BeyeNETWORK
Global coverage of the business intelligence ecosystem
US EDITION |
Login | Become a member!
search

Home
+ Channels by Expert
+ Channels by Industry
+ Channels by Topic
Articles
Web Seminars
Research Library
Events
+ Blogs
Spotlights
Podcasts
Expert Round Table
Videos
SearchDataManagement.com
SearchBusinessAnalytics.com
BeyeUNIVERSITY
BeyeRESEARCH
BeyeBLOGS
BeyeCONNECT
News
Content Archive
+ Resources
+ Stay Informed
+ About BeyeNETWORK

Accelerating Big Data Analytics with Vector Processing: A Q&A with Peter Boncz of Actian

by Ron Powell

ORIGINALLY PUBLISHED SEPTEMBER 5, 2014

This BeyeNETWORK article features Ron Powell's interview with Peter Boncz, professor at Vrije University in Amsterdam in the Netherlands. He is also the chief technical advisor to Actian. Peter and Ron talk about vector processing and what it means for analytics.

PRINTER-FRIENDLY
EMAIL TO A FRIEND
EMAIL TO MYSELF
COMMENTS

Peter, I understand that you are considered the father of MonetDB and vector database processing. Can you tell us a little bit about that?

Peter Boncz: MonetDB is considered a column store pioneer. It is one of the first systems specialized for analytical data processing, and it was the result of my PhD research at CWI in the Netherlands. Subsequently to MonetDB, I did a follow-up system and project directing a group of students and myself, which became VectorWise. VectorWise is emerging proof of MonetDB, and VectorWise became a spinoff company as well, which was later acquired by Actian.

Through this relationship with VectorWise being acquired by Actian, I started working with Actian and I am still working with Actian now in a research relationship.

What is vector processing and what makes vector processing different from other database approaches?

Peter Boncz: Vector processing is a way to organize a query processor that is redesigned to take the best advantage of modern hardware, specifically of modern CPUs. Modern CPUs are very good at parallel execution of operations. And they have been really tuned to do so, actually not for database purposes or analytical database purposes, but for things like video. So people may have heard of SIMD instruction sets, which are capable of executing many operations in single clock cycle. These kind of video instruction enhancements that were introduced into common CPUs, like Intel CPUs and AMD CPUs and PCs, were not really used very well by database systems. But by vectorizing query processing, database systems can take advantage of them and therefore become much faster.

Use of vector instructions is just one aspect. There is actually much more to this, but I will keep my answer short here.

When you talk about speed, what type of performance testing or benchmarking has been done to showcase the speed of vector processing?

Peter Boncz: Well, I can say that irrespective of the kind of customer that has tried VectorWise, they have always come back extremely excited about its performance, particularly beating their previous solution by at least a factor of ten. So, this is evidence directly from customers that already experienced significant value from vector processing.

Subsequently, we have also done industry benchmarking, specifically the TPC benchmarks. There is an analytical database benchmark called TPC-H, and VectorWise has been leading the performance charge for that since its existence by quite a wide margin. When people look at these standard industry benchmarks, they will see the product name as VectorWise, but going forward they will see the product name Actian Vector because the product was rebranded earlier this year.

What types of applications or use cases are ideal for vector-based analytic databases?

Peter Boncz: Any analytical database usage scenario plays well into the capabilities of vector processing, so this comprises reporting and also forecasting, data mining and trend analysis. So all situations where you have complex queries that go through huge amounts of data are where the vector technology shines.

Can you share some specific examples of how organizations are using the Actian Analytics Platform and specific use cases benefiting from vector processing?

Peter Boncz: I am not actually a sales engineer or somebody who usually goes to customer accounts. I am a professor, and that would be work on the development side. But we have heard indirectly many use cases. Typically customers use Actian Vector for reporting, forecasting and analysis. These customers come from either finance or insurance, social networks – any kind of organization that has a large amount of customers they don't know personally but want to address in a more personalized and adapted fashion. Telecommunications is another example. I must say that analytical data processing has really broken through and there are now many different usage scenarios in which Actian Vector has been involved.

You mentioned the social side of the world. We hear how more organizations are adopting Hadoop as a low cost way to store big data, but most are struggling to provide users with SQL access to Hadoop data. Have you been able to extend vector processing to work on Hadoop to provide SQL access?

Peter Boncz: Recently I was at the Hadoop Summit in San Jose. Actian announced the Actian Analytics Platform - Hadoop SQL Edition which brings analytics and vector processing to the world of clusters, specifically Hadoop clusters. So by far fastest single server database technology is now available for use on Hadoop clusters, which is the standard in exploiting the power of cheap compute clusters. I think that's very exciting and the audience at the Hadoop Summit thought so as well.

Peter, why is this approach better than other SQL on Hadoop solutions we hear about?

Peter Boncz: Well, first of all I think it is the invention of vector processing technology, which is just an extremely high performance technology that leads the performance charts. So now this fastest way of processing analytical queries has come to Hadoop. That's point one.

The second advantage is that other vendors are building SQL on Hadoop solutions by using an open source system like Postgres or Apache Derby, which was not developed for analytical applications, and then create middleware around it to make it a clustered Hadoop solution. So basically you've got Postgres and old database components sitting in a Hadoop solution. But these old database components don't perform well for analytics. They are ten to one hundred times slower on analytical queries than Actian's SQL in Hadoop. That's why these other solutions are not very competitive in that sense.

