Announcement Review

IBM's New Moves in Big Data: Acceleration, Optimization, and an Open Source Alternative

IBM Almaden Labs, San Jose, CA, April 2, 2013
In a major, multi-faceted announcement, IBM announced yesterday that it will soon deliver "speed-of-thought" analytics using DB2 "BLU" acceleration, improvements in its Big Insights and Stream products, and a tuned-for-Hadoop PureData System. IBM also announced time series improvements in Informix database line.

What do all of these announcements mean?

From a big picture perspective, IBM is doing the following:

- Dramatically accelerating the performance of its analytics products;
- Simplifying the deployment and operations of analytics systems (turnkey analytics systems combined with usability improvements in software);
- Creating an "enterprise-class" alternative to open source Hadoop offerings;
- Differentiating itself in the smart meter and sensor data analysis markets; and,
- Increasing competitive pressure on Oracle and Teradata in the database and analytics markets.

The remainder of this report looks at IBM's announcements more closely.

Dramatically Accelerating Performance
Probably the biggest news in IBM's extensive announcement is the announcement of an analytics accelerator known as "BLU" acceleration. Although classified as an accelerator, to us, BLU acceleration is more like a process that uses a series of steps to analyze a very large database ─ delivering results in seconds or less as opposed to hours or days.

Here's how BLU acceleration works:

- Start with a large database (for instance, 10 TB of data);
- The first action that BLU acceleration takes is to compress that data to 1 TB in memory using unique encoding techniques.
- IBM then uses its unique memory management facilities to place data in memory, thus putting that data close to the processor. (The closer data is to a processor, the faster it can be processed);
- The next step involves reducing that 1 TB database to 10 GB using column processing which is built directly into the DB2 database kernel – making this compression phase extremely fast. (In recent years IBM has become a master at compressing data ─ saving its customers from having to buy additional storage ─

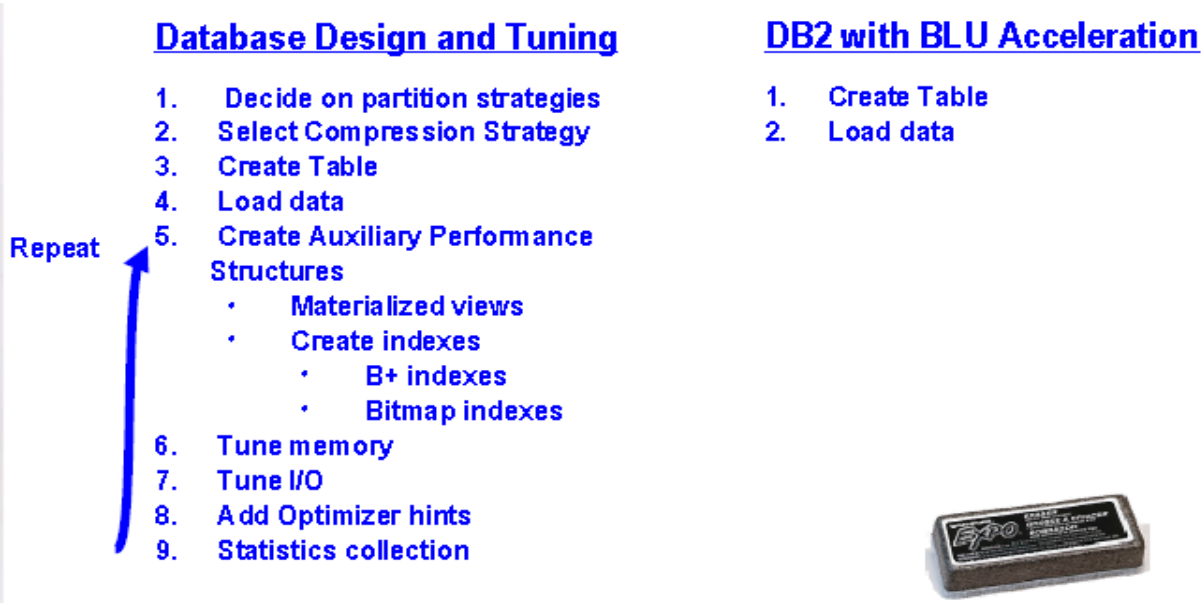
and in this case, enabling its customers to move compressed data into memory to speed analytics processing);

- A "data skipping" process is then used to reduce that data to 1 GB of sifted data. This skipping process filters out non-relevant data ─ presenting the processors with only relevant data to analyze;
- This data is then turned over to the processors (each performs a linear scans). Each core uses a vector processing technique to scan data in parallel at speeds up to 8 MBpS using a SIMD (single instruction, multiple data) parallel processing approach. What this means is that the processors are taking the data from memory, placing it in cache, and scanning that data in parallel at extremely high speeds.

