Data Integration and Warehousing using the Data Vault

DAMA Calgary Chapter - May 25, 2016

Bruce McCartney
bruce.mccartney@dbinfosystems.com

DBIS

Agenda

- Introduction
- The Business Problem – challenges
  - Data Explosion – Big Data
  - Constant Change (ETL eats your lunch)
  - Complexity – dependencies cause exponential chaos
  - Business Domain missed
  - Agility – need to see it and change it
- Enter the Data Vault
  - What is a Data Vault?
  - When do you use a Data Vault?
  - Introduction to Key aspects
- The Business Solution

Introduction

- My background
  - Data guy
  - Oracle guy
- How I found Data Vault
- Purpose of this presentation
  - Introducte concepts
  - Nothing to Buy☺
  - Structure of this talk:
    - challenges, concepts to solution

What is a Data Vault?

- According to Data Vault Inventor (Dan Linstedt)
  - "*A detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business.*"
  - A Modeling approach, DV 2.0 now include complete Architectural Blueprint
- When do you use a Data Vault?
  - Enterprise Data Warehouses – Inmon or Kimball Style
    - Bill Inmon: "*The Data Vault is the optimal choice for modeling the EDW in the DW 2.0 framework*"
  - Data Integration/Migration Projects
    - Merger/Acquisitions requiring data *alignment*
    - Data migration projects – upgrades/migrations
    - Master data management initiatives

Challenges in Data Architecture

- Getting it right – “the truth”
- Integration
- Compliance
- Time dependency
- Modeling

The Truth- Your Business Rules

- There is no truth, only facts as they were at the time
  - Truth is subjective and changes over time with the application of **business rules**
- Modeling things *in advance* forces you to do two hard jobs:
  1. Mind reader
     - Assume user meaning and business context
  2. Fortune teller
     - Predict future
       - Business requirements, additions etc.
       - Relationship meaning/aging
- Yahoo vs. Google Analogy
  - Taxonomy (arbitrary structure) vs. Use Frequency (actual state) for **search**
- Automation and (deep) machine learning (Rise of the Robots – Martin Ford)
  - Predictive analytics etc.
- "We have come to trust our screens"(Future Crimes by Marc Goodman)
- "Bad data is a business process problem, not an information technology problem" – Dan Linstedt

DBIS

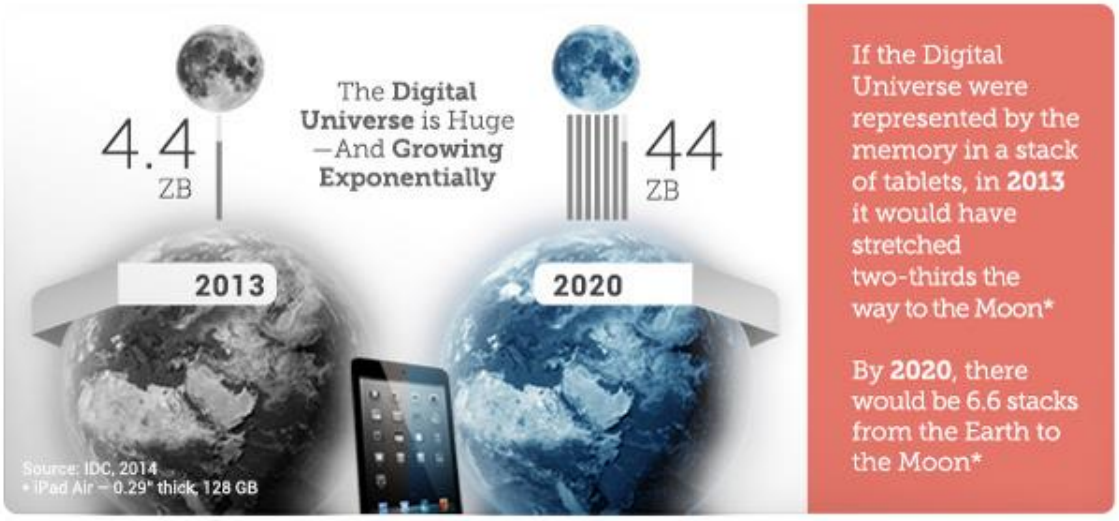
Integration – Big Data

- Business Keys
  - Unique? global?
  - Use of “smart keys”
  - Multiple systems carry different and same parts of data objects
- Quality varies by source system
- Timing
  - Business Cycles
  - Global Enterprises
- EAI vs. EII Architecture
  - Are we integrating ***process*** or ***data***?
- 4Vs of Big Data
  - volume, variety, velocity, and veracity