Invariably, if you have an old technology like Postgres put into this new Hadoop environment, there are many constraints. Very often these solutions cannot update data. This is because the HDFS file system of Hadoop is immutable. So you can only append data – you cannot modify data. These old database systems are not capable of that. It happens to be the case that the Actian vector technology has a patented way of updating data without writing in place. This is called Positional Delta Trees. This Positional Delta Tree technology actually enables Actian's SQL in Hadoop capability to support update queries. So, it supports mixed workloads, and this is a unique capability in the Hadoop world.

Ten to a hundred times performance gains with vector processing – are you seeing that from Actian's SQL in Hadoop capability as well?

Peter Boncz: Yes. If we now compare Actian's SQL in Hadoop with alternative products like Hive or Impala, we do see that. For instance, Impala – which claims to be 10 times faster than Hive – is at least 15 times slower than our SQL in Hadoop using the chosen benchmark by Impala, which is a subset of the TPC-DS benchmark. So, yes, indeed we see that Actian's SQL in Hadoop is the fastest solution for providing SQL access to Hadoop data.

What key capabilities should organizations look for when evaluating SQL on Hadoop solutions?

Peter Boncz: There are a number of issues there. One thing to look at is performance, and we have already covered that, specifically performance on analytical database queries because that is the prime use scenario for SQL on Hadoop database systems.

The second thing to look at is how well these systems are really integrated in the Hadoop ecosystem. This means that a SQL on Hadoop system should integrate with YARN and it should operate and integrate with HDFS.

If an organization decides to standardize its cluster hardware on Hadoop, using Hadoop as the standard software layer, and this is happening, then one of the reasons for doing so is consolidation of hardware. So you have a single cluster and you can use it for all demanding tasks. This also means that they're like traditional MPP databases much like the traditional cluster database systems that were the only users of their own cluster hardware. In a Hadoop system, you may have multiple workloads that use the cluster at the same time. And Hadoop has a component called YARN, which is its resource manager, which makes sure that in the case where there are multiple users of the same cluster, the resources in the cluster are divided in a good way among the different users. But that's why it's important to integrate a SQL on Hadoop system with YARN, so you can play well with the other users of the cluster and not thrash, not compete and destroy each other's performance due to the fact that everybody is fighting for the same core or the same memory or the same disk. So that's number two.

So I had performance, I had YARN, and the third thing is update capabilities. Many workloads are not purely read only. And this is where Actian's SQL in Hadoop really distinguishes itself. So these are three things I would insist on as criteria to evaluate systems.

What is on the horizon for vector databases and what problems are you looking to tackle next?

Peter Boncz: Well, we are really excited about deploying the Actian Vector technology in Hadoop. And we think it's the start of a whole new development. We are actually taking a deep look at the way people are using Hadoop and are bringing industrialized, mature data infrastructure into Hadoop. Hadoop is a powerful, extensible framework where you can easily write programs that run over the entire cluster and allow you to attack complex problems without wizard programmers. In fact, the Actian platform provides a visual data science workbench and data flow engine enabling users to quickly build and run any type of analytics without having to code. One of the interesting aspects of Hadoop is that there are interfaces that allow applications to easily read various formats like Parquet or ORC files, which are the formats used in Hive and Impala, which are all SQL-only solutions. We think that Actian's SQL in Hadoop data format is significantly better because it is more compact and yet easier to read and write. And the HDFS integration of Actian's SQL in Hadoop is very interesting, I think, for all kinds of uses – not only for uses by Actian Vector but possibly also by other components in the Hadoop space. So we are actually looking to expand and publish the Actian SQL in Hadoop file format for programmers and contribute to the Hadoop ecosystem. And the Actian Analytics Platform – Hadoop SQL Edition, which includes the SQL in Hadoop capability, was just released and there is already strong interest and demand. People are looking for mature, proven technologies to analyze Hadoop data and provide high performance SQL access without the need for an army of Hadoop programmers. Stay tuned. There is a full roadmap of feature enhancements that will follow specifically exploiting YARN, better workload management and elasticity.

Well, we look forward to the future. Thank you, Peter, for discussing how we can accelerate big data analytics with vector processing.

Ron Powell
Ron, an independent analyst and consultant, has an extensive technology background in business intelligence, analytics and data warehousing. In 2005, Ron founded the BeyeNETWORK, which was acquired by Tech Target in 2010. Prior to the founding of the BeyeNETWORK, Ron was cofounder, publisher and editorial director of DM Review (now Information Management). Ron also has a wealth of consulting expertise in business intelligence, business management and marketing. He may be contacted by email at rpowell@wi.rr.com.

More articles and Ron's blog can be found in his BeyeNETWORK expert channel. Be sure to visit today!

Recent articles by Ron Powell

- SAP BW – Why Semantic Intelligence Matters: A Q&A with Lothar Henkes of SAP
- Data Streaming Essential for Real-Time Big Data Applications: A Q&A with Neil McGovern of SAP by Ron Powell - BeyeNETWORK
- The Guinness World Record for the Largest Data Warehouse: A Q&A with Tom Traubitz of SAP
- Data Modeling in the Big Data Age: A Q&A with Neil Buchwalter of ERwin

Comments

Want to post a comment? Login or become a member today!

Be the first to comment!

Comprehensive resources for business intelligence and data warehousing professionals.

NETWORK WEBSITES: BeyeUNIVERSITY | BeyeRESEARCH | BeyeBLOGS | BeyeSURVEY | BeyeNETWORK | SearchDataManagement.com | SearchBusinessAnalytics.com
PARTNER SITES: TDAN.com | DSSResources.com | DataGovernance.com | Decideo.fr

BeyeNETWORK

Twitter LinkedIn YouTube

Press Releases | About Us | Contact Us | Content Archive | Privacy Policy | Media Kit