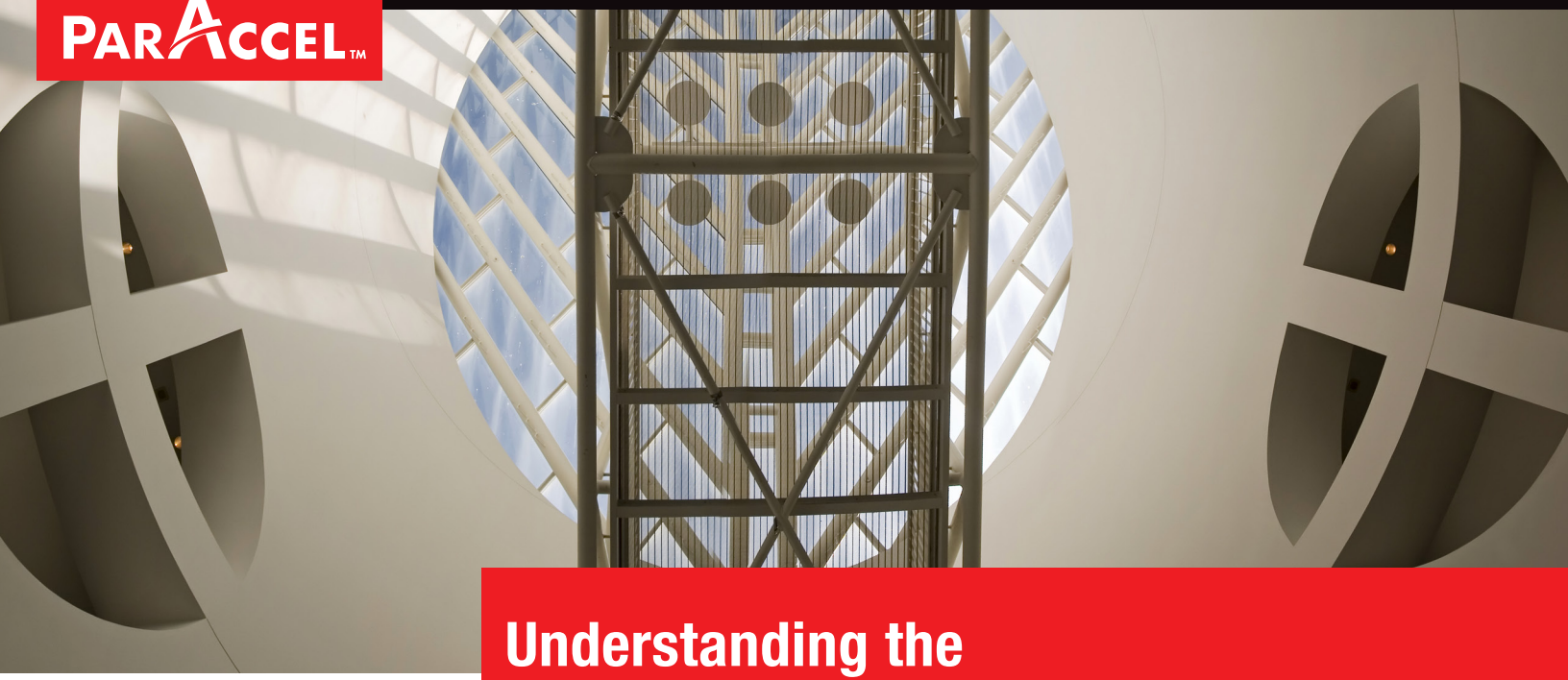
WHITEPAPER



Understanding the New Cooperative Analytic Processing Architectures

Driving Value from Big Data with Analytic Platforms and the New Cooperative Analytic Processing Architectures

The desire for new big data analytics capabilities is driving a wave of expert recommendations for cooperative analytic processing architectures, where processing is shared between several integrated systems. Experts have also identified a new breed of system, the analytic platform, which coordinates analytic processing across these next-gen architectures. In this paper, learn more about experts' new architectural recommendations, read criteria for analytic platforms – and find out why ParAccel's analytic platform is the leading example of this next wave of analytic systems.

ParAccel™

Introducing the New Cooperative Analytic Processing Architectures

For more information, visit ParAccel.com/technology or contact us at 866.903.0335.

The need for a new generation of analytics is well-established in the business community today, however the ability to technically deliver these capabilities is lacking in many organizations. There are compelling reasons to become an analytics-driven enterprise, though: Those that can effectively apply analytics to create competitive advantage within their markets or industries are "more than twice as likely to substantially outperform their peers," according to a 2011 MIT *Sloan Management Review* study[1]. Yet, many organizations focus primarily on collecting, storing and managing data, without the analytic infrastructure to extract real value from it. The big data trend is no exception. Many technologies are focused on collecting and processing data, only to fall short when it comes to analyzing that data to get any real value. There are no rewards for the highest data volumes. Analytics is the true difference-maker, giving businesses the intelligent edge they need to accelerate, innovate and compete.

But, getting value from data analytics is a difficult proposition today. Rather than a few controlled data sources, companies must now contend with an entire datasphere that extends beyond their own firewall to the cloud, partners, providers, machine sensors and beyond. Many are intrigued by the potential of big data, but this class of data by definition has greater volume, velocity and variety than the typical corporate data that most data warehousing architectures were designed to handle. Compounding this challenge are new competitive requirements for complex analytics on large data sets and new needs for integrated analytics across data sources. Then, there are demands from data scientists and business analysts for sophisticated ad-hoc capabilities that will help them to do analysis with no technical or data limitations.

The vision is ***unconstrained analytics***. Too often, analysts can't run all the queries they want to, because they can't access all the data they need or processing takes too long or they can't get "time on the system" for ad-hoc activities. Too often, innovative ideas from analysts or business users fall by the wayside due to technical limitations – and this holds companies back from effectively using analytics to meet their goals. But the IT challenge is significant: What is the best way to deliver all of these analytical capabilities to the business so they have truly unconstrained analytics?

As analytic innovators can attest, there is no single answer. It is unrealistic to expect a single system to handle the new requirements of unconstrained analytics. The needs of today's complex analytics are simply too varied to be effectively handled by a single engine, according to experts and early adopters. Monolithic data warehouses are already straining to deliver the required performance and agility for business intelligence (BI) and reporting needs, while new options like Hadoop don't have the analytic capabilities (such as SQL support), the performance or the robustness to be the primary enterprise analytic system. These systems will likely continue to play a role in many analytic architectures going forward, but more is needed.

> Cooperative Analytic Processing Architectures leverage the analytic capabilities of different systems to process complex queries quickly.

Experts and practitioners agree: the only way to adequately address the wide variety of new analytics requirements is with tightly-integrated systems, each doing what they do best. This concept is driving a wave of specific expert recommendations for next-generation architectures to support the new requirements of big data and other advanced analytics. Experts have slightly different takes on exactly how these architectures should look (more below), but all are examples of ***cooperative analytic processing architectures***, leveraging the analytic capabilities of different systems to process complex queries quickly and efficiently, minimizing data movement.

Experts have also identified the need for a new breed of system, often referred to as an ***analytic platform*** or ***analytic relational database management system [analytical RDBMS]***. These technologies orchestrate this cooperative analytic processing, reaching across the datasphere to efficiently process complex queries. Analytic platforms help organizations mask the complexity of these new architectures to business users, providing a single interface for unconstrained analytics across an integrated ecosystem.

[1] Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., Haydock, Dr. M., Analytics, the Widening Divide. (Boston: MIT Sloan Management Review and the IBM Institute for Business Value, 2011.)

Understanding the New Cooperative Analytic Processing Architectures

Many experts have published their own architectural recommendations and criteria for cooperative analytic processing architectures. Noted visionary Colin White published *A Next-Generation Analytics Ecosystem for Big Data*[2], including technical criteria for an analytical RDBMS, aka an analytic platform. White's approach focuses on how companies can combine analytic platforms, existing enterprise data warehouses (EDWs) and data management components to meet the new requirements of big data analytics.