The end result of this BLU acceleration process is that data can be analyzed faster than ever before ─ or as IBM likes to put it: data can be analyzed "at the speed of thought". But, exactly how fast is the "speed of thought?" IBM is reporting that BLU acceleration can lead to increases of 1-25X in reporting and analytics activities. And IBM has even intimated that it has seen over 700X performance improvements in some analytics queries in beta tests.

One of the elements that we like the most in this BLU acceleration process involves the creation of BLU tables. The traditional way of structuring data for analysis is to use a classic row-structured table. Database administrators go through nine distinct steps to set up their data and then perform analysis on it. These steps include making partition decisions, compressing data, creating tables, loading data, creating auxiliary performance structures, tuning memory and input/output, adding optimizer "hints" and collecting statistics. With BLU acceleration database administrators simply need to create their tables, load them, and execute the analytics process (these steps are known as "create, load, and go" ─ see Figure 1).

Figure 1 ─ How BLU Acceleration Speed Data Analysis Preparation



Source: IBM Corporation, April, 2013

It is also noteworthy that these BLU tables can coexist next to traditional row-oriented data structures ─ or these traditional structures can all be converted easily to BLU tables. So database administrators can change all structures to BLU tables, or convert their databases incrementally.

Simplifying the Deployment and Operation of Analytics Systems

Only a year ago IBM announced a new type of systems design: the IBM PureSystem. Since then IBM has created several specialized off-shoot PureSystems designed specifically to process analytics and transaction workloads. These analytics/transaction-oriented designs are known as IBM PureData Systems. These PureData Systems include specialized models for analytics and reporting, for operational analytics, for transaction processing ─ and now a specialized model for Hadoop Big Data processing (more on this design in the next section). All of these designs are turnkey, hardware/software implementations that have been designed by experts for easy deployment and tuned for high performance.

From a software perspective, IBM announced that it has simplified the interface to its BigInsights product offering (while also improving its GPFS file system security and increasing reliability). And IBM announced that it has simplified application development for its Streams product (and that it has taken steps to simplify large scale deployment and integration with Streams enhancements). The way we see this Streams announcement is that IBM is further extending its lead in streams analysis ─ *none of IBM's leading competitors are even close to IBM in terms of streams analysis capabilities*.

> ***The reason that simplification is important is that the business analytics market is undergoing extremely rapid growth — and skilled individuals are needed to drive continued growth. The more complexity that can be taken out of the deployment of analytics systems — and the less complex analytics software can be made to use — the faster the market will grow. IBM recognizes this and is strongly focused on simplifying systems deployment through expert integration, and simplifying software interfaces.***

Creating an Enterprise-class Hadoop

Apache Hadoop is a software library that serves as a framework for processing large data sets (Big Data) across clusters of distributed computers using simple programming models. The use of this library and approach has greatly accelerated over the past few years and, accordingly, Hadoop is becoming the preferred approach for dealing with the analysis of very large databases.

Hadoop is an open source movement. And typically, Hadoop is deployed on x86 clusters. So the buying trend in the Hadoop marketplace has been to acquire Hadoop open source software and to deploy and integrate that software on x86 hardware. IBM recognizes this trend ─ and has contributed code to the Hadoop open source effort and sells x86-based hardware for Hadoop deployments.

