
BIG DATA: HANDLING LARGE DATA SOURCES WITH DATABEACON

A DATABEACON.COM WHITE PAPER

Data volumes are growing, partly from the increased monitoring of today's complex business processes, but also from the explosion of e-business and the new performance metrics that need to be collected and understood.

Companies now regularly measure their critical data in terms of megabytes, gigabytes and terabytes. Operational managers and strategic decision-makers need to gain maximum insight from these data assets, and to be competitive, they need to do it in Web time.

EXECUTIVE SUMMARY

Operational managers and their I.T. staff look at Databeacon and immediately grasp the benefits of applying Web Reporting and Data Analysis to business decisions. In particular, they see that the analytic reporting features of Databeacon offer an attractive method to quickly gain insight into the mountains of data that now deluge their organizations.

The purpose of this document is to provide guidance in how to best treat large volumes of data, and to help correctly manage expectations. This document is written for Databeacon Collaboration Edition, and assumes the reader has some understanding of the dbPublish Designer, dbAccess Builder and dbInsight Viewer components, dimensional modeling, and RDBMS structures.

TABLE OF CONTENTS

BIG DATA	1
THE USER EXPERIENCE	2
INFLUENCING SIZE	2
FILTERING	3
INTEGRITY	4
SPARSITY	4
COMPRESSION	4
PREBUILT OR BACKGROUND CUBES	4
CUBE-TO-CUBE	5
MICROSOFT ANALYSIS SERVICES	5
SUMMARY	5

INTRODUCTION

Operational managers and their I.T. staffs look at Databeacon and immediately grasp the benefits of applying Web Reporting and Data Analysis to business decisions. In particular, they see that the analytic reporting features of Databeacon offer an attractive method to quickly gain insight into the mountains of data that now deluge their organizations.

The purpose of this document is to provide guidance in how to best treat large volumes of data, and to help correctly manage expectations. This document is written for Databeacon Collaboration Edition, and assumes the reader has some understanding of the dbPublish Designer, dbAccess Builder and dbInsight Viewer components, dimensional modeling, and RDBMS structures.

BIG DATA

Data volumes are growing, partly from the increased monitoring of today's complex business processes, but also from the explosion of e-business and the new performance metrics that need to be collected and understood. Companies now regularly measure their critical data in terms of megabytes, gigabytes and terabytes. Operational managers and strategic decision-makers need to gain maximum insight from these data assets, and to be competitive, they need to do it in Web time.

Databeacon is well suited to these needs. It delivers the analytical reporting capabilities necessary for operational managers to explore trends, patterns and anomalies in their data. It allows them to interact with personalized data, and do so through an intuitive interface all delivered over the Internet through a standard browser.

To accommodate the need for data delivery in the form of self-serve explorable reports, Databeacon first applies a multidimensional hierarchical structure to the data, and then compresses it into a small-footprint file for delivery, viewing, and ongoing analysis at the operational manager's PC.

However, when the source data gets to be extremely large there is a risk that the

compressed data file, also called an OLAP cube, will get too large to accommodate direct delivery. So yes, you can build an Online Analytical Processing (OLAP) cube from a 100 million record set, but are your users willing to wait for it to appear? The other question that could be asked is, "Did you really need to see all that data at once anyway?"

When working with extremely large data sets, I.T. staff must very often find a balance between the generalized delivery of high-volume data and the personalized delivery of targeted data. They need to review:

- what data is to be delivered
- how it gets delivered, and
- what the user experience will be as operational managers first request then explore the information.

WHAT DATA IS DELIVERED

While data can be sourced from many diverse locations, large data sources are most often RDBMS structured. They are usually the collection of automated data points, OLTP-style transactions, individual billing records, or other atomic level information. Data warehouses can also be a source of big data, again offering detailed information, but further extended over significant time intervals.

These sources provide real value when used with adhoc query tools, managed and parameterized reports, but usually the data is too detailed for traditional OLAP analysis. In these cases, managers and analysts explore higher-level trends and patterns as they try to resolve strategic business issues. The needed data requires summarization, a higher-level abstraction that inevitably results in a much smaller set of records being used.

