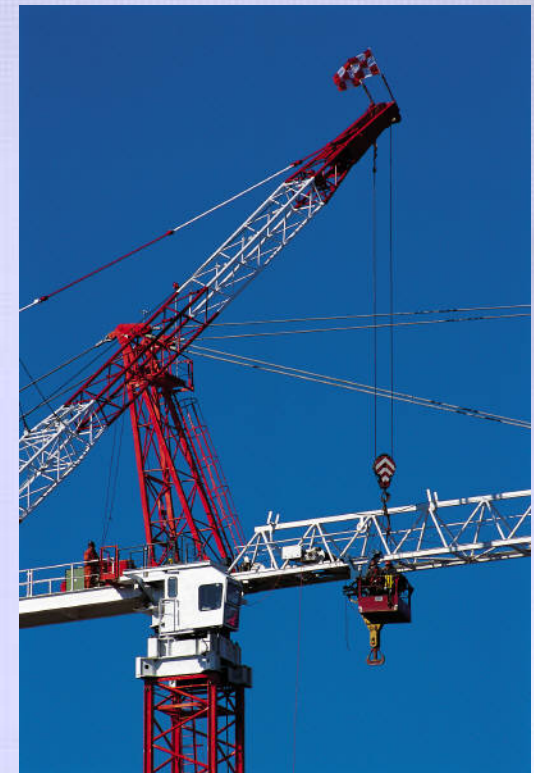
User Grouping Link Structure

The User Grouping Link, allows users to “state” how they want roll-ups to occur – in situations where source data doesn’t exist.



Agenda

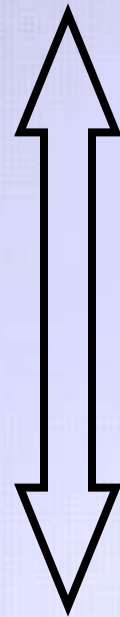
- What and Why?
 - Evolution of the Data Model
 - Business and Technical Justification.
 - Modeling Pros and Cons
- Components of a Data Vault
- Constructing a Data Vault
 - Business Case Modeling
 - Dimension to a Data Vault
 - 3NF to a Data Vault
- Conclusions and Q&A



The Modeling Process

The Business Case

All Models are
Tightly Integrated



Business Processes

Business Logical Model

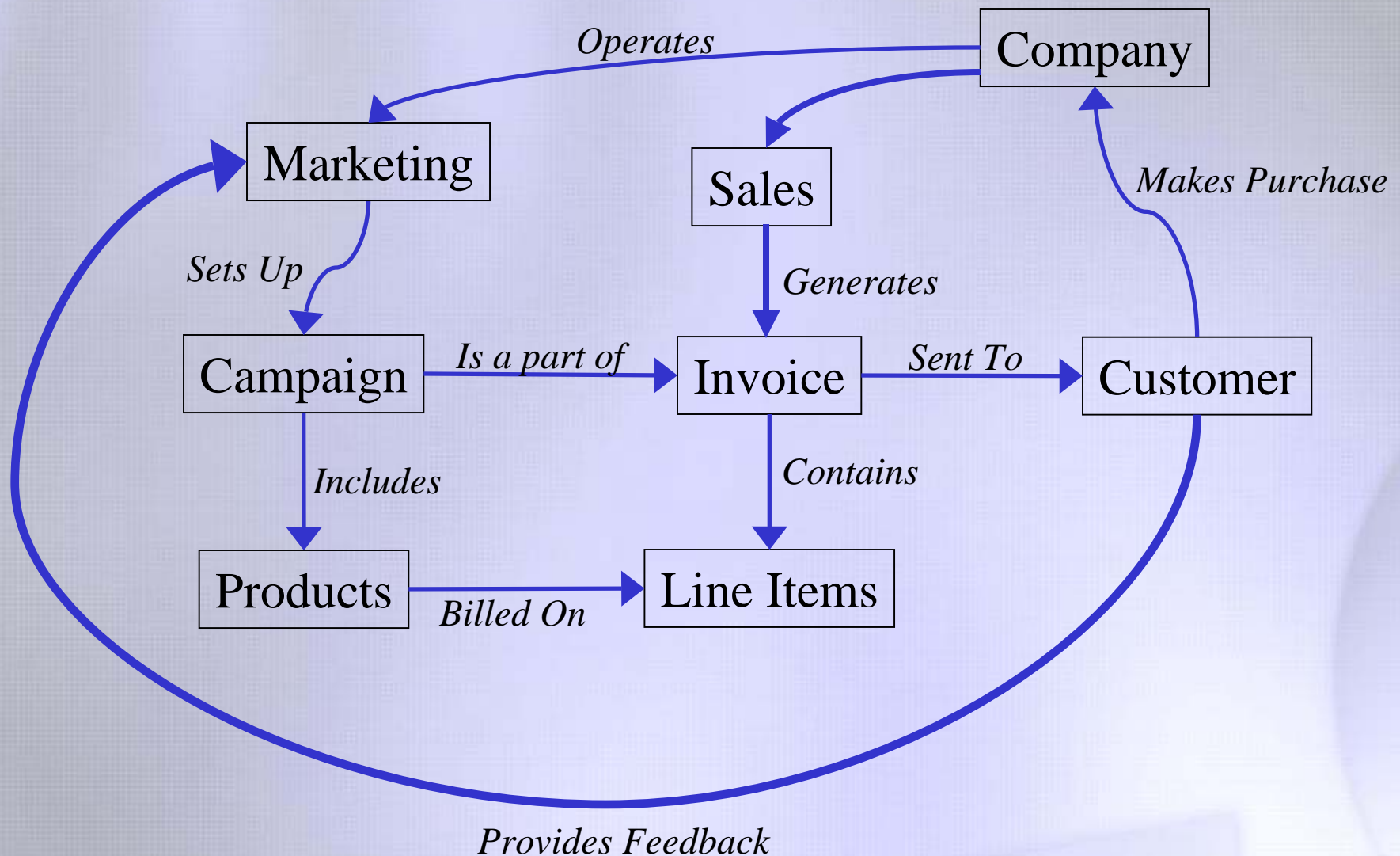
EDW Logical Model

EDW Physical Model

What does the Logical Model Look Like?

