Making the Best Selection and Utilization of New IT Tools for Data Warehouse Systems

Lee Bowling
National Agricultural Statistics Service (NASS), United States Department of Agriculture (USDA)
1400 Independence Ave. SW, Room 4833 South Bldg.
Washington, DC 20250
USA
Lee.bowling@nass.usda.gov

Taizhu Zhou
National Agricultural Statistics Service (NASS), United States Department of Agriculture (USDA)
1400 Independence Ave. SW, Room 4833 South Bldg.
Washington, DC 20250
USA
Taizhu.zhou@nass.usda.gov

DOI: 10.1481/icasVII.2016.f39c

ABSTRACT

There are many types of databases. For this topic the focus will be on the data warehouse. While a data warehouse can be defined in a wide variety of ways, for this discussion the focus will be around a data warehouse based on the Inmon concept of having one enterprise data warehouse which serves as the source for all other data based systems in an organization. On line analytical processing, or OLAP, will also be the targeted type of system for the purpose of this discussion. In short, the data system is optimized for rapid data retrieval and analysis.

Most entities recognize the utility of data and its retention. As these data stores grow, more and more resources are needed to hold the data, make backup copies of it, create new copies for reporting, and many other uses. Planning for the best methods of accessing and coordinating data have always been of paramount importance. Many organizations are still working on making maximum use of their data for adding value to all business processes. In many places even with efficient planning for how to use data, however, a point has been reached where the amount of data preserved can be problematic for retrieval with the IT tools which have been available in the past.

The National Agricultural Statistics Service has used the same data warehouse system for more than 17 years. There have been upgrades to hardware and software, along with needed structure changes, but no major shift in types of processors or software vendors. The system has served well, but recently the agency has seen more and more need to schedule certain analytical queries for 'off' hours when there would be few users on the system. One of the most basic reasons for having a data warehouse is the ability to analyze data and make use of it. Large queries accessing every known table and row could take up to five hours to run on the system as it was designed. It was also found that the software provider was not planning to make more substantive upgrades to the system, but would instead put resources into other products they felt were more in keeping with current trends.

Any decision for changing the data warehouse would affect literally hundreds of in-house agency applications. Along with considerations of cost, support, and integration, there was recent research into new massive parallel processing systems which could yield dramatic increases in query speeds. This presentation will detail the planning, areas of consideration, and comparison of features available in newer systems which lead to the purchase of a new data warehouse appliance for the agency and potential decreases in query time from hours to minutes or even seconds.

Keywords: Data Warehouse, Massive Parallel Processing

1. Introduction

1.1 An agency within the United States Department of Agriculture (USDA), the National Agricultural Statistics Service (NASS) conducts hundreds of surveys every year and prepares reports covering virtually every aspect of agriculture in the United States. Our mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. The surveys and related work are conducted mainly in twelve Regional Field Offices and five call centers across the United States.

1.2 NASS has used the same data warehouse system for more than 17 years. There have been upgrades to hardware and software, along with needed structure changes, but no major shift in types of database or software vendors. The system has served well, but recently the agency has seen more and more need to schedule certain analytical queries for 'off' hours when there would be few users on the system.

1.3 One of the most basic reasons for having a data warehouse is the ability to analyze data and make use of it. The largest queries, accessing every known table and row, began taking more time to process, taking up to five hours to run on the system as it was designed. The software behind the database engine had been purchased over time by a number of vendors. It was announced that the most recent software provider was not planning to make more substantive upgrades to the system, but would instead put resources into other products they felt were more in keeping with current trends. This was a concern with a steady progression of the hosting hardware and operating systems possibly leading to more errors or speed problems with the warehouse system which would not be upgrading.

1.4 Any decision for changing the data warehouse would affect literally hundreds of in-house agency applications. Along with considerations of cost, support, and integration, there was recent research into new Massive Parallel Processing (MPP) systems which could yield dramatic increases in query speeds. This presentation will detail the planning, areas of consideration, and comparison of features available in newer systems which led to the purchase of a new data warehouse appliance for the agency and potential decreases in query time from hours to minutes or even seconds.

2. Background

2.1 There are many types of databases. For this topic the focus will be on the data warehouse. While a data warehouse can be defined in a wide variety of ways, for this discussion the focus will be around a data warehouse based on the Inmon concept of having one enterprise data warehouse which serves as the source for all other data based systems in an organization (Inmon, 1993). On line analytical processing, or OLAP, will also be the targeted type of system for the purpose of this discussion. In short, the data system is optimized for rapid data retrieval and analysis.

2.2 Most entities recognize the utility of data and its retention. As these data stores grow, more and more resources are needed to hold the data, make backup copies of it, create new copies for reporting, and many other uses. Planning for the best methods of accessing and coordinating data have always been of paramount importance. Many organizations are still working on making maximum use of their data for adding value to all business processes. In many places even with efficient planning for how to use data, however, a point has been reached where the amount of data preserved can be problematic for retrieval with the IT tools which have been available in the past.

2.3 The variety of data available has also increased. The internet has virtually exploded with new and varied data and data sources. While there are still differences in internet access across the globe, the coverage and availability is growing at an increasing pace, with over 1000 percent growth in some areas of the world since the beginning of the century (Bell, 2011). Multi-media data and Big Data concepts are available if one has a suitable system, with enough speed potential and analytical tools, to take advantage of them (Beyer & Edjlali, 2015).

2.4 NASS was a comparatively early adopter of data warehouse concepts and has gathered a great deal of historic data over time. Most of the agency data is structured and uses detailed metadata. These characteristics affected the field of choices in warehouse products, making those with more open formats or ‘no sql’ choices more problematic in our case. There were early concerns about any cloud offerings, as well. Most of this was regarding either real or perceived security concerns. The selection team was open to possibilities but concerned that the timing might not yet be right for a cloud product.

2.5 More and more vendors are offering data warehouse appliances. Instead of purchasing a variety of hardware and software separately and then doing the integration within your own organization, there are now viable choices for systems which have already been designed and tested for optimum performance from a combination of hardware and software. Queries that might have taken hours on self-assembled systems can potentially take only seconds on pre-configured appliances (Beyer & Edjlali, 2015).

3. Steps Taken

3.1 With the current system in place for nearly two decades, there was some expectation at NASS that whatever new system was selected should be something that could grow with the agency and with the changing needs. The current system had over 11 billion rows of data, grew from over 250 surveys’ data annually, and was accessed by several hundred in-house software application systems and two primary Commercial Off The Shelf (COTS) software packages for business analytics and statistical analysis. The agency staff were particularly familiar with one COTS business intelligence analytical tool considered the standard tool for its purpose.

3.2 Investigation and planning were needed, but also some degree of speed in the decision. In our case ‘speed’ meant the investigation/selection project should take less than one fiscal year in order to take advantage of funding that could not be guaranteed in subsequent years. Even with time concerns, however, the agency would follow the three historic ‘pillars of progress’ which had been observed in the past NASS data warehouse implementation: 1) Focused Direction; 2) Sound Evaluation and Development; 3) Solid Implementation (Yost, 1999).

3.3 A seven person team was formed for the purpose of investigating options for replacement of the data warehouse system. Members were drawn from among the database administrators, metadata specialists, contract database support, the application software architect, and data analysts from outside of the IT division. The IT Division Head was the executive sponsor and the IT Division Senior Project Manager provided support for budgetary and procurement concerns.

3.4 From the literature available and based on the original cost of the current solution, a very broad estimate of purchase cost was put forward. This was needed at the beginning to plan and help in discussions with senior management. If the upper managers were put off by the cost or the upcoming degree of effort then any project could be considered defeated before it began. By setting the general expectations early in the project senior manager support was fully behind the process. This helped in a wide variety of ways, including recruiting team members and building favorable support.

3.5 The general timeline for the project was:

Dec. 2014:	Establish initial project plan
	Collect market information
	Reach out to business users
	Determine criteria for further investigation and selection
Jan. -Feb. 2015:	Request contract specialist from procurement staff on team
	Selection of pool of vendors based on criteria for products
	Prepare Proof Of Value (POV) trial criteria for vendors
Feb – Apr. 2015:	POV with vendors (four total)
May - June, 2015:	Report of findings to Data Services Branch staff
	Report of findings to NASS Enterprise Architecture Council
	Report to Senior Executive Service
June – Sept. 2015:	Work with contract specialist to procure recommended product

3.6 Communication was key. The project team included statisticians from the business community, who were also heavy/frequent users of the data warehouse. These same people helped to spread information and interest for the project. Other business side users were invited to help in developing the POV tests to be used with the various vendors. Evaluation results and periodic status messages were communicated directly among the team, with the entire Data Services Branch staff, the NASS Enterprise Architecture Council and the Senior Executive Team. All of this helped in planning and to promote acceptance and a favorable climate of acceptance for the general direction and product choices of the team.

3.7 A separate project team was also formed to review choices for a new analytical software tool. That effort was conducted in much the same way as the analytical database team's project, with both teams communicating with each other.

3.8 Instead of creating all of the selection criteria from scratch, the team chose to use a third party authoritative source to describe the features desired. In our case the publicly available Gartner ***Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics*** report was an excellent resource. An abridged version was advertised and available via the web and we received the entire report via a subscription service (Beyer & Edjlali, 2015). This had several benefits. The research was current, in-depth, and forward looking. In this case it helped narrow the field considerably when it was found that only four of the vendors in the 'Leaders' quadrant also produced appliances. And a major benefit in our situation was the ability to provide a third party list of characteristics and rating to our procurement office in support of the team's methods and ultimate recommended selection.

