

A Technical Overview of *DataPulse* Technology

By Jim Luisi

Scope and Purpose

The goal of this paper is to provide a technical description of the *DataPulse* database technology, which presently supports, a data warehousing and a direct marketing application, with a data mining application to be released in the near term. As an entirely new DBMS paradigm that specializes in decision support capabilities involving many tera-bytes of data, we will discuss some of the concepts that comprise the advanced architecture of the database, and the magnitude of its benefits, both technical and business. Although targeted for Fortune 100 companies, this technology can accommodate the pocketbook of Fortune 2000 companies, supporting multi-lingual capabilities for data types typical of relational database management systems. It positions companies to perform all categories of analysis on all categories of production data, regardless of quantity.

The background for this includes examination of the following elements:

- *DataPulse* Architecture – What Makes It Unique
- Application of *DataPulse* Technology
- Installation Process of the *DataPulse* Technology
- Business Advantages of the *DataPulse* Technology

***DataPulse* Architecture**

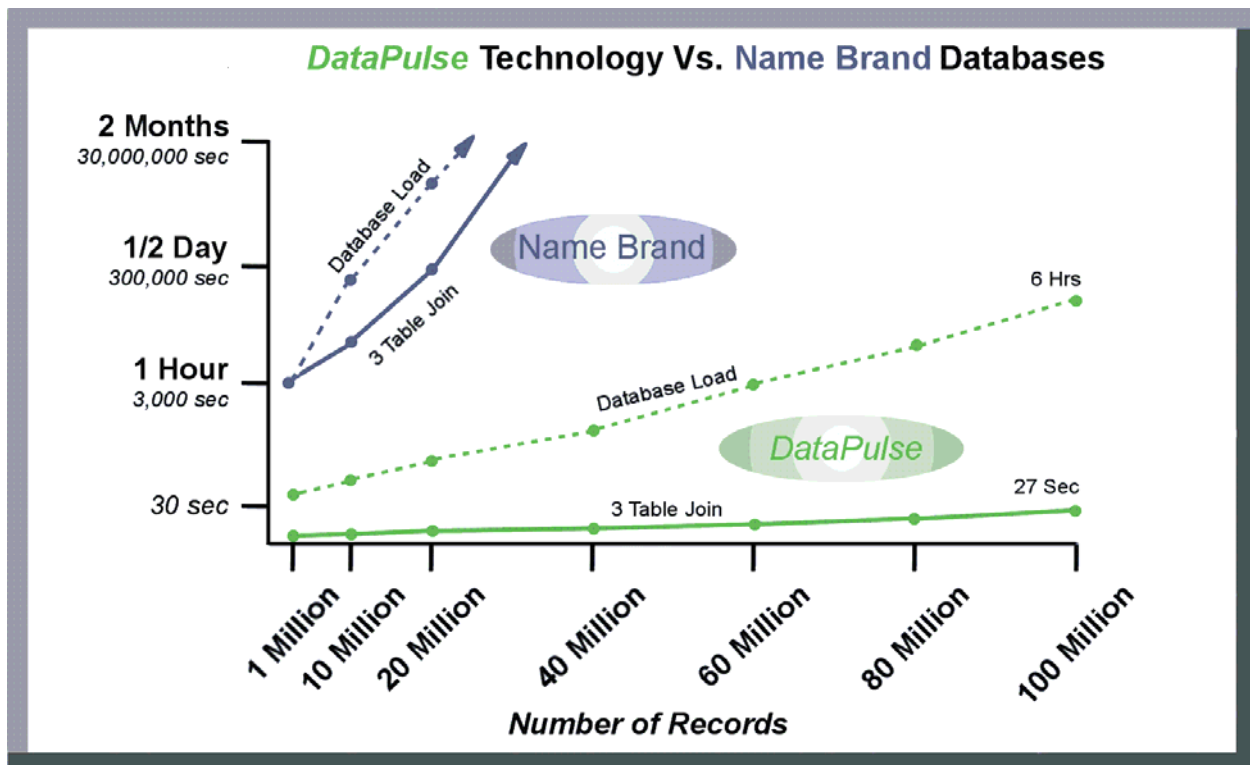
What Makes It Unique –

First of all, unlike most new technologies, one should be pleased to note that the developers of this new database paradigm did not create a new set of terminologies. Unlike the discipline of neural network technology which refers to the act of populating a record with data, 'semantic case frame instantiation', *DataPulse* uses all of the standard terminologies employed by relational database technologies. As for its components, users and database administrators alike, will readily recognize the basic concepts of databases, tables, attributes, and prime keys, as well as more advanced features, such as mirrored databases and hot backups.

