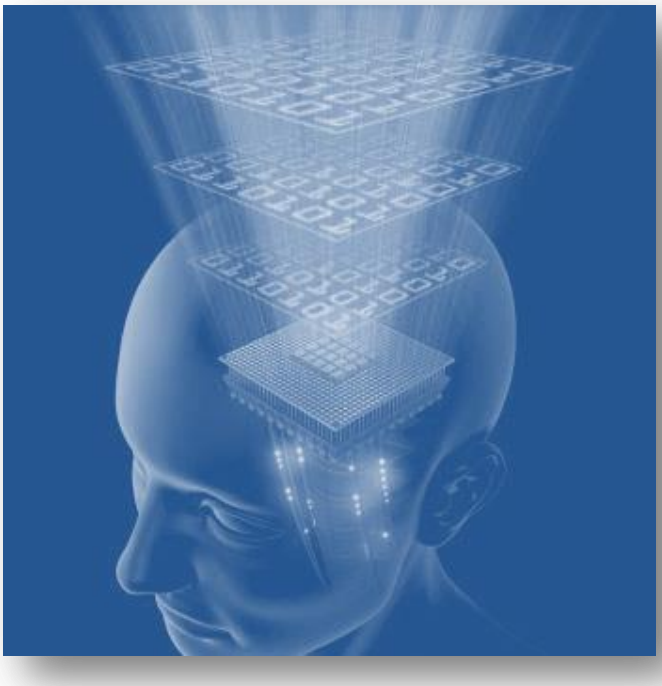
March 7, 2014

Enterprise Business Intelligence & Data Warehousing:

The Data Quality Conundrum



A study by RTTS

For Business and IT Professionals

By Bill Hayduk

Enterprise BI & Data Warehousing: the Data Quality Conundrum

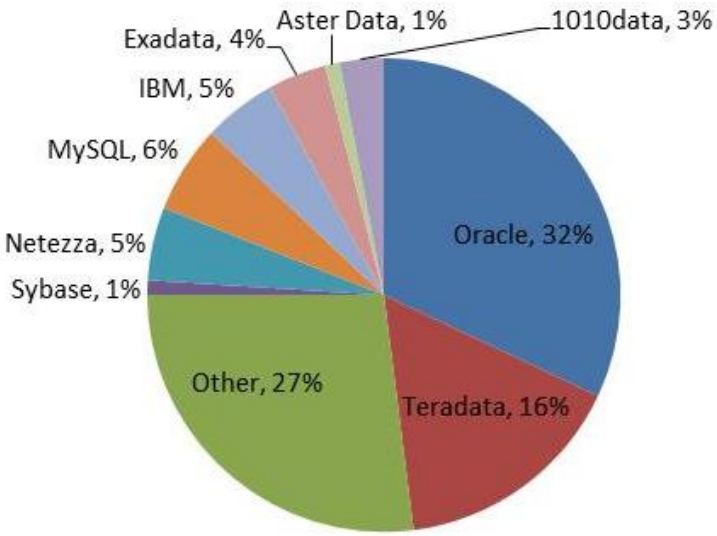
Poor data quality is a huge and exponentially growing problem. Big Data is causing massive increases in the volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources) of data. Therefore, concern for poor data quality or 'bad data' is now a critical issue. According to Gartner, "the average organization loses $8.2 million annually through poor data quality, with 22% estimated their annual losses resulting from bad data at $20 million and 4% put that figure as high as an astounding $100 million". And InformationWeek found that "46% of companies cite data quality as a barrier for adopting Business Intelligence products".

We recently performed a study that included responses from over 200 companies interested in improving the data quality in their Business Intelligence, ETL software and Enterprise Data Warehouse implementations. Below are our findings, along with context around these results.

SECTION I.

Our first section polled customers on their architecture: specifically on their data warehouse, ETL and business intelligence software and vendors.

Enterprise Data Warehouse Software
The top data warehouse vendors are Oracle (plus MySQL and Exadata) as number one and Teradata (and Aster Data) as number two and every other one far behind. This is in sync with research by most analyst firms that track this platform. Oracle and Teradata dominate the space with both their loyal customer base and their innovation.

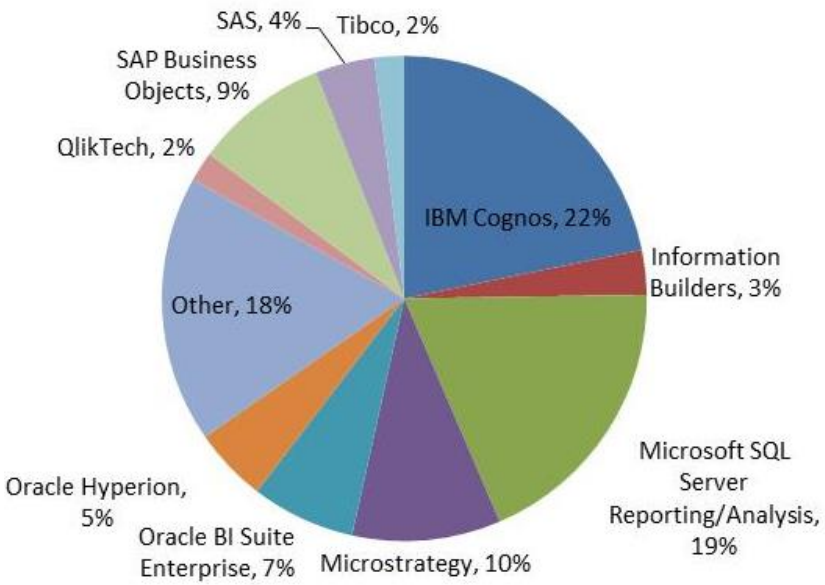


Analyst firm Gartner, in its 2013 Magic Quadrant for Data Warehouse Database Management Systems report, projected a 10% growth in the database management system market and it pinpointed a significant increase in organizations seeking to deploy data warehouses for the first time.

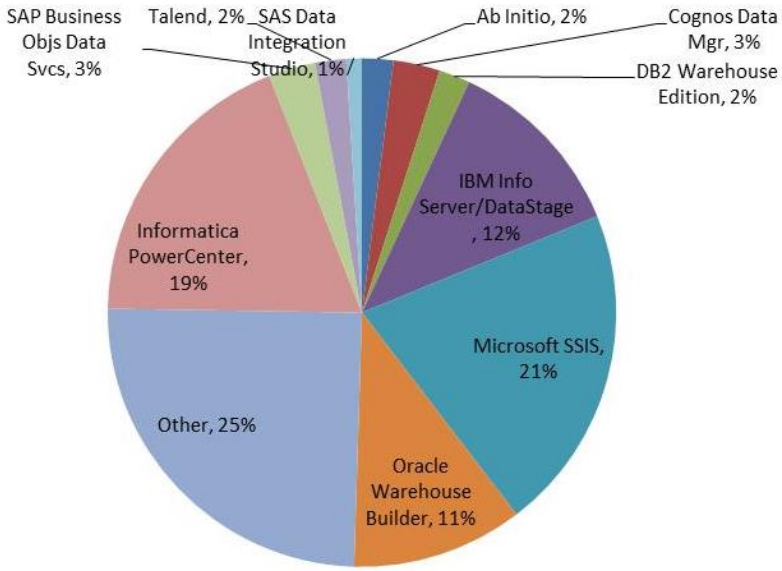
Business Intelligence Software

IBM leads the BI space with Cognos, followed by the surprising performance of Microsoft at second and Oracle's combined offering at 3rd. Unexpectedly, 'other vendors' were chosen 18% of the time, meaning there is still room for growth in the BI space.

Analyst firm IDC stated that the market is now forecast to continue to grow at a 9.8% compound annual growth rate through 2016. One key observation they made was "the media attention on Big Data has put broader business analytics on the agenda of more senior executives."



