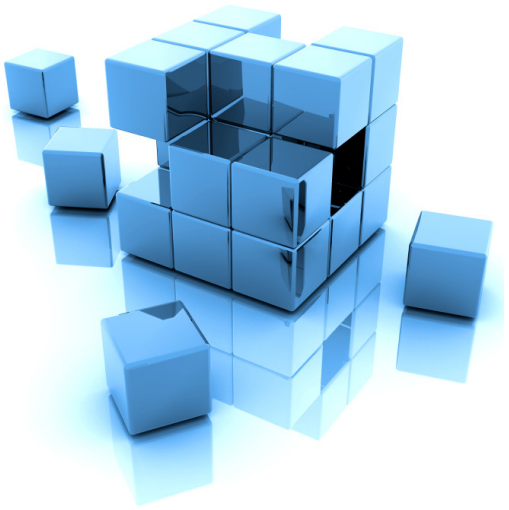
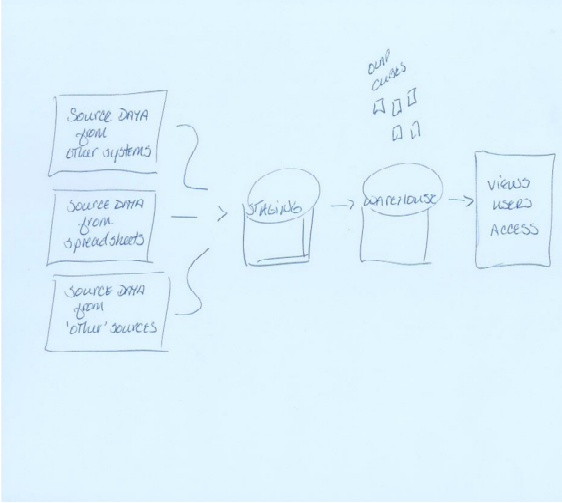


Data Testing on Business Intelligence & Data Warehouse Projects

Karen N. Johnson

Construct of a Data Warehouse



A brief look at core components of a warehouse.

- From the left, these three boxes represent the "source" data that is brought into the warehouse.
- Next there is often a staging environment.
- And then finally data is loaded from Staging to the data warehouse.
- Users typically do not access a warehouse directly but instead access data through non-updatable views.

ETL

"ETL: a workflow process used when transferring data from one system to another, specifically moving data to a data warehouse. Typically used to describe the process of acquiring source system data, manipulating it based on the data and business rules, then populating a data warehouse."

Source: The Data Asset, by Tony Fisher

The "E" of ETL

Data Extraction

- What data?

  Not all the data from a source is brought into a warehouse. What's extracted? What's not extracted?

- What parameters?

  Are there time boundaries around the data selected? Other boundaries?

- What about the quality of existing production data?

  When we "pull" or extract existing data we find mistakes of the past.

- What are the business rules of the data?

  When we understand the rules around each data element, then we can plan the boundary conditions to test.

The "T" of ETL

Data Transformation

- How is data is being altered?
- What business rules does that data need to adhere to?
- Are the data transformations custom or "out of the box" transformations from say, Microsoft's BI SSIS?

The “L” of ETL

Data Loading

The initial loading of a data warehouse begs the following questions:

- Is all the data there?
- How can we be certain?
- What about performance?
- What about security, access and permissions?
- What about subsequent data loads?

The OLAP cube from the Business Perspective

A common example of an OLAP cube includes data for:

- Products
- Cities
- Time

When data such as this is loaded into a warehouse, business analytics can address questions such as:

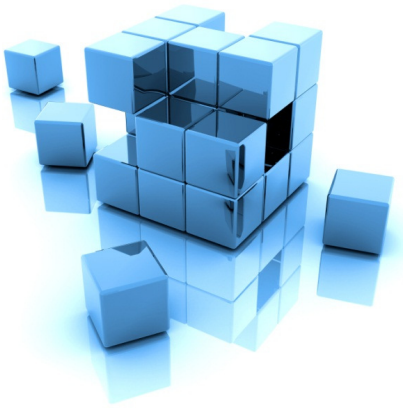
What are the top-selling products in each region over the past 12months?

