

## > Data Vault Case Study

>Raphael Klebanov, *Customer Experience at WhereScape USA*

>**Data Vaults** have been gaining huge attention in recent years all over the planet. The Data Vault, invented by Dan Linstedt, is a detailed, historically oriented, uniquely linked set of normalized tables that support one or more functional areas of business. This “hybrid” approach encompasses the best of breed between the third normal form and the Dimensional models - the design is flexible, scalable, consistent, and adaptable to changing business needs. This makes Data Vaults an optimal way to build enterprise data warehouses.

>**WhereScape** an industry leading information systems software company, has two products on the international market:

>**WhereScape 3D** (Data Driven Design) is a data warehouse planning tool. 3D aids in exploring and understanding a data warehousing project at the time you need it most - the beginning.

>**WhereScape RED** is a data warehouse productivity (automation) software tool that promotes building data warehouses faster and more accurately. The resultant warehouse is easier to support, change and extend. Out of the box, WhereScape RED builds target systems on Teradata, Microsoft, Oracle, DB2, Netezza, Greenplum and other Big Data platforms.

## >WhereScape and Data Vault

At WhereScape, we started to build Data Vaults (DV) in 2009 (IPC/Subway, USA). Currently, WhereScape has over 45 Data Vault projects underway worldwide - from mature to expansion to development. Those DV projects span four continents - Europe, Asia, North America, and Australia and three database platforms - SQL Server, Oracle and Teradata.

A Data Vault consists of Hubs (business keys), Links (relationships) and Satellites (descriptions) plus some auxiliary objects. Data Vaults can be implemented using standard WhereScape RED objects.

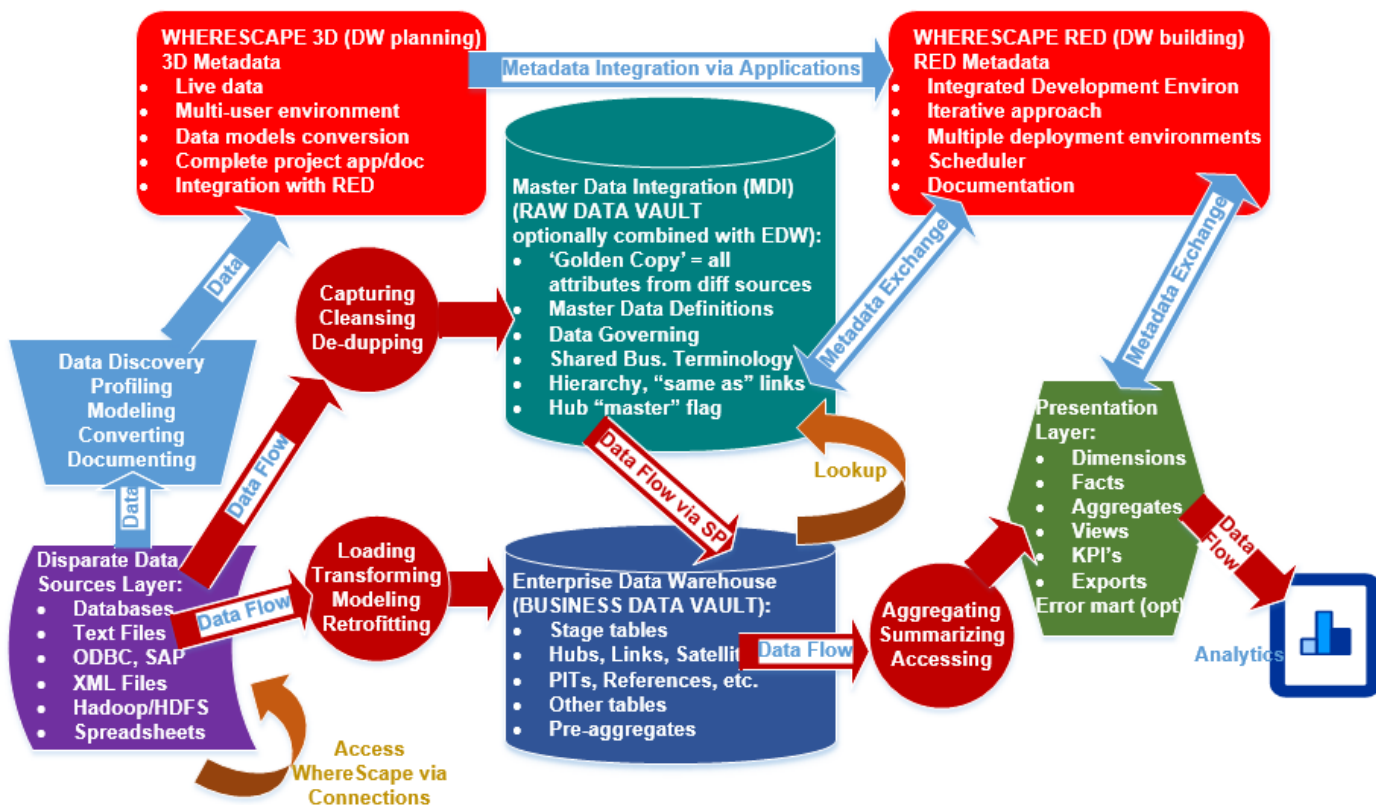
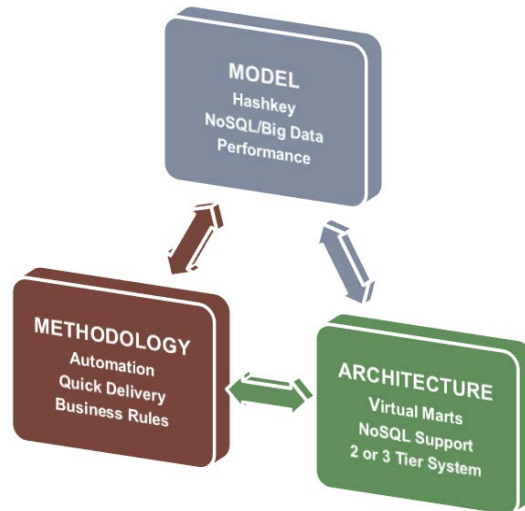


Figure 1 Data Vault Model in WhereScape Environment

## >WhereScape and Data Vault 2.0

As described by Dan Linstedt: "Data Vault 2.0 is the evolution of the standards for Data Vault..." (<http://danlinstedt.com/datavaultcat/a-short-intro-to-datavault-2-0/>)

"DV 2.0 is a system of data warehousing and business intelligence that is comprised of three major components..." ([http://vuymebsl.metroblog.com/the\\_business\\_of\\_data\\_vault\\_modeling\\_ebook](http://vuymebsl.metroblog.com/the_business_of_data_vault_modeling_ebook))



### Model:

#### RED

- Use of hash keys in stage tables for Hubs and Links
- Indexing strategy, partitioning, compression
- Business key replication in Links and Satellites (opt.)
- Parallel loading
- Models retrofitting and diagramming
- Data lineage capturing

#### 3D

- Creation of conceptual, logical and physical models
- All modelling patterns: DV, 3NF, Dimensional, and Hybrids Models
- Hash keys in transformations and conversion rules
- Auto-converting "to" Data Vault and "from" Data Vault
- Modelling standards and best practices
- Supports modelling in Hadoop/Hive environments

### Architecture:

#### RED

- Full compliance with the Data Vault Architecture
- Supports all data warehouse flavors, including raw DV and business DV
- Metadata, Drag-n-Drop, Wizard-driven operations
- Access layer in Views (optional)
- Beta: stage area in Hive Managed tables
- Beta: Unstructured data in Hive "External" tables

#### 3D

- Enterprise Data Warehouse Conversions: 3NF→DV; DV→Stage/Load tables
- Mart Conversions: DV→Star Schema; DV→Views
- Auto-generated Hubs, Satellites, Links, Point in Time, other ancillaries
- CLIENT↔DB or CLIENT↔APP↔DB architectures
- Auto-generated DB/non-DB objects based on business rules
- Connections to Hadoop/HDFS data

### Methodology:

## RED

- Auto-generated data warehouse objects, indexes and procedural code
- Auto-generating documentation, diagrams, reports and maps
- Flexibility: Quickly responding to business changes
- Stability and Performance
- Consistent delivery of working software
- Project-driven approach implements standards, rules, best practices

## 3D

- Fully configurable discovery, profiling and model conversions
- Supporting multi-user environments and self-organizing teams
- Shared responsibility and customer collaboration
- Data warehouse projects broken down by subject area or tasks
- Wizards for different types of objects depending on data warehouse model
- An iterative approach to 3D development

## >DV2.0 Principles for Information Systems

### 1. Agile Methodology:

- Estimation templates, business rule application
- Testing best practices on every stage of the software development life cycle
- Reputability, adaptability to business changes
- Provides complete tracking back to system of origin
- Raw, Business, Virtual Data vaults; Metadata



Figure 2 Agile (Iterative) Approach in WhereScape Environment

### 2. Automation Solution:

- Automation / Generation tooling (such as WhereScape)
- Dynamic structure adaptation of the Data Warehouse
- Performance tuning, partitioning standards, indexing strategy, referential Integrity
- Pattern based, standardization

