

AGILE DATA WAREHOUSING / STATE OF THE ART FOR 2016

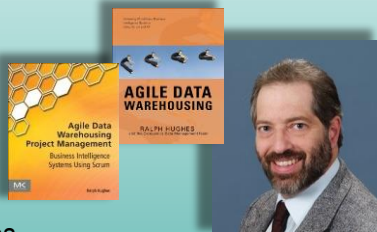
Adaptive Data Modeling Techniques

MAXIMIZING THE WORK NOT DONE

Ralph Hughes, MA, PMP, CSM
ralph.hughes@ceregenics.com

Ceregenics proprietary information

Presenter's Background



Ralph Hughes

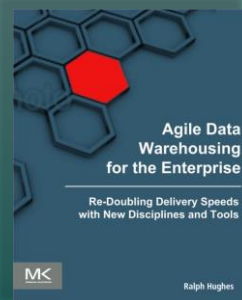
- ▶ 30 years solutions architecture, ETL & BI development
- ▶ MA, PMP, CSM
- ▶ Author of three agile methods books
- ▶ Member of DW/BI advisory boards and best practices panel
- ▶ Frequent keynote speaker & instructor DWBI conferences

Website: www.Ceregenics.com

Email: ralph.hughes@ceregenics.com

www.linkedin.com/in/ralphhughesadw

Twitter: [@ceregenics](https://twitter.com/ceregenics)



2016

Ceregenics proprietary information

Today's Topics



Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

- ▶ “Surface Solutions”
- ▶ End-User Hadoop
- ▶ Document data stores
- ▶ Hyper modeling
 - Hyper normalization
 - Hyper generalization
- ▶ Agile value cycle



Agile EDW Works Fabulously



- ▶ Major Healthcare Clinic (2014)
- ▶ Ceregenics joins an agile team that wasn't getting traction during Iteration 7
- ▶ Best practices accelerates project by 3x to 10x, depending upon units of measure



First 6 iterations with a Scrum master only, no AEDW best practices

(Formerly) Outrageous Statements

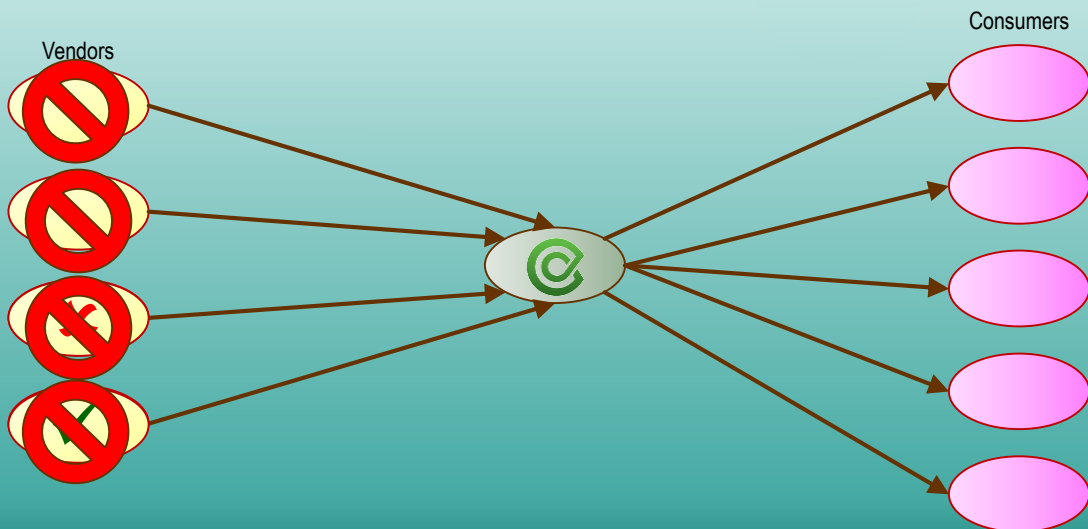


The problem of enterprise data warehousing has been solved:



- ▶ If you're not using 80/20 specifications, you're taking 5x longer to get started
- ▶ **If you're coding by hand, you're wasting 90% of your programming**
- ▶ **If you're thinking only RDBMS, you're building at least 3X more than you need**
- ▶ Without automated testing, you're missing over half of the defects

Presenters Must Remain Tool-Agnostic



Today's Topics

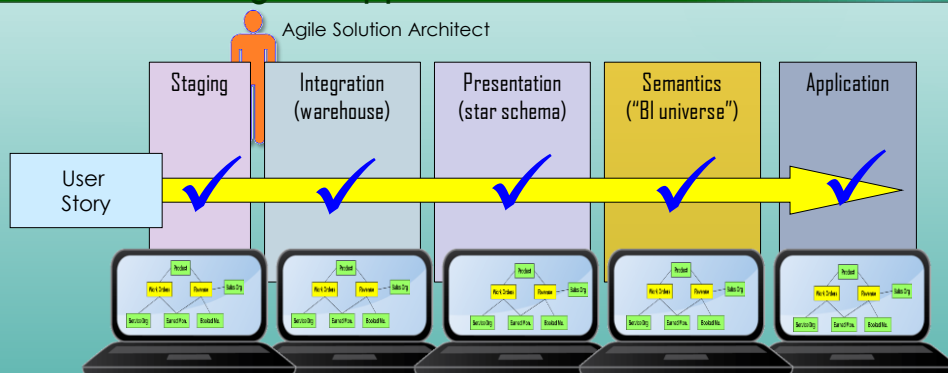


Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

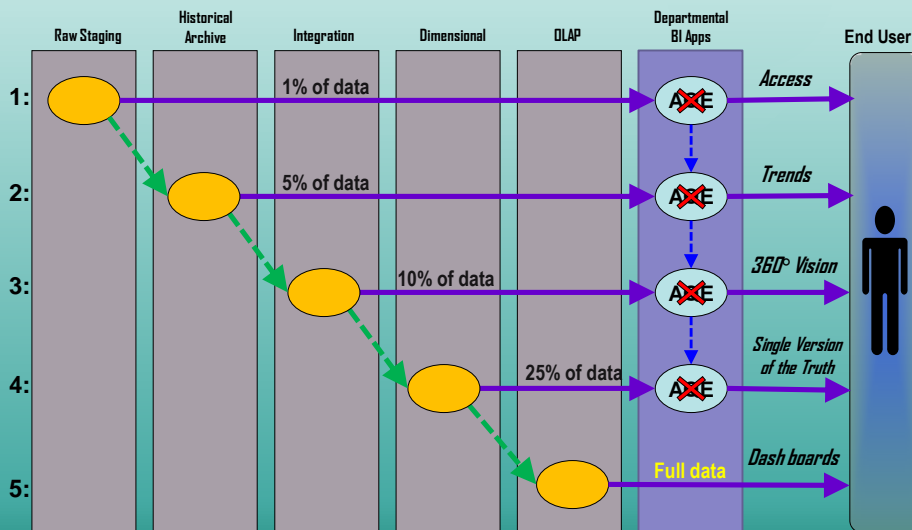
- ▶ “Surface Solutions” ←
- ▶ End-User Hadoop
- ▶ Document data stores
- ▶ Hyper modeling
 - Hyper normalization
 - Hyper generalization
- ▶ Agile value cycle

EDW is Like Building 5+ Applications at Once



- Each architectural layer has different purpose and constraints
- Why approach them all with the same techniques and tools?
- Provisional value available long before application layer, so why wait?

Surface Solutions Possible Even with RDBMS



Ceregenics proprietary information

10

Today's Topics



Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

- ▶ “Surface Solutions”
- ▶ End-User Hadoop
- ▶ Document data stores
- ▶ Hyper modeling
 - Hyper normalization
 - Hyper generalization
- ▶ Agile value cycle