As for its key discriminator, *DataPulse* is remarkably faster than other database architectures, and hence, is capable of remarkable functionality. The magnitude of its difference in performance, however, is multi-faceted, as performance relates to the speed of loading data, the retrieval of data, and data modification. Additionally, within the aspect of data retrieval, there is the performance of retrieving wide versus narrow tables, as well as, simple versus complex joins.

With respect to joins alone, while the most powerful mainframe database technologies are significantly challenged when performing two and three table joins involving one million records per table, the *DataPulse* technology is not appreciably challenged when performing thirty table joins involving hundreds of millions of records per table.

The chart below describes the difference in performance for loading and retrieval of data. Note that the 'y-axis' represents an exponential increase in factors of ten.



The numbers represented in this chart are based upon tables of similar width and characteristics. The **Name Brand** product was operating on a typical multi-processor mainframe platform and the **DataPulse** product was operating on a standard NT server 4.0 environment with a single processor Pentium 400, without the use of RAID or fiber SAN solutions.

DataPulse Architecture –

One of the key architectural differences of the **DataPulse** database technology is that it relies upon a completely different set of database internals. The way that data is organized and stored in **Name Brand** database products has not changed from the inception of database technology, over twenty years ago. In summary, data is stored on pages, usually 4K in length, with each page having a header and a footer record. In the old paradigm, the data belonging to a table is physically stored as a group of attributes that belong to an occurrence of a specific row. Depending upon the width of the tables and the number of tables that are stored within the particular address space, the number of times that a given attribute may be stored within a 4K page may vary significantly. Hence, in order to retrieve a handful of occurrences of a particular attribute, it may be necessary to retrieve a large number of 4K pages.

In **DataPulse**, on the other hand, even the bits that comprise a given attribute need not be stored contiguously on a page.

Another important distinction of the **DataPulse** database technology is that it relies upon an advanced form of data compression, achieving a higher compression ratio. Unlike the older technologies, which compress only the non-key attributes, the **DataPulse** technology compresses key data as well, and can perform queries and arithmetic expressions without decompressing the data.

While some of the more proprietary features of the technology involve the way that various types of joins are supported, the end result is a product with the look and feel of an RDBMS, with performance commensurate with the new breeds of technological advancements.

Application of **DataPulse** Technology –

Although the performance profile of the **DataPulse** technology for use in transactional systems is comparable to products that have been specifically designed to support transaction processing, its ability to support decision support system processes is remarkably superior. As a result, the **DataPulse** technology is particularly adept at supporting data warehousing applications, quite easily supporting complex queries involving unlimited quantities of information.

Additionally, with the I/O power of the **DataPulse** technology combined with the calculation power of the newest Intel processors, the ability to support statistical analysis on large quantities of information in a VLDB data warehouse is significantly greater than expensive mainframe platforms.

Perhaps the most significant application developed thus far, however, is **DataPulse's** ability to support multi-dimensional OLAP processing, without the aid of summarized data. This new paradigm makes all forms of summarized data, such as star schemas and cubes, typically used in products such as Essbase, Cognos, and Microsoft's OLAP Services 7.0, obsolete.