3.9 A special comparison was made with a leading vendor cloud offering. This was in anticipation of trends in the industry and potential questions from stakeholders. NASS is a statistical agency dealing in confidential data protected by law. Even the appearance of a compromise in data security can lead to lower response rates to agency surveys. The Agricultural Advisory Committee (composed of public and private stakeholders in statistical reporting) had even made previous recommendations against some cloud systems housing our data. In addition to any concerns about public opinion, there is also an official accreditation process for federal government systems. There are several federal programs underway specifically for the purpose of certifying cloud based offerings for various purposes. The Department of Agriculture also has some internal initiatives with the goal of providing cloud services. Ultimately it was decided that cloud offerings were not yet at a stage of acceptance either internally or externally in terms of confidential data security or the perception needed for our data security. As government certification programs advance, this may change and cloud offerings may be a more viable option in the future.

3.10 Market research consisted of internet sources and searches (including the previously mentioned magic quadrant report), vendor meetings, demonstrations, attendance in public conferences on the topic, and internal stake holder meetings. The team compiled a list of other government agencies doing similar work to our own. Calls and meetings were organized to enquire about these agencies' own data warehouse solutions, planning direction, and general satisfaction with their current systems and vendors. General comparison criteria during the market research included:

- Architecture.
- Scalability
- Reliability
- Performance
- Compatibility (with existing environment and code in place)
- Administration tools (availability, and ease of use)
- Price (a formal decision was based on best overall value, not simply lowest purchase price)

3.11 Our investigations led to a special focus on two characteristics. The systems using both preconfigured appliances and described as using massive parallel processing (MPP) presented the potential for the greatest speed increases. Simply put, MPP means dividing the work for different parts of an application or data retrieval among multiple processors. In the majority of the vendors we reviewed, the solutions included specialized processors and a set number of specific disc drives associated with any one processor. It was difficult to get a precise estimate ahead of time of what the speed increase would be without testing with a specific data structure and volume. Estimates ranged from 10 to 100 times faster (Lopes, 2015). All of the vendor systems reviewed showed significant speed increases over the current NASS data warehouse. The system ultimately selected returned the longest running known queries roughly 100 times faster. Queries that would typically take over five hours, returned results in roughly three minutes.

3.12 Once the field of potential vendors that excelled in our initial criteria was established at four, the Proof Of Value (POV) trials were begun. Vendors were allowed to conduct the tests off site from our own facility. All were given a group of representative queries for benchmarking times. All were given the same test data which had been approved by our agency Security Staff and statistical staff to ensure there was no release of confidential data and to prevent any confusion leading to any appearance of releasing confidential data. The team visited each vendor and took along specific queries and use cases for testing. Each vendor was asked to specifically demonstrate the ability of their system to be accessed by one COTS statistical software package and one COTS analytical software package. Both of these are considered standards and are in wide use in the agency. The broad criteria used in the evaluations included:

- Benchmark and test case performance.

- Workload of database migration
- Post migration work on in-house applications and programs.
- Backup and restoration
- DBA technical requirements
- Administration tools (availability, and ease of use)
- Price (a formal decision was based on best overall value, not simply lowest purchase price)

3.13 The project team created a project plan and documented the steps taken. The team also created a slide presentation which could be used to communicate the process and give a comparison of the vendors involved showing all of the criteria in side-by-side comparisons. The same groups which were consulted at the project initiation were again presented with results and recommendations. The presentation and approval sessions included a dry run within the team to ensure there was consensus and record notes and details for items that generated discussion. The findings were then presented to the IT Division Head, the Data Services Branch staff, the NASS Enterprise Architecture Council, and to the Senior Executive Team. All of this sought to ensure alignment within our agency planning and with the USDA capital planning and investment direction. Once any lingering questions were answered we sought final approval from the IT Division Head.

3.14 Working through the federal government procurement process was the final step. The team had worked diligently to be sure a contract specialist was engaged months before sending the final recommendation. The contract specialist worked with the team to ensure that all the required forms and steps were followed and completed. Because this person's expertise was in federal contracting and not in IT or data warehousing, there were many opportunities to help in their understanding of terms and clarification of concepts and requirements.

3.15 The size of the procurement in terms of monetary value was also a consideration. In our purchasing system the higher the value, the longer the time period is for notice of the potential contract to vendors, and the more demanding the process for documentation of the salient characteristics of the intended system. Purchasing something costing over one million dollars requires more rigor in documenting the selection process and can also increase the potential for vendor protests. The artifacts developed by the team proved to be invaluable as documentation in the actual contracting process. It was used to demonstrate thoroughness, objectivity, and as a rationale for the recommended purchase.