Gartner Inc. proposes its Logical Data Warehouse model[3], a reference framework that combines various analytical systems and repositories with management and integration layers to deliver an optimized platform for BI and analytics. Enterprise Management Associates recommends its Hybrid Data Ecosystem[4], combining new analytical components (including Hadoop/NoSQL databases) and legacy systems into an integrated data architecture. Expert and author Wayne Eckerson has put forth the BI Delivery Framework 2020[5], addressing the need not only for delivering business intelligence (BI) and analytic intelligence, but also continuous intelligence from event-based systems and content intelligence from unstructured data. The 451 Group promotes its concept of Total Data Management[6] and an associated architecture, which accounts for BI, business analytics, big data, Hadoop, other NoSQL systems and the emergence of "private data clouds." Even more recommendations and best practices will likely emerge as more organizations take on this game-changing challenge and move toward unconstrained analytics.

While each of these recommendations has its own nuances, they have core philosophies in common:

- **Analytic workloads must migrate to most efficient platform.** A single system that "does it all" with no constraints simply does not exist for today's increasingly sophisticated analytic requirements. Experts agree that different analytic workloads should be processed on the platform that gives best performance and price-performance metrics.
- **The next generation of analytics requires a cooperative, tightly-integrated architecture, not a single monolithic system.** Not only is it unrealistic for one system to process the variety of analytic workloads required today, a single machine is also a single point of failure for analytics, which are an increasingly mission-critical endeavor. Instead, analytic architectures should include high-performance integration between analytic platforms, legacy data warehouses, other analytic systems, integration and data quality platforms, master data management, and other databases. Then, data and analytic processing can be easily shared as needed, ideally with the help of an intelligent analytics platform to orchestrate cooperative analytics efficiently.
- **Data movement must be minimized and optimized, due to growing data volumes.** It is more important than ever to bring analytical processing to the data whenever possible, instead of moving data over already-stretched networks. This is a concept Gartner calls 'distributed processing,' and other experts endorse similar models.
- **The new architectures must allow for IT agility, and have appropriate management controls.** For accurate analytics on large volumes of data, and in order to provide unconstrained analytics enterprise-wide, IT and data teams must be able to set and maintain SLAs – and must develop flexible architectures that can change with shifting business priorities.
- **Seamless access for business users and analysts is critical.** The new architectures must provide a single interface to access to all data and analytical capabilities, regardless of physical locations and delivery platforms. Business users should not have to be concerned with the specific data source or platforms – they should be able to focus just on analysis.
- **Cost-effective components are essential to success.** Using "the right tool for the right job" means using the system with the best performance at the lowest cost, without compromising on requirements. With the right analytic platform and integrations, this can easily be achieved while still providing a single interface to business users.
- **ROI, value and new business capabilities must be created.** These new architectures must be focused on delivering measureable value from analytics, empowering analysts with access to more data and high-performance tools for unconstrained analytics. IT must arm the business with new analytic capabilities that can be used immediately to accelerate innovation and competitive advantage.

[2] White, C., *A Next-Generation Analytics Ecosystem for Big Data*, BI Research, 2012.

[3] Beyer, M., Edijlali, R., *Understanding the Logical Data Warehouse: An Emerging Practice*, Gartner, Inc. 2012.

[4] Rogers, S., *The Rise of the Analytic Platform in Big Data, Enterprise Management Associates*, 2012.

[5] Eckerson, W., *Analytic Architectures: Approaches to Supporting Analytics Users and Workloads: BI Delivery Framework 2020*, BeyeNetwork, 2012.

[6] Aslett, M., *Total Data: Data Management Approaches in an Era of 'Big Data'*, 451 Research, 2012.

Challenges of Designing Analytic Architectures

These ideas make sense on paper, but there are many challenges inherent to delivering on this vision, as Colin White points out in his whitepaper. First, educating the business side on the possibilities afforded by a new analytical architecture, finding the right use cases and building a business case that will deliver ROI and value may be surprisingly difficult for some organizations.

Further, as evidenced by the variations in expert recommendations, many IT and data teams will be challenged to accurately understand, design and select the right components for their own architectures. While many companies want to be analytics-driven enterprises, they often don't realize the complexity or up front investment involved. Analytic advocates may need to get creative, discovering and forging new organizational partnerships to develop the right business case.

For those pursuing Hadoop or other new non-relational systems, simply dealing with the programming requirements and relative immaturity of these systems will present a new learning curve. Since SQL, the lingua franca of analytics, cannot be used directly on Hadoop data, this adds another potential layer of complexity to the equation. Related, with growing data volumes, and more data sources, many organizations will face another hurdle in determining the level of data and metadata integration required – and in figuring out the best way to minimize data movement across already stretched networks. Then, there are data governance and data quality management issues to consider, especially when it comes to big data and other previously ungoverned data sources.

And, determining how to deploy functionality will clearly be a big decision –considering organizational readiness for the cloud (or not), existing data center standards and available resources for technical implementation and support. Organizations will need to carefully consider the technical capabilities of new analytic components, to choose what will best complement their existing environment.

The final challenge, White says, are the skills required to make analytics a successful venture. He sees "lack of skills for enabling data science and investigative computing requirements" as a significant barrier to success. Organizations will need to choose where to invest in training and development, which may mean focusing on analytical skills over pure technical skill. For many, this means hiring new resources or developing the skills to use analytical platforms effectively rather than support them. That's why experts recommend choosing a platform with minimal administration and maintenance requirements – so analysts can focus on higher-value activities.

Experts recommend choosing a platform with minimal administration and maintenance requirements so analysts can focus on higher-value activities.

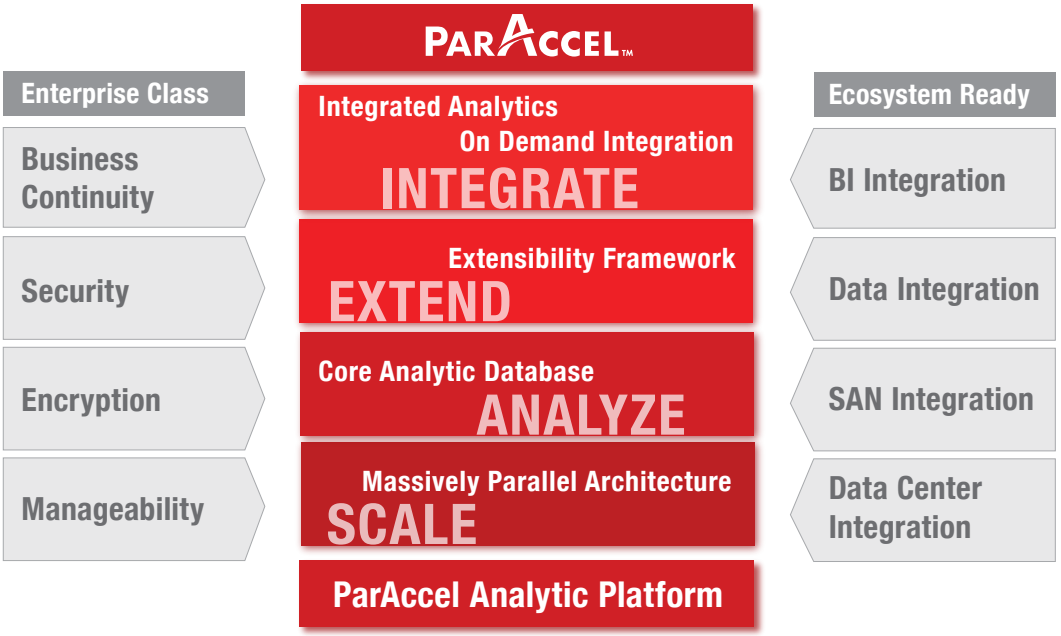
This all raises an overarching challenge in building the next-gen foundation required to support new analytical capabilities: Business will not stop while an overhaul is completed, and few will fund architecture development without a plan for delivering real, measurable value. Practically speaking, most organizations will iterate their existing architectures toward this goal throughout a series of targeted projects. Organizations must carefully consider each new component as part of an integrated architecture, rather than rushing into black-box solutions or short-term fixes. Each incremental component added to the architecture should deliver real value, while supporting the greater long-term vision of unconstrained analytics.

The Rise of the Analytic Platform – Far Beyond the EDW or Legacy RDBMSs

Luckily, technology is evolving to support this vision, as experts agree that the current options are not adequate. Neither data warehouses nor legacy RDBMSs are suited to handle the new demands of analytics today. Both are good at certain activities, but lack the overall intelligence and processing power for new complex workloads. Even more problematic are their limited integration capabilities, if any. Quite often, early adopters have found that their legacy data warehouses just cannot scale to handle the demands of the new analytics. As White writes, "Optimized systems are often required to handle the extreme workloads involved. This is especially the case when the processing involves complex analytical models and algorithms."

Enter the analytic platform, a technology endorsed by many leading experts, albeit under a few different names (including analytical RDBMS, the term favored by White). Analytic platforms include powerful analytical engines and flexible integration capabilities built around a high-performance database. This enables them to handle complex analytical processing of any data, while integrating seamlessly into existing and future architectures.

These evolved systems are far more than just analytical engines or RDBMSs. Analytic platforms must be able to make intelligent decisions about how to process queries to deliver the most efficient performance. They must have in-database analytics capabilities, and be able to determine when to move the analytical process to the data – or the data to the process. They must have the open architecture and capabilities for high-speed, parallel integration with any other analytic data source, including legacy systems, Hadoop or custom systems. And, they must do all this while being easy and cost-effective for IT to deploy, configure and manage.



PARACCEL™

The New Analytic Platform Criteria

Analytic platforms are decidedly different than the data warehouses or legacy RDBMSs that preceded them, and as such, new criteria must be used for evaluating them. Here is a summary of the expert-recommended criteria for analytic platforms.

Intelligent processing:

Perhaps the biggest differentiator between analytic platforms and legacy systems is the embedded intelligence throughout the stack, which determines the best way to process queries, negates the need for indexes and ensures the highest possible query performance for the most complex analytics.

Ability to ingest and transform new data types:

Analytic platforms embed functions for ingesting and transforming data from any data source – and a schema-neutral database means no data modeling or indexing is required up-front (a major difference from many RDBMSs). This means that it's easier than ever to bring in new data to enhance analytics.

Single interface to all data and processes, open interfaces for integration:

Analytic platforms are all about working in a cooperative architecture, fitting seamlessly into existing ecosystems. They provide a single, seamless interface to all data and platforms, maintaining knowledge of metadata and data locations.

High-speed data and process integration:

One of the most important aspects for next-gen architectures are high-speed integration capabilities, with bi-directional, parallel connections to other platforms and data sources wherever possible. This enables analytic platforms to be tightly integrated with other elements of the analytic ecosystem, so they can quickly share data and processing with limited impact on performance.

In-database analytics:

As the name implies, analytic platforms are purpose built to achieve the performance required for ad-hoc, iterative and complex analytics. An "in-database" processing engine moves data analysis as close to the data as possible – into the actual database. This is in stark contrast to the old way of moving data from an RDBMS into another analytical or BI tool for processing, which adds to network overhead and processing time.

Analytic extensibility:

The ability to extend analytics to incorporate 3rd party or custom analytic functions also set analytic platforms apart from legacy RDBMSs. Analytic platforms should be able to take advantage of industry-specific algorithms and specialized analytic functions developed by third parties – ideally embedding analytic functions as close to the data as possible for maximum processing and performance benefits.

Intelligent workload and storage management:

Due to the growing volumes of data and increasing complexity of analytic workloads, these new platforms must be able to store data and manage workloads with extreme efficiency. This requires intelligent workload management, to ensure that analytics are prioritized and processed in the most efficient manner possible coupled with storage systems designed for fast data retrieval.