## RED:

- Auto-generated DB objects and procedural code for all info systems models, major platforms, heterogeneous data sources with history support
- Easy adding of new DB objects, schema changes, and procedure code to an existing data implementation
- Building and deployment of the subject areas (GUI or command line)
- Drag-n-drop operations through wizard-driven development
- Customizable code generator behavior naming conventions and storage management
- Auto-generated reports, documentation and diagrams

## 3D:

- Auto-profile, explore, and capture source systems; auto-generate project documentation and diagrams
- Massively configurable discovery, profiling, and modeling conversion methods
- Design, model, profile, and test any schema using live source data, complete source-to-target mapping
- Wizards to derive foreign keys, primary keys, and data lineage
- View, manipulate, and associate conceptual, logical, and physical views of the information system
- Auto-generate tables, attributes, and indexes based on business rules, standards and predefined entities

### 3. Quick Delivery

- With 2 to 3 week delivery cycles, hence lower total cost of ownership, lower risk
- DV can be easily broken down by Subject areas for delivery
- Knowledge sharing between IT and Business
- Scalability and accommodation of growth; refactoring
- Virtual marts, Views

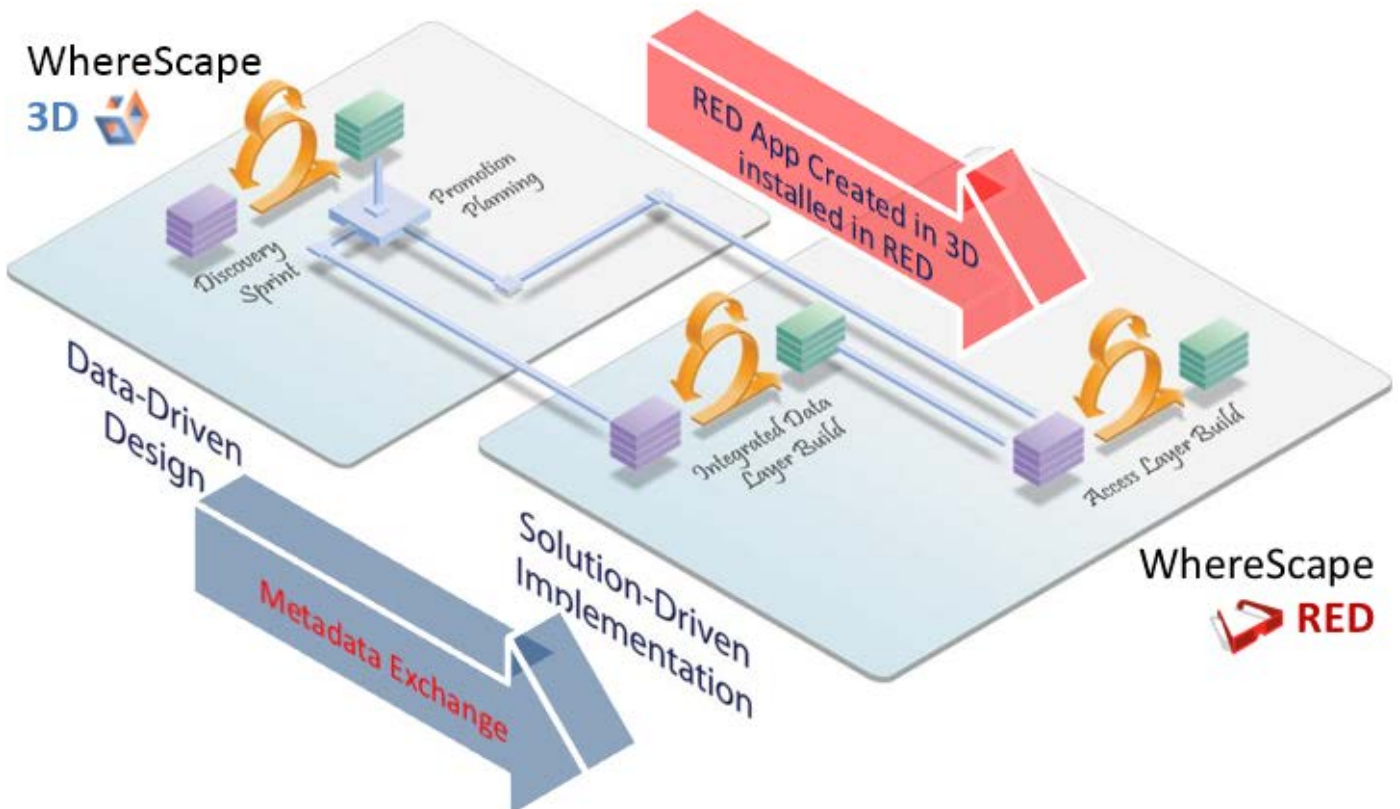


Figure 3 Automation Solution by WhereScape

## >DV2.0 New Big Data Warehouse from WhereScape

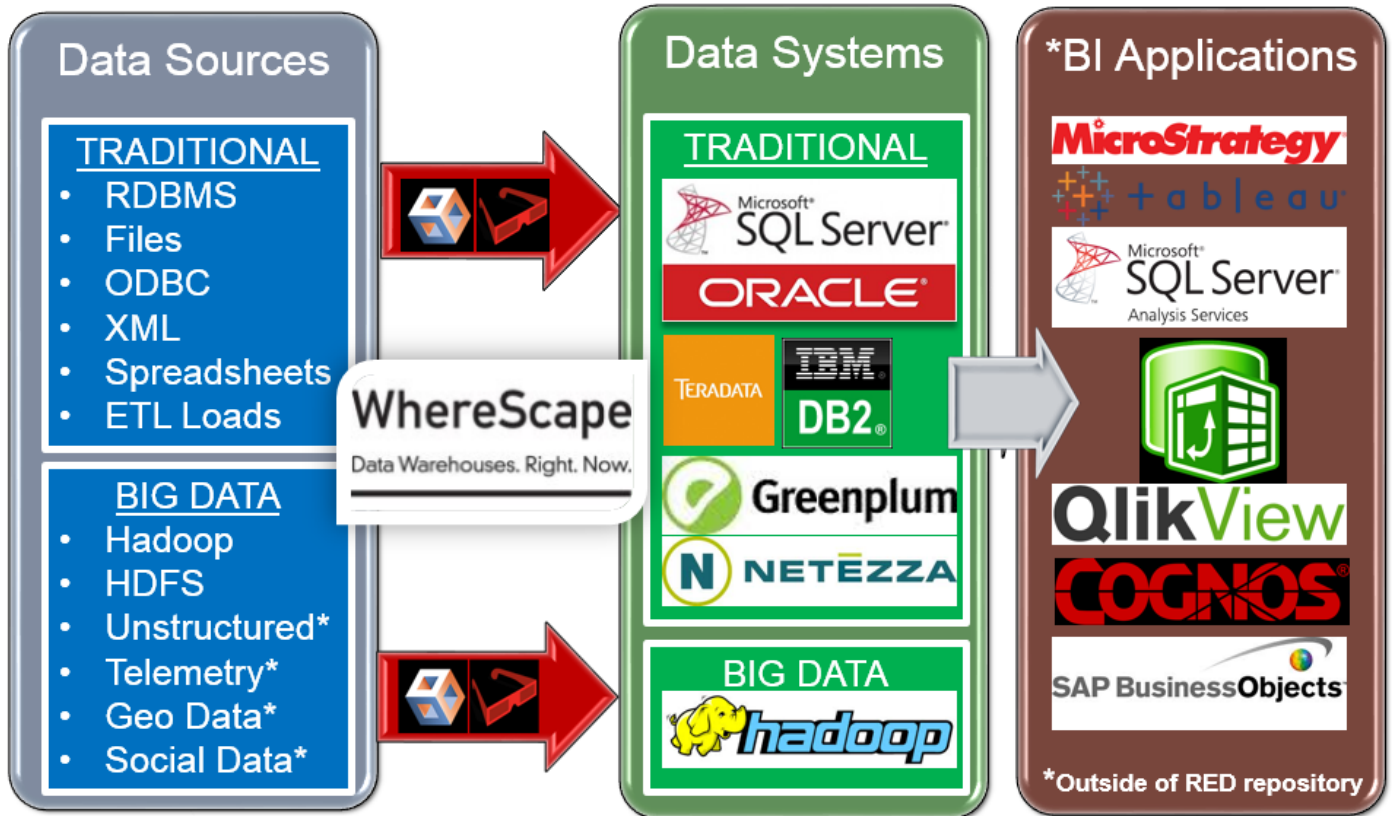


Figure 4 Big Data Approach in WhereScape Environment

### RED

- Support of MPP, MSPDW, Ext columns (>512); "big tables"
- Connect to Hadoop HDFS and Hive using ODBC or Hadoop Connector®
- Build a DW iteration in any of the supported target platforms (SQL Server, Oracle, DB2, Teradata, Greenplum, Netezza)
- Extend RED to support Hadoop HDFS and Hive environments as "target"
- ELT processing on both enterprise data warehouse and Hive in single tool.
- Both-way transfer of data between enterprise data warehouse and HDFS or Hive