The OLAP cube from the End User Perspective

The end users first contact with the data from the data warehouse is often a report or a dashboard.

If the data on the reports or dashboard is not accurate, the entire BI project can be at risk if the users lose confidence in the data.

The OLAP cube from the Tester's Perspective



Data reconciliation is the term often used to refer to the process of confirming the data in a warehouse has been loaded completely and accurately.

Determining what is *complete* and what is *accurate* is necessary and may not be obvious.

Testing Data Reconciliation

As is often the case in testing, there are more questions than answers at the start of the testing process.

- How do we know all the data is loaded?
- Why can't we use SQL counts to confirm loading?
- Why can't we test "all" the data?
- Why is some data from production outside of the business rules and the application rules?

Fully understanding the questions and potential issues is a good start for testing.

On one BI project, I started by discovering the data flow; every step and every process the data moved through.

Instead of SQL counts

At the onset, it would seem that data loading could be reconciled by counting. 100,000 rows get extracted, 100,000 rows get loaded, right? Because of data transformations and stored procedures straight counting may not work.

What are alternative methods?

- Trolling production records to match test conditions and "hand" verifying those rows through the process. Trolling refers to what is more properly known as data profiling.
- Using SMEs to identify where the "ugly" data is and verifying those specific rows of data.
- Building test data that tests business rules and data transformations in both positive (confirming) and negative (challenging) ways.
- Harvesting all three of these methods to build and fortify a rigorous test data set.
- Advocate for error logging as well as automated unit tests.
- Don't forget to explain the rigor behind this process to the business users.

Source to Target Mapping

- Where is the data is coming from?
- Do you understand all the attributes of the data such as – data type, field length and business rules?
- Where is the data being loaded – what is the target?
- Do you understand all the attributes of the "destination" for the data such as the data type, format, rules, etc.
- The data dictionary is often where source to target mapping is documented. That dictionary is invaluable to testing (as well as for data governance). Internally "selling" the importance and need for the dictionary can be an uphill battle.

Explaining Data Reconciliation Testing to the Business

- In order for the business to be confident of the testing efforts, maintain details of testing. This might include:
- A list of the parameters and business rules of the data being extracted from source systems.
- Building data sets that challenge the parameters and rules. Maintaining and sharing those data sets for current and future testing.
- Identifying the date/time of data loads from source to target used for data reconciliation.

Invisible “influences” on Data

In addition to the more obvious data transformations that take place through ETL jobs, there are other less visible processes that may alter data. Including:

- Stored procedures
- Triggers
- Indexes
- Views
- Permissions
- Batch runs

Ask if these exist. Investigate what may need to be tested.

Testing Business Rules

- Business rules might preventing the loading and use of specified data. Regardless of the specifics of any rule, it is not the data that has the rule, but the requirement or a restriction around the data that is set.
- Once the business rules are defined, the developers translate those parameters into code to prevent loading of particular data. Those business rules may also refine the data in the cubes.
- It's imperative to test the business rules as the rules have considerable impact into what data is loaded and how data is presented to the consumers of the data.

Testing Reports & Dashboards

- On one BI project, I tested data from reports back to the source. The concept of this testing was to “follow the data.”
- There are tools to track lineage. Lineage can also be done manually although it is not practical for each data element, multiplied by all of the processes that may take place.
- Samples of data lineage can help business users understand how data can be transformed.
- Report testing should include multiple iterations and different sets of data, such as different regions and over different periods of time including month end and year end reporting.
- Reports are not just a byproduct of a BI project but are often the primary need for the business users.

A Few Reasons for Data Discrepancies

- Permissions and access to data is one (often unexpected) reason why the data appearing is not what may be expected.
- Views that may be out of date is another reason by data appearing on reports may not "match."
- Aggregate data and derived data - both forms of data are not easily "tied back" to a source to reconcile.
- Sample test data sets can be a useful way to track down issues.