Multi-dimensional Analysis Performance Benchmark		9/2000
MicroSoft's OLAP Services 7.0		DataPulse 8.0
Steps:	Beginning with a Data Warehouse <ol style="list-style-type: none"> 1. Predict the user's business questions 2. Potentially create a DataMart 3. Design the cubes to support the questions 4. Predict usage to determine the aggregation level 5. Calculate summary data to populate the cubes 6. Formulate the syntax to support the query 	Beginning with a Data Warehouse <div style="border: 1px solid blue; padding: 5px; width: fit-content;"> <i>Powerful enough to act upon all of the detail data, not requiring DataMarts or precalculated summary data.</i> </div> <ol style="list-style-type: none"> 1. Point & click to build any business question
Hardware:	(2) Servers w/ four 450 MHz Pentium III Xeon Processors 8,192MB RAM (8 GB) 512MB Hard Drive Cache Fibre-channel, RAID 10 With five mirrored sets 100 MB Ethernet Network	(1) Server w/ one 400 MHz Pentium II Processor 256MB RAM (1/8 GB) No Hard Drive cache 7200RPM IDE Hard Drive 10 MB Ethernet Network
Database:	One Table of 13 MM Records <div style="border: 1px solid blue; padding: 2px; width: fit-content;"> <i>Only One Table - No Joins Required</i> </div>	Household Table 200,000 Records Transaction Table 13 MM Records Product Table 14 Records
Performance:	Hours of Data Preparation	A moment to get comfortable in the chair
Queries of:		
One Dimension w/ 14 Permutations -	Take up to 11 minutes	Take up to 18 seconds
Two Dimensions w/ 4 Permutations -	Take up to 5 minutes	Take up to 16 seconds
<div style="border: 1px solid blue; padding: 5px; margin: 10px auto; width: 80%;"> <i>Important Note - DataPulse technology remains real-time in situations with significantly more joins, records, dimensions, and aggregates.</i> </div>		
(MS OLAP Services 7.0 statistics from SQL Server Magazine, 09/00, Optimizing Cube Performance, Pgs. 53-58, and http://www.unisys.com/sq17/)		

It should be noted that the OLAP Services 7.0 benchmark employed a special version of the NT operating system from Microsoft to support 8 Gigabytes of RAM. Additionally, the approximate hardware cost for the OLAP Services 7.0 hardware configuration would have been one million dollars, whereas the significantly faster **DataPulse** OLAP product was operating on hardware priced below \$3,000 using a standard version of NT server 4.0.

Based upon the advantages that the **DataPulse** technology offers multi-dimensional analysis, it is reasonable to expect comparable advantages in the upcoming Data Mining and Forecasting product set, based upon the same technology.

Installation Process of the *DataPulse* Technology –

Installation of the *DataPulse* Technology is quick and easy, using a phased implementation approach that provides rapid availability to end-users. The *DataPulse* load utilities are highly advanced, allowing users to identify the source and integrity of every attribute. The *DataPulse* data administrator may designate a field as having multiple sources, each with its own rules.

For example, the administrator may designate customer birth date from Production System “A” as highly reliable, designating that the value of that attribute should always be accepted, thereby overlaying the contents of the data warehouse. Similarly, the administrator may designate customer birth date from Production System “B” as somewhat reliable, thereby accepting the value only when the data warehouse does not have a known value. Finally, the administrator may also designate customer birth date from Production System “C” as highly unreliable, thereby refusing to accept the value under any circumstances.

Installation may require an iterative analysis process to accurately identify the information to be included in the data warehouse. While the analysis process itself can be time consuming, implementing the results is typically quick and easy. The initial installation process can usually be performed within six to eight weeks.

Business Advantages of the *DataPulse* Technology –

Although the *DataPulse* technology is extremely powerful, special care has been put into the design of the user applications to make them extremely easy to use. The entire Data Warehouse is exposed to the user in a tree structure referred to as the *Available Attribute Window*. The administrator can organize and name the tables and attributes displayed to suit the business community. Similarly, the user can construct extremely complex queries using a graphical representation referred to as a *Query Picture*. *Query Pictures* eliminate the need for users to develop programming skills like SQL. *Query Pictures* support all Boolean operations, as well as, embedded arithmetic expressions in order to achieve any level of sophistication.

Aside from its ease of use, the primary business advantages of the technology stem from its ability to analyze significantly larger volumes of data in real-time.

For example, since the product can easily support the volumes associated with time series data, a user engaged in direct marketing can receive monthly demographic overlays of consumer or business prospects, thereby allowing them to focus their marketing to just those prospects with an important new characteristic. This permits companies to significantly reduce their direct marketing costs by limiting contact to the few extremely valuable prospects, several months before their competition.

Summary of the *DataPulse* Technology –

Just as the Measure of Work in physics is not determined by the force that is applied, similarly, obsolete database technologies push hardware to its limits while only producing small results. *DataPulse's* ability to achieve new levels of analytics can provide insight into data that was previously promised by other technologies, but is now possible to obtain.