ETL (Extract/Transform/Load) Software

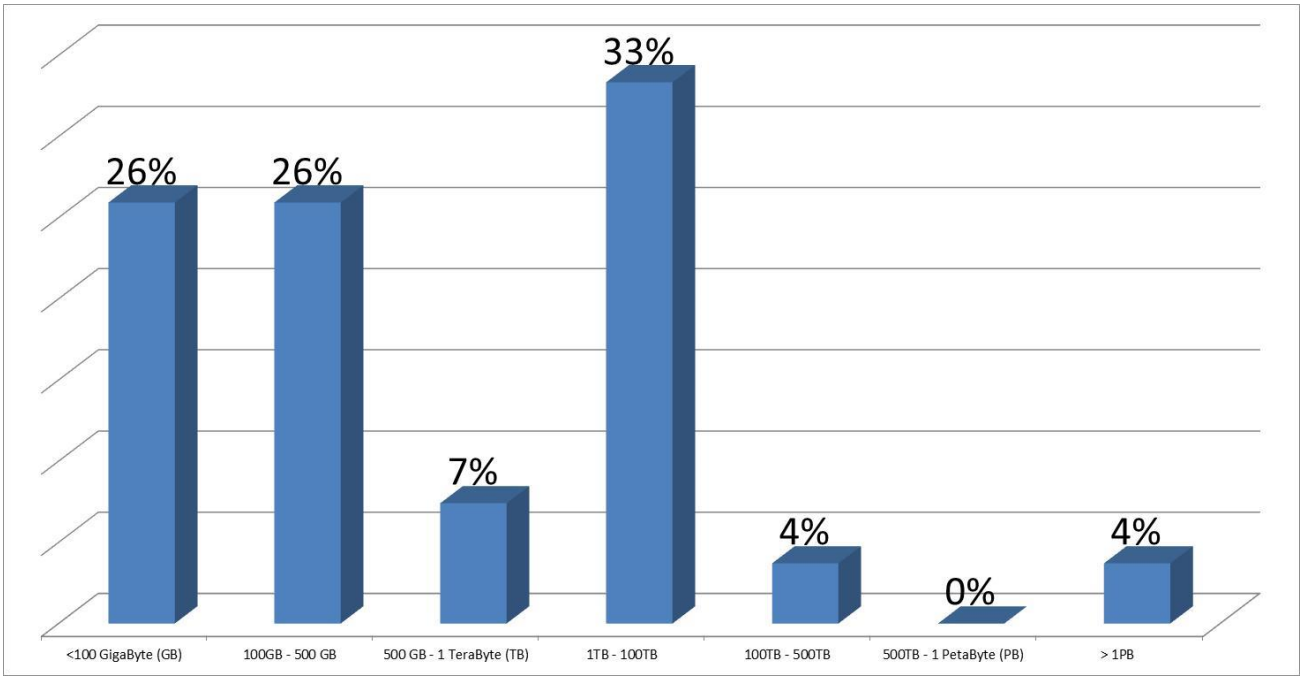


Here we see another surprising result of our survey. Microsoft finished first in the ETL Vendor software survey. Informatica PowerCenter, the most widely known vendor finished second, closely followed by IBM's combined offering at third. It is interesting to note that 'other' came in first above Microsoft. There is still a large contingent of companies using home-grown and open source software in the marketplace.

According to analyst firm Forrester, "The enterprise ETL market continues to grow at a healthy pace as more enterprises replace manual scripts with packaged ETL solutions. This migration… toward ETL tools is driven by the need to support growing and increasingly complex data management requirements."

