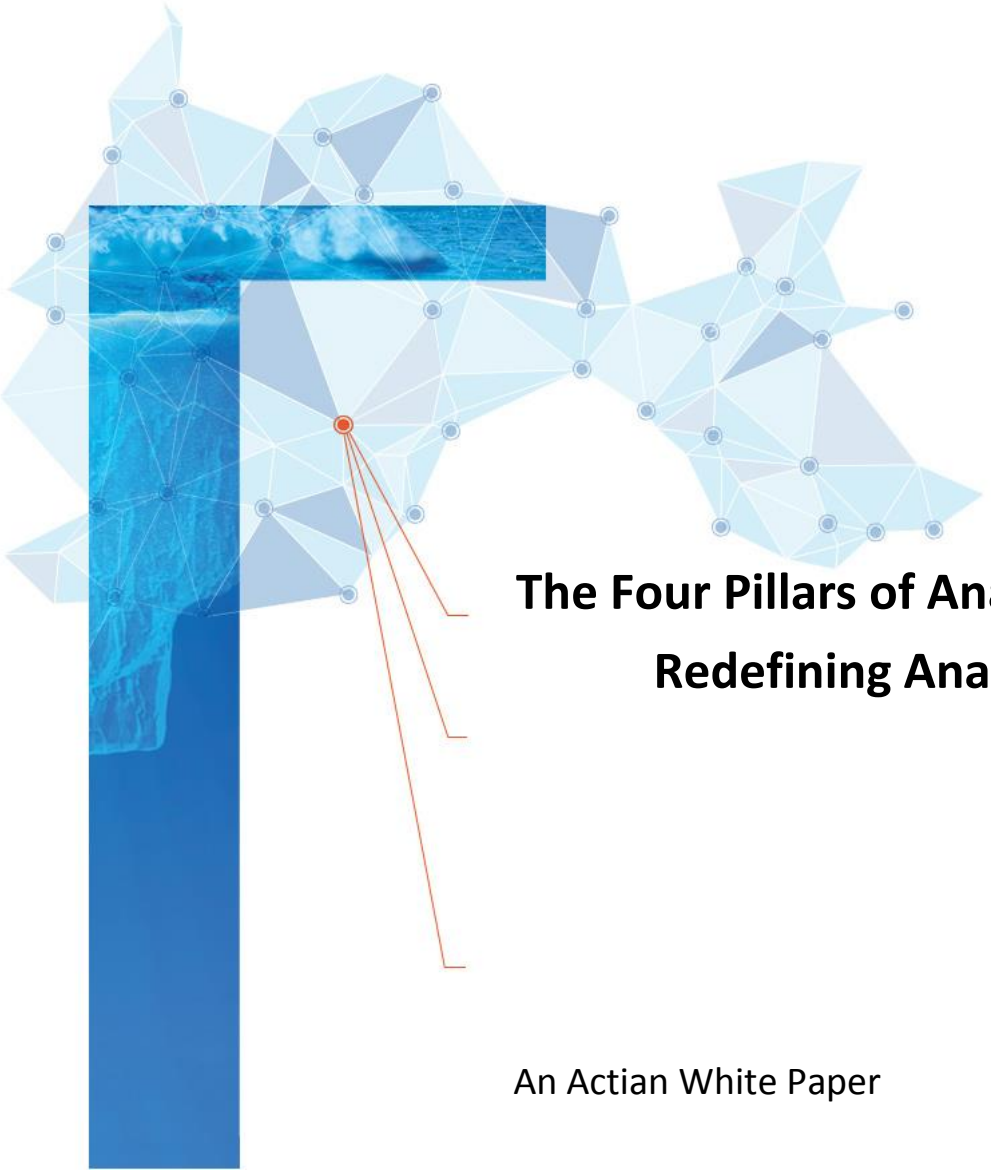


The Four Pillars of Analytics Technology: Redefining Analytical Speed

An Actian White Paper

"Speed equals value as long as you get the answers right."
– Robin Bloor, The Bloor Group

Overview

Data analysis projects to improve customer relationships, fraud detection, cyber-security, and a variety of other critical objectives are making a massive impact on the way modern companies do business. Understanding and predicting markets, trends, threats, and customer behavior dramatically can increase a company's ability to acquire and retain customers, to create or stock the products or services that customers will buy, to avoid potentially devastating pitfalls, and to increase wallet share and profitability of key accounts. In order to implement these valuable analytics, companies are faced with a dizzying array of analytics choices, data storage and data processing tools, frameworks, and platforms. Knowing what really matters when selecting analytics technologies and how to put it all together can be daunting, to say the least.

Adding to the confusion is that every analytics tool on the market claims to be fast. While we agree that speed is an essential element of analytics power, possibly the most crucial element, it's important to note that speed means different things to different people. This is why so many claims sound alike, "Our platform is the fastest," and yet, they don't result in fast business value. Each kind of speed is essential to analytics success in its own way; speed of development, speed of data processing, speed of deployment, and speed of response.

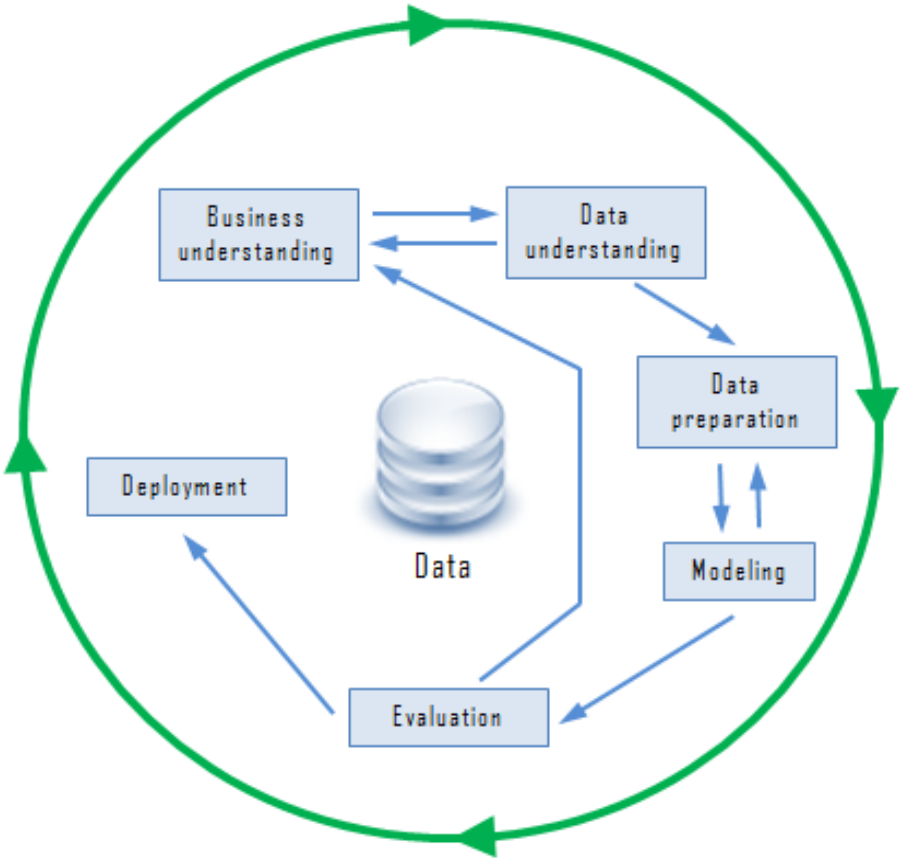
Each type of speed is a strong pillar to build on, but all four are needed to build a genuinely transformational analytics infrastructure.

"Data analytics is not an activity. Data analytics is a multi-disciplinary end-to-end process."
– Robin Bloor, The Bloor Group

Four Definitions of Fast

Analytics project don't just happen. There is a process, from initial concept to final production execution and action, which every analytics project follows. In his white paper, "Minimizing the Five Latencies in the Analytics Process," Dr. Robin Bloor discussed how, at each stage of this process, delays can be compounded. An analytics platform that only speeds up one stage of that process can leave the other stages stalled. Different analytics platforms and tools claim to speed up analytics but, often, they speed up only one aspect of analytics, not the whole process.

Advantage in the competitive marketplace is about analytics speed and agility overall. This includes the ability to experiment, discover, iterate, and refine for high accuracy, and then deploy with ease. Understanding the requirements of analytics success, and what people really mean when they say an analytics platform is fast, can be the key to building an analytics technology foundation that will boost revenue, customer satisfaction, and market share over time.



Data Analytics Process (CRISP-DM)[1]

"You can't simply deploy a static model. Signals and patterns change. Models must constantly improve, re-learn, and update to keep their value."
– Laks Srinivisan, Opera Solutions

Development and Discovery Speed

Many analytics platforms that claim they're fast often take months, or even years, to take a new analytics project from conception to implementation. The software can't be used by the data experts in the company; it requires expensive, hard-to-find programming and data science skills to operate. This type of "slow, expensive development, fast execution" platform can be poison for businesses that want agile, responsive analytics. Development, design, and discovery speed is the first and arguably the most important aspect of efficient customer analytics software. When you hear about big data analytics projects still in the proof-of-concept stage after years, a lack of development speed in the technology is often the issue.

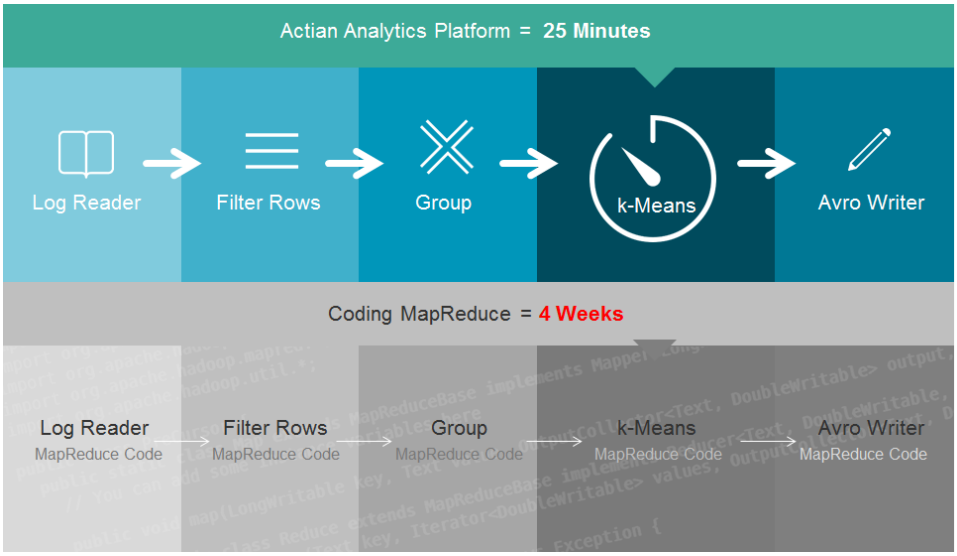
[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step Data Mining Guides*.

Analytics processes require dozens or even hundreds of iterations to get them right. Any time saved in one design iteration is saved again and again, rapidly reducing months to days, making an efficient, easy-to-use design-time environment invaluable. This time is saved again every time an in-production analysis needs to be re-trained, re-designed, or tweaked. Compressing the iteration cycle can be the single most important key to analytics agility.

Expertise in the fundamental nature of customer and business interactions and requirements can be hugely valuable at this development stage, but many analytics tools simply are not usable by the people who have that knowledge. Tools that make analytics design faster, easier, and accessible to people who have expertise in the business can provide the kind of speed that saves months every time a new analytics project is needed, or an existing analysis needs to be modified.

Development and Discovery Speed Example

As a demonstration for a potential client, a consultant built an end-to-end analytics workflow with the Actian Analytics Platform using the KNIME visual interface and deployed it on a Hadoop cluster in under 30 minutes. This workflow duplicated one that a team of MapReduce coders had spent more than two weeks coding. The output from the two was nearly identical, except that the predictive results were slightly more accurate from the Actian job.