But IBM also recognizes that open source Hadoop deployments are haphazard ─ saddled with deployment and tuning complexities. Accordingly, to streamline Hadoop deployment, as well as to harden Hadoop with enterprise-class features, IBM has created a PureData System for Hadoop. This system deploys 8X faster than a customary open source

deployment effort; it uses built-in visualization to accelerate insight; and it provides built-in analytics accelerators that can be used to accelerate Hadoop performance. Further, IBM's PureData System for Hadoop uses a single system console for management; it allows rapid maintenance updates, and it can be data load ready in hours instead of days or weeks. This implementation is the only Hadoop system on the market with built-in archiving tools. And PureData Systems also have robust security extensions and have been architected for high availability.

What is important to note about IBM's new Hadoop system is that IBM is supporting open source Hadoop while also competing against open source Hadoop with an enterprise-class turnkey Hadoop implementation.

Differentiating in Smart Meters and Sensor Data Analysis Performance

Sandwiched in amongst IBM's DB2, PureData system, BigInsights, and Streams announcements was the announcement of a new version of IBM's Informix database (version 12.1). What is important to note about this announcement is that IBM has added TimeSeries acceleration to Informix ─ enabling Informix customers who need faster reporting and analytics on sensor-driven data to get results up-to 100X faster than in previous versions.

Enterprises collect a lot of data from sensors ─ and much of that data goes un-analyzed do to the sheer volume of this data. The ability to accelerate timestamp analysis means that IBM Informix customers can analyze more of the data that they have already captured, leading to new insights and operational improvements. (Note: some benchmarks show that Informix is outperforming its competitors in time stamp performance by a factor of five; while using 1/5 of the system resources).

Accelerating Its Lead over Oracle and Teradata

We've seen estimates that the analytics market could generate as much as $50 billion in revenue by 2015. IBM wants to become the leading force in analytics systems ─ and is investing heavily in hardware and software solutions to capture a large share of this marketplace. To date IBM has spent $16+ billion on analytics software acquisitions ─ a trend that we expect will continue as IBM seeks to acquire new analytics algorithms for new markets. And IBM is also spending billions building highly-integrated analytics systems ─ another trend that we expect to continue as IBM builds out its analytics offerings with specialized servers and appliances designed to process specific analytics tasks. Further, we are also aware that IBM is spending big on training its field and professional services sales forces in order to drive analytics sales and to support analytics customers.

NOTE: Last year we reported that IBM's goal was to drive about $15 billion in revenue by 2015. But IBM has recently raised its expectations ─ and now plans to drive $20 billion by year-end 2015. This exhibits IBM's growing confidence in customer uptake of its analytics messages as well as IBM confidence in the strength of its analytics solutions. And we think IBM's confidence is growing due to lack of competitive response to IBM's foray into analytics markets. I don't see this level of investment and this breadth/depth of product offerings from any of IBM's leading competitors in analytics (specifically Oracle and Teradata).

IBM's New Moves in Big Data:
Acceleration, Optimization, and an Open Source Alternative

Summary Observations

A year and a half ago *Clabby Analytics* started to write a book on workload optimization. Our theory was that vendors were working on "software pathing" ─ a concept that we created to describe how vendors could tune various aspects of program execution in order to achieve exponential improvements in performance. We expected these performance improvements to occur by finding paths through the infrastructure layers, to the processor layers (where specialized instructions could be exploited to increase performance), and out to the database where kernel level instructions would make applications perform even faster. Unfortunately, we were never able to prove-out our software pathing theory ─ so we canned the book (and took whatever useful information that was created and loaded it onto a Website that we call www.workloadoptimization.com).

As we look this series of IBM announcements, however, we are finally seeing some evidence of our software pathing theory in the announcement of IBM's BLU accelerator. A closer look at IBM's BLU accelerator shows how compression and optimization techniques are used to streamline the execution of a query. This is very sophisticated tuning that compresses data, manages memory, and exploits underlying processors to the max. BLU is exactly what we meant when we were trying to describe software pathing (it accelerates application/database performance). In the future we expect to see even more examples.

As we look back over this extensive Big Data announcement, what we see is a company that has recognized a major market opportunity (analytics) — is making all of the right moves to expand marketshare before its competitors wake up. With a series of products that offer industry leading performance, with simplicity baked into the deployment and operations processes, and with advanced technologies (such as IBM's BLU), IBM is going to be extremely hard to beat in analytics from a competitive perspective over time.

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