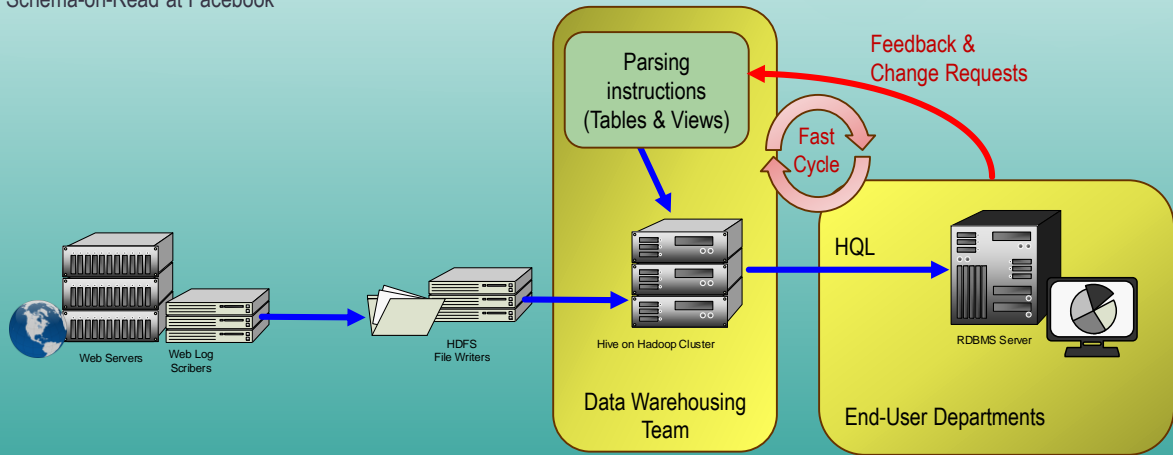
Ceregenics proprietary information

11

Option 3: Agile Big Data

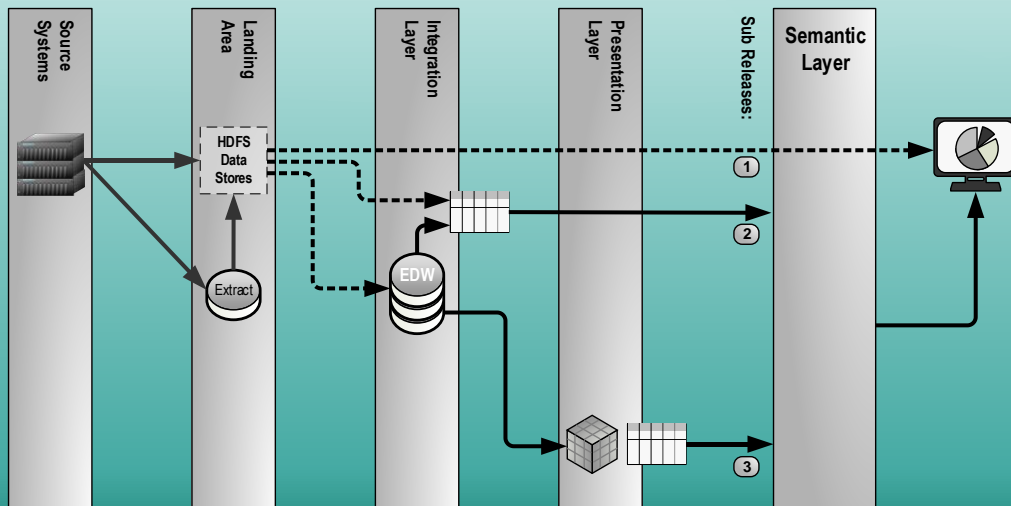


Schema-on-Read at Facebook



Ceregenics proprietary information

Evolving Surface Solutions Using Hadoop



Ceregenics proprietary information

Today's Topics

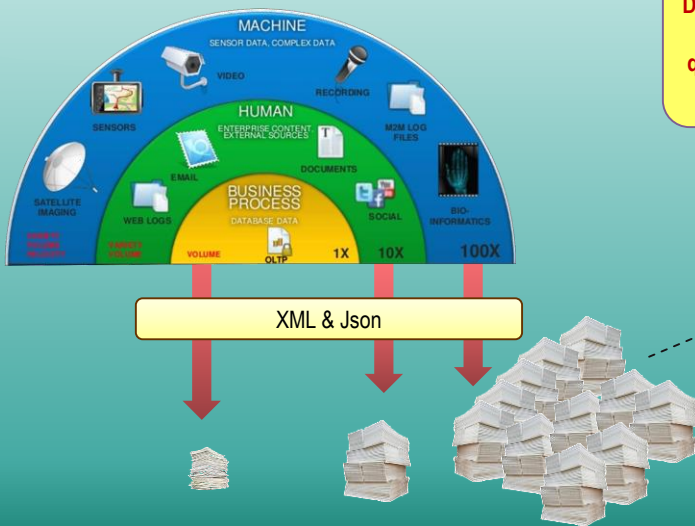
Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

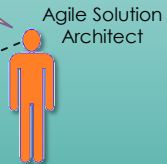
- ▶ “Surface Solutions”
- ▶ End-User Hadoop
- ▶ Document data stores ←
- ▶ Hyper modeling
 - Hyper normalization
 - Hyper generalization
- ▶ Agile value cycle

Option 4: Document Database

Increasingly More Big Data are in XML and Json Documents

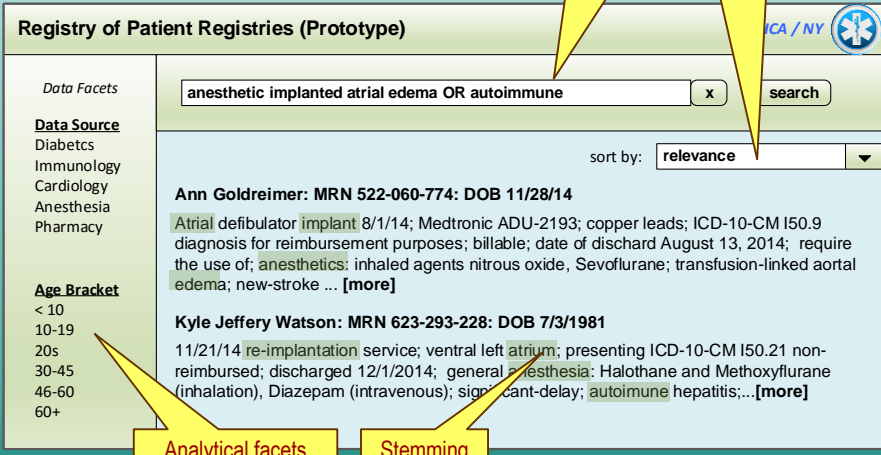


Document data stores will let us explore and build quick solutions with very little programming



Powerful Search with Little Programming

Out-of-the-Box Delivery Tool – Requires only changing the query text and the indexing of the documents



Registry of Patient Registries (Prototype)

anesthetic implanted atrial edema OR autoimmune x search

sort by: relevance

Ann Goldreimer: MRN 522-060-774: DOB 11/28/14
 Atrial defibrillator implant 8/1/14; Medtronic ADU-2193; copper leads; ICD-10-CM I50.9 diagnosis for reimbursement purposes; billable; date of discharge August 13, 2014; require the use of; anesthetics: inhaled agents nitrous oxide, Sevoflurane; transfusion-linked aortal edema; new-stroke ... [more]

Kyle Jeffery Watson: MRN 623-293-228: DOB 7/3/1981
 11/21/14 re-implantation service; ventral left atrium; presenting ICD-10-CM I50.21 non-reimbursed; discharged 12/1/2014; general anesthesia: Halothane and Methoxyflurane (inhalation), Diazepam (intravenous); significant-delay; autoimmune hepatitis;...[more]

Skills needed:

- Some HTML
- A little CSS
- Some XML or Json
- Some Xquery / XPath or Java script