Development Speed: Drag + Drop vs. Coding Strategies for Development and Discovery Speed

Unfriendly, programmer-centric interfaces for analytics applications cause a variety of problems. For example, finding and retaining specialized skillsets in a highly competitive market can be virtually impossible. When those skillsets reside in one person or team, and the business acumen and knowledge of the business need resides in a different person or team, communication issues and conflicts are common. Some will understand and seek answers to business problems, such as how to optimize a marketing campaign, or how to prevent customer churn.

Others will understand how to write applications that tease those answers out of data, and possibly a whole other set of people will understand how to hook those applications into the corporate data infrastructure so that the answers are accessible when needed.

The more separation that exists between the people asking the questions and the people able to find answers, the more delay and disconnect will happen. Like the children's game of Telephone, the answer found at the end may not bear any resemblance to the answer requested at the beginning.

There are strategies within organizations that can make a big difference in mitigating this problem, but the solution starts with the technology. Software built with user-friendly visual interfaces always will be more accessible to business experts than software that requires hefty programming expertise.

Data discovery largely has been done via SQL for decades, and a huge number of data-oriented business experts understand how to use this powerful data interface. A standard SQL interface can make analytics software much more accessible and usable than a non-standard SQL interface, a crippled pseudo-SQL interface, or a complete lack of SQL interface.

In many cases, data will be diverse and will not easily fit in standard snowflake or star schemas or cubes. Analytics databases that are flexible enough to analyze data at speed regardless of schema can take a huge burden off the analytics development team. The ability to do some exploration and analysis on large datasets where they sit, such as in Hadoop data lakes, also can be a big speed boost.

Often, extracting, merging, joining, and other data preparation work that must be done before the actual analysis requires the majority of time and resources used on the project. If the analytics expert realizes during the design iteration process that she needs another column, or another minor data transformation added to the data preparation routines, it is a huge help if that is something she can handle herself. An accessible, self-documenting visual high-speed ETL interface can be very beneficial in that case. The analyst may not want to build the entire data preparation process alone, but being able to make a quick adjustment to the data pipeline leading into the analysis, without having to wait for IT to get around to it, can collapse the iteration cycle considerably.

Development and Discovery Speed Strategies in a Nutshell:

- Visual interfaces
- SQL interfaces that are fully ANSI compliant
- Schema flexibility
- Visual ETL

"Doing away with sampling would be a massive boon to analytics. The more datasets grow over time, the more sampling is required, and the more radically sampling skews the results."
– Krishna Roy, 451 Group

Data Processing Speed

Another essential type of speed for analytics power is the ability to crunch through large amounts of data in a short amount of time. There is a cost component to this type of speed since, if you had unlimited funds to buy hardware, of course you could crunch through as much data as you wanted very quickly. This type of speed is about price/performance; preparing and analyzing today's massive, diverse data sets without spending so much that it costs more than the analytics insights are worth.

The rapid rise in popularity of Hadoop is due to this kind of data crunching power per dollar, but there are other analytics platforms now that leave standard Hadoop MapReduce in the dust as far as data processing speed. MapReduce-based analytics are infamous for being painfully slow in the design and deployment phases.

Routine sampling, especially of very small percentages of the available data, is a common symptom of using analytics technologies with poor data processing speed. In some cases, sampling data is a sensible practice, but in far more cases today, tiny fractions of available datasets, as small as 2%, are used for analytics when the full dataset is available and would provide far more accurate answers. Data analysts routinely sample as a workaround for the limitations of analytics software to the point where other options are not even considered.

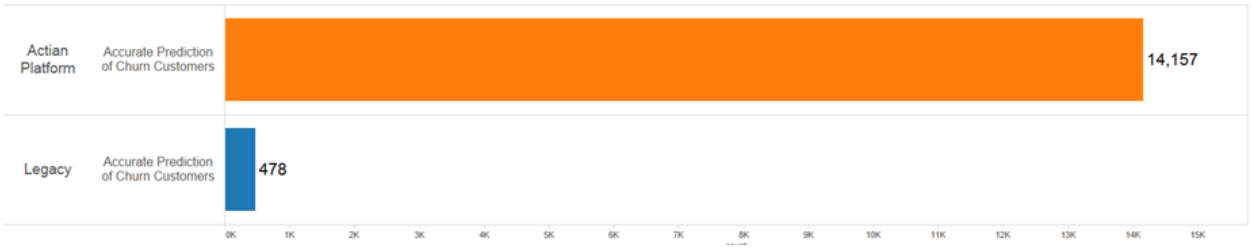
Data processing speed also is about the ability to handle all kinds of data, not just processing one kind of structured data, but joining, de-duping, and jointly processing all structured and unstructured data sets that could add to the accuracy and ROI of an analytics project. Many analytics platforms lack the capability to merge diverse datasets entirely, or have only a very limited capacity. This can be a big stumbling block, requiring complex, custom-built solutions to fill that gap.

If the analytics software only can handle a maximum of 5% of a currently available dataset, the last thing a data analyst would consider is what additional value might come from analyzing additional datasets. As the example below shows, a limitation in this area can cause a tremendous loss of revenue and customers.

Data Processing Speed Example

A machine learning algorithm predicted customers likely to churn from a Telecommunications company's call detail record data. With the limited sample size that legacy software could process, only 19 columns of the data could be analyzed. Using what seemed to be the best 19 indicators, only about 500 potential churn customers were identified. With the Actian Analytics Platform, a far greater sample size could be analyzed in the same amount of time on the same hardware, allowing the data scientist to include 15 more variables from other data sets.

This new data, combined with the original 19 variables, allowed the algorithm to accurately identify about 14,000 customers likely to churn. In particular, the additional variables were very valuable for eliminating possible false positives with surety. At an average cost to a Telecom company of $500 per customer acquisition, the 14,000 churning customers not accurately identified by the previous technology could mean a cost of as much as $6.75 million.



Number of Correctly Identified Churners

(Done on a known test dataset with 16,000 true positives. Dataset available upon request.)

Strategies for Data Processing Speed

Many analytics platforms have some surprising limitations around data processing speed, despite being widely accepted and well-known software. Limitations on the amount and variety of data that software can process have existed for so long that many data analysts assume they are the norm. New, cutting edge software has been built with today's massive and ever-growing data sets as the expected norm.

Strategies like in-memory processing and parallel processing have made huge advances. The advantages of these strategies are well-known. But even more advanced strategies bring data processing speed to the next level. For example, data pipelines that take advantage of parallel hardware to perform multiple operations simultaneously often can increase data processing speeds by an order of magnitude.

Other strategies include taking advantage of modern CPU memory cache, which is an even faster way to process data than using RAM exclusively. This requires a sophisticated vector-based data processing paradigm to take full advantage of today's high-powered hardware.

Anyone who has done data analytics knows that data preparation takes up the lion's share of time and effort in most analytics projects. Fully parallel data merging, de-duplication, data transformation, filtering, and cleansing are not optional with large and diverse data sets. Without a high level of data processing speed in data preparation, the entire analytics process bogs down before the data science heart is even reached.

Combining data sets from different sources in order to enrich analytics requires another set of important strategies. Doing this right requires full joining capabilities, even between data sets with completely different formats and structures. The ability to use sophisticated fuzzy matching techniques to identify how data should merge and eliminate duplicate entries is crucial.

Data Processing Speed Strategies in a Nutshell:

- In-memory processing
- Chip caching
- Pipeline processing
- Vector processing
- Parallel processing
- Disparate data joining and merging
- Parallel ETL
- Fuzzy matching

Deployment Speed

Deployment is the point where a great many corporate analytics projects stall out. At this point, the design and testing phase has been completed, the value of customer analytics clearly has been proven, but that value only can be gained when the new analytics process is hooked into production systems where it can make a difference. Proprietary analytics tools that require their own specific hardware or operating systems, or don't integrate well with existing corporate architecture, can delay project deployment for months. Many times, it even stalls the projects completely before any value is realized.

In far too many cases, the deployment phase involves either a complete rebuild of the entire end-to-end —data connection to action—workflow in a new technology. It also can mean a massive rip and replace of existing systems when new analytics platforms are incompatible with old. It can mean having to purchase expensive, specialized hardware that doesn't scale well over time. None of these options are fast or efficient. Keep the deployment phase in mind from the beginning. The technology you choose for design and development should be the same technology you will deploy in production. Ensure analytic portability right from the beginning. Don't work into a corner with non-portable or non-scalable technology, where everything has to be re-done at deployment time.