Integration – Big Data

- Internet of THINGS

Like the physical universe, the digital universe is large – by 2020 containing nearly as many digital bits as there are stars in the universe. It is **doubling in size every two years**, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes.



http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Compliance – Simple and Consistent

- ISO 9001 – Quality/Process
- SEI Capability Maturity Model (CMM) Level 5 (Repeatable, consistent, redundant architecture)
- Manage and Enforce Compliance to Sarbanes-Oxley, HIPPA, and BASIL II in your Enterprise
- Require Audit history including before/after values
- Need to be able to reconstruct source data at any point in time
  - Prove it!

DBIS

Time dependency – agile and automatic

- Scaling loads to near real-time
  - EAI and SOA
  - Event based paradigm
  - Log based architectures
    - LinkedIn (https://engineering.linkedin.com/distributed-systems/log-what-every-software-engineer-should-know-about-real-time-datas-unifying)
- Source systems often don't remember old values of data, change keys etc.
- Near real-time OLTP mixed with Batch 'header file' updates
  - Eg. Bank ATM Transactions without customer file built
- There are 3rd party tools to **automate** DV build
  - BI Ready, WhereScape, AnalytixDS, Quipu
  - Home grown, Dan had a tool called RapidACE (SaaS)

Modeling - Adaptability

- **Need to Adapt** some of these:
  - 3NF
    - Rework and inflexible
  - Star Schema Structure
    - Type-2 Dimensions
    - Aggregation and help tables
    - Snowflakes
  - Anchor Model
    - More tables instead of attributes grouped
- Operational Data Store requirements
  - Audit ability
  - Scalability
- Query
  - Performance
  - Flexibility
- Lots of material comparing methods for modelling

What is a Data Vault?

- According to Data Vault Inventor (Dan Linstedt)
  - "*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 3NF and Star Schemas.*"
  - A Modeling approach, DV 2.0 now include complete Architectural Blueprint
- When do you use a Data Vault?
  - Enterprise Data Warehouses – Inmon or Kimball Style
    - Bill Inmon: "*The Data Vault is the optimal choice for modeling the EDW in the DW 2.0 framework*"
  - Data Integration/Migration Projects
    - Merger/Acquisitions requiring data *alignment*
    - Data migration projects – upgrades/migrations
    - Master data management initiatives

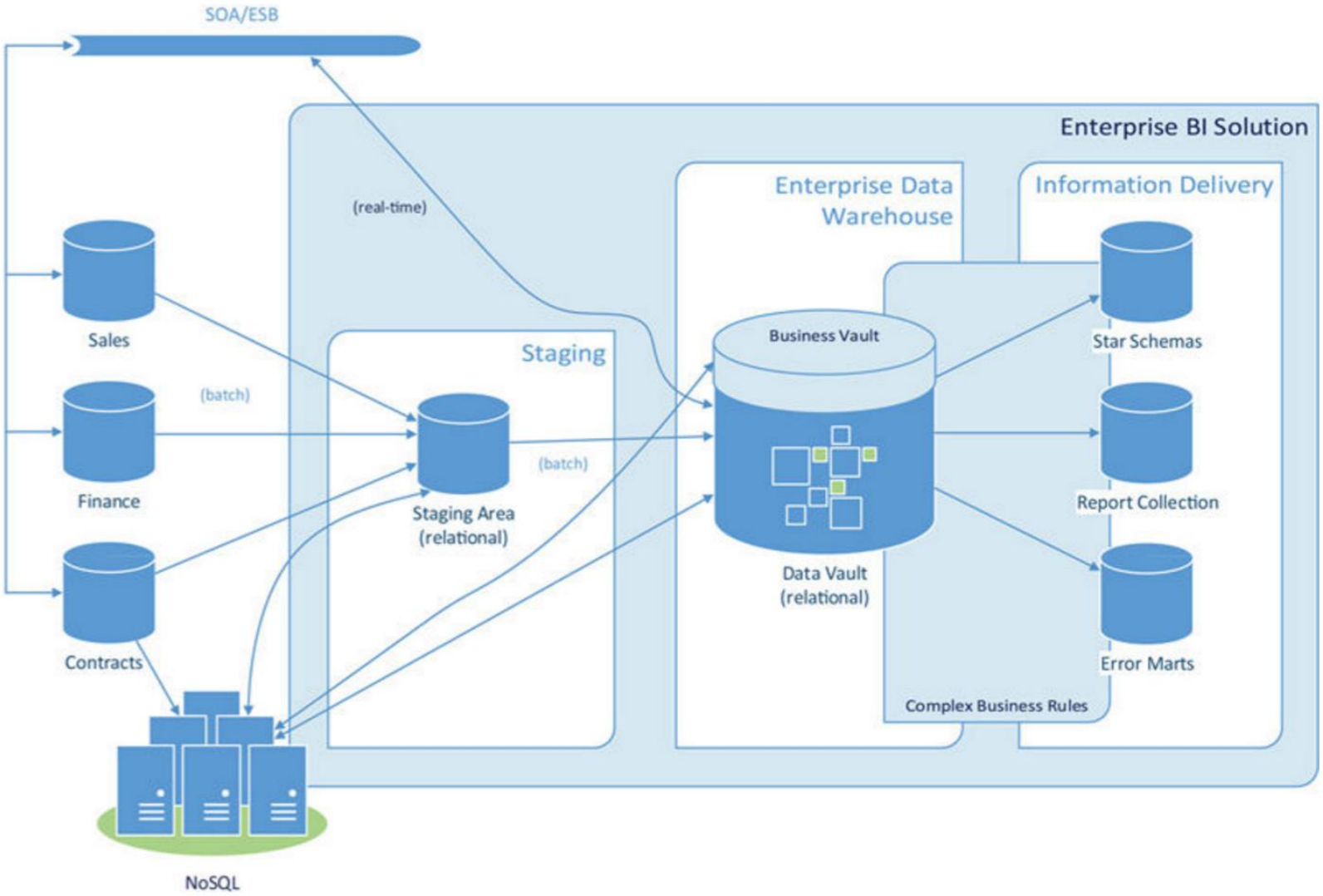
DBIS

Data Vault Explained

- Dan Linstedt developed in 1990
- Partnered with Hans Hultgren with RapidACE (software) and Genesee Academy (teaching)
- I obtained Certification in 2007 from Dan/Hans
- Data Vault 2.0 (2014 -new book)
- 3rd Global Conference this week in Vermont (wwdvc.com)