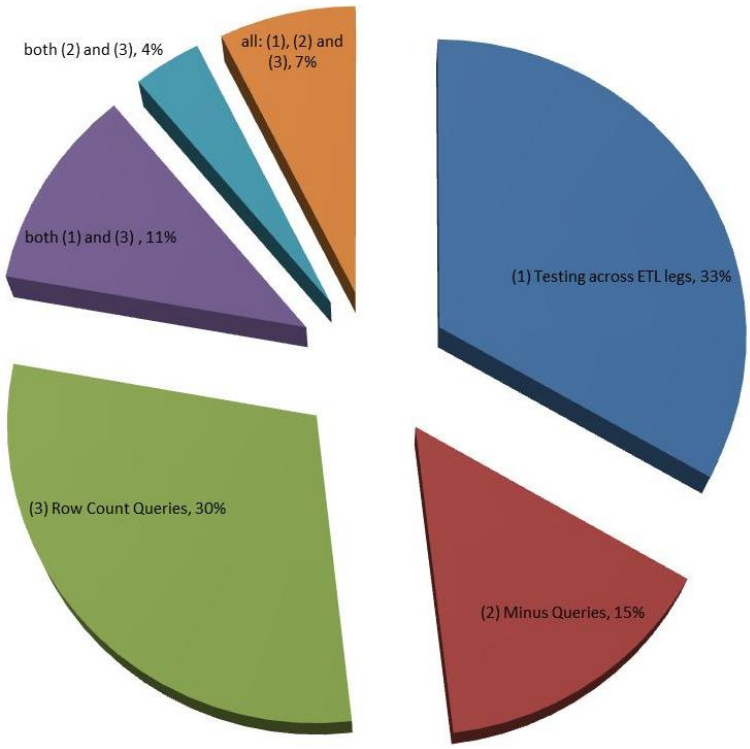
Current Data Warehouse Size

We inquired as to the current size of the data warehouse implementations. We discovered that 91% were less than 100 Terabytes and 52% were less than 500 Gigabytes. Interestingly, the largest sector (33%) of implementations was between 1 Terabyte and 100 Terabytes. This is a significant increase in data size when the largest sector in our poll 2 years ago was measured in Gigabytes.



SECTION II.

In Section II, we analyzed the current quality situation of firms to determine their effectiveness.

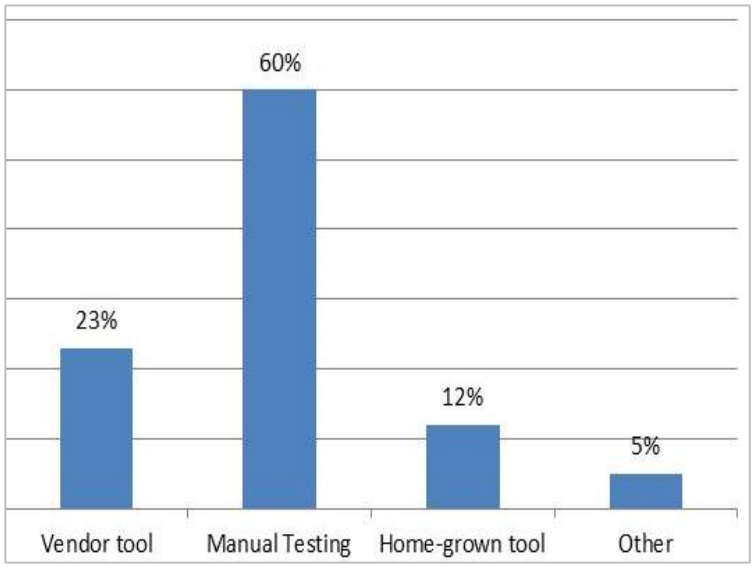


Current Testing Strategy

We inquired as to which test strategy was being implemented: (1) testing across ETL legs (source-to-target DWH, DWH to data mart, etc.), (2) utilizing Minus Queries (see full definition of minus queries here http://bit.ly/13lmp8N), and/or (3) comparing row counts from source-to-target.

The preferred strategy is to utilize a combination of the three, depending upon the role (tester, ETL developer, operations). But if one were going to choose a single strategy to check for data quality, it would be (1). We found that 30% of companies polled only verify row counts and only 7% implemented all 3 as their preferred strategy. Many singled out a lack of automated testing and a lack of testing resources as reasons they did not deploy a more rigorous testing strategy.