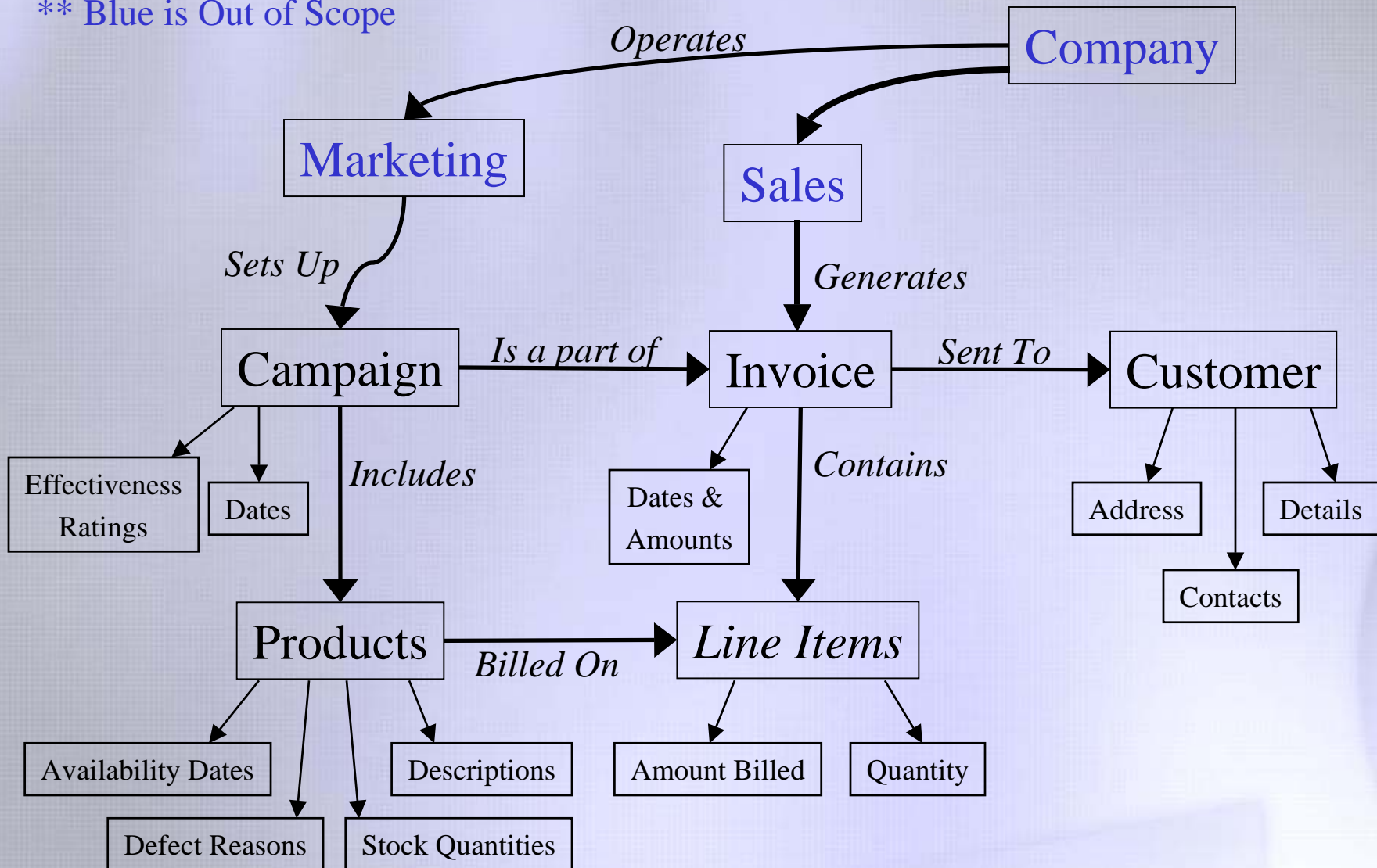
- The Business Case looks like the business logical model, which looks like the logical data model, which will look exactly like the physical model (Except of course for added attribute definitions).
- The Data Vault modeling technique focuses on aligning the business to the physical data model, so that it's easily adapted and maintained as the business changes.