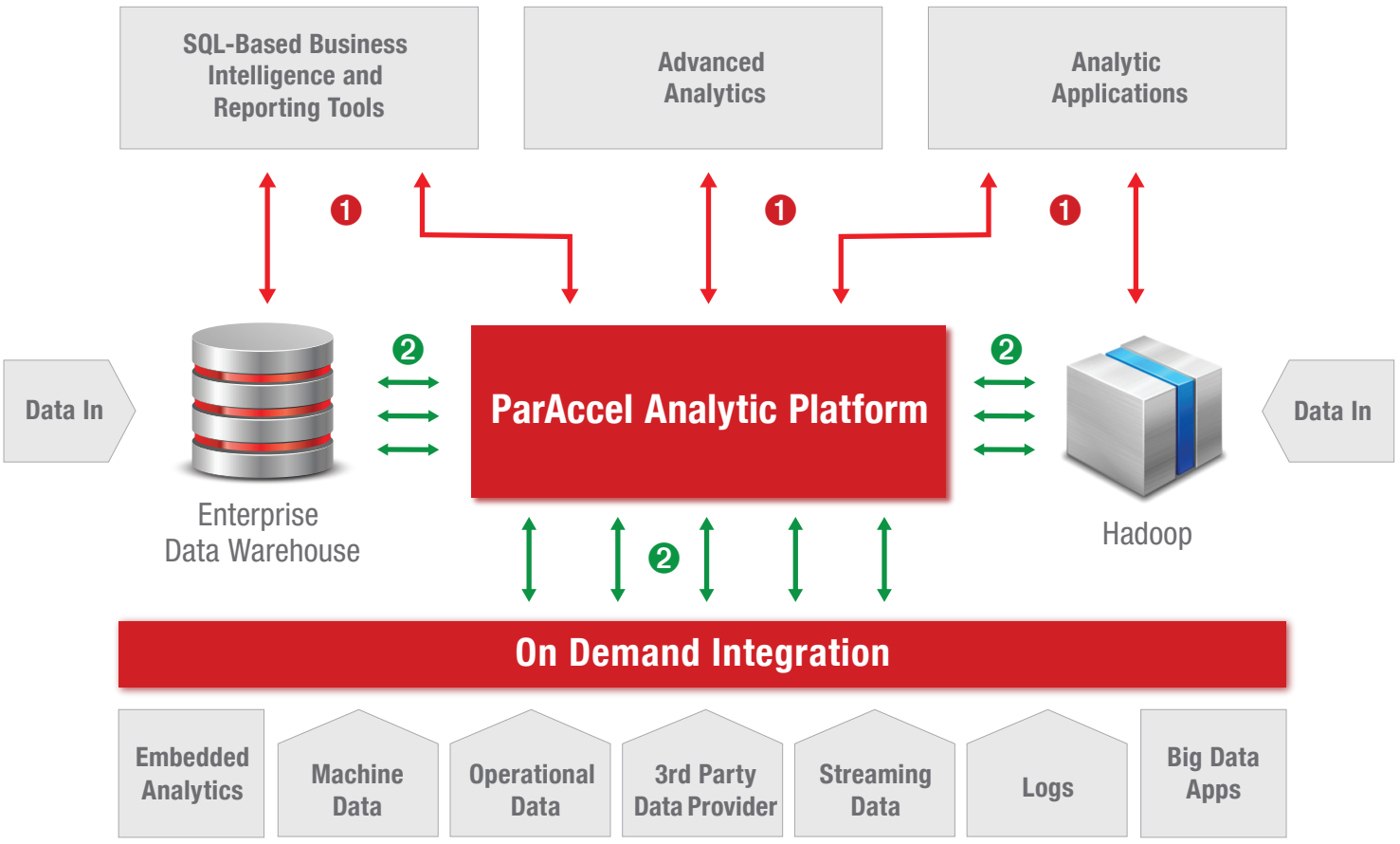
PARACCEL™

The Analytic Platform and the Cooperative Analytic Processing Architecture

The ***analytic platform*** is a fundamental part of what ParAccel calls the ***Cooperative Analytic Processing Architecture***, a concept inspired by the many next-gen architectures recently recommended by experts.

Analytic Platform: An analytic platform is designed for complex, ad-hoc queries, as well as integrated analytics requiring data or analytic results from multiple source systems. A fundamental requirement of analytic platforms is high-performance integration with any other data sources, as well as the speed, performance and embedded intelligence to process queries as efficiently as possible.