Extensibility is key here, too. If one small bit of functionality is missing from the analytics platform and can't be added easily, that can slow deployment down to a crawl. Speed and ease of deployment can mean the difference between a highly profitable project and complete failure.

Deployment Speed Example

Deployment speed is important for every company, but especially is essential to companies whose business depends on their ability to deploy powerful analytic solutions for customer after customer after customer, each with their own widely varied technology landscape, some with data warehouses, some with Hadoop, some with both, some with SOA infrastructures, some with scattered data marts.

Opera Solutions, one of the largest analytics consulting firms in the world, faced that challenge on a daily basis. Opera Solutions already prided itself on its impressively fast deployment, an average of **two to three days** for each new client. Then, the company employed the Actian Analytics Platform to create a clever re-usable architecture with highly parallel ETL, a flexible data storage layer to adapt to whatever current data storage the client had, and a rapidly installable, extremely low latency analytics database at its heart.

The Actian Analytics Platform improved execution time on a particular risk analysis solution by 90%, reducing Opera Solutions' customers' exposure time from three days to a few hours. The development time reduction, due to the framework automatically handling the parallelization aspects, was a nice bonus. But the true benefits to the company appeared when it came time to deploy the solution to many customers, and to deploy other solutions that had been built on the same framework.

Now, Opera Solutions takes an average of **two hours** to deploy a new analytics solution. This has allowed the company to accept far more clients while still meeting its SLAs, and significantly increasing overall corporate revenue. Technology choices can make a huge difference.

Strategies for Deployment Speed

Most businesses will not have to deploy data analyses on top of widely varying technology landscapes multiple times a day, like a large analytics consulting firm. However, technology that makes deployment a matter of a few hours, rather than days, weeks, or months, is advantageous every time any new analytics project comes up in any business.

There are a lot of technology strategies to make deployment smooth. The most valuable feature for deployment speed is scalability. Design, iteration, and proof of concepts rarely are done on massive production scale systems. Analytics technology should be capable of scaling down to work on normal work station hardware for design phase, testing, and iteration, and then ramp up to large scale production hardware at deployment time, without significant work. If the technology only works at small scale, or only at large scale, this will make moving from proof of concept to production a project in itself.

A huge advantage for deployment speed is software that works equally well, regardless of hardware, operating system, or—in a Hadoop environment—distribution. One thing that can save money, as well as time, is software that takes good advantage of ordinary commodity servers, rather than requiring high powered boxes to function. OS and distribution agnosticism provides more freedom and flexibility in deployment, avoiding conflicts before they can become a problem.

The ability to interact with, import from, export to, integrate with, execute, call, and otherwise cooperatively work with other applications is a huge advantage for ease and speed of deployment. In many cases, analytics processes work to a large extent, but have specific bottlenecks, choke points, or problem areas that need to be addressed to improve business outcomes. Rather than having to rip entire systems out and replace them, if new technology provides easy cooperative touch points, the problem areas can be addressed rapidly. Other points in the analytics workflow may need to be replaced at a later date, but the ability to only replace what is needed at the time is a huge advantage.

In many cases, specific analytics components, such as a key R algorithm or a SAS model, may be a genuinely essential and useful aspect of an old analytics workflow. The rest of the workflow may be dated and bogging down and need replacing. If the new replacement technology can call or import or otherwise make use of those essential old components, it gives the deployment option of not "throwing out the baby with the bathwater". Predictive Model Markup Language (PMML) import and the ability to call other technologies inside a workflow or SQL request can be key.

Similarly, new technology should itself be amenable to interaction with other applications and cooperative development. Workflows and analytics models created in the new technology should be re-usable, and PMML export should be an available option.

Having a built-in, high-speed, parallel integration capability also can be a huge advantage to deployment speed, as long as it has a straightforward visual interface and can handle the amount of data that needs to be processed for the analytics requirements at a good speed. No matter how comprehensive an analytics platform is, it always will need to connect and interact with other systems. Having high speed, easy to design integration readily available vastly shortens this process.

No matter how comprehensive an analytics platform is, there may be some very specialized bit of functionality that your business needs that is missing from that platform. A ready-to-use, user-defined function or other extensibility utility can mean the difference between a quick addition and having to build or buy a whole other application to fill a tiny gap in functionality.

It pays to think about these deployment-specific technology advantages before you get to the deployment phase of the project. A little planning ahead can save a lot of time, money, and effort.

Deployment Speed Strategies in a Nutshell

- Scalability
- Hardware agnosticism
- Operating system and Hadoop distribution agnosticism
- Easy extensibility
- Interoperability with other applications
- Reusable, portable analytics workflows and models
- Parallel integration

Response Speed

The first thing people think of when someone talks about “fast analytics” is the speed of response. This can be the analytics performance speed at time of production execution of an analytics process, or the speed of response to interactive queries. Response speed can mean the difference between getting an answer at the right moment, and not getting an answer in time to matter to the business. Good response speed may save hours, or even just seconds, but it does it every single time the in-production analytics process is run or a data discovery interactive query is initiated, compounding that savings by thousands. Clearly, this runtime execution speed is important, but for many analytics technologies, it is their sole claim to fame. Runtime performance speed is table stakes in the world of analytics.

Low latency query response on large datasets still is a competitive area in which many technologies are fighting to be first. This is valuable but, when looking at this type of speed, be sure to pay attention to what queries the technology can handle. A full range of SQL capabilities is even more important than speed. Plus, all of the data that is needed for answers needs to be in range of the query, not a tiny subset. Speed of response is meaningless if you can’t get answers to the questions you most want to ask, or you can’t query the data that has the best answers.

Response Speed Example

OfficeMax needed to move much of its retail business online. The company was in an aggressively competitive and unstable market, and had a highly complex business model with more than 2200 stores, 25 distribution centers, and 55,000 SKUs. It had slow, tired systems running core market basket analysis and shrink processing analytics. This resulted in sluggish response to market and customer shifts that cost the company dearly.

In order to modernize and survive, OfficeMax needed to provide rapid analytics responses to business users across all 2200 stores worldwide. It needed to analyze every basket, every day, across multiple hierarchies, including product, time, planogram, and store. Each basket represented a unique and complex ad hoc query with multiple self joins.

The Actian Analytic Platform allowed the company to reduce response times from multiple hours, or even days in some cases, down to seconds. The shrink processing query, for example, that once took 46 hours, now runs in 30 seconds. The base market basket analysis reporting query used to take seven hours, now runs in under two minutes.

This provided OfficeMax with standard market basket analysis benefits—such as determining optimal pricing to drive basket profitability, identifying cross-sell and up-sell opportunities, optimally arranging products within stores and catalogs, and online—at far more useful speeds. This means that business users can query to answer important questions as they need them. No more waiting days to find out what they need to know right now.

Strategies for Response Speed

Responses need to be not only fast but useful and accurate. In many cases, response speed is related to interactive SQL queries. Fully functional, ANSI SQL capabilities are essential to being able to do the kind of data exploration and discovery that data analysts have been accustomed to for decades. Many current analytics platforms offer only a limited subset of SQL functionality. Getting answers fast won't help if you can't get the answers you need.

In order to get fast query response, a lot of different strategies can be helpful. Different strategies might be needed in a cluster-based database, or in Hadoop, versus a single server database environment. In some cases, similar strategies are useful in both configurations. In a cluster environment, a well-designed query optimizer that can break queries down into parallel components and execute across the nodes is an essential strategy to keep SQL responses within reasonable limits.

Sophisticated vector-based data processing and utilizing CPU chip memory for faster-than-RAM processing provides lightning fast responses in both environments. TPC-H benchmark hardware speed record holders often use analytic databases with vector processing and chip caching capabilities to show off response speed levels. This type of processing takes full advantage of modern hardware's advances to provide the best speed possible on that machine.

Response Speed Strategies in a Nutshell

- Query optimization,
- Full SQL support
- Vector processing
- Chip caching

Response Speed Benchmarks

TPC-H – Top Ten Performance Results – non-cluster

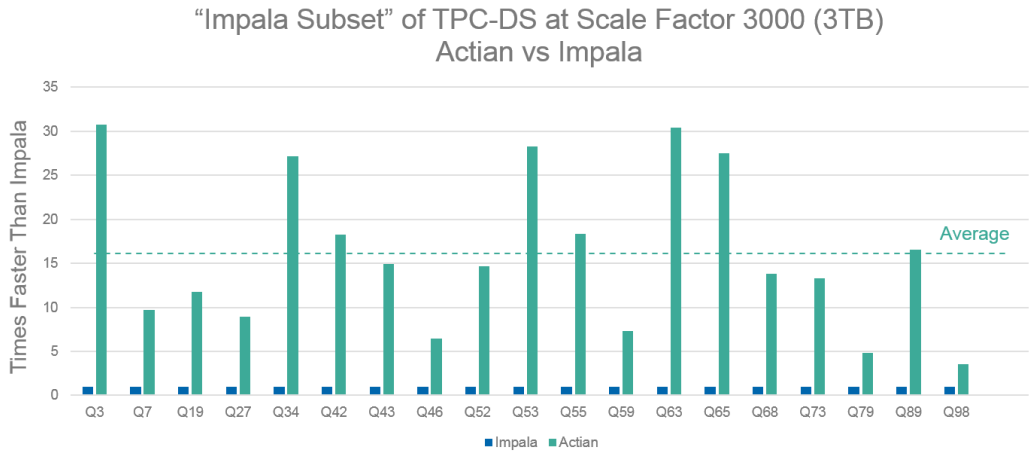
Http://www.tpc.org/tpch/results/tpch_perf_results.asp?resulttype=noncluster&version=2%&currencyID=0

SQL query response speeds for open source cluster technologies and Amazon Redshift
https://amplab.cs.berkeley.edu/benchmark/

Radiant Advisors Benchmark – SQL on Hadoop
http://radiantadvisors.com/q1benchmarkdownload/

Highest Performing SQL in Hadoop

Up to 30X Faster Than Impala



Both Executed on the Same Hardware and Software Environment:
5 Node Cluster with 64GB of RAM per node and 12x2TB Hard Disks.

Background to "Impala Subset "of TPC-DS benchmark can be found here:
http://blog.cloudera.com/blog/2014/01/impala-performance-dbms-class-speed/

Performance Comparison: Actian Analytics Platform VS Impala

"The impossible is now possible. What would you attempt to do if you knew you could not fail?"
– Laks Srinivisan, Opera Solutions

The Need for Speed

All four types of speed always should be considered when deciding which analytics platform will do the job best. Don't just run a response speed race and pick the platform that crosses the finish line first. There's more to analytics than final execution. When choosing software, balance the needs of your business against the strengths of analytics technology in ALL phases of the analytics life cycle. Each type of speed provides a different business and analytic advantage:

1. Development and iteration speed provides accessibility to business experts and far shorter time-to-value.
2. Data processing speed provides the ability to look at ALL data of value for greater analytic accuracy.
3. Deployment speed provides a boost over the last hurdle to gain business value from analytics projects.
4. Response speed provides low latency interactive answers and just-in-time actions.

When choosing an analytics technology architecture, focusing on execution speed alone, with no attention to price per performance, or ability to handle widely varying data sets, or design ease, or compatibility with other enterprise systems, is a recipe for disaster. Any analytics platform that focuses on only one kind of speed at the expense of the others will let you down at a key point in the analytics process. Development and iteration speed, data processing speed, deployment speed, and speed of response are all essential to gaining business value from customer analytics.

Actian: Accelerating Big Data 2.0™

Actian transforms big data into business value for *any* organization. Actian drives revenue growth while mitigating risk with high-performance, low-latency in-database analytics, extensive connectivity, and data preparation. Using off-the-shelf hardware, the Actian Analytics Platform empowers users to connect, analyze, and take near real-time action on big data. Actian in Hadoop provides high-performance data enrichment, visual design, and SQL analytics. Follow Actian at www.actian.com, Facebook, Twitter, and LinkedIn.

For more information, contact us info@actian.com