DBIS

Where does the Data Vault fit in?



Data Vault Key Concepts

- Everything is MANY-TO-MANY
- Time dependency on everything
- Uses Relational DBMS – can extend to NOSQL
- Late BINDING for data – the LINK
  - Closer alignment to schema-on-read

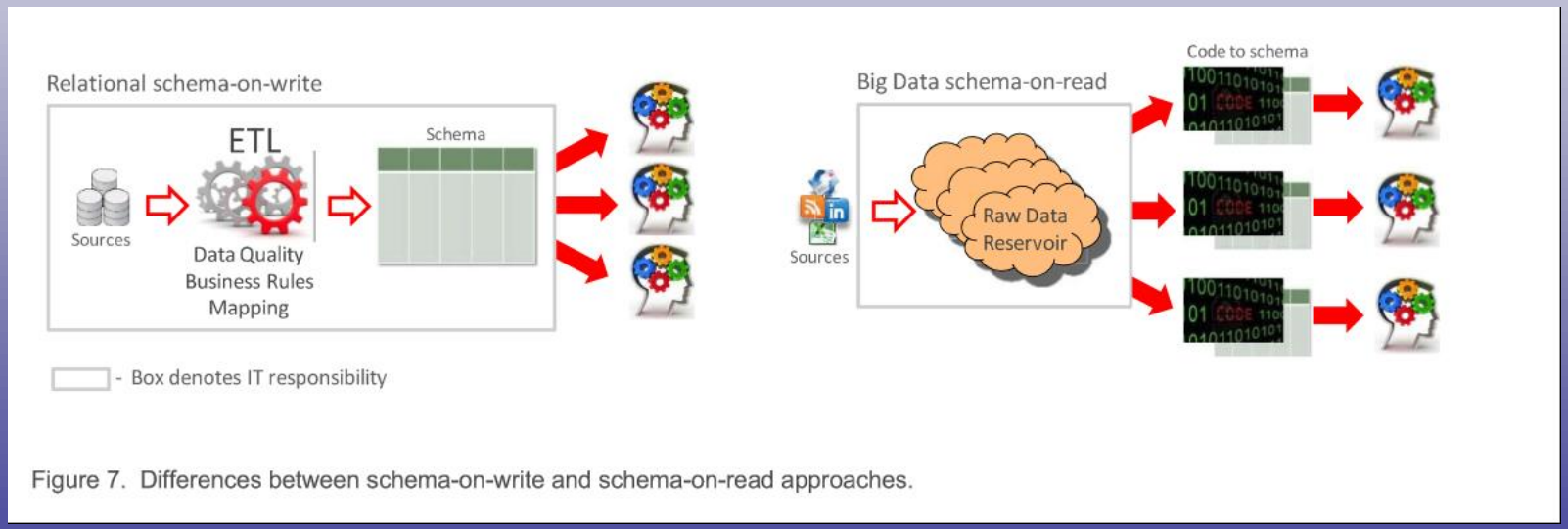
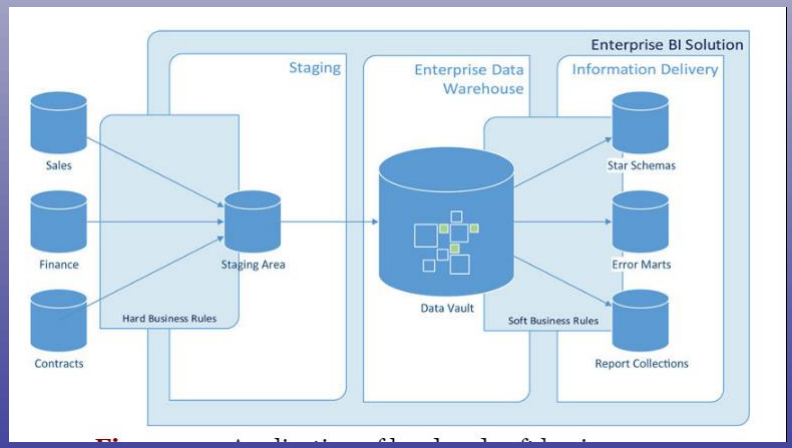


