



## The Speed of ParAccel's Data Warehousing Solution Changes the Economics of Business Insight

Analyst: Anne MacFarland

### Management Summary

Economic constraints, paired with innovation imperatives, can increase the costs of being wrong for any business. Even the pursuit of a sub-optimal strategy, when acquiring companies or addressing new markets, may be not just a waste of time and money but, more importantly, also a more critical waste of opportunity. As a result, the discipline of analysis has greatly expanded its domain.

There is plenty of information to analyze. As their processes and even their surveillance goes digital, businesses are drowning in information. There are also increasingly diverse but relevant external sources of information available that were inconceivable a few years ago. The rub, for most organizations, is their heterogeneity. For all we hear about smarter, more effective large systems, one must remember that most of them were designed as a system. By contrast, a business' information landscape, while it may have well-designed systems in it, also may have many quagmires of stale data and incomplete data, as well as some gleaned but undocumented information. If the business has grown by acquisition, it may have many disparate versions of a customer, as well as disparate systems trying to support the many relationships.

Businesses now seek to leverage their capabilities in ways that in the past might have been considered unlikely – think of today's routes to market and product placements such as compute elements in automobiles. This opportunism makes the queries that are needed to do business well more complex. Where operational strategies, in the past, might have been analyzed with the information in the companies' back end database, these days it is often beneficial if more information can be brought to bear. When addressing new market opportunities, the size of the market (often glibly mandated in round numbers) is not as important as the details of its character, and the specific consequences that character imposes on the sales and marketing activities that are needed to address it. Complex scenarios must be analyzed to support a timely go or no-go decision.

At a more prosaic level, analysis of the operational detail enabled by digitization can reveal unanticipated bottlenecks or redundancies whose mitigation might save time and money. Any process, when examined in detail, becomes far more complex than just its inputs and outputs. Web operations involve still more variables. To meet the needs of both familiar and first time patrons, you need to understand more about how your processes intersect with their expectations.

Where, in the past, you may have been able to depend on experts, these days an expert system may be both an accelerant and a permanent resource. ParAccel, located in Cupertino, has developed a data warehousing solution and ancillary tools focused at the need for both speed and effectiveness in business analysis. Its columnar, massively parallel, share-nothing, schema neutral, scalable environment can support complex analytics in all their considerable ugliness. For more on what this all means to business, please read on.

### IN THIS ISSUE

➤ The Changing Nature of Decision Support .....	2
➤ Six Important ParAccel Analytic Database Characteristics .....	2
➤ Conclusion .....	4

## The Changing Nature of Decision Support

The manipulation of information has spread over time from spreadsheets to database tables to cubes using well-understood disciplines that were well supported by technology, as long as the goal was timely and accurate traditional reporting, i.e., retrospective reporting. Using these traditions to support tactical operations and the what-if scenarios needed for prompt decision-making adds some new requirements.

Analytics moves beyond the arithmetic *how many* of financial reporting – though results still tend to come in numbers. Data set operations such as joins, while mathematical, are not strictly computational. Computation has simple rules and computers produce seemingly instantaneous results. Data set operations, if badly designed, can take hours or days. They are more like the set theory you learned in school, except that the business rules involved and particularities of the field and data definitions mean that combining data without losing quality is far less straightforward than overlapping the mathematical abstracts of *Set A* and *Set B*.

More data sources, instead of being a pain, become critical success factors in making better decisions. Information from partners and other external sources is usually highly relevant. This begs the question of data quality. The old rubric was that data quality plummeted as more sources of data were added. This is still true for financial data, but can be less important in operational analysis as long as the queries are designed well. Adapters and other forms of just-in-time data transformation support operational analysis in many situations.<sup>1</sup>

The business decisions that organizations are seeking to clarify are often innovative in nature, and do not come with a roadmap. There are no templates of expected results (like last year's financials) to indicate when something is dreadfully unlikely – though a query taking forever to run may indicate something is amiss.

When looking to optimize analysis with

technology, performance and ease of ownership are the two major dimensions of the challenge. Throwing a lot of memory at the situation adds costs as it adds speed. Using indexes and other intermediary forms of information can speed up specific analytic operations – but they also take up space and add ongoing overhead (for updating, managing, etc.).

The following characteristics of ParAccel are highly relevant to business operations, and must be understood in business terms as well as technical terms if the full value of this kind of business tool is to be realized. (See also Exhibit 1, on the next page.)

## Six Important ParAccel Analytic Database (PADB) Characteristics

### 1. Architecture

Architecture for analytics tends to be massively *scale-out*. With the PADB, there is a *leader node* and then a mass of *scale-out nodes*. This is more “horizontal” than the cellular architectures of archiving, because the focus is not just on fast retrieval but on the data manipulations, joins, and intermediate states that underlie presentation of precise results, drawn from lots of data, in a useful way.

The nature of parallel database operations is that, in doing joins,<sup>2</sup> intermediate results move. This involves many-to-many communications – a process that often can be gated by physical links and system interrupts. ParAccel uses a custom protocol because TCP/IP is too prone to packet loss. The custom approach is not a security feature. However, the interconnection is private to the cluster of blades, and no other computers see the traffic.

This architecture may sound like a map-reduce or Hadoop scenario – the algorithm popularized by Google – but, in fact, ParAccel uses other more robust algorithms. Hadoop is more presentation oriented than other algorithms that have been developed over decades.

### 2. Column-Based Physical Storage Strategy

To get the fast retrieval, ParAccel uses column-based approach to storage. The range of values in a column is narrower than in rows, so physically storing information is more manageable and more compressible than that stored

---

<sup>1</sup> The values used for financial reporting have higher standards of consistency, supported by practices including master data management and the use of XBRL tagging. Master Data Management (MDM) is an approach to data quality that is front of mind for many organizations, but it takes considerable effort to accomplish and keep current. As a data warehousing provider, ParAccel is a supporting player in the realm of Master Data Management but does not do it and has no immediate plans to go in that direction. Instead, it focuses on leveraging data as-is.

---

<sup>2</sup> A join is a way that two tables of data that share a common element can be aggregated. There are many kinds of joins, including self joins that can be done.

## Exhibit 1 — The ParAccel Analytic Database (PADB)

### Architecture

- A minimum of two computer blades /servers (clustered for high availability) are to be connected by Ethernet using multiple independent subnets. One will be the master to which the others are attached. A custom protocol is used between the other analytic blades for greater efficiency than TCP/IP provides.
- ParAccel supports both bulk and trickle down data loading – both parallelized. The incremental approach gives a more atomic time stamp.
- The latest release of the PADB supports encryption.
- ParAccel reference architecture for its Scalable Analytic Appliance, supporting the optional blended scan, is based on *Clariion CX4 240/280*, with its *FLARE* operating Environment and *Navisphere* management components (*SnapView*, *MirrorView* (Asynchronous and Synchronous), *SANCopy* and *NQS* (Navisphere Quality of Service Manager). ParAccel has models for 2, 4, 6, 8, and 10 TB databases.

### Targeted Opportunities

- Sophisticated Data Transformation
- Merchandising and advertising
- Internet traffic analysis
- Credit Card Fraud Detection
- Customer insight

### Pricing

- List pricing is \$75,000 per TB of searchable data. ParAccel also offers a subscription license starting as low as \$5,000/month for smaller operations. At present, it is offering a *faster or free* guarantee.

### Partnerships

- EMC ParAccel is a Velocity Partner of EMC and is available through EMC. It also has OEM agreements with companies such as Arrow, Autometrics, and Fusion-io, and SI partnerships with companies such as Baseline Computing, DB Architects, Solution Forge, and White Oak Technologies.

### Pending Certifications

- SAP, including Business Objects
- IBM, including Cognos
- MicroStrategy, Inc.
- Information Builders

Source: ParAccel

as rows. A columnar orientation is helpful when more columns are to be appended (in outer joins), or when data is sparse (with null values<sup>3</sup>) ParAccel does not use column orientation at a logical level, nor does it use tokenization<sup>4</sup>.

For business, the ability to optimize data placement on disk for expected use pays back in

<sup>3</sup> Tables with NULL values can be problematic because NULL values, by their nature, match nothing. If the join conditions are not explicit as to their treatment, a Cartesian product operation may greatly expand the size of the data set.

<sup>4</sup> This contrasts with SAND Technology. (See [The Clipper Group Navigator](http://www.clipper.com/research/TCG2006072.pdf) entitled *SAND/DNA Gives Tools to Temper the Hugeness of Data*, dated August 16, 2006, and available at <http://www.clipper.com/research/TCG2006072.pdf>.)

both data center operations and telecommunications costs. These cost savings can be significant.

### 3. Schema Neutrality

To be *join focused*, it is useful to be *schema neutral*. Since ParAccel compiles queries, it recognizes the relationships documented by the schema, but compiles to best accommodate the characteristics of the data and the nature of the joins. Some people have moved to star schemas or de-normalization to avoid joins. The former slows data loading and the latter adds copies of data, which increases the overhead. With ParAccel, you do not have to change your schemas to get good performance. This simplifies the

analytic process and lets customers work with more data sources with less preparation work.

#### 4. *Locking Avoidance*

Locking is a process that avoids data corruption, but it also limits the performance of many queries on the same data. ParAccel avoids locking by internal versioning of the data at a table level, using a transaction ID that is invisible to other ongoing queries. “Think of it this way,” advises CTO Barry Zane. “Each row is associated with a transaction. Using snapshot isolation, ParAccel ensures that each query sees the data as it was when the transaction occurred.”