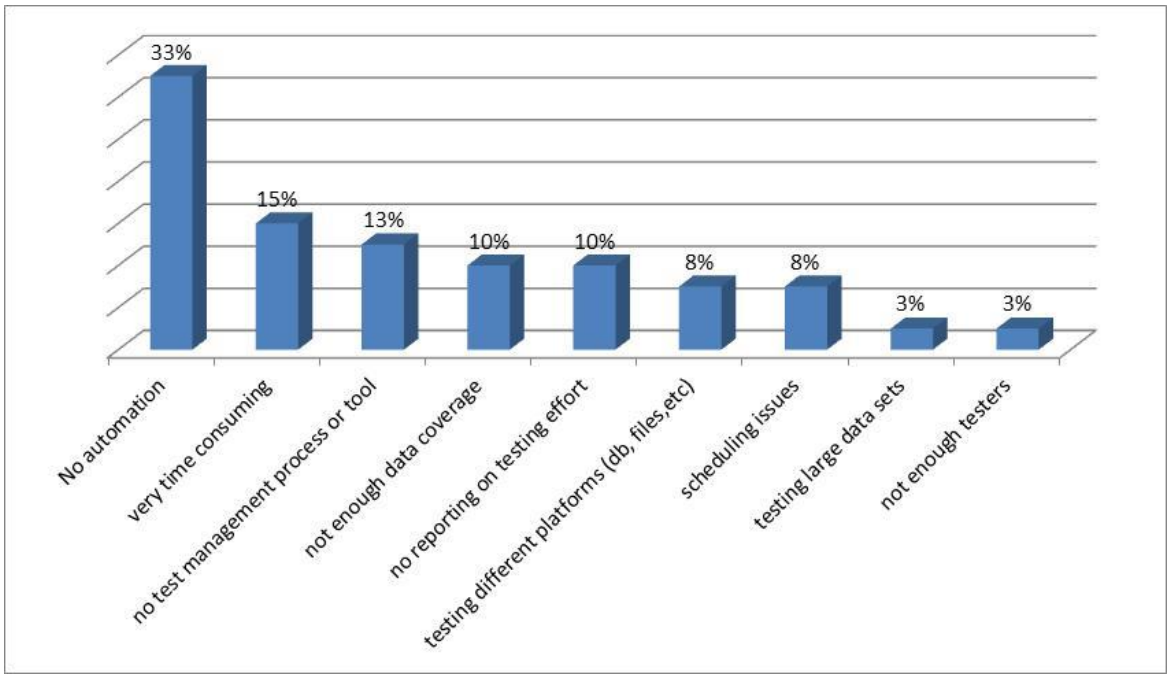
Current Test Execution Method



When surveying customers on their current test execution method, we found that 60% of testing is currently performed manually. Manual testing consists of extracting data from the source databases, files and XML and also extracting data from the data warehouse after it goes through the ETL process and then comparing these data sets manually, by eye. This is quite extraordinary when considering that the average data warehouse is measured in gigabytes and typical tests return millions of rows and upwards of hundreds of columns, meaning millions of sets of data to compare. Therefore, testers can only sample the comparisons for practical purposes. Vendor tools come in second and a home-grown finished third.

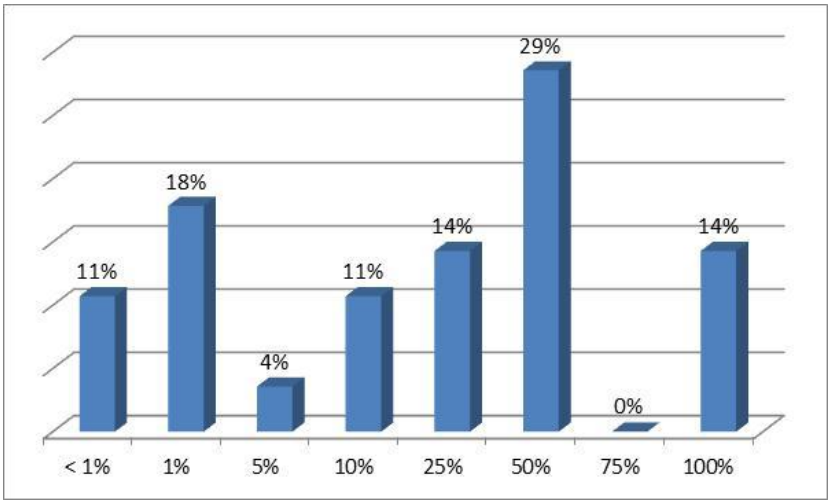
Data Quality and Testing Challenges

When customers were asked about their biggest challenges when it came to testing the data for accuracy, the clear top choice was 'No Automation'. This goes hand-in-hand with the 2nd biggest challenge – that testing manually is very time consuming. This was followed by the lack of a test management process and/or tool, not enough data coverage and no reporting on the testing effort.



