Data Integration and Warehousing using the Data Vault

Alberta Data Architects
Calgary November 8\textsuperscript{th} 2017
Bruce McCartney
bruce.mccartney@dbinfosystems.com
• Introduction
  – about me
  – What is the Data Vault?
• The Business Problem – challenge in getting data right
• Enter the Data Vault
  – What is a Data Vault, really?
  – When do you use a Data Vault?
  – Introduction to Key Concepts
• The Business Solution
  – Challenges revisited
Evolution of Bruce’s career

- 1980: Programmer
- 1990: DBA
- 2000: DV Certification
- 2010: Data Architect
- Independent Consultant
- Architect DV2.0 Instructor
What is a Data Vault?

• According to Data Vault Inventor (Dan Linstedt)
  – DV 1.0 - A Modeling approach
    • “A detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business.”
  – DV 2.0 now include complete Architectural Blueprint and Methodology
    • “A System of Business Intelligence containing the necessary components needed to accomplish enterprise vision in Data Warehousing and Information Delivery.”
When do you use a Data Vault?

• Enterprise Data Warehouse (Inmon or Kimball Style)
  • Bill Inmon: “The Data Vault is the optimal choice for modeling the EDW in the DW 2.0 framework”
  • Kimball – Persistent Staging Area

• Data Integration/Migration Projects
  • Merger/Acquisitions requiring data alignment
  • Data migration projects – upgrades/migrations
  • Master data management initiatives
Challenges in Data Architecture

• Getting data right – “the truth”
• Integration of data (rather than applications)
• Compliance/Auditability
• Modeling Paradigm
• Agility
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*
- Two people can look at the exact same situation and come up with completely different ‘truths’
- “We have come to trust our screens” (Future Crimes by Marc Goodman)
- Closer Facts ~ Information
  - Less “technical debt” resulting
- More data → less rules
  - IOT, Internet, Automation, AI and deep learning
  - Rise of the Robots – Martin Ford
  - Monkeylearn Taxonomy Classifier

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.

[Image: Digital Universe growth chart]

If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon*

By 2020, there would be 6.6 stacks from the Earth to the Moon*


• 4 “V”s of Big Data
  • volume, variety, velocity, and veracity
Data Integration

• “Business Key” Alignment
  – Unique? global?
  – Use of “smart keys”
  – Multiple systems carry different and same parts of data objects
  – Cross platform integration

• Timing – dependencies
  – Where to put you business rules -conforming
  – Optimize Business Cycles
  – Global Enterprises 24x7 Integration in real time

• EAI vs. EII Architecture
  – Are we integrating process or data?
• **Prove** the information is the facts as they were with business rules applied
• Dataucracy – data lakes, data governance, data quality
Data preparation accounts for about 80% of the work of data scientists.

What data scientists spend the most time doing:

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Data Scientist’s Dilemma

What's the least enjoyable part of data science?

- Cleaning and organizing data: 57%
- Building training sets: 10%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Lack of Agility

• IT record in data warehousing and integration not good

• Why?

1. ETL eats your lunch (up to 70%)
   - Business rule applied on the way in to DWH
   - Think Data Warehouse Automation

2. Development Paradigm
   - Waterfall, Big Bang, Do Over
   - Lack of Agile BI adoption (Scott Ambler - Discipled Agile Development)
     • Agile BI Manifesto (Agile Analytics Ken Collier)

http://www.datamartist.com/making-rapid-prototypes-for-data-warehouse-etl-jobs
• **Needed to Adapt** some of these for data warehousing:
  – 3NF
    • Rework and inflexible
  – Star Schema Structure
    • Type-2 Dimension Conformance complexity
    • Aggregation and help tables
    • Snowflakes
  – Anchor Model
    • More tables instead of attributes grouped

• Lots of material comparing methods for modelling
• Boils down to “Model last” vs. “Model first”
Where does the Data Vault fit in?
Data Vault 2.0 Key Concepts

• KEY CONCEPTS
  – Data
    • Data separation into business key, relationships and context (ensemble)
    • Immutable raw data (non volatile) – Inmon definition of DWH
    • Decoupled from information created through execution of business rules
    • Focused on Integration by BK as opposed to “process” orientation
  – Architecture
    • Layered Logical Architecture – implementation agnostic (relational/Big Data)
  – Methodology
    • Agile Build (incremental, automatable approach)
Data Vault Model 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.