Cooperative Analytic Processing Architecture (CAPA): This architecture takes advantage of the data storage and analytical processing capabilities of other components in the ecosystem – such as an enterprise data warehouse, RDBMS or Hadoop. At its heart is an analytic platform, which adds the intelligence to orchestrate complex queries across the architecture. Thanks to tight integration between systems, an analyst can pull data or analytic results from other systems into their query in real-time.



1 Analyze Any Way, with Any Tool.

Use any front-end tool to analyze data from any source system, with the help of ParAccel's analytic platform.

2 Analyze Quickly, with Cooperative Processing:

The ParAccel Analytic Platform leverages its On Demand Integration modules for bi-directional, parallel connections with other analytics systems (such as Hadoop and EDWs) to share data and processing for high-performance analytics.

PARACCEL™

ParAccel vs. the New Analytic Platform Criteria

As evidenced by the new criteria below, the analytic platform represents the a critical evolution of the RDBMS, with new intelligence and integration features making it an extensible platform for managing all enterprise analytics.

ParAccel's Analytic Platform has been endorsed by many experts as the leader of this new category. The platform has been architected from the ground up to deliver on the sophisticated analytical requirements emerging for businesses today. Below, see how ParAccel exceeds all of the criteria required to be the core analytic platform component of the new analytic architectures recommended by experts.

Intelligent processing	The patented, ParAccel Omne Optimizer understands data properties, resource costing involved for queries and other important details about data – and uses these to develop highly optimized query plans which the execution engine runs with industry-leading performance. ParAccel easily handles complex queries with 1,000 way joins, adjusts for concurrent queries, supports correlated sub-queries and other sophisticated analytic techniques. Omne Optimizer and other performance-enhancing features throughout the stack ensure the most efficient query processing.
Ability to ingest and transform new data types	ParAccel On Demand Integration (ODI) modules establish high-performance connections to common data warehouses (Teradata, Oracle), Hadoop, streaming data, and other sources. These modules allow users to easily bring in data from other systems, transform the data, or leave the data on the original system. ParAccel can also initiate analytic jobs as part of complex, multi-source queries. On Demand Integration Services enable organizations and third-party developers to build their own custom integrations with other data sources and analytic systems.
Single interface to all data and processes, open interfaces for integration	ParAccel supports SQL as the top level language for queries executed across the cooperative analytic eco-system, including Hadoop. ParAccel works with both relational and non-relational data using our extensibility framework and On Demand Integration modules enable using the most appropriate languages (e.g. Java, R and Pig) for data access in other systems. SQL remains the top-level language, making it straightforward for analyst access and able to be a SQL-based interface for Hadoop and other systems lacking SQL support. In addition, ParAccel's Analytic Platform integrates with common dictionaries and metadata repositories from other platforms using our On Demand Integration technology. The platform works with HCatalog (for Hadoop users) and dictionary tables from other database management systems. The ParAccel platform maintains the properties and locations of data throughout the architecture to make intelligent decisions about how to best process queries.
High-speed data and process integration	ParAccel is built on a massively parallel processing (MPP), shared-nothing architecture, with a high-speed interconnect network and custom communications protocol for fast data exchange between nodes. All of the ParAccel On Demand Integration modules are fully aware of this parallelism. As a result, these modules can establish high-speed, bi-directional, parallel connections to other platforms and data sources wherever the connecting platform is capable or open enough to handle parallel interactivity.
Analytic extensibility	The ParAccel Extensibility Framework provides open access to easily bring in data from many different sources, leverage external analytics, and quickly assimilate new functionality critical to handling different kinds of analytic workloads. This approach opens the extreme performance of the core ParAccel database to new data and analytic sources through our On Demand Integration modules. In addition, ParAccel's Analytic Platform enables organizations to access embedded, in-database libraries of hundreds of analytic functions, both custom or off the shelf, which can be shared across the organizations. Our philosophy of deep integration, down to the table level, drives high performance and ease of use.
In-database analytics	ParAccel's Analytic Platform was built with four combined in-database analytics layers, the most comprehensive in-database analytics functionality in the market. First, advanced functions are stored in a library and execute within the database. Second, advanced functions can be written to utilize the ParAccel Extensibility Framework and execute at the table level for even better performance. Third, queries using these advanced functions are compiled and executed within the database. Fourth, all optimizations including query, planner, and execution engine, are aware of the unique parallelism and pipelining available for each advanced function.
Intelligent workload and storage management	The innovative workload management function in the ParAccel platform ensures the appropriate prioritization of queries, based on requirements and available resources. This is important for maximum efficiency of queries in a high-performance analytics environment. When it comes to storage, ParAccel's platform is based on an MPP, columnar database – with adaptive compression for maximum data storage capabilities. The technology is schema-neutral, and does not require indexes or views, which take up additional space. This gives the platform unprecedented storage capabilities. One customer's ~10 terabyte Oracle database compressed to only 800GB in ParAccel.