HOW IT GETS DELIVERED

To give operational managers robust analytical capabilities while maintaining a low total cost of ownership for I.T., Databeacon extracts data and restructures it into highly compressed multidimensional data files called OLAP cubes. The OLAP cubes are then sent on request to the operational manager's PC for the duration of an analysis session. Once delivered, processing

TO ACCOMMODATE THE NEED FOR DATA DELIVERY IN THE FORM OF SELF-SERVE EXPLORABLE REPORTS, DATABEACON FIRST APPLIES A MULTIDIMENSIONAL HIERARCHICAL STRUCTURE TO THE DATA, AND THEN COMPRESSES IT INTO A SMALL-FOOTPRINT FILE FOR DELIVERY, VIEWING, AND ONGOING ANALYSIS AT THE OPERATIONAL MANAGER'S PC

THE BYPRODUCT OF THIS PERSONALIZED DELIVERY APPROACH IS THAT DRAMATICALLY SMALLER AMOUNTS OF DATA ARE BEING SELECTED FOR ANALYSIS, LARGELY INDEPENDENT OF THE ORIGINAL RAW DATA SOURCE SIZE

is limited solely by the speed of the PC, eliminating many performance issues found in typical client-server architectures. However, the cube's initial delivery time is key to user satisfaction, and will be impacted by the compression rate and the bandwidth available.

Databeacon delivers highly compressed cubes. As examples, a 6.5MB 120,000 record healthcare file becomes a 514KB cube; a 133MB 300,000 record financial portfolio file becomes a 1.5MB cube; and a 1 million record sales transaction file becomes a 5.4MB cube. There are a number of factors that impact cube size, and the ratio of compression will vary for each file. Even so, most cubes in a standard office environment are less than 1MB, and take only a few seconds to deliver.

If you are at home, working with a 56Kbps modem, then a 1MB file download might appear troublesome. However, most organizations have some level of high-speed Internet connection, and are easily able to handle file transfers of 5MB, even 10MB. Many MS PowerPoint files and e-mails already fall into this size range and are used without question. Databeacon cubes perform similarly, and as mentioned above, a 5MB cube could easily represent 1,000,000 records or similar SQL found set from some multi-terabyte data store.

THE USER EXPERIENCE

Demands for larger and more complex OLAP cubes are becoming a common occurrence as organizations grow and expand. But so too is audience reach and involvement. Traditional client-server OLAP products attempt to do all of the processing at the server. This works well in that you have a lot of processing power available to handle a big data cube, but this architecture falls short in two important ways:

- 1) the ability to economically serve large volumes of people, and
- 2) the ability to deliver each operational manager a customized or personalized view of the data.

In effect, client-server architectures impose a "cube to rule them all" solution, and then struggle with issues of infrastructure, incremental builds and partitioning as they try to bring efficiencies to the process.

Unfortunately, new and larger audiences of managers, partners, suppliers, consulting collaborators and even citizen stakeholders – all with diverse and individual needs – are only increasing the problems connected to a client-server approach to delivering self-serve explorable reports.

Databeacon applies a more intelligent, efficient approach, leveraging both the power of the Internet and the processing power available at each person's PC. With Databeacon, I.T. staff can create cube designs, called profiles, which can offer a choice of data for inclusion into the cube.

When operational managers select such a profile through a Web interface, they can then select only the data they wish to explore. Databeacon will extract their selection from the data store, create a personalized cube on the fly and deliver it to their PC for near-real-time analysis. In this approach, each person controls the data discovery experience, seeing only data relevant to the job at hand.

The byproduct of this personalized delivery approach is that dramatically smaller amounts of data are being selected for analysis, largely independent of the original raw data source size. Once dimensionalized and compressed, the data is easily delivered over the Internet, maintaining the overall low total cost of ownership for I.T. while ensuring broader user adoption and satisfaction.

INFLUENCING SIZE

When we say "big data", how big is big? In a Databeacon environment, the answer is determined or influenced by a number of factors, including cube size, bandwidth, and an operational manager's patience. While the size of the original source is important, the real answer lies in the size of the cube and the length of time it takes to create it. The following describes some of the factors that help influence the size of data.