Ceregenics proprietary information

17

Today's Topics

Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

- ▶ “Surface Solutions”
- ▶ End-User Hadoop
- ▶ Document data stores
- ▶ Hyper modeling
 - Hyper normalization ←
 - Hyper generalization
- ▶ Agile value cycle

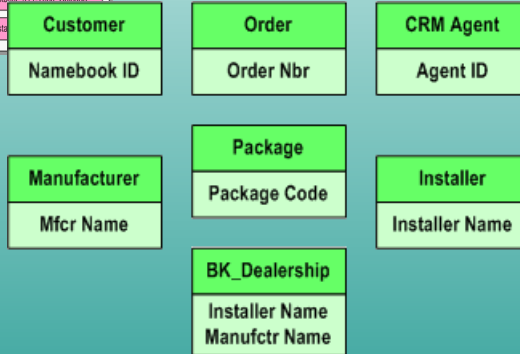
Ceregenics proprietary information

19



Step 1: Identify Business Keys

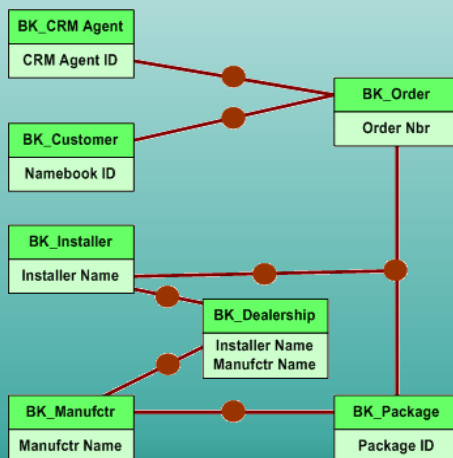
Online Sales Order # 5432			
Order DTM: 5-Jan 11:23	Namebook ID: paula_g	Customer: Westwood Rec Center	City: Westwood, CA
CRM Agent: asmith	CRM Name: Adam Smith	eBiz Site: www.onlinedepot.com	eSegment: Small Office
Item 1: Pkg ID: VPH	Package 1: VOIP Phone	Qty: 3	Mfg: Winsome
Item 2: Pkg ID: STV	Package 1: Satellite TV	Qty: 2	Mfg: Longlife Products



Ceregenics proprietary information

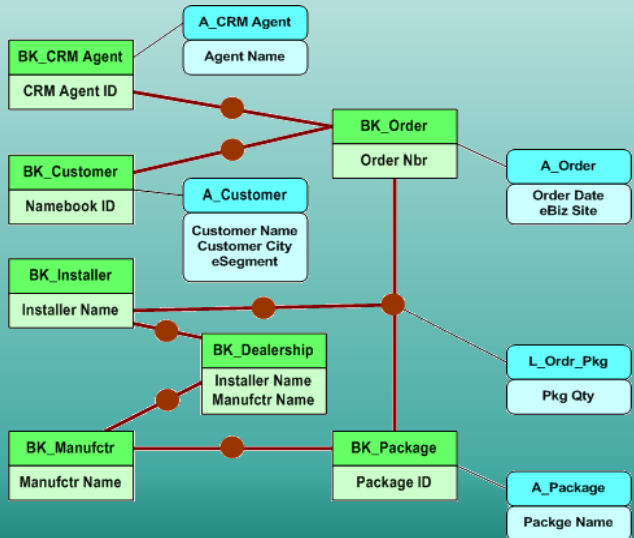


Step 2: Create M-M Links



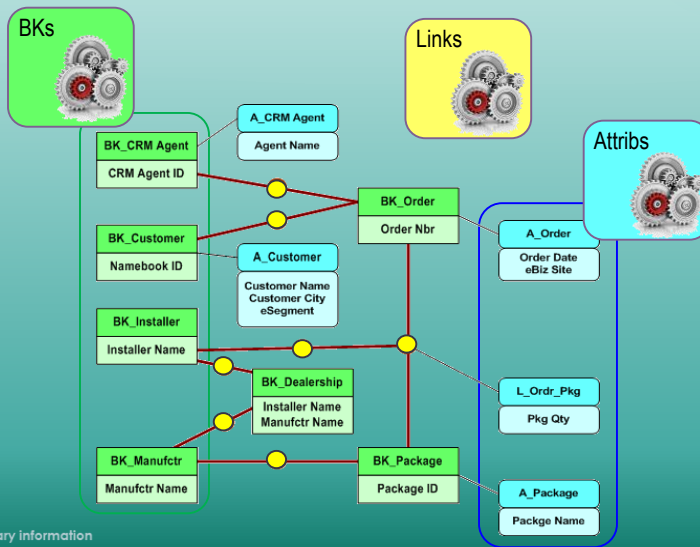
Ceregenics proprietary information

Step 3: Add Attributes



Ceregenics proprietary information

HNF Makes Re-Usable ETL Straightforward



Ceregenics proprietary information

Parameter-Driven ETL



Load_BK (target, source, natural key column list)



Load_Link (target, source 1, src 1 natural key cols, source 2, src 2 natural key cols)



Load_Attribs (target, source, exclude column list)



“Cookie-cutter ETL”

Ceregenics proprietary information

Today's Topics



Agile = quick & continuous delivery of value to the customer

Agile EDW achieves this goal through:

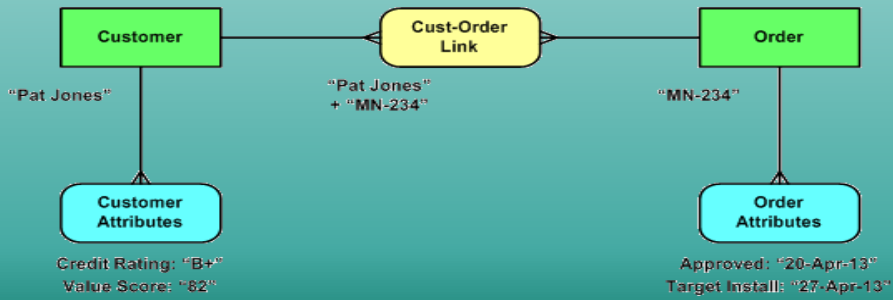
- ▶ “Surface Solutions”
- ▶ End-User Hadoop
- ▶ Document data stores
- ▶ Hyper modeling
 - Hyper normalization
 - Hyper generalization ←
- ▶ Agile value cycle

Ceregenics proprietary information

25

From HNF to HGF

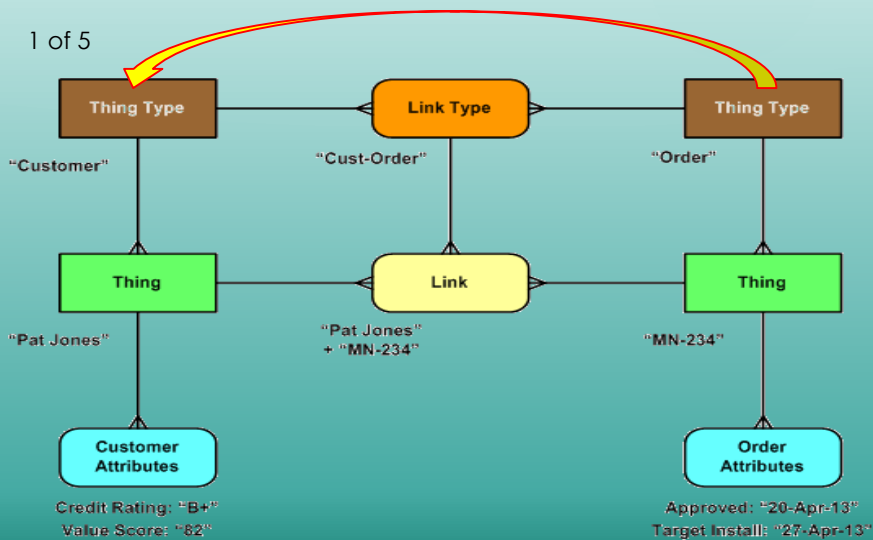
0 of 5



Ceregenics proprietary information

Convert to Metadata to Distinguish Instances

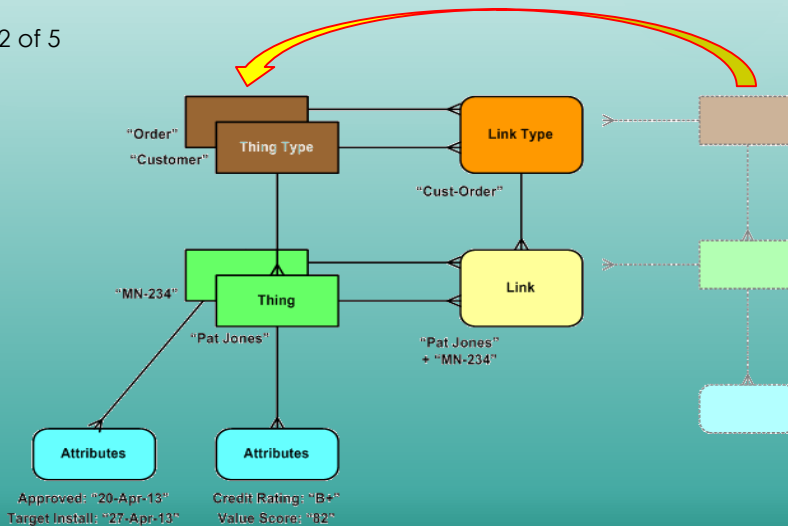
1 of 5



Ceregenics proprietary information

“Fold” the Model to Eliminate Separate Tables

2 of 5

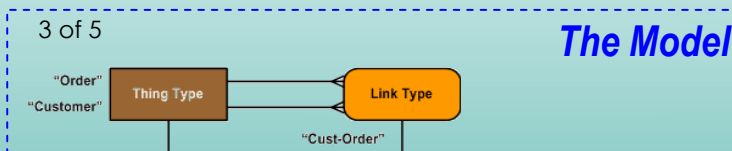


Ceregenics proprietary information

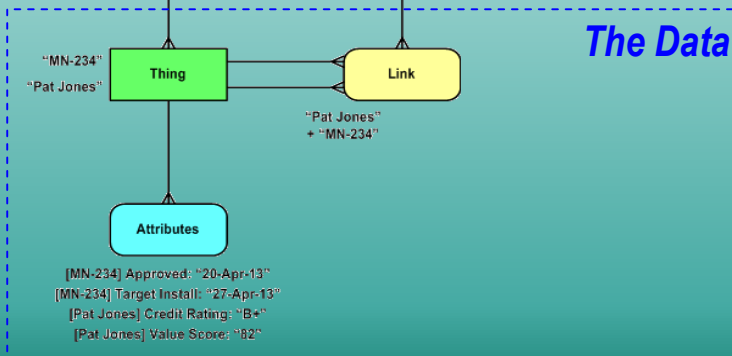
Combine Tables with Equivalent Function

3 of 5

The Model



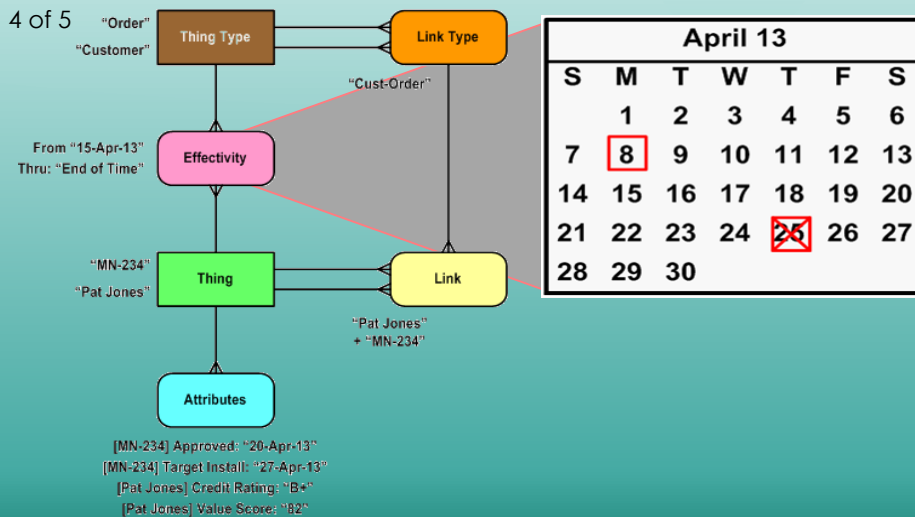