As an example, in tracking the cadence of commercial operations, time-of-day comparisons across channels are important. However, it is possible either to micro batch commits at a cadence – say every 30 seconds – or to leave the load stream live and use an automatic commit. It all depends on the nature of the environment being analyzed.

ParAccel can support full transactional integrity (ACID)<sup>5</sup>, where it is required. It can also specify a commit cadence in the load logic, where full transactional integrity documentation is less important.

#### 5. *ParAccel’s OMNE Optimizer*

Optimization is needed to get results in the timeframes demanded by contemporary operations. ParAccel has included a *Postgres* parser/finder/optimizer, which is adequate for many kinds of analytic operations. Recently, ParAccel introduced its *OMNE Optimizer*. This is a patented columnar execute compiler built from scratch by ParAccel. Omne Optimizer can handle many-to-many table joins across multiple databases, varied process queries, correlated sub queries, and other abstruse analytic elements. This further disencumbers the limitations of analytics operations.

#### 6. *Blended Scan*

In ParAccel’s optional, patent-pending *Blended Scan* paradigm<sup>6</sup>, the RAID mirror of data is not in an appliance but in a SAN (Storage Area Network). This allows backup and other administrative operations to be done unobtru-

sively against the remote mirror copy at a convenient time

In blended scan, ParAccel distributes queries across both the SAN and local disks to optimize performance. Together, local (internal disk) and SAN comprise what ParAccel calls a *distributed mirrored pair*. This lets the data center leverage more of its infrastructure in its analytic operations. Where needed, only local disk and/or in-memory information can be used. However, for high availability, the option leveraging a SAN-based “distant pair” of replicated data is very useful.

#### Conclusion

As you can see, 21<sup>st</sup> Century use of business information is a brave new world – but it also has some enduring characteristics. Latency is still a factor. The bulk of data will still be expensive to keep and expensive to move. Time to results will continue to be a challenge. The need for optimization of queries will only increase as businesses seek to innovate.

Petascale computing is in the future of most organizations. Like 20<sup>th</sup> Century labor-saving devices, analysis of operational and market detail can expose the character and drive the insights that let a business avoid risk – or accept it on reasonable terms. Analysis lets managers see patterns and unusual behavior that lets them anticipate trends and avert crises. However, analysis is only helpful if it adequately reveals what you want to know. A “surround” of corroborative detail will not makeup for shortcomings of information about the crux of the problem.

The base requirements for this new kind of high value analytics operations include simplicity and speed. ParAccel provides both, and lets its customers embrace the future with new paradigms of data leverage. Consider your constraints and requirements. If scope-of-inquiry and time-to-usability is high on your list of critical success factors, consider ParAccel.



<sup>5</sup> ACID properties are Atomicity, Consistency, Isolation, and Durability. They are essential in transactional environments, but may be less necessary in some kinds of analytic use of data.

<sup>6</sup> When PADB is running in a cluster without SAN, the blended scan does not apply and the RAID 1 mirror is of DAS (internal storage).

### **About The Clipper Group, Inc.**

**The Clipper Group, Inc.**, is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- **The Clipper Group can be reached at 781-235-0085 and found on the web at [www.clipper.com](http://www.clipper.com).**

### **About the Author**

**Anne MacFarland is a Senior Contributing Analyst for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group in 2001 after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- **Reach Anne MacFarland via e-mail at [Anne.MacFarland@clipper.com](mailto:Anne.MacFarland@clipper.com) or at 781-235-0085 Ext. 128. (Please dial “128” when you hear the automated attendant.)**

### **Regarding Trademarks and Service Marks**

**The Clipper Group Navigator, The Clipper Group Explorer, The Clipper Group Observer, The Clipper Group Captain's Log, The Clipper Group Voyager, Clipper Notes,** and “*clipper.com*” are trademarks of The Clipper Group, Inc., and the clipper ship drawings, “*Navigating Information Technology Horizons*”, and “*teraproductivity*” are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

### **Disclosures**

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

After publication of a bulletin on *clipper.com*, The Clipper Group offers all vendors and users the opportunity to license its publications for a fee, since linking to Clipper's web pages, posting of Clipper documents on other's websites, and printing of hard-copy reprints is not allowed without payment of related fee(s). Less than half of our publications are licensed in this way. In addition, analysts regularly receive briefings from many vendors. Occasionally, Clipper analysts' travel and/or lodging expenses and/or conference fees have been subsidized by a vendor, in order to participate in briefings. The Clipper Group does not charge any professional fees to participate in these information-gathering events. In addition, some vendors sometime provide binders, USB drives containing presentations, and other conference-related paraphernalia to Clipper's analysts.

### **Regarding the Information in this Issue**

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.