4. Current State and Next Steps

4.1 NASS procured two appliance systems to be placed in physically separate USDA data centers; one for production and one for disaster recovery (DR). We also procured two development blade servers for testing purposes that will reside in our headquarters location. The purchase was made in September, 2015. The equipment was delivered in early calendar 2016. USDA has been undergoing continuing consolidation of data center resources and deployment of the NASS data warehouse system had to be scheduled between many of these activities. The primary appliance was put in place in March 2016 and the DR appliance was put in place in April 2016. The selection team has turned the project over to the Data Services Branch Staff for setup, population and integration of the units.

4.2 The fundamental plan is to parallel test the new system through the next year's survey cycle, incorporating all of the same data and inputs, leading to creating the same outputs. Where differences are found, further work will be needed to document and change either the new or old system if needed. This is no small task due to the release of over 400 statistical reports and the existence of hundreds of systems/applications involved in the annual process.

4.3 One large challenge is the integration of all of the feeder systems through the Extract, Transform, & Load (ETL) process. There are discussions under way about the best ways to adapt the current procedures to the new products and potential paradigm shifts to an Extract, Load, and Transform (ELT) process. With the speed potential of the new system, it may often make sense to do data transformations within the appliance rather than at the feeder system or within the ETL programming currently in place. This potential speed increase also points toward the efficiency of doing other analytical operations within the database system rather than extracting data with another tool and pulling that over the network to the analyst's workstation. While this may seem like an obvious improvement to systems integrators it has proven to be a persistent problem with agency employees.

4.4 The agency has a public facing web system based in the same data warehouse technology. This 'Quick Stats' system provides the public with the latest estimates. The new data warehouse platform must also be integrated into Quick Stats with little or no interruption to public service.

5. Lessons Learned

5.1 Lessons learned include:

- Do preliminary research on technical possibilities and price ranges
- Think strategically and look to the future
- Secure senior management buy-in early
    - Include budget considerations
- Communicate, communicate, communicate
    - Include entities outside of IT
    - Include your procurement staff early in the process
- Where ever possible use third party and objective sources in your market research
- Plan the work and document as you go
    - Keep the documentation of comparisons as objective as possible
- Test the options as much as possible before committing to any particular solution
    - Document the testing
    - Document why a particular direction was chosen
        - In-house vs. cloud
        - Structured vs. unstructured data
- Consider all of the other systems and data interactions
    - Can they work with any new solution
    - What changes will be needed?
    - Will there be any required paradigm shifts (ETL vs. ELT)

6. Conclusion

6.1 NASS and other statistical organizations are 'data factories'. Our agency mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. It can be difficult at times to remember that we are *not* IT systems integrators. Everything we do should be put in place to support the business mission. Our data warehouse replacement team had to consistently remind ourselves of that as we went through the process. That reminder promoted some of the decisions to move toward a pre-engineered appliance. There were advantages in not only speed of queries and results but in ease of maintenance and reduction of workload in areas where we did not need to focus. When one buys a refrigerator, you do not have to assemble the compressor and charge the Freon unit. It is delivered, plugged in to power, and generally works.

6.2 Communication takes time but proved to be an overall time saver in the end. By including both our business partners and our procurement staff, we promoted understanding, minimized 're-telling' of the story, and negotiated a good business deal for the product and services. We also found that some advance favorable opinion of what was being done had preceded the team to a variety of meetings. When the time came for final approvals, stakeholders reached a consensus fairly quickly because of their involvement in the process.

REFERENCES

Bell, F. (2011). Connectivism: Its place in theory – Informed research and innovation in technology-enabled learning. *International Review of Research in Open and Distance Learning.* Retrieved from http://files.eric.ed.gov/fulltext/EJ920745.pdf

Beyer, M. A., Edjlali, R. (2015). Magic quadrant for data warehouse and data management solutions for analytics. *Gartner.* Retrieved from www.gartner.com

Inmon, W. H. (1993). *Building the data warehouse.* New York, NY: John Wiley & Sons, Inc.

Inmon, W. H., Hackathorn, R. D., (1994). *Using the data warehouse.* New York, NY: John Wiley & Sons, Inc.

Lopes, S. (2015). Massively parallel processing database. *BI4ALL.* Retrieved from http://www.bi4all.pt/node/141

Nealon, Jack and Elvera Gleaton, "Consolidation and Standardization of Survey Operations at a Decentralized Federal Statistical Agency," Journal of Official Statistics Vol. 29, No. 1, 2013, pp. 5–28, DOI: 10.2478/jos-2013-0002.

Yost, Mickey (1999), Historical Data Warehouse 3 Pillars of Progress. Retrieved from http://www.advancedatatools.com/BrioUserGroup/mtg200404/3_pillars2.pdf