Example Business Case



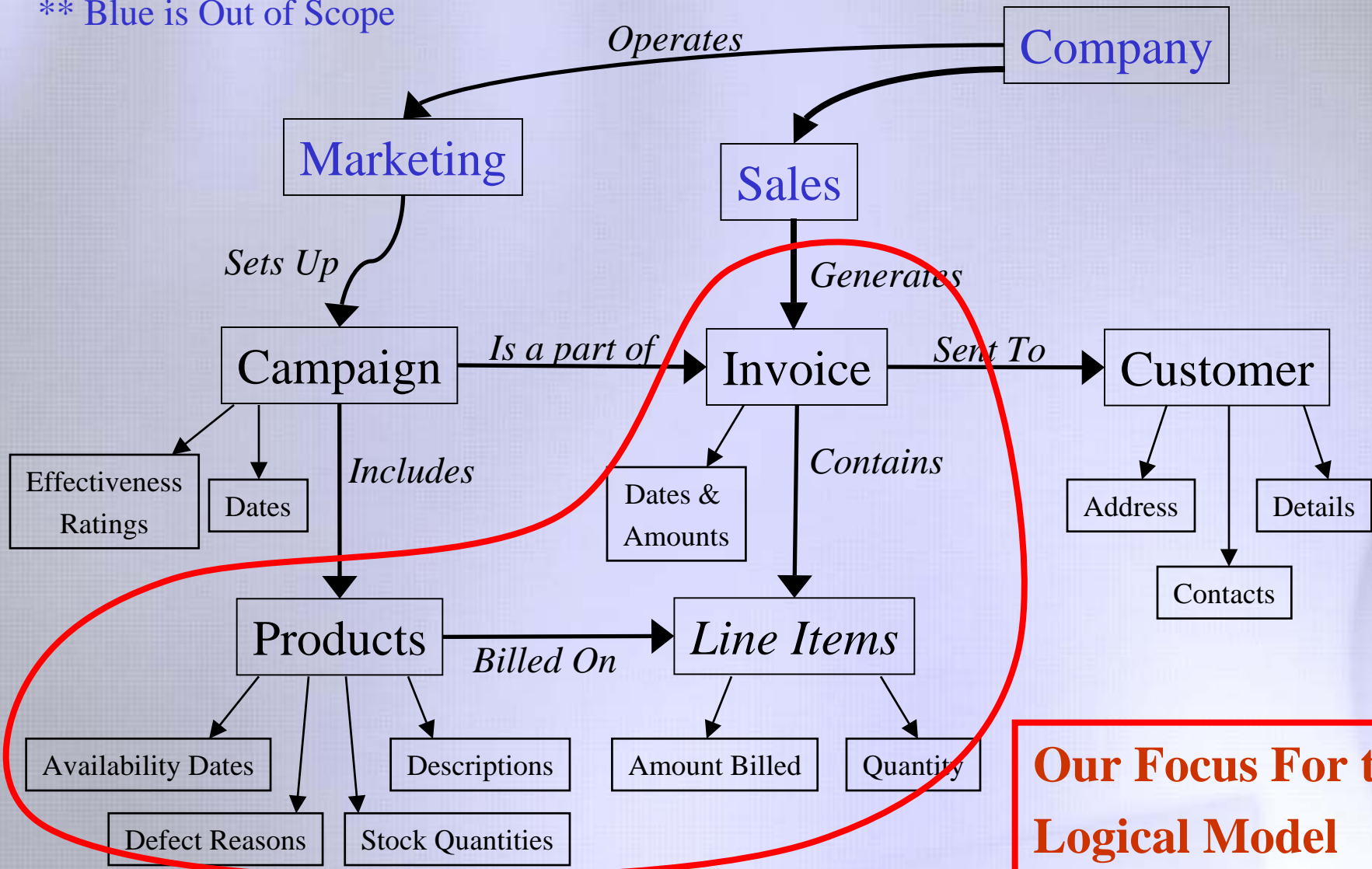
Business Logical Model

** Blue is Out of Scope

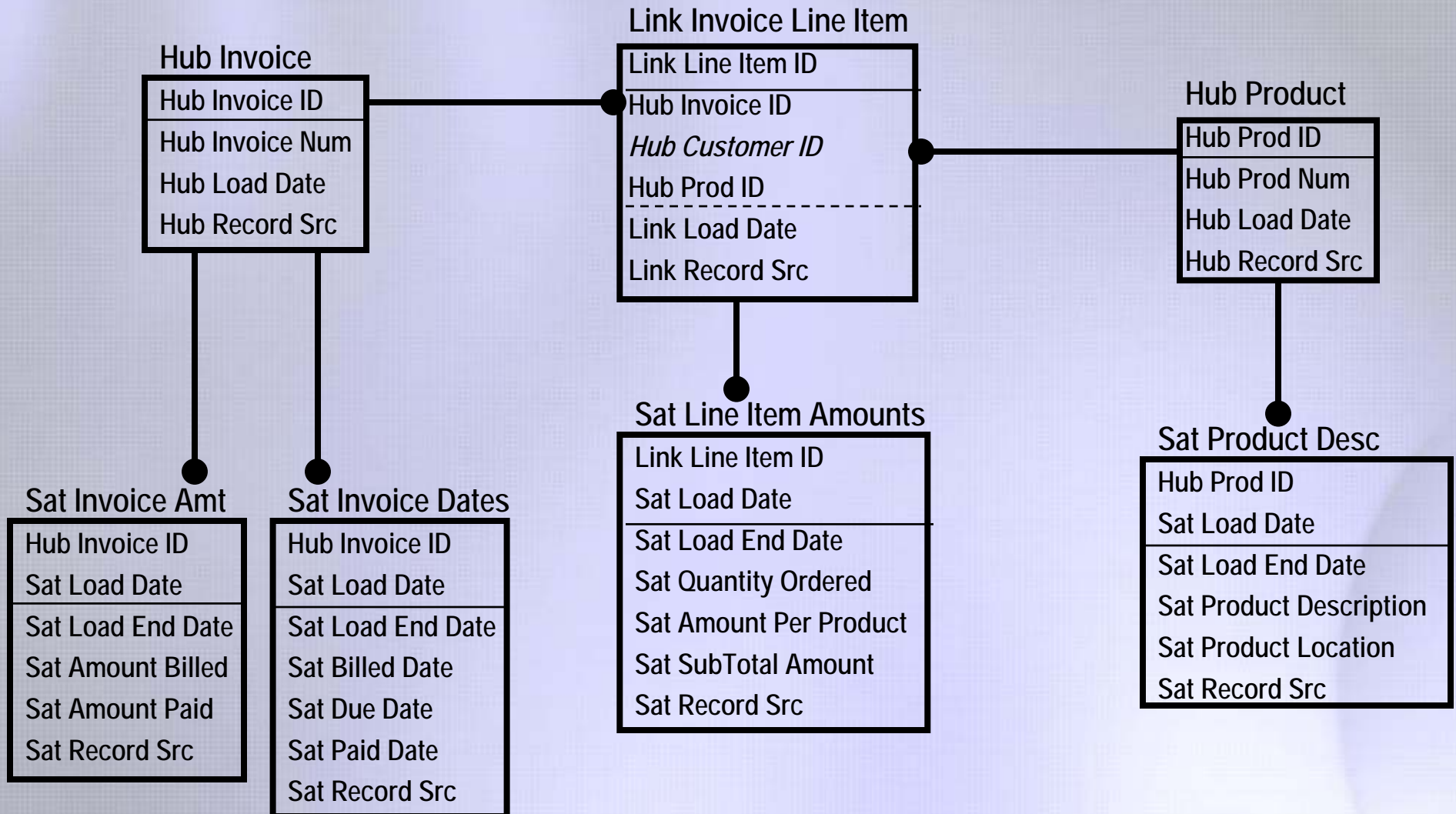


Business Logical Model

** Blue is Out of Scope



1st Cut Logical Model



Items to Note

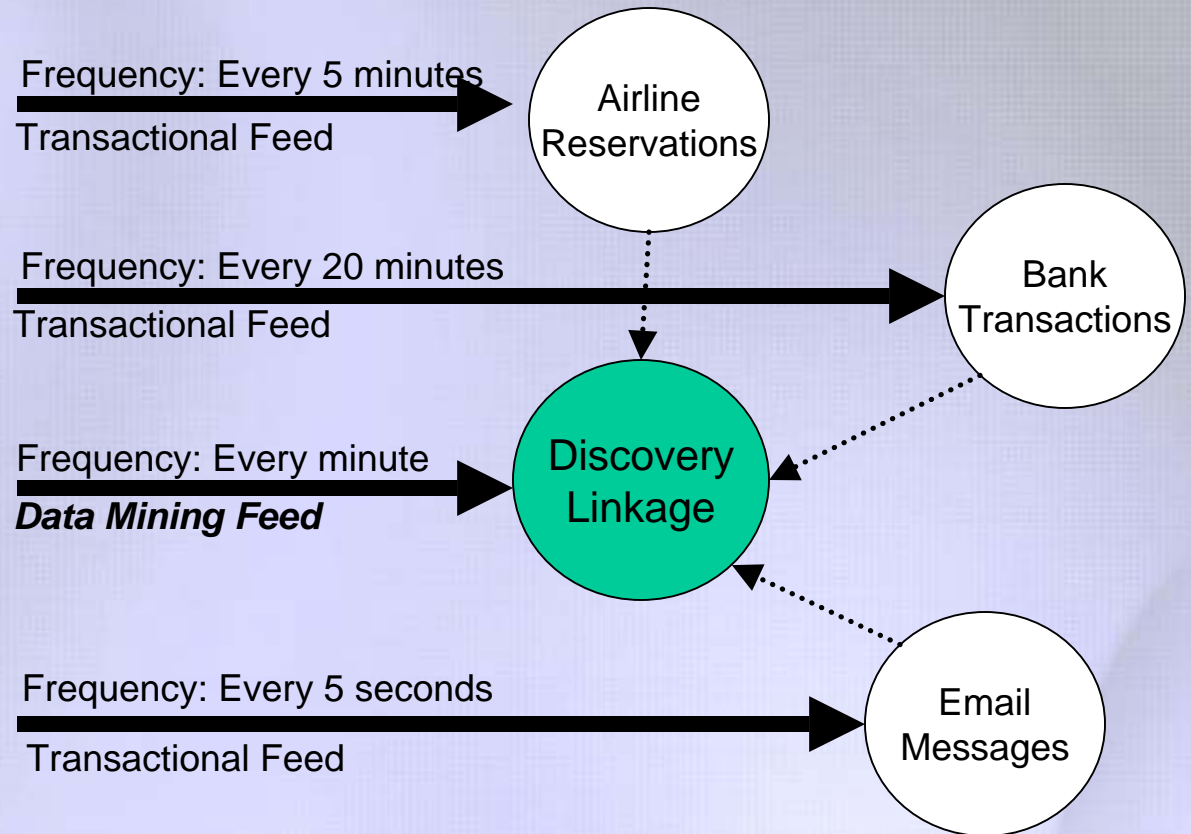
- Naming Conventions – Prefixes Used (optional)
- Load Dates are always a part of the Satellite Keys
- Load End Date utilized, if no PIT table
- Record Sources are provided for traceability
- Link Tables appear to be detailed transactions
- Surrogate keys to Business Keys are a 1 for 1 mapping.
- Satellites house only descriptive data.
- Satellites are split by rates of change of data.

Loading the Data Vault

Each set of elements is independent of the others.