GRANULARITY

When building a cube, I.T. staff often feel the need to include data at its highest level of detail granularity. For example, cubes often get built containing date detail to the level of day, hour or minute. However, when viewed and explored, the most valuable information is gained at some higher level of aggregation – year, quarter or month. By limiting the number of hierarchies to the level required by an operational manager, and thus cutting off some of the leaf levels in the data tree, you can significantly reduce the size of the cube, and the length of time needed to create it.

Databeacon Designer provides the capability to summarize the detail records as they are processed into an OLAP cube, but if the source is an RDBMS, the summarization might be best done using the GROUP BY clause in the SQL query. In this case, a 100 million source record file might yield a 5 million row found set, enabling fast processing and a small cube.

DIMENSIONALITY

Be careful to recognize the natural parent-child relationships in the data, and to take advantage of the hierarchical structures that can be defined in Databeacon. Instead of Country, State and City being three separate categories, define a Location category with Country, State and City being hierarchical subcategories. This will reduce the cube size, and your audience will see increased performance, not to mention a better data semantic to work with.

BREADTH

There is a tendency to put as much data into a cube as can possibly fit, but this does little to help out the ultimate operational manager who has to investigate the result. In fact, studies have shown that analysis works best when the number of dimensional columns is limited to seven, plus or minus two. Beyond this number, I.T. staff are encouraged to design alternate cubes, and to employ cube-to-cube capabilities where and if needed.

With this in mind, I.T. can gain further processing efficiencies when the data only contains the information required for the model. I.T. staff can create applications that

first interact with the operational manager to determine what subset of categories are required for a particular analysis session, and then dynamically construct the cube definition accordingly, followed by dynamic cube creation and deployment. Not only will the cube be smaller, but also operational managers will receive exactly the information they need to solve their business-related questions, and a response time that they are personally comfortable with.

FILTERING

Does the operational manager need to see all the data at once, or just a segment of the data? There might be 100 million rows of source data, but if the operational manager needs only to see their product line, by region or by a particular time slice, then the found set of records would be far less, and again a smaller, faster cube would result.

I.T. staff can parameterize their queries as part of a Databeacon integration process. When operational managers select the cube profile, an appropriate data segment will be selected in building the cube. So for example, an application could reference a single cube profile definition serviceable for any hospital administrator, and through parameter substitution, an administrator responding or being identified as from Hemingway Memorial Hospital would see different data than an administrator from other hospitals. A hospital board executive, however, would see the results from all hospitals.

Unlike traditional OLAP tools that first build a large cube and then segment the view, Databeacon provides a new dynamic approach to Web Reporting and Data Analysis, allowing operational managers to quickly iterate on smaller, targeted data volumes, and work with near-real-time data. It is an approach best suited to meet the diverse needs of an expanded business audience.

This approach also works well in applications where “segmentation” is core to the solution. Consider an Electronic Bill Presentment & Payment (EBPP) system as a typical example, where transactional data stores can easily be in

WHEN WE SAY ‘BIG DATA’, HOW BIG IS BIG? IN A DATABEACON ENVIRONMENT, THE ANSWER IS DETERMINED OR INFLUENCED BY A NUMBER OF FACTORS, INCLUDING CUBE SIZE, BANDWIDTH, AND AN OPERATIONAL MANAGER'S PATIENCE

the hundreds of millions of records and occupy terabytes of space. As a system subscriber, a login response could be “Databeacon”. Under program control, the login response would be validated through some security layer, and then substituted for a parameter in the cube build process. The modified query would access the 150,000 or so records attributable to Databeacon, build the OLAP cube and deliver it to the PC of the subscriber – all in a matter of seconds!

INTEGRITY

While not normally a problem, particularly in data warehouse installations, a lack of data integrity and cleanliness can lead to larger cubes and increased processing time. Some care should be taken to review the data, and to determine whether preliminary work needs to be done to assure quality information for distribution. Databeacon Designer offers some capabilities in this area, but in many cases it may be necessary (or desirable) to pre-stage the data or employ Extract, Transform and Load (ETL) technology to improve its value. Previously, this paper describes granularity in the context of a data hierarchy, but in a similar way, granularity can be mixed at a category level. For example, an operational manager might want to explore the top 100 selling products, and have all others grouped as OTHER. The grouping or clustering could be done at the data source, reducing the number of records that go to building the cube.