The Data



Ceregenics proprietary information



Allow Re-Classifications of Instances

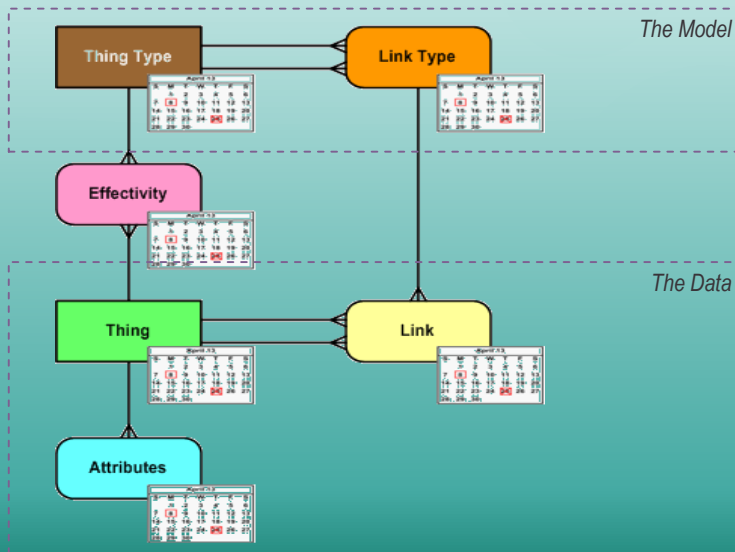


Ceregenics proprietary information



Completely Temporal Data Warehouse

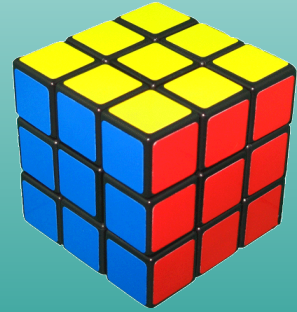
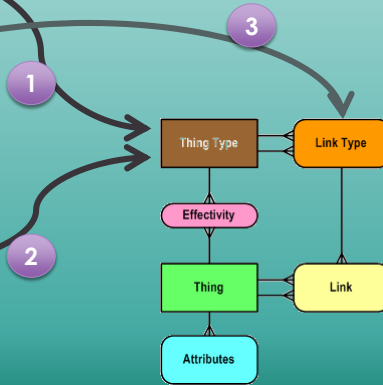
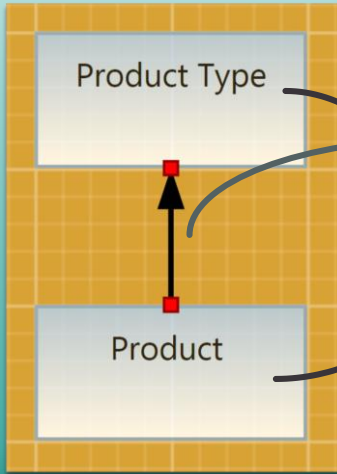
5 of 5



Ceregenics proprietary information

Model Objects Map to Meta Data Entities

1. "Product Type" class exists
2. "Product" class exists
3. Product rolls up to Product Category

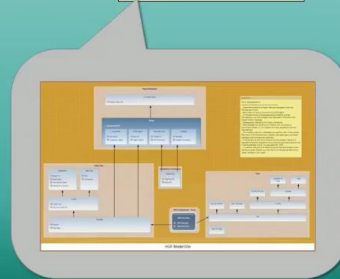


Ceregenics proprietary information

Computer-Assisted ETL Programming

Parts Data Mgt Extract 160213	
Product Nbr	→
Product Name	→
Packaging Code	→
UPC	→
Product Type Nbr	→
Prod Type Name	→
Extraction Date	→