Percent of data coverage by current test process

When determining the amount of data coverage that companies' current test process provides, it is clear that there is not much data coverage at all. Of those companies surveyed, 84% had less than 50 percent data coverage, 58% had less than 25 percent coverage, 33% had less than 5 percent and 29% of companies had less than 1 percent. The 14% who had 100 percent coverage had data warehouse implementations less than 500GB in size and were interested in finding a data warehouse testing tool that could speed up the testing cycle. Also, the coverage represents the amount of data brought back by SQL queries, not the amount compared. Since comparisons (as noted above) are typically performed by visually reconciling the 2 data sets, it is impractical for more than 5-10% of the data to be compared. Our estimate is that in reality, far less than 1% of data is actually explicitly verified by these companies.



Ratio of Developers to Testers

Average ratio:	2.1 to 1
Median ratio:	1.7 to 1
Highest ratio:	20 to 1
Lowest ratio:	2 to 3

We found that, for the most part, the ratio of developer to tester was standard in principle. The issue that most firms surveyed stated was that they could not get enough data coverage. The reason: ETL developers are utilizing some form of tool to make their jobs faster and more efficient while most ETL Testers are not.

Effects of Bad Data

We surveyed customers for their anecdotal responses on the impact of bad data on their organizations. Of the ones who answered, 100 percent said that they experienced some form of bad data in their data warehouses. Below are samplings of their free-form answers on the effects that bad data caused them.

- "Incorrect business intelligence reports"
- "Poor delivery quality & customer dissatisfaction resulting in re-work"
- "Missing revenue opportunities"
- "Critical business decisions rely on underlying bad data"
- "Bad quality of projects"
- "Negative business Impact"
- "SLA Issue with our customers"
- "Major embarrassments to our team"
- "Long working hours to fix bad data"

Conclusion

Many companies are using Business Intelligence (BI) to make strategic decisions in the hope of gaining a competitive advantage in a tough business landscape. But Bad Data will cause them to make decisions that will cost their firms millions of dollars. It is clear from the results of this survey that companies are not providing the level of data quality that C-level executives need to make reliable strategic decisions. Most firms test far less than 10% of their data by sampling the data and the comparisons. Therefore, at least 90% of data remains untested. Since bad data exists in all databases, firms need to test closer to 100% of their data and guarantee that this critical information is accurate.

There is no practical way for testers to verify this level of coverage without the use of automated testing tools. An automated testing solution will speed up the process, provide much more data coverage, perform comparisons automated, and provide reports for audit trails. An example of an automated testing tool is QuerySurge (www.QuerySurge.com) It is clear that as data grows exponentially, a more complete solution is needed to keep enterprise-level data clean.

About The Author

Bill Hayduk
founder, CEO
RTTS

Bill founded software and services firm RTTS in 1996. Under Bill's guidance, RTTS has supported over 600 projects at over 400 corporations, from Fortune 500 to midsize firms.

Bill holds an MS in Computer Information Systems from the Zicklin School of Business (Baruch College) and a BA in Economics from Villanova University.

About RTTS

RTTS is the premier software and services organization that specializes in providing software quality for critical business applications. With offices in New York, Philadelphia, Atlanta and Phoenix, RTTS has been serving Fortune 500 and mid-sized companies throughout North America since 1996.

*RTTS draws on its expertis*e utilizing its proven methodology, expert test engineers and the industry's best-of breed tools to provide the foremost end-to-end solution that ensures application functionality, reliability, scalability and availability. For more information, visit www.rtts.com.

References

- *Gartner: Magic Quadrant for Data Warehouse Database Management Systems (January 31, 2013)*
- *IDC: Worldwide Business Analytics Software 2012-2016 Forecast and 2011 Vendor Shares (June 2012)*
- *The Forrester Wave™: Enterprise ETL, Q1 2012 (February 27, 2012)*
- *InformationWeek: 2012 BI and Information Management Trends (November 2011)*
- *RTTS 2013 Client Survey on BI and Data Warehouse Quality*