One element common to many big data stores is in the use of descriptor fields. As companion to the encoded fields that support fast retrieval and indexing, descriptor fields provide a clear identification of a product, person, et cetera, and can extend to 30 characters or more. These fields are invaluable in bringing clarity to most reports, but are most often unnecessary to the operational manager doing advanced OLAP analysis. Using brief descriptors, if any, will help reduce the size of the cube and increase performance.

SPARSITY

Merriam-Webster refers to sparsity as being “of few and scattered items”. When we use the term in referencing source data, we are

referencing the variability and breadth of the data that feeds the cube build process. Is each record reasonably distinct, or is it repeated many times? Sparsity is the single biggest determinant in the size of a cube. Unfortunately, it is not easily measured and is wildly variable.

For example, consider a simple retail transaction file that tracks revenue and quantity sold for 50 products in 5 colors at 8 stores. Independent of the number of records in the file, the total number of combinations that we would need to track is $(50 \times 5 \times 8)$ or 2000. If the source data is dense, then each product is being sold in all stores in almost all colors, and most of the 2000 combinations will be filled. But if only a few of the products are carried in each store, and in only one or two colors, then the data would be sparse as only a fraction of the 2000 combinations are filled. Your source data may contain a million records, but if sparse and only 150 combinations are filled, you will end up with a very small cube.

COMPRESSION

Databeacon employs advanced compression techniques to ensure that the cube is as small as possible, facilitating quick delivery to the operational manager.

STILL TOO BIG

Occasionally, the combination of source data size, cube size, delivery time and an operational manager’s viewing need adds up to a requirement that exceeds the standard Databeacon deployment. In these cases, try the following options.

PRE-BUILT OR BACKGROUND CUBES

When initiating a Databeacon Insight viewer session, the wait time to delivery is composed of three parts: extraction of data, building the cube, and finally, delivery to the PC. For big data, the sum of all three may exceed the operational manager’s expectation or comfort. In these cases, it may be best to run the extraction and build processes in background (or on a scheduled basis) and advise the manager when ready. This means that the operational manager sees a far faster startup to their analytical reporting session.

CUBE-TO-CUBE

Another approach is to divide and conquer. Instead of one large and unmanageable cube, design and build several independent cubes that maintain a common crossover dimension. By grouping similar data into each cube (similar to the concept behind data marts), operational managers can answer most questions in any individual cube, but when necessary, use the cube-to-cube capability of Databeacon to quickly move to an associated cube.

As an example, a large survey company that specializes in global Internet-use metrics employed this technique to manage their big data volume, and to ensure that their subscribers would have little difficulty in accessing the data. They created a single summary cube that provided high-level aggregations across all months, and then a separate detailed cube for each month. Subscribers would first access the summary view, and only when they wanted more detail would they then use cube-to-cube to go to the specific month. The strategy supported the subscriber, but also allowed I.T. to limit the cube creation to two cubes each night – the summary and the current month.

MICROSOFT ANALYSIS SERVICES

A final option is to use Microsoft's Analysis Services and SQL Server to define and build the "cube", and then use Databeacon Insight viewer to see it. While this option allows an operational manager to explore larger data volumes, it is a client-server architecture and will have inherent limits as to user scalability, and overall matched functionality. That said, it allows I.T. staff to deploy a mix of Databeacon and Microsoft Analysis Services analyses, all viewed through a common interface. Extreme data can be viewed by smaller numbers of operational managers through Microsoft Analysis, while most others can enjoy the personalized delivery of targeted information through Databeacon's native Web Reporting and Data Analysis technology.

SUMMARY

To properly address the issues of big data, I.T. needs to review what data is to be delivered, how it gets delivered, and what the individual operational manager will experience as they request then explore the information. Databeacon provides a highly scalable solution to these issues, and is best at the personalized delivery of targeted information to meet individual analytical reporting needs.

**DATABEACON
PROVIDES A
HIGHLY SCALABLE
SOLUTION TO
THESE ISSUES,
AND IS BEST AT
THE PERSONALIZED
DELIVERY OF
TARGETED
INFORMATION TO
MEET INDIVIDUAL
ANALYTICAL
REPORTING
NEEDS**