Add/Mod Instance PRODUCT	
<input checked="" type="checkbox"/> Product Nbr	←
Product Name	←
Packaging Code	←
UPC	←
<input type="checkbox"/> PRODUCT TYPE	←
Trans Date	←



Ceregenics proprietary information

Automation Surrounds Us

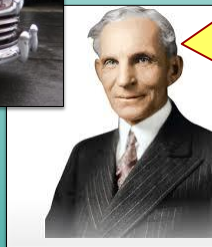
- ▶ Computers build our goods...
- ▶ ...land our planes...
- ▶ ...will soon drive our cars.



Why are we still building data warehouses by hand?

Henry Ford Considers a Tesla

1947 Ford Coupe



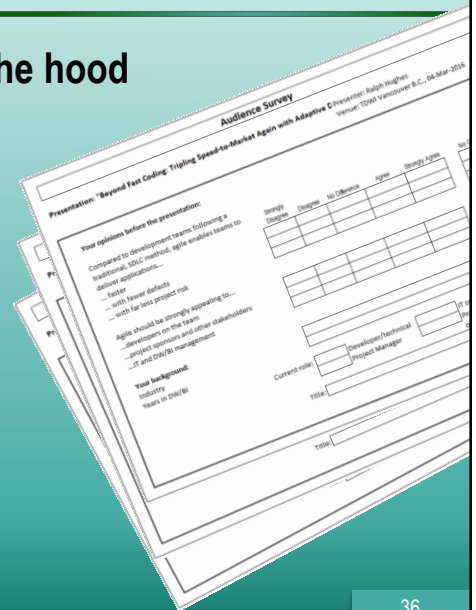
Where's the carburetor?
 ... the transmission?
 ...the radiator?
 I can see all kinds of problems with this car. What a hoax!



Why are we still building data warehouses by hand?



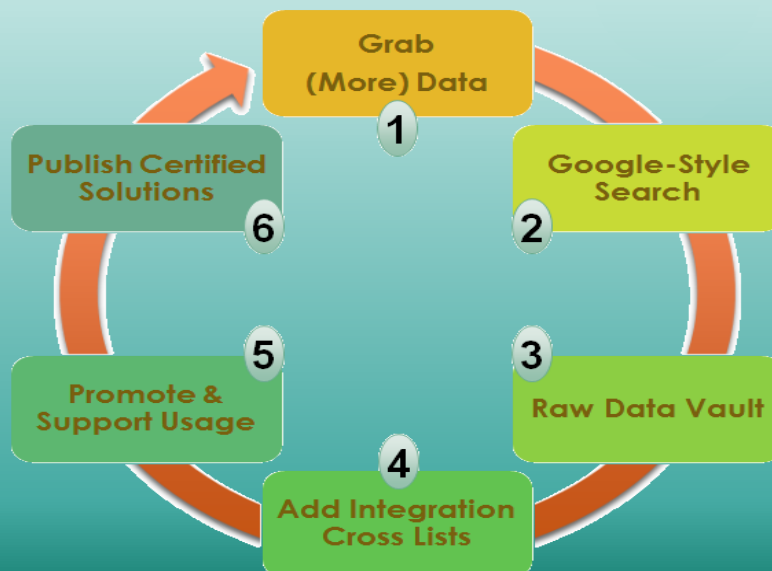
- ▶ Don't understand what's going on under the hood
- ▶ Don't want to give up data modeling
- ▶ Don't want to be the first one to do it
- ▶ Too much invested in 1990s technology
- ▶ Don't want to eliminate people's jobs
- ▶ Don't want to eliminate my department
- ▶ Staff can't handle learning another tool