One version of the truth

In computerized business management, **svot**, or **Single Version of the Truth**, is a technical concept describing the data warehousing ideal of having either a single centralised database, or at least a distributed synchronised database, which stores all of an organisation's data in a consistent and non-redundant form.

source Wikipedia

Star schema

- A star schema consists of fact tables and dimension tables. Fact tables contain the quantitative or factual data about a business--the information being queried. This information is often numerical, additive measurements and can consist of many columns and millions or billions of rows. Dimension tables are usually smaller and hold descriptive data that reflects the dimensions, or attributes, of a business. SQL queries then use joins between fact and dimension tables and constraints on the data to return selected information.

- The facts--what is being analyzed in each case--are revenue, actuals and budgets, and shipments. These items belong in fact tables. The business dimensions--the by items--are product, market, time period, and line item. These items belong in dimension tables.

source:
http://publib.boulder.ibm.com/infocenter/rbhelp/v6r3/index.jsp?topic=%2Fcom.ibm.redbrick.doc6.3%2Fwag%2Fwag32.htm

Snowflake schema

A **snowflake schema** is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake in shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions.

source: Wikipedia

Slicing and Dicing

- *Slice*: A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.
- An example, the sales figures of all sales regions and all product categories of the company in the year 2004 are "sliced" out the data cube.

- *Dice*: The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).
- An example, the sales figures of a limited number of product categories, the time and region dimensions cover the same range as before.

source: Wikipedia

Drilling Up & Down

Drill Down/Up: Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).

source: Wikipedia

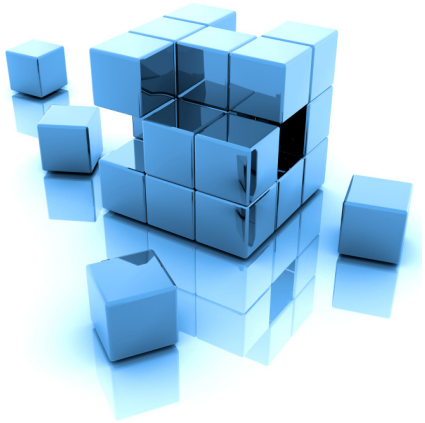
BI terms

- Facts and measures
- Granularity, the grain
- OLAP, ROLAP, MOLAP
- Star schema
- Snowflake schema
- Entities
- Data model
- Cardinality
- Inmon vs. Kimball

Data terms

- Data federation
- Data profiling
- Data steward
- Data cleansing
- Data lineage
- Data governance

BI Testing Skills



- SQL
- Excel including Pivot Tables
- MDX
- BI purpose
- Data Warehouse Architecture
- The ability to read stored procedures and triggers.
- The ability to “talk” to business users.
- Tenacity
- Curiosity
- Visual tools such as Tableau
- Confidence in testing without a GUI; testing that is not visual or easily tangible.

Data Warehouse Resources

The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset
by Joy Mundy, Warren Thornthwaite and Ralph Kimball

The Data Warehouse ETL Toolkit: Practical Techniques
by Ralph Kimball and Joe Caserta

Data Strategy
by Sid Adelman, Larissa Moss and Majid Abai

Data Quality: The Accuracy Dimension
by Jack E. Olson

Dimensional Data Warehousing with MySQL: A Tutorial
by Djoni Darmawikarta

Business Metadata: Capturing Enterprise Knowledge
by William H. Inmon, Bonnie O'Neil and Lowell Fryman

Fast Track to MDX
by Mark Whitehorn, Robert Zare and Mosha Pasumansky

Business Intelligence with Microsoft® Office PerformancePoint™ Server 2007 by Craig Utley

Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence
by William H. Inmon and Anthony Nesavich

Professional SQL Server 2005 Integration Services
by Brian Knight, Allan Mitchell, Darren Green and Douglas Hinson

Questions?

Thank you for your time.

Email: karen@karennjohnson.com

Site: www.karennjohnson.com