### 3D

- Connect to Hadoop and Hive using JDBC
- Discover structures of Hive tables and HDFS files
- Profile data in Hadoop and HDFS
- Convert to any of the data warehouse models
- Integrate Big Data sources with others into 3D model
- An iterative approach to 3D development

## >Future Big Data Features in WhereScape

- Extended Hive as a target
- DB-specific loaders for Hadoop to RDBMS (Oracle/Teradata/Greenplum)
- Moving of data between RDBMS and HDFS or Hive.
- Processing of Extract Load and Transform (ELT) on Hive for standard object types - Load, Stages, Dimensions, Facts, Aggregates, etc.
- New "File" object = 'HADOOP'. Load tables created from HADOOP file directly.

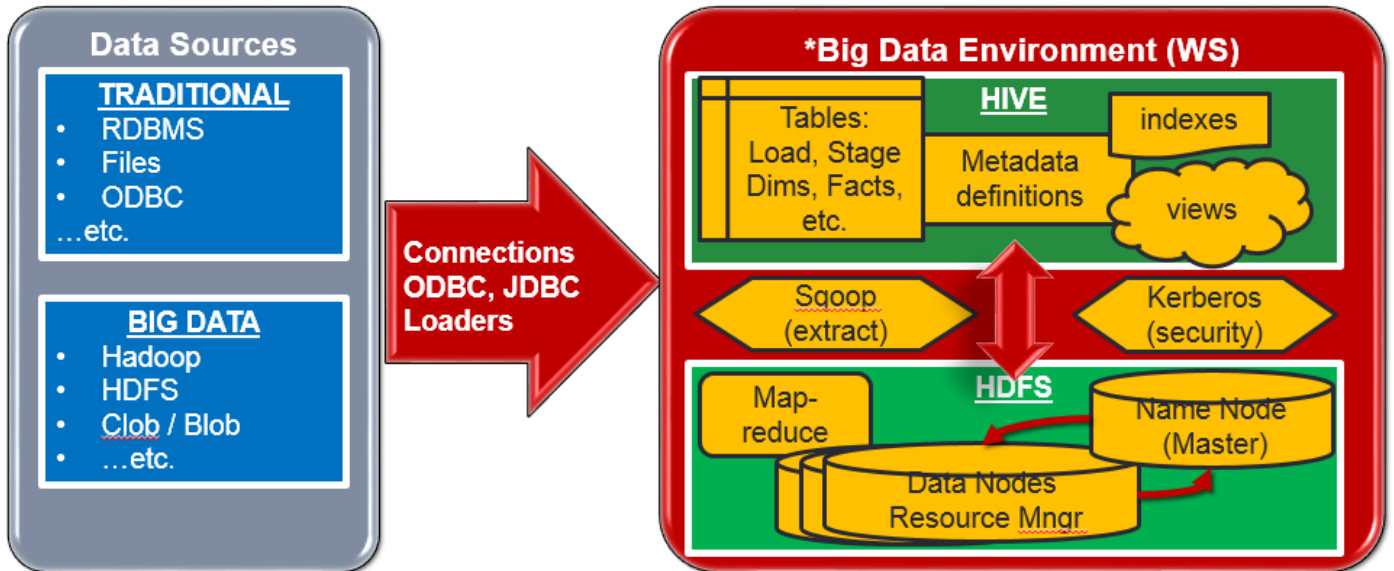


Figure 5 Big Data Solution by WhereScape

## >Conclusion

Utilizing WhereScape 3D - data warehouse planning tool - in conjunction with WhereScape RED - data warehouse building tool - provides a complete Agile solution for discovering, profiling, planning, building, deploying, managing and renovating data warehouses and data marts

WhereScape fully supports the DV 2.0s in all components: Conceptually, Structurally, and Procedurally. The synergy between DV 2.0 methodology and WhereScape implementation provides the analytical results for your decision-makers in hours and days instead of weeks and months.

## >Contact Information: *Raphael Klebanov* -- WhereScape USA, Customer Experience

**Mobile:** +1 303 968 0703

**Email:** [Raphael.klebanov@wherescape.com](mailto:Raphael.klebanov@wherescape.com)

**Skype:** raphael\_ws

**Website:** <https://www.wherescape.com/>

Contact me for scheduling 30-60 minutes no-obligation demonstration of the WhereScape RED and/or 3D including Data Vault implementation with either or both tools. Let me know of specifics of your environment so I can deliver an on-target demo. I am here to be a resource to you, so do not hesitate to call, email or skype