Figure 7. Differences between schema-on-write and schema-on-read approaches.

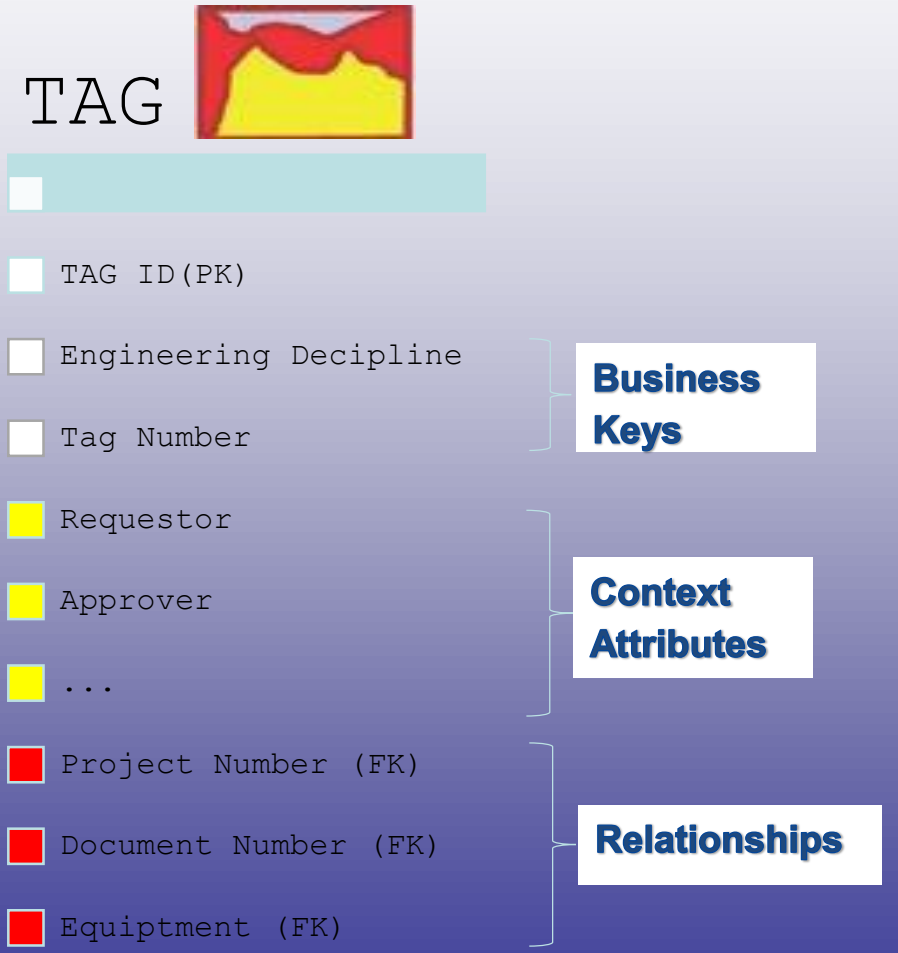
Data Vault Model Components

- "Table types":
  - *Hub* = List of Business Keys
  - *Satellite* = Descriptive Information
  - *Link* = Describes Relationship Between Business Keys
- Colors of the data vault (https://www.youtube.com/watch?v=kRoDRlj8_YU)
  - Hans Hultgren of Genesee Academy
- RAW and Enriched data
  - Business Vault



Unified Decomposition

In Consolidated Raw database, we load and decompose data into 3 areas.
For example TAG:

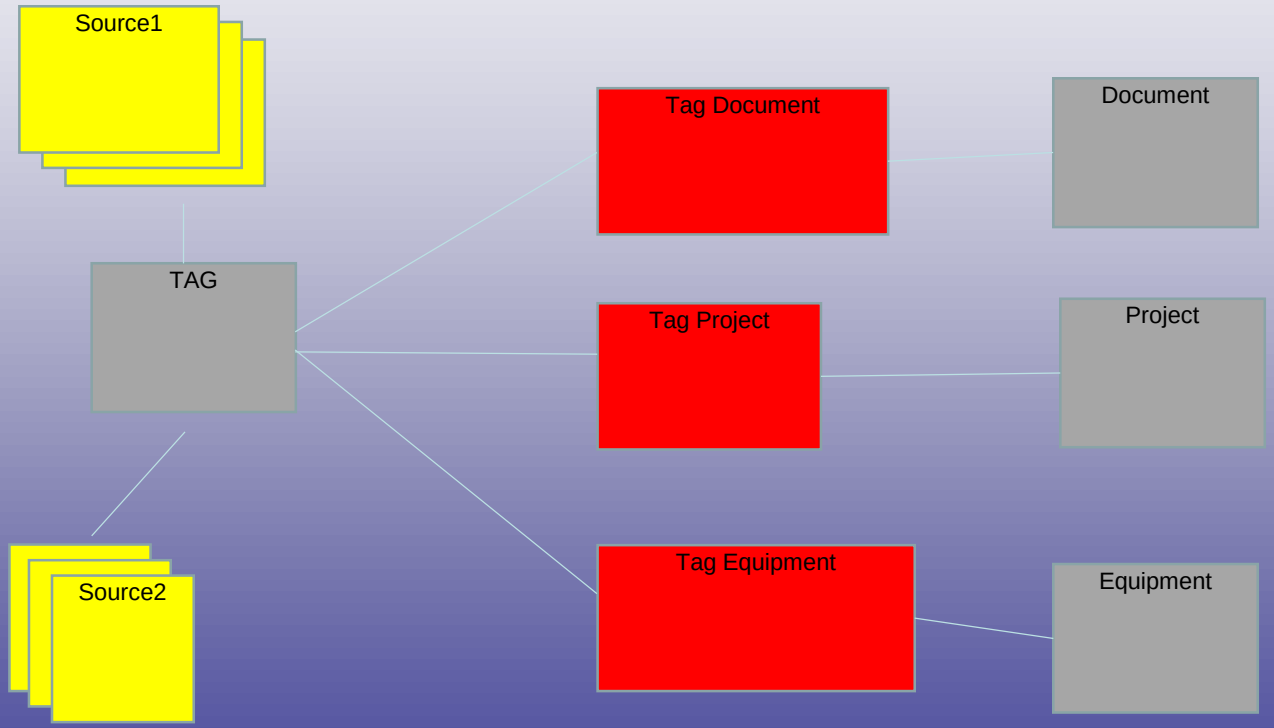


Hans Hultgren:
https://www.youtube.com/watch?v=kRoDRlj8_YU
Book: http://www.amazon.com/Modeling-Agile-Data-Warehouse-Vault/dp/061572308X