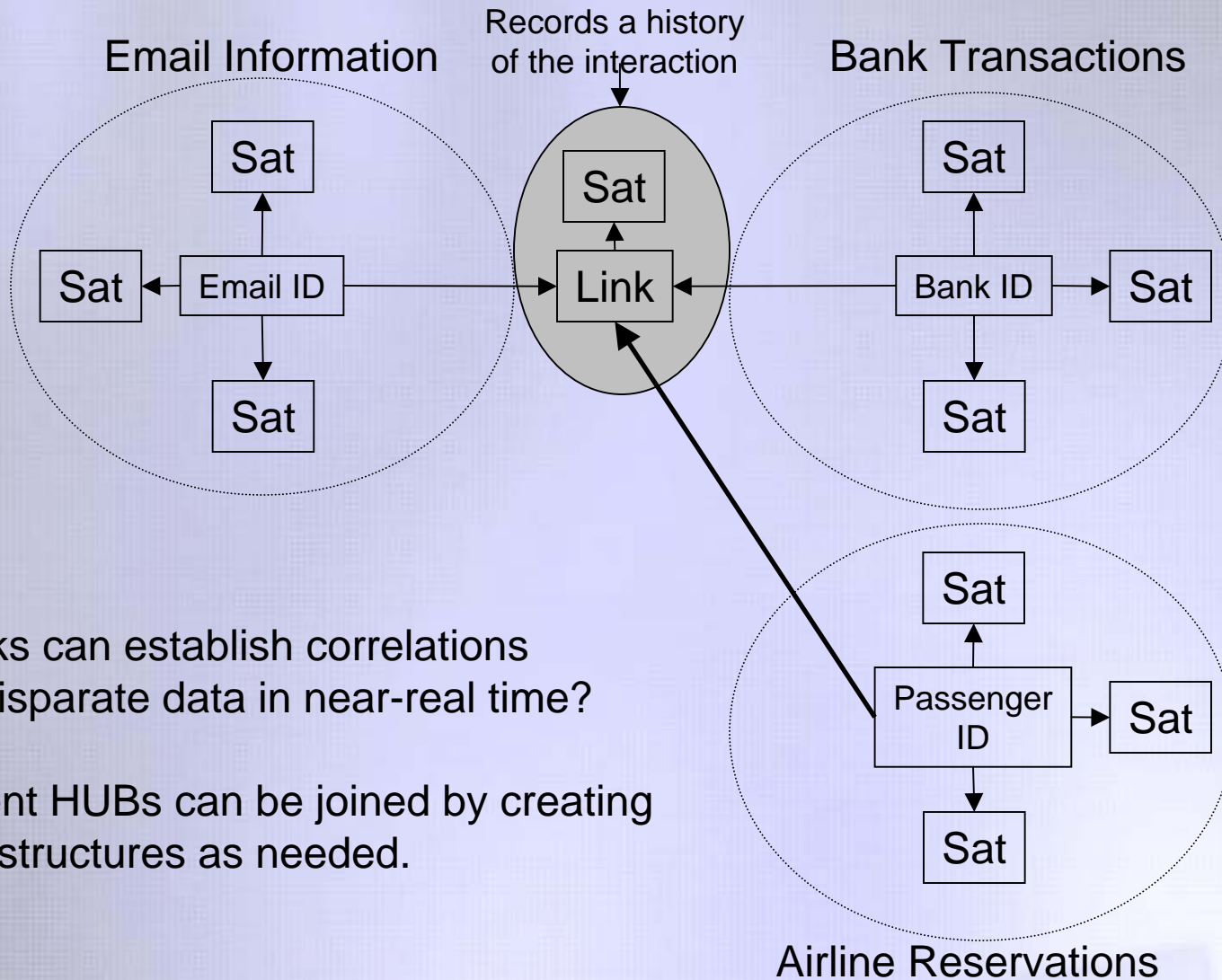
This allows the frequency of loading to vary without impact across the warehouse.

It also isolates the growth patterns to the necessary data.



The architecture is built for enterprise data warehousing, and is capable of storing massive volumes of information over time.

Dynamic Integration of Information



What if links can establish correlations between disparate data in near-real time?

Independent HUBs can be joined by creating new LINK structures as needed.

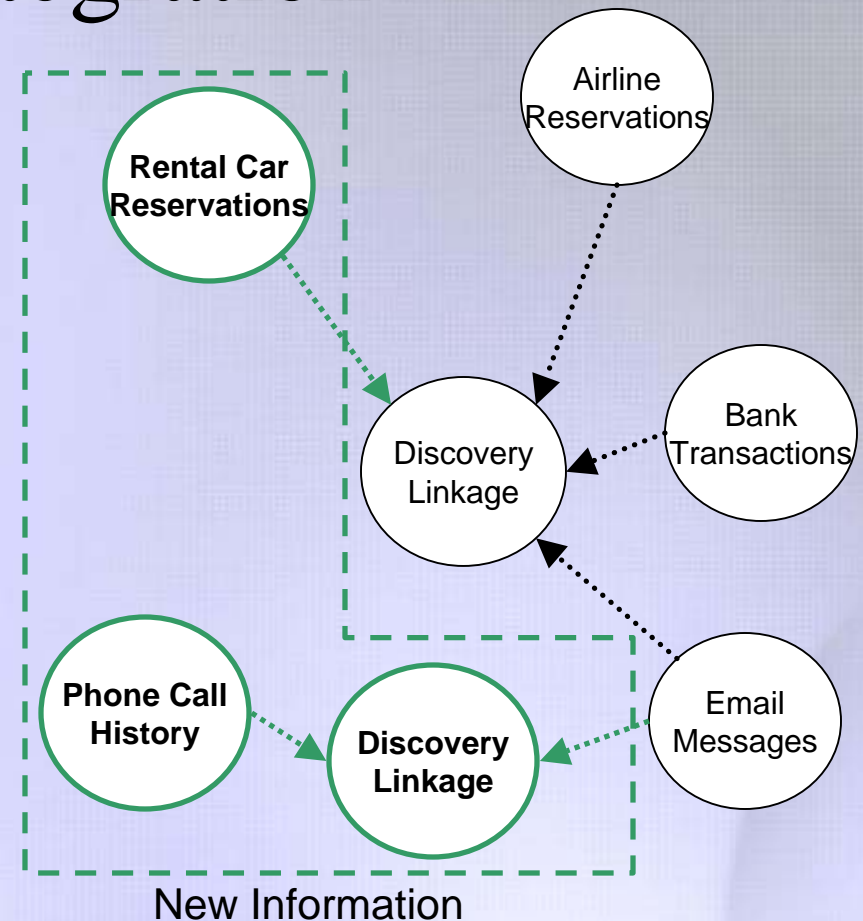
Dynamic Integration

If it were discovered that Rental Car Reservations were important, it would be easy to add the elements to store historical data, and then build the link to the other information in the discovery linkage.

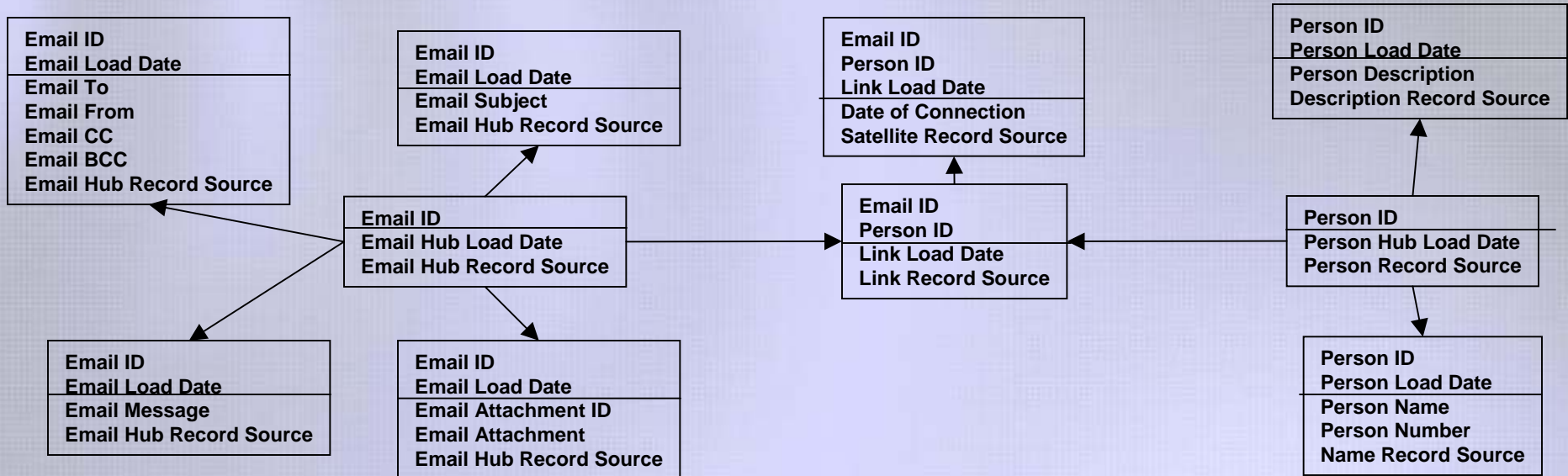
If it were discovered that phone calls had a relevance to Emails, or maybe the decision is to find out if there's relevancy to emails – the addition of a second discovery linkage and a phone call history element is easy to do without disturbing the rest of the information.