Ceregenics proprietary information

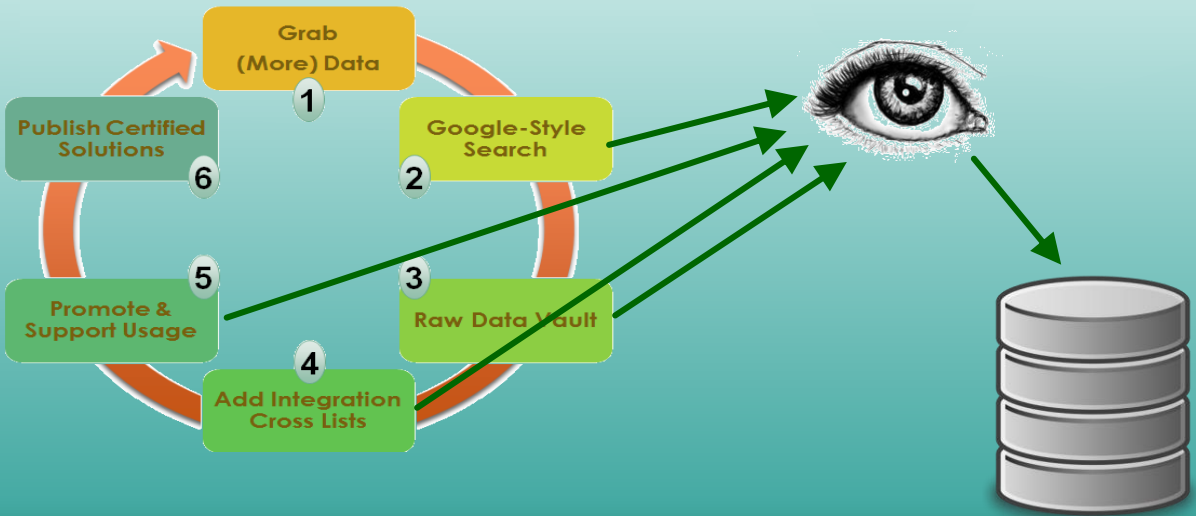
36

Tools for the Business Opportunity Cycle



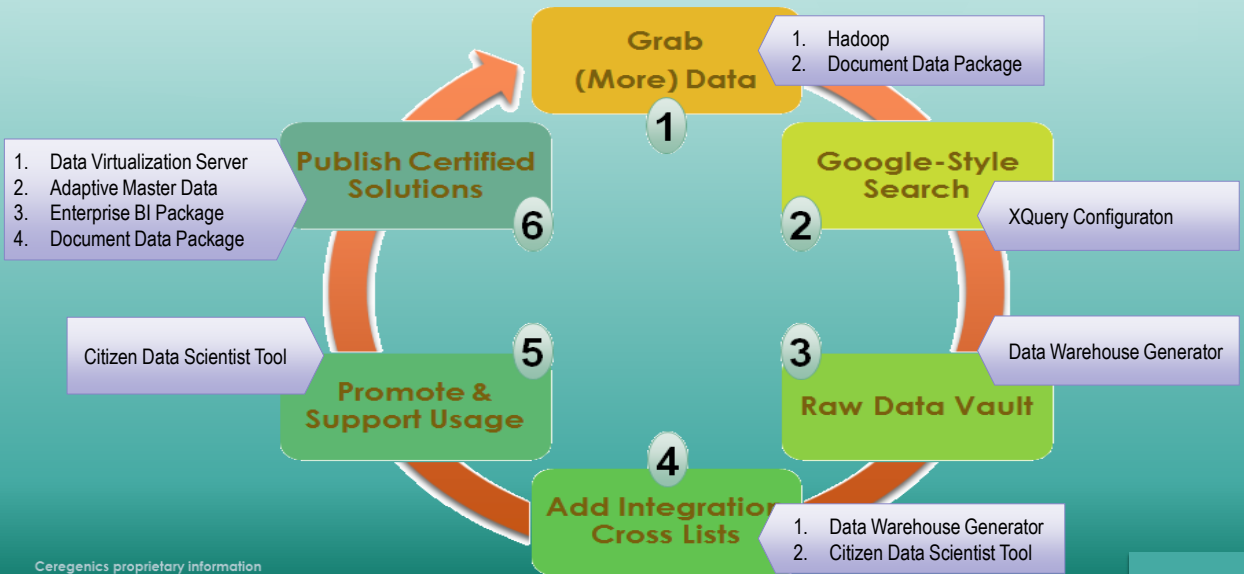
Ceregenics proprietary information

Automated Monitoring for Faster Requirements

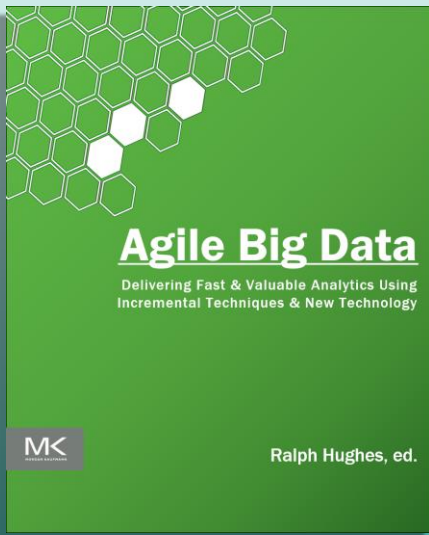


Ceregenics proprietary information

Tools for the Business Opportunity Cycle



Ceregenics proprietary information



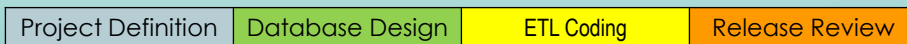
Find Your Voice & Help Others to Find Theirs

– Stephen Covey’s “Eighth Habit”

- ▶ Call for contributors
- ▶ Write a chapter, sidebar, or a section
- ▶ Focus: theory & case histories that blend
 - Disciplined agile & EDW methods
 - Hadoop, M/R, Spark
 - Textual and triple data stores
 - Empowering citizen data scientist

Long Design will Still Delay Value

Traditional Methods



Agile Approach + Productivity Tools



- **“Surface Solutions”**
- **End-User Hadoop**
- **Document data stores**
- **Hyper normalization**
- **Hyper generalization**
- **Agile value cycle**



999 18th Street, Suite 3000
Denver CO 80033
303.274.9101
www.Ceregenics.com



Hyper normalization: [//www.youtube.com/watch?v=3QOS0eN8vcY](http://www.youtube.com/watch?v=3QOS0eN8vcY)
Hyper generalization: [//www.youtube.com/watch?v=aNtUoVkeq_Q](http://www.youtube.com/watch?v=aNtUoVkeq_Q)

Ceregenics proprietary information