PARACCEL™

ParAccel's Analytic Platform Powers Cooperative Analytic Processing Architectures

As evidenced above, ParAccel has all of the capabilities required to be the core analytic platform in a Cooperative Analytic Processing Architecture. It delivers the functions required for both quick-start analytics projects today and future needs, particularly:

- High-speed integration capabilities, which enable efficient data movement between ParAccel and any other system in the datasphere (including legacy data warehouses,Hadoop/NoSQL databases or operational systems.)
- Embedded "in-database" advanced analytic functions, allowing over 500 complex analytic functions to be run as close to the data as possible for maximum efficiency.
- Intelligent query optimization and embedded "in-database" analytics functions that minimize data movement and maximize complex query performance.
- Flexible configuration options and instrumentation for effective IT management and, importantly, best-in-class price/performance metrics for cost-effective deployments.
- The world's fastest MPP database at its core, built on columnar technology with adaptive compression to maximize storage space.

By implementing the versatile ParAccel Analytic Platform, organizations can expect immediate value creation, while delivering new capabilities for unconstrained analytics. In the short term, ParAccel helps organizations quickly solve urgent analytical problems (see sidebar). This is achieved through an efficient installation process, fast "load and go" analytics capabilities (with no indexing required), and extremely high-performance processing for the most complex queries. Further, since it is often implemented as a software-only solution on commodity hardware or the cloud, it can be quickly deployed without a lengthy data center certification process and scaled up as required without a large up-front investment.

This means that ParAccel customers realize returns within hours or days, rather than months or years. For example, one ParAccel customer had a query that took almost two days (41 hours) to run on a legacy platform. They moved it to ParAccel and with no tuning or customization, the same query took only one minute to run.

In addition to solving analytical challenges like that one, ParAccel's approach creates additional value for companies in other ways. Employees who used to spend time preparing data for analysis through modeling, indexing and other data management duties can now be freed up to work on higher-value projects. The flexible platform can be deployed for a variety of projects, enabling IT to address a backlog of analytics requests quickly. And, since the platform can be implemented cost-effectively on commodity hardware or the cloud, it can start small and scale up with available project budgets.

ParAccel's Analytic Platform is the leading example of this new breed of system, architected from the ground up to deliver on the complex analytical requirements emerging for businesses today. It represents the new direction of analytic architectures based on cooperative processing, leveraging the capabilities of all available analytic systems to deliver results quickly and efficiently every time.

Popular ParAccel "Quick Win" Projects

Analytic Offload Solutions to Extend the Life of Legacy Data Warehouses: Offload to ParAccel any analytic workloads that cannot be processed effectively by a legacy data warehouse. High-speed integration with Teradata, Oracle and other data warehouses make this a efficient solution for processing complex queries.

Big Data Analytics with (or without) Hadoop: Use ParAccel to do complex SQL analytics on data stored in Hadoop via bi-directional, parallel integration. Analytics may be run in Hadoop or ParAccel. ParAccel is the perfect complement to Hadoop, extending it with new capabilities for big data analytics.

High-Performance Data Marts for Analytic Applications: Quickly deploy new analytic applications with ParAccel. With embedded in-database functions and supporting any SQL-based BI and analytics tools, ParAccel enables new applications to get up and running quickly.

For more information, visit www.ParAccel.com, or call your local sales director.

ParAccel can also be reached at 866.903.0335.