The modeling technique capable of dynamic addition of information without losing history.

*The architecture allows addition of information at any time, and the linking of the information to be tested for relevancy – if the information linked is not relevant (or no particular significance is discovered) **the link can be dissolved without losing history of the information.***



Granularity Of Information

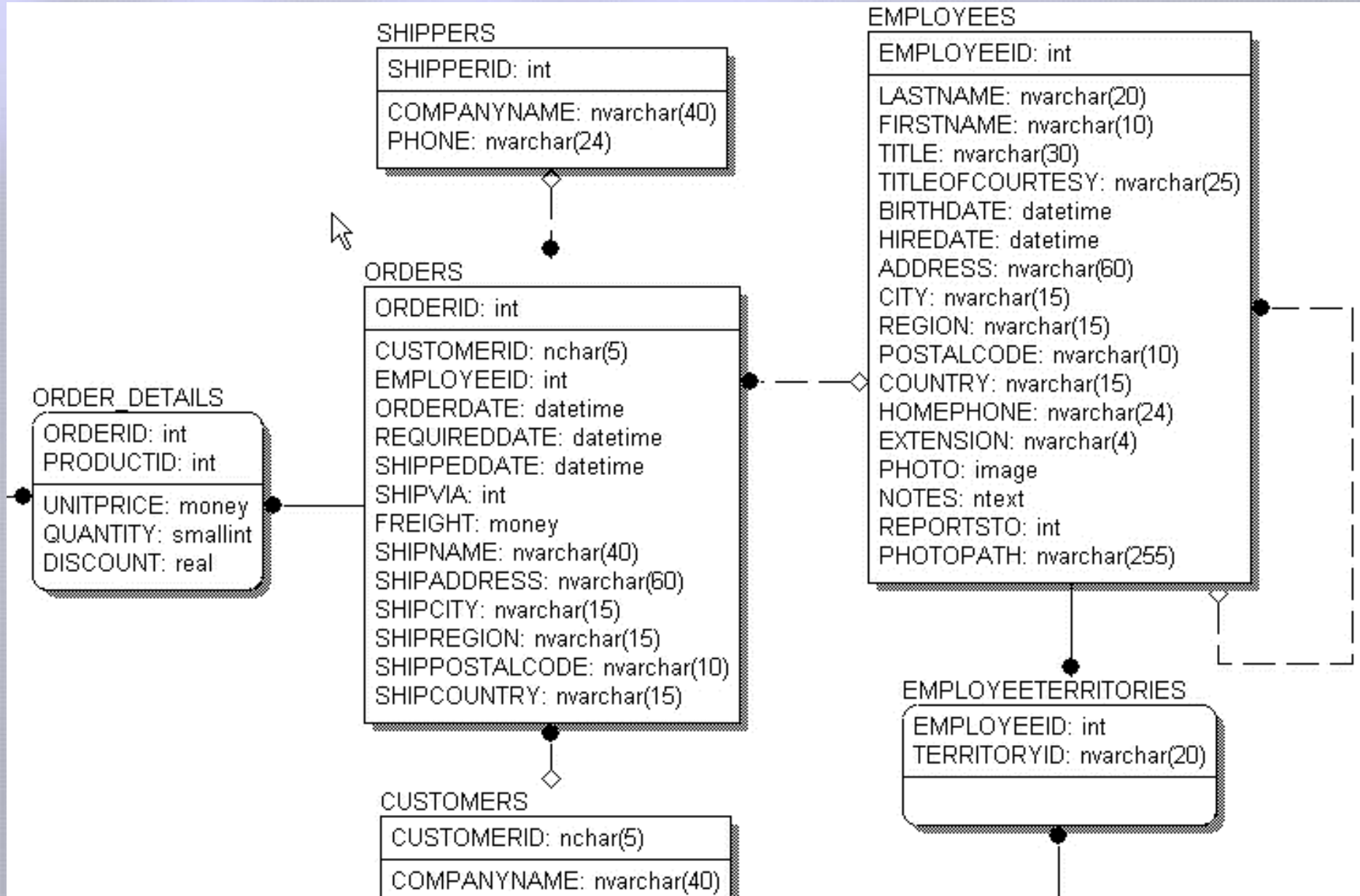


- The granularity of this information can change across the links. This allows each component (such as email and person) to store their own unique level of granularity.
- The Data Vault offers storage of information at any granularity with the least amount of data redundancy possible.
- In this case, person is linked with emails, numbers of emails, date of emails, types of emails can be grouped together in another link off the email-person linkage.
- The satellites' content is separated by: rate of change and type of information.
- **The data model is mathematically based to take up the least amount of physical space over time (compared to 3rd Normal Form and Star Schema).**

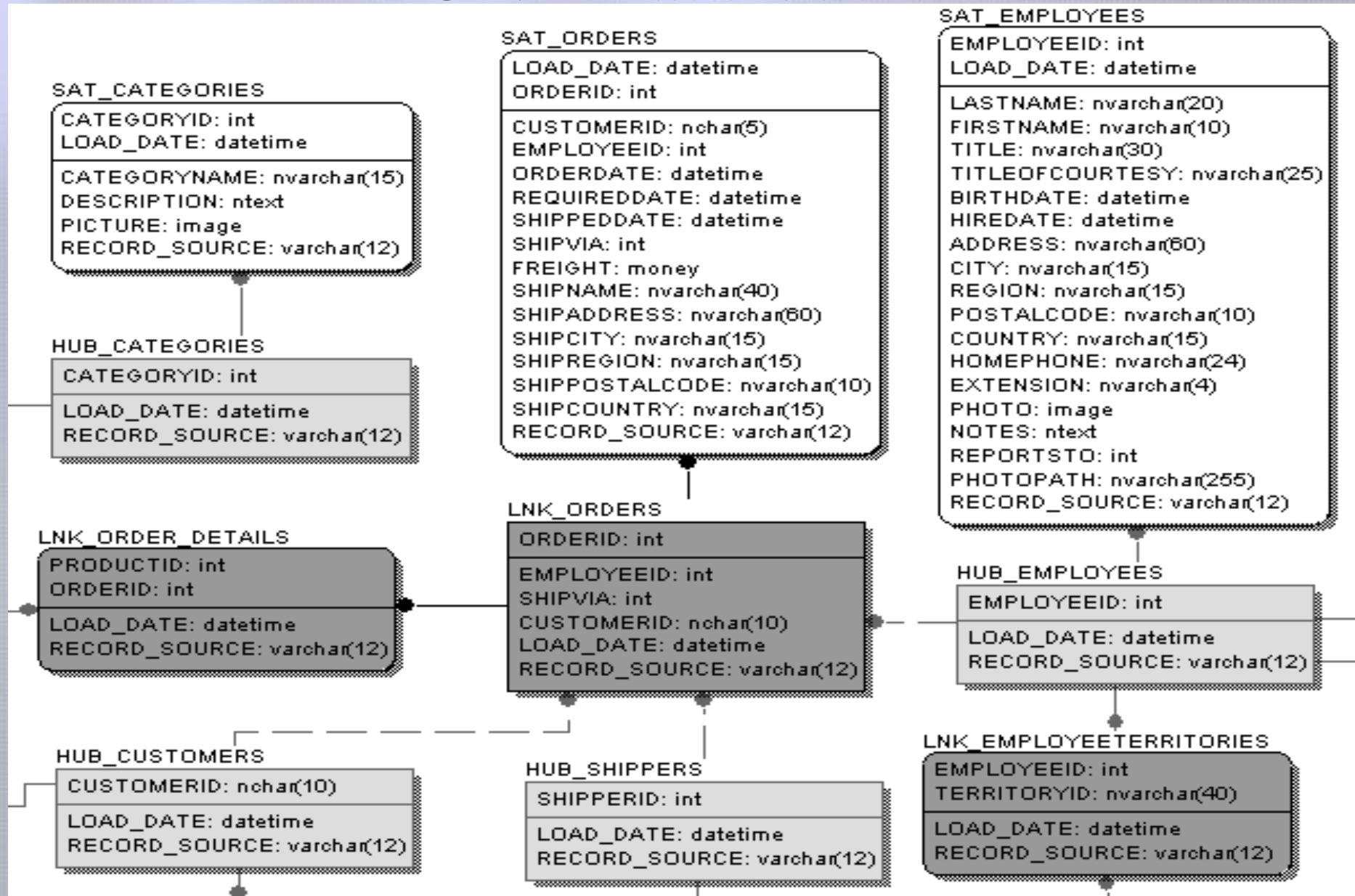
Steps To Building a Data Vault

1. **Model the Business Case**
2. **Model the Hubs** - identify key business attributes that stand on their own – place into Hubs.
3. **Model the Links** – Identify relationships between these key elements of business, and represent them as links between the elements. Link tables cannot “stand on their own”, only hubs can.
4. **Model the Satellites** – Identify the additional descriptive information you want to have to describe the business keys.
5. **Add the Point-in-time tables** to those Hubs that have more than 1 satellite.

3NF to a Data Vault



3NF Data Vault



Dimension to a Data Vault

Customer_DIM

Customer_Key: INTEGER
Customer_acctnum: VARCHAR(20)
customer_addr1: VARCHAR(35)
customer_addr2: VARCHAR(35)
Customer_state_code: VARCHAR(2)
customer_zip5_code: VARCHAR(5)
customer_city_name: VARCHAR(50)
customer_province: VARCHAR(50)
customer_country: VARCHAR(16)
phone_prefix_code: VARCHAR(3)
subscribe_date: DATE
customer_ssn: VARCHAR(11)
income_group_num: INTEGER
profitability_score_num: INTEGER
customer_last_name: VARCHAR(50)
customer_first_name: VARCHAR(50)
customer_billing_address: VARCHAR(35)
customer_billing_address2: VARCHAR(35)
customer_billing_city: VARCHAR(50)
customer_billing_state: VARCHAR(2)
customer_billing_zip: VARCHAR(5)
customer_billing_province: VARCHAR(50)
customer_billing_country: VARCHAR(16)
customer_billing_phone_num: VARCHAR(11)
customer_works_for_company: VARCHAR(45)
customer_represents_company: VARCHAR(45)

1. Define primary key
2. Split into Multiple / Like Groups
3. Define by Rate Of Change as well as type of information

Normalizing helps contain the volumes as well as defining scalability

Customer Data Vault

Customer_DIM

Customer_Key: INTEGER
Customer_acctnum: VARCHAR(20)
customer_addr1: VARCHAR(35)
customer_addr2: VARCHAR(35)
Customer_state_code: VARCHAR(2)
customer_zip5_code: VARCHAR(5)
customer_city_name: VARCHAR(50)
customer_province: VARCHAR(50)
customer_country: VARCHAR(16)
phone_prefix_code: VARCHAR(3)
subscribe_date: DATE
customer_ssn: VARCHAR(11)
income_group_num: INTEGER
profitability_score_num: INTEGER
customer_last_name: VARCHAR(50)
customer_first_name: VARCHAR(50)
customer_billing_address: VARCHAR(35)
customer_billing_address2: VARCHAR(35)
customer_billing_city: VARCHAR(50)
customer_billing_state: VARCHAR(2)
customer_billing_zip: VARCHAR(5)
customer_billing_province: VARCHAR(50)
customer_billing_country: VARCHAR(16)
customer_billing_phone_num: VARCHAR(11)
customer_works_for_company: VARCHAR(45)
customer_represents_company: VARCHAR(45)

Customer_Address

Customer_key: INTEGER
Load_Date: DATE
isa_billing: INTEGER
Address_1: VARCHAR(35)
Address_2: VARCHAR(35)
Address_City: VARCHAR(50)
Address_State: VARCHAR(2)
Address_Province: VARCHAR(50)
Address_Country: VARCHAR(16)
Record_Source: VARCHAR(6)

Customer_Name

Customer_key: INTEGER
Load_Date: DATE
Customer_First_Name: VARCHAR(50)
Customer_Last_Name: VARCHAR(50)
Customer_SSN: VARCHAR(11)
Record_Source: VARCHAR(6)

Customer_Detail

Load_Date: DATE
Customer_key: INTEGER
Phone_Prefix_Code: VARCHAR(3)
Subscriber_Date: DATE
Income_Group_Number: INTEGER
Profitability_Score_Number: INTEGER
Record_Source: VARCHAR(6)

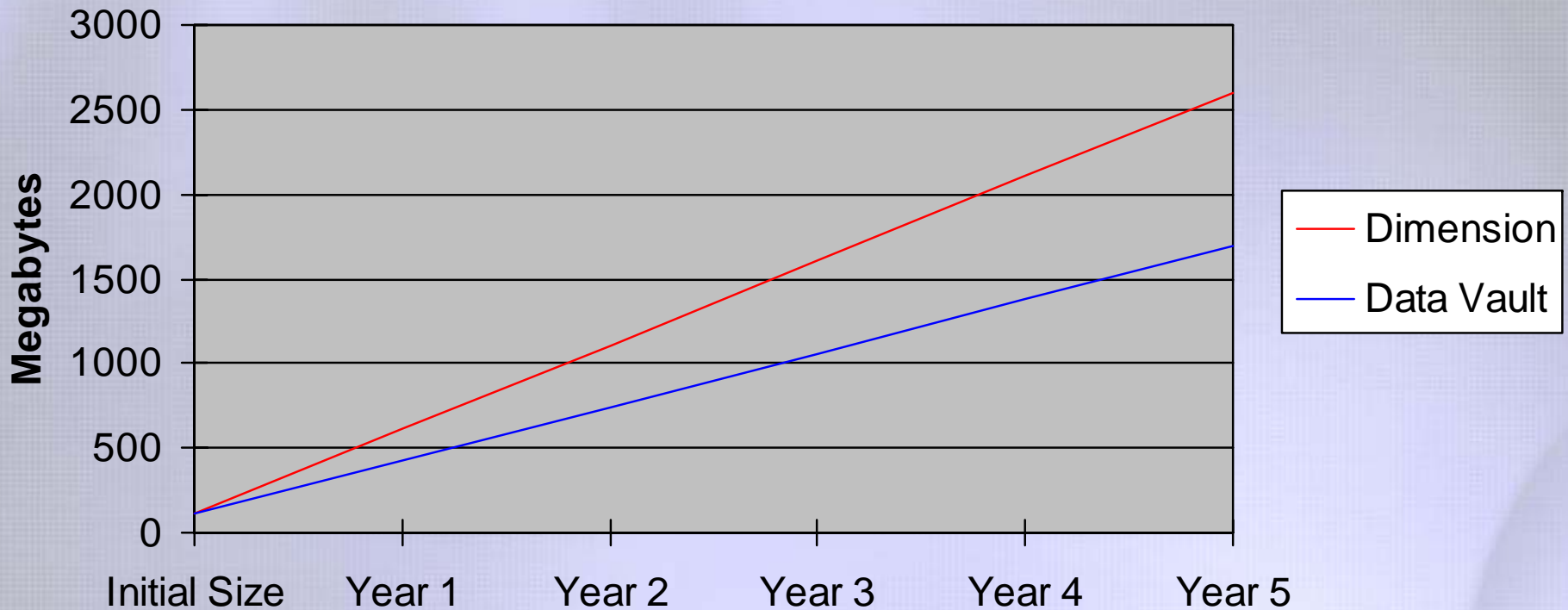
Customer_Hub

Customer_key: INTEGER
Customer_Acctnum: VARCHAR(20)
Record_Source: VARCHAR(6)
Load_Date: DATE