Oracle – Information management and big data reference architecture
Unified Decomposition

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

- **Business Keys**
- **Associations or Relationships**
- **Details/Context/Attributes**

Example TAG:

- Engineering Discipline
- Tag Number
- Requestor
- Approver
- Project Number (FK)
- Document Number (FK)
- Equipment (FK)

Hans Hultgren:
https://www.youtube.com/watch?v=kRoDRlj8_YU
Data Vault Model Components

- “Table types”:
  - **Hub** = List of business keys
  - **Satellite** = Time dependent descriptive Information
  - **Link** = Describes relationship between business keys

- Raw and Enriched data
  - Raw Data Vault
  - Business Vault
Case Study Business Problem

- Manage “Asset Information Lifecycle”

www.linkedin.com/pulse/new-research-90-oil-gas-projects-delayed-due-handover-pateman-jones
Motivation: The Value Chain

Value Levers
- Improved Uptime (Asset Availability)
- Operations Cost Reduction (OPEX)
- Capital Cost Reduction (CAPEX)
- EH&S

Measurements
- Higher Availability
- Higher Utilization
- Higher Throughput
- Maintenance Labor
- Maintenance Materials
- Indirect Labor
- Major Projects
- Sustaining Projects
- Project Turnover
- Higher Regulatory Compliance
- Lower of $MM Fines
- Asset Information related ILP

Values
- Details Omitted

CAN$86M/yr
Enrichment Database (eDB)
Data is stored in HUBs (Keys), Links (Relationships) and Satellites (Time dependent Context attributes)
Data Model – Incremental/Agile build

Additional Data from logical model added over time
Goals

- Program Level
  - Data consolidation, integration, and enrichment.
  - 95% automation & accuracy

- Technical Level
  - Automate migration build
  - Automate verification and enrichment using rules, and humans as necessary.
  - Produce enriched data set that is audible and traceable
  - Have approach repeatable for the next 13 facilities.
Automate Raw Data Vault

- Generate DDL and load procedures using metadata
- Excel → SQL Server and TSQL
Business Vault Use Case

- Driven “change of data” in raw data vault
- No requirement to grain shift data
  - (i.e. normalize, summarize or allocate)
- “Getting Data Right”
  - Schema last vs. schema first
  - Do everything you can automatically – with machine learning and statistics, ask for help only when necessary
- Time required to enrich not a factor
Business rules engine decision

- Looked at SSIS DQS, JBOSS, DROOLS, AnalytiX DS, Talend, WhereScape and Tamr
- Decided to **build**, due to funding restrictions, skill-set, political situation, and agility
Tenets of our Business Vault

- Business vault regeneration possible “at any time”
  - using business rules valid at that time

Back to the future… and forward to the past
Tenets of our Business Vault

• If there is nothing to do, … do nothing
  – Rules run once unless data or rule changed since last run

• Support for “hard” and “soft” business rules
  – Hard rule example: Business Key override resulted in SAME-AS Links

• Provide lineage/auditability for each business rule application
  – Data Vault gives you this for free. Versions of satellite data tagged with rule id

• Failures in validation sent to "work queue" for human analysis.
  – Data fix and reprocess; or new transform rule / override
Challenges revisited

With Data Vault 2.0:

- Getting it right — “the truth”
- Data Integration,
- Compliance/Auditability,
- Resilient, Scalable,
- Agility

are all “baked in”
Getting it right – “the truth”

- with Data Vault
  - *All we want are the facts*, ma’am
  - Decouple facts (hard rule) from information (soft rules)

- Information becomes built for purpose

\[
\frac{1}{2} \text{ truth} + \frac{1}{2} \text{ truth} \neq \text{ Truth}
\]
Data Integration

• with Data Vault
  – Alignment by Business Key
  – Master data management application
  – Relational and Hadoop
  – Incremental
Compliance/Auditability

• with Data Vault
  – Built-in features
    • Consistent repeatable, generatable logic
    • Time dependent versions of data
    • Immutable facts
  – Capabilities possible
    • Rapid rework of information
    • Regenerate source data at a point in time (best practice for testing)
    • Regenerate information from a point in time with current rules
    • Regenerate information from a point in time with rules from a point in time using a time dependent rules engine
Modeling

• with Data Vault
  – Decouple data from information
  – Ensemble hub/spoke
  – Integration by BK
  – Information Marts
    • Virtualization
    • Flexibility (Star Schema, Graph, Wide tables)
  – Resilient Scalability
Agility

• with Data Vault
  – Incremental build
    • Maximize work not done
  – Generation/automation
    • Metadata/model driven
    • Data Warehouse Automation tools
  – Disciplined Agile Development (http://www.disciplinedagiledelivery.com)
Data Vault is “more better” way to do data warehousing and integration
ME:  
bruce.mccartney@dbinfosystems.com  
https://www.linkedin.com/in/mccartneybruce/

Google: “Data Vault”

Dan’s Book:  
https://www.amazon.ca/Building-Scalable-Data-Warehouse-Vault/dp/0128025107

User’s Group:  
http://www.DataVaultUsersGroup.org

Calgary Meetup Nov 21st  
https://www.meetup.com/preview/Data-Vault-Meetup-Calgary

CIPS Calgary BI SIG:  
http://www.cips.ca/node/3417