DBIS

TAG decomposition completed

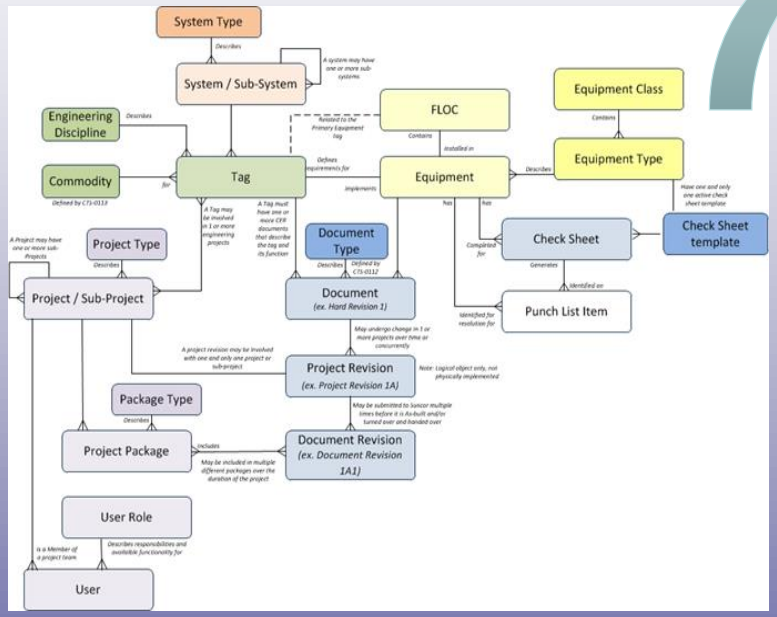
Data is stored in HUBs (Keys), Links (Relationships) and Satellites (Time dependent Context attributes)

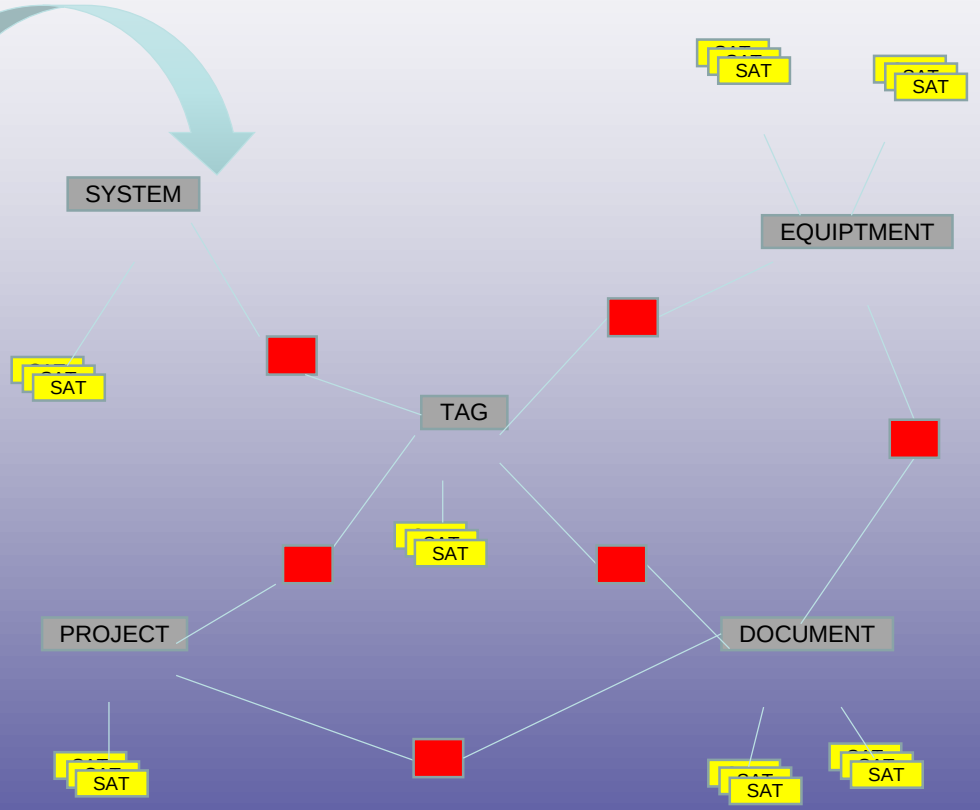


DBIS

Data Model – Incremental/Agile build

Additional Data from logical model added over time





DBIS

Challenges revisited

Challenges

- Getting it right – “the truth”
- Integration
- Compliance
- Time dependency
- Modeling

With Data Vault

- Facts as they were
- Business Key Alignment
- Model driven
- Incremental/agile build
- Scalable, adaptable

Questions?

bruce.mccartney@dbinfosystems.com

Google: Data Vault