Customer_Company

Customer_key: INTEGER
Load_Date: DATE
Works_For: VARCHAR(45)
Represents: VARCHAR(45)
Record_Source: VARCHAR(6)

Data Vault versus Dimension Growth



	Initial Size	Year 1	Year 2	Year 3	Year 4	Year 5
Dimension	114	611.16	1108.32	1605.48	2102.64	2599.8
Data Vault	109	426.03	742.06	1059.09	1376.12	1693.15

How does the extensive growth rate affect queries?

Security and Access

- Highly controlled access is standard. No end user query allowed.
- Frequently views, or processes provide an abstraction layer between the vault and the data marts (i.e., virtual data mart).
- Vaults should not allow updates.
- Vaults should not allow deletes. All rows are typically set to a status of deleted instead.

Additional Thoughts

- A Data Vault can handle near-real-time loads, and keep the data synchronized for reporting/mart purposes.
- A Data Vault can handle in-database data mining operations, which can assign weights of relevance to the associations (link tables) between hubs.
- A Data Vault can handle terabytes of information load without breaking the architecture.
- A Data Vault can handle loading BOTH batch and Near-Real Time at the same time, just not to the same table.

Typical Topic Areas

- Employees
- Parts
- Sales Orders
- Billing Items
- Customers
- Contacts
- Addresses
- Phone Numbers
- Manufacturing Orders
- More specific and focused around business keys
- Less focused around Subject Areas.
- Granular, capable of handling trickle feed updates.
- Capable of generating Audit Trails where none existed before.

Suggested Satellites

- **Status** = Holds either the status of the business key, or the status content delivered from the source system(s)
- **Quantity** = Holds all quantity values (for example: # shipped, # ordered, # on back order, # defects, etc..) All integer or numerical values over time
- **Dates** = Holds start/stop, beginning/end, effective/end, and any other types of delivery dates regarding the business key.
- **Schedules** = Holds schedule dates information. It is separated from DATES, because the schedules for a particular business key change quite frequently.
- **Descriptive** = Holds descriptions, short & long, textual, drawings, binary information, pictorial in nature, or other kinds of “slowly” changing information (meta-data as it were) to the business.

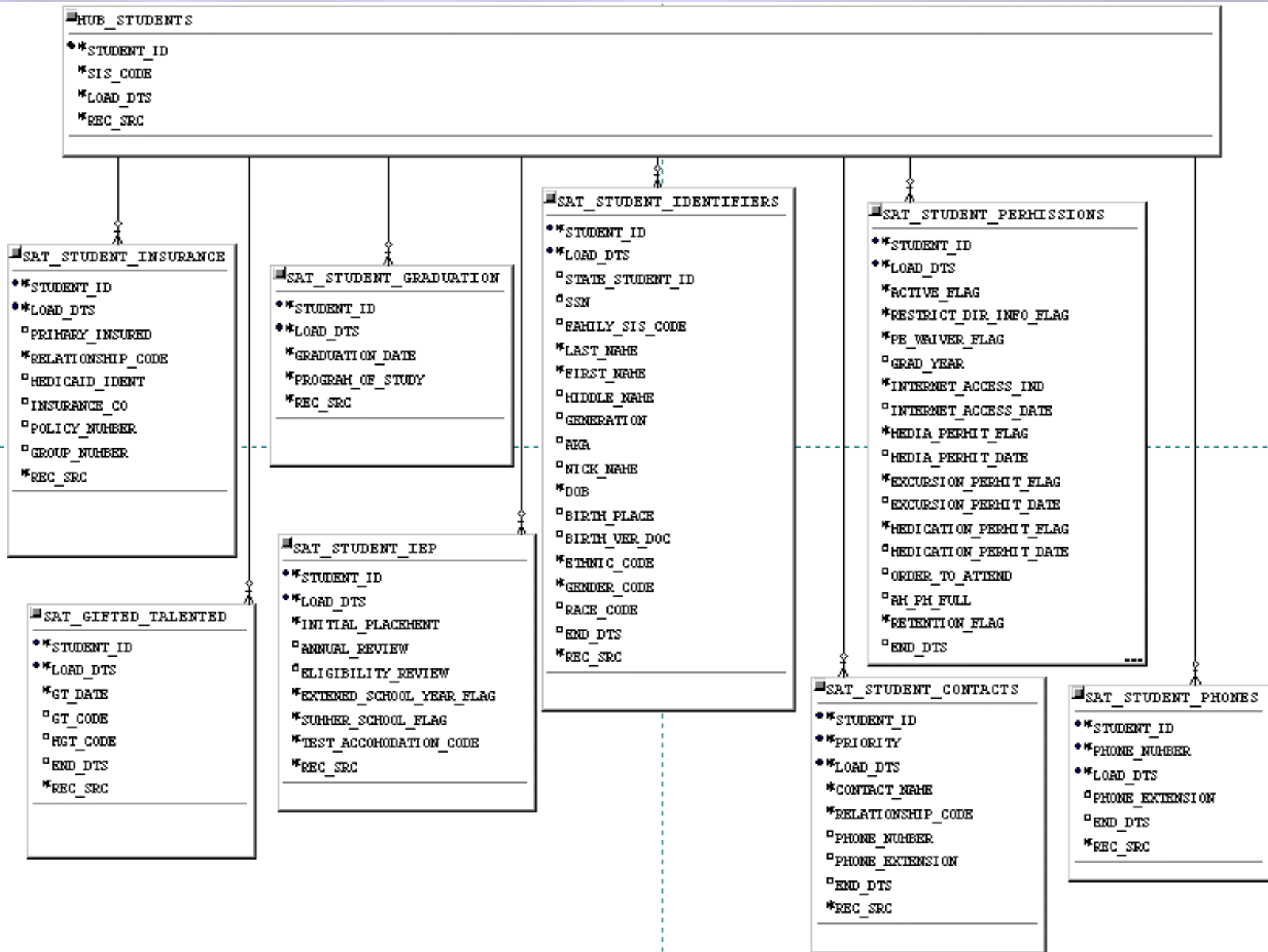
Additional Suggested Satellites

- **Item Sequence** = Holds the sequence of the item in a hierarchy. For instance: Line Item # on a Customer Bill.
- **Reason Codes** = Reasons why certain changes to this information takes place, and the dates it takes place on. If provided by the business, can be vitally important going forward for explaining certain “changes” that were stored over time in the data sets.
- **Codes & Names** = Table linkage to company standard CODES and descriptions, single satellite usually linked across different hubs.
- **Errors** = Holds errors for each row loaded, can detail for the user what are the problems (according to the business) with the data.
- **Hours** = Holds hours, decimal, or time card information for the topic.
- **Geographical** = Spatial location data, large complex floats, or other numeric types.

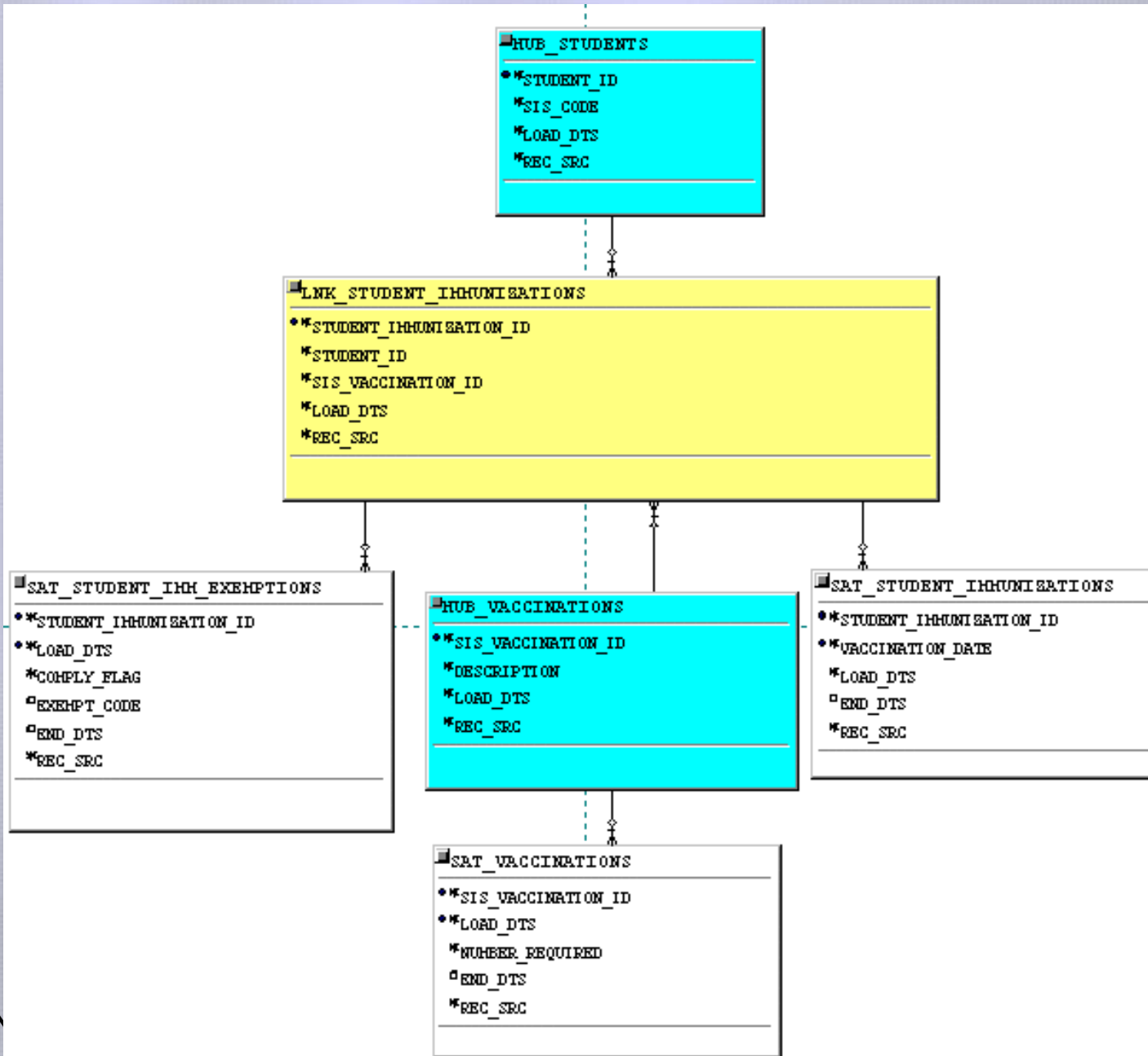
Data Vault at DPS

- Hub_codes
 - Sat_codes
- Hub_students
 - Sat_student_immunizations
- Hub_employees
 - Sat_employee_dates
 - Sat_employee_names
- Hub_schools
- Lnk_school_enrolments
- Lnk_teacher_schools

DPS Data Vault



DPS Data Vault



Industry Quotes About Data Vault

Bill Inmon: “The Data Vault is functionally strong and a viable architecture in implementing your Enterprise Data Warehouse.”

Kent Graziano, Denver Public Schools: “We believe that this data modeling technique is the best suited for designing a central, historic data repository because of its flexibility to easily add new subject areas and attributes. This will allow us to grow the EDW in an organic manner, over time, as we discover new requirements.”

Clive Finkelstein: “This should be called the "Foundational Warehouse Model", and it looks to be a solid implementation paradigm that's highly scalable.”

Stephen Brobst, CTO Teradata: “The Data Vault is foundationally strong and exceptionally scalable architecture.”

Doug Laney, META GROUP: “The Data Vault™ is a patent-pending technique which some industry experts have predicted may spark a revolution as the next big thing in data modeling for enterprise warehousing....” (Wilshire Conferences, Enterprise Data Forum Brochure, November 4-7, 2002),

Questions?

The Home of the Data Vault

[HTTP://www.DanLinstedt.com](http://www.DanLinstedt.com)

Implementation Specialists

www.coreintegration.com

Free White Papers

www.TDAN.com

Customers Using the Data Vault:

Denver Public Schools

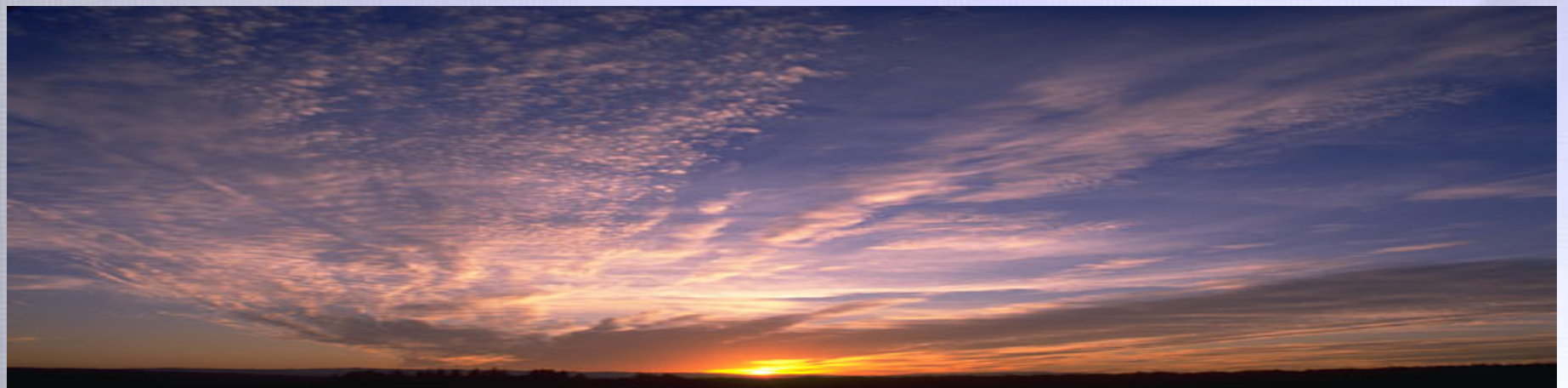
Colorado Dept. Of Corrections

State Court of Wyoming

Federal Express

US Dept. Of Agriculture

Anthem Blue-Cross Blue Shield



Desktop 2005 -

You're Virtually There!

February 15–17

All you need is a computer and an internet connection to attend the most comprehensive Oracle virtual conference of the year.

- Live Demonstrations
- Chat Rooms with Oracle Experts
- Case Studies
- Best Practices
- Tips and Techniques
- Virtual Exhibit Hall



Sponsored By

ORACLE®

To register visit www.odtug.com

My Contact Information

- Kent Graziano
 - Kent_graziano@dpsk12.org