MUST READ: DDoS attacks are becoming more prolific and more powerful, warn cybersecurity researchers

MapReduce and MPP: Two sides of the Big Data coin?

To many, Big Data goes hand-in-hand with Hadoop + MapReduce. But MPP (Massively Parallel Processing) and data warehouse appliances are Big Data technologies too. The MapReduce and MPP worlds have been pretty separate, but are now starting to collide. And that's a good thing.

By Andrew Brust for Big on Data | March 2, 2012 | Topic: Big Data Analytics

When the Big Data moniker is applied to a discussion, it's often assumed that Hadoop is, or should be, involved. But perhaps that's just doctrinaire.

Hadoop, at its core, consists of HDFS (the Hadoop Distributed File System) and MapReduce. The latter is a computational approach that involves breaking large volumes of data down into smaller batches, and processing them separately. A cluster of computing nodes, each one built on commodity hardware, will scan the batches and aