CLABBY ANALYTICS

Advisory

VelociData: Dissolving Transformation Bottlenecks for Enterprises with Big Data Challenges

Executive Summary

Should your organization use general purpose servers to process Big Data workloads? At *Clabby Analytics*, we would argue that general purpose computers are excellent for processing certain types of data-intensive analytics workloads (for instance, ad hoc queries and pre-defined reporting). But are they the right systems to use to process tens of millions of rows of structured and unstructured data? Maybe – or maybe not …

Over the past few years *Clabby Analytics* has witnessed the rise of new accelerated hybrid systems designs. These systems make use of several different types of integrated circuits (microprocessors) to analyze very large volumes of data at lightning speed. The processors within these "*accelerated systems*" filter data extremely efficiently and accelerate parallel processing – making these systems particularly well suited to analyze large volume, complex analytics workloads (such as processing streaming real-time data, or to process deep and advanced analytics workloads that use very large databases and varied data sets).

VelociData is a maker of one such accelerated system. The company offers a purpose-built micro-supercomputer enclosed in a 4U form factor that can analyze streams of structured and unstructured data at line speed. VelociData appliances feature multiple types of processors (commercial central processing units [CPUs], graphics processing units [GPUs], and field programmable gate arrays [FPGAs]) that sustain very high throughputs and process data in a highly parallel, lightning-fast fashion.

> ***The primary benefit of accelerated system designs is speed – these systems can be used to greatly reduce the amount of time it takes to process very large volumes of data (Big Data). The VelociData appliance reduces the time it takes to transform data (a time consuming element of the extract, transform and load [ETL] process). By speeding the transformation process, VelociData can greatly reduce ETL costs while also dramatically decreasing the time it takes to achieve computing results on large volumes of data.***

In this *Research Report*, *Clabby Analytics* takes a closer look at VelociData's accelerated system design. We describe the market situation; we examine the VelociData technology; and we also take a look at other accelerated system designs. Further, we share some of the insights that we gathered when discussing VelociData deployments with customers.

> ***After evaluating VelociData and listening to customer feedback, we conclude that enterprises looking to very significantly lower the cost of analyzing Big Data databases – and looking to very significantly decrease the time it takes to process Big Data queries – should take the time to evaluate VelociData's accelerated system appliance.***

The Situation

To process Big Data (very large volumes of structured and unstructured data) many systems architects offload data from centralized mainframe databases onto distributed systems for processing. They do this because they believe that they can reduce their MIPS processing costs (MIPS stands "millions of instructions per second" – a way to measure mainframe processing cost) by moving data to less expensive distributed systems.

> ***In our opinion, however, this belief is simply not true (see our report on ETL costs here). Yet many IT managers persist in attacking the problem of processing large volumes of data by moving data to distributed systems.***

Some of the steps involved in ETLing data include:

1. Partitioning data and parallelizing that data;
2. Transforming data from source data formats to the destination data formats;
3. Cleansing and quality checking that data for invalid or malformed values;
4. Sending that data to a destination (sending large amounts of data over the network can in a lot of network latency and overhead);
5. Receiving that data (and the receiver needs to acknowledge receipt – again creating network overhead); and
6. Duplication (copies are made of the data must be made for reliability purposes).

> ***Notice how more systems and storage are needed to handle and host ETL transfers. Also notice that additional labor is needed to manage the ETL process.***

The Hidden Costs of ETLing Data

In addition to ETL costs, there are ancillary costs related to the administration and management of data; the acquisition and maintenance of additional systems used to perform ETL functions (the hardware needed to transfer, transform, validate, host, and store the data that is being sent); and power usage (including the power consumed by additional servers as well as the power needed to cool those servers). Many IT managers fail to weigh these additional costs when they are considering moving their data from system to system.

ETL Costs Are NOT Trivial

Clabby Analytics has access to data that shows *in one case it cost over* ***$8 million in systems related costs and over $223,000 in administrative costs*** *over a 4 year period to ETL 1 TB of data per day (creating one copy and three derivative copies) from a centralized mainframe to a group of distributed systems*. These costs included:

- Mainframe data extract and send;
- Data transformation and cleansing;
- Distributed systems receive and load;

- Network switching;
- Mainframe storage;
- Additional distributed systems storage;
- Mainframe and administrative ETL-related management costs; as well as
- Distributed storage management costs.

Is there a better way to offload and process data without having to pay extreme ETL penalties? The answer is "yes" – use appliances specially designed to lower ETL costs by accelerating data transformation processing speed, while reducing costs related to having to purchase additional system cores and related storage in order to ETL large volumes of data.

This type of hardware-accelerated transformation appliance is what VelociData has brought to market.

A Closer Look at VelociData's Accelerated Appliances

The way that accelerated systems achieve faster processing is to assign different types of computing tasks to different compute resources in order to get the optimal use out of every system component and to leverage massive and pipelined parallelism when practicable. For instance, standard general-purpose commercial CPUs are used handle data flow mechanics and to prepare data for sending to co-processors; GPUs can leverage thousands of internal cores for parallel blocks of data processing; and FPGAs are used for performing massively parallel functions that transform, filter, or quality check data. Working in concert, accelerated system components can deliver dramatic performance increases over traditional general purpose server environments.

How VelociData Transforms Data

VelociData can stream large blocks of data from various sources (including mainframes) and in various formats including COBOL EBCDIC, packed decimal, and others, XML, flat files, and relational databases. During the brief window of microseconds that structured and unstructured data are moving through the appliance, *VelociData can convert data between different formats*, including XML to columnar, mainframe to ASCII, and altering data layouts. In the same trip, without adding any latency or slowing the data down, VelociData can also perform a wide variety of additional computations and tasks including:

- Computation of hash keys to use as surrogates for join operations;
- Performing massive data lookups and reference table joins;
- Mask or encrypt data fields without changing their format (e.g. credit card numbers retain 16 digit ASCII formats);
- Validating input fields for type, content, format, etc.;
- Validating and correcting USPS mailing addresses;
- Performing data aggregations; and
- (soon) Sorting data in real-time at 10Gb network rates.

Data can be moved through the system at wire rates: as fast as it comes in, data can be transformed and delivered. The right compute resource is applied at the right time in the data flow. Since the data is flowing directly through the appliance, no other extraneous processing that would otherwise slow down the stream (such as indexing, locking or database transaction handling) is done along the way. Further, no extra copies of the data need to be made anywhere within the VelociData appliance. And finally, processes can be chained together without compromising throughput or latency.

How Do Workloads Interface with the VelociData Appliance?
To take advantage of the processing speed of the VelociData appliance, the company has created a set of tools that can be installed or called from a variety of differing environments. When installed on an ETL server they can be integrated into workflows through direct interfaces or as "scripts". When installed on any other server in a customer's environment these tools look like local data processing utilities. *Whether they appear as scriptable components on an ETL server, or as local utilities, they transparently handle the high performance data movement to and from the appliance*. Users do not "push" or "pull" data from VelociData, they simply declare the workflow with appropriate sources and destinations and VelociData appliances facilitate the data movement.

VelociData solutions snap directly into existing IT environments, so IT managers can start reaping its benefits quickly without costly, disruptive re-engineering. Simple APIs and CLI tools make it easy to configure VelociData for specific needs. And data center operations staff can monitor and manage VelociData just as they would any other SNMP-enabled server.

> ***Technology research analysts Robin Bloor and Rebecca Jozwiak of the Bloor Group explain the differences between CPUs, GPUs and FPGAs in greater depth in their own VelociData report entitled "The Road to Real-Time BI". But what we like best about this report is that they hit the nail right on the head when then describe VelociData's product offerings as being "deployed primarily to extend the life of corporate infrastructure by offloading the heavy duty ETL, data flow and database workloads". In other words, one of the things that the VelociData accelerated system platform does best is it helps dramatically improve batch performance while reducing ETL costs.***

Applications and Partnerships
Offering fast hardware is only part of the solution when building Big Data analytics facilities. Having analytics software in place that can analyze data and produce reports from fresh data quickly is the ultimate outcome that IT buyers look for. VelociData works closely with industry leading ETL tools to offload the most painful performance challenges. To this end, VelociData has partnered with Informatica Corporation to meet growing customer demand for affordable hyper-scale/hyper-speed Big Data analytics solutions. By virtue of its partnership with Informatica, for example, a task can be created that moves structured and unstructured data between Informatica and VelociData seamlessly, and involves no user coding. VelociData can also utilize a variety Informatica mapping definitions to define their processing workflows.

IBM is another partner that is working to integrate VelociData's accelerations into its technology stack. A similar integration also exists with IBM Data Stage, in that workflows can be integrated and driven completely through a customer's existing tools.

In cases where customers do not already have an ETL tool in place, VelociData utilities can be run in a standalone mode to perform data transformations. VelociData operations can be driven from Linux, AIX, and Windows systems. VelociData also offers a visual tool that can be used to create workflows and configure components. Even without an existing visual ETL environment, customers will only need to write a small amount of code to utilize VelociData accelerations.

The Informatica Partnership

By partnering with Informatica, VelociData has enabled its users to take advantage of Informatica's data integration platform. By coupling with VelociData's fast underlying platform users can achieve analytics results on varied and massive databases in near real-time. The most common integration approach is to pre-process data with the VelociData appliance and send the output directly into the Informatica Data Integration server. As VelociData completes its initial integration with Informatica, ETL developers will be able to transparently use the Informatica GUI to design mappings that call Velocidata directly. Tighter integration will allow VelociData to more effectively move data to and from Informatica to yield the fastest accelerations. VelociData provides a great complement to Informatica technology components. Informatica provides a simple user interface along with a powerful metadata repository and a wide breadth of transformation capabilities and data plug-ins, and VelociData offers the horsepower to enable the entire system conform to service level requirements.

The Pricing Model

The VelociData engagement model is very different from other competitors in that VelociData prices its servers on a monthly subscription as compared to the traditional approach of selling hardware, related licenses, and maintenance. In this way, the acquisition of a VelociData platform is much like purchasing an on-premise "cloud service".

Customers pay one subscription fee that covers on premise usage, maintenance, support and new service releases– independent of the amount of data processed – thus creating a Big Data processing environment that allows customers to avoid capital expense (CAPEX) outlay and instead operate on a fixed-cost basis (OPEX). VelociData's subscription approach enables the company to keep its service uniform and up-to-date across its entire installed base, thus eliminating having to support multiple product versions simultaneously.

> ***As VelociData puts it, our "engagement model allows you to see first-hand just how much our solution will speed your Big Data operations before you commit to a full deployment in your production environment. That's why VelociData offers a subscription model that allows you to avoid capital spending and effectively manage your long-term OPEX. Often, our three-year TCO saves millions."***

To make the VelociData platform even more enticing, the company offers reduced subscription costs for test environments, for data recovery and for other "non-production" instances. Finally, the company also offers fixed-price initial implementation/installation services (instead of time-based deployments where costs can be uncertain).

Contrasting VelociData to Other Accelerated Systems

There are other accelerated systems on the market – so how is VelociData different? In short, because it is "purpose-built" to improve data integration performance.

As part of our mainframe coverage, *Clabby Analytics* collaborated with fellow technology research firm *Enterprise Computing Associates* to examine how the mainframe handles various types of analytics applications. In this joint report we found that IBM tightly couples an accelerated system formerly known as Netezza (but newly named "IBM PureData System for Analytics") to offload the mainframe from having to run certain complex analytics tasks (IBM calls this tightly coupled environment "IDAA" or Integrated DB2 Analytics Accelerator solution). *This is a specialized appliance solution that focuses on streamlining SQL database operations*.

By comparison, *VelociData focuses on speeding data transformation and data flow*. As Chris O'Malley, VelociData's chief executive officer, puts it: "VelociData works to dissolve data transformation bottlenecks for companies with large data volume, velocity and variety requirements and an urgent business need for faster analytics. A qualified prospect for us is a company with table sizes greater than 10 million rows with service levels to supply transformed, high quality, secured data in less than a few hours."

O'Malley also stressed the performance/cost differences between Teradata (a leading data warehouse/analytics vendor), IBM, and VelociData. For Fortune 1000-size Big Data environments, O'Malley stated that VelociData "will be 80% less expensive than alternatives such as the 'push down optimization' on Teradata and IBM's PureData System for Analytics – and tremendously less expensive than throwing more cores at an existing ETL tool. *What costs $10M on a typical database machine to improve data integration performance by 8X, costs less than $100K/month using a VelociData solution – and improves performance by 1000X.* The reason we have such an extreme advantage over these options is that we are purpose built for the job".

Customer Feedback

As part of our VelociData research we spoke with two very large enterprise customers about how they use their VelociData appliances. (The names of these customers have been withheld by request).

The first customer, a healthcare benefits provider, uses VelociData in its labs as well as in a production environment. In the labs the customer is experimenting with using VelociData to manage encryption/key token generation, and is experimenting with large Hadoop databases. In the production environment, the VelociData appliance is deployed in a high availability configuration (using a tightly coupled VelociData appliance for live redundancy). This configuration is used to perform code set conversions and to pull data from many distributed systems into a common database.

When asked about VelociData performance, this customer stated that VelociData can take workloads that previously took 24 hours to execute, and instead execute those workloads in about 2 minutes! As for cost, this customer had not done a formal cost study, but pointed out that without VelociData they would have needed to acquire several large Unix servers, numerous software licenses – and then they would have had to coordinate the ETL effort.

This customer observes that "VelociData costs only a small fraction" of what they would have had to pay for hardware, software and management needed to analyze Big Data databases.

The second customer, a financial services provider, described several tests that his company had conducted with the VelociData appliance – and he also described one pilot project that the company has underway. Starting with deployment, this customer described how easy it was to deploy a VelociData server ("it took us about two hours to configure and start running this device"). In one of the first proof of concept tests that this company performed, they found that it took about one minute for the VelociData appliance to transform and analyze a 15GB database (9 million records of varying lengths). This was a complex copybook structure with 15 record types and over 10,000 redefines. As this customer put it after the test: "we were surprised – no one expected this".

Because this company has tested other devices in the past, testers were a bit wary about the extraordinary performance of the VelociData appliance had shown in tests so far. In the past, other systems that had been evaluated had been optimized for performance – but at the expense of the ease at which analytics changes could be made. So testers made modifications to the test environment and reran the test. And, yet again, the VelociData appliance executed its analytics in seconds. This customer concluded that "making changes can be expensive – but not with VelociData".

As was the case with the first customer, this VelociData customer had not performed a formal cost study. But in one case they believe that they can save a particular department about $500K annually by eliminating the need to process a Big Data Hadoop workload on the mainframe. These cost savings come from not burning mainframe MIPS; from not having to ETL data (instead, transformations and data quality checks would be off-loaded to the VelociData appliance); and from reduced management costs.

Also worthy of note is the testing this customer had performed from an encryption perspective. "We saw 2.5 million encryptions on 100,000 records (100 fields with 25 fields marked for FPE) done in 1 second" stated the tester.

When asked about VelociData's market positioning, this customer responded "I think the current positioning is spot on. The acceleration of data transformations, in near real-time, is exactly how we see VelociData bringing us a completive advantage".

Summary Observations

As we discuss in this *Clabby Analytics* report ("The ETL Problem"), ETLing (moving and transforming) data can greatly increase processing costs because additional systems resources are required to handle ETL overhead (accordingly, numerous processing cores are wasted performing ETL functions) – and related network latency issues tend to slow performance. These additional systems resources also burn more power and require more cooling than better optimized accelerated solutions.

VelociData helps lower ETL costs by accelerating the data transformation/data flow processes using a less expensive platform (as compared with traditional general-purpose servers) to accelerate Big Data analytics processing on very large databases. VelociData is also able to greatly reduce overall systems costs through comparatively low cost, cloud service-like subscription service pricing.

In addition to lowering analytics processing costs, VelociData accelerates time-to-result by assigning the right types of work to the right processors within its hardware platform. CPUs are used to manage serial tasks and task assignment to other components; GPUs are used to provide greatly accelerated parallel processing; and FPGAs are used to filter data and speed communications. The way that workload are optimized on VelociData systems can result in achieving results 1000X faster than traditional general-purpose architectures and provide customers with real-time capabilities.

IT executives who need to process large volumes, varieties and velocities of data in a reliable fashion that constantly meets service level expectations – and who are looking for low cost processing solutions for analyzing Big Data – should be looking closely at using accelerated systems solutions. VelociData, with its super-fast micro-supercomputer, should be at the top of the list of accelerated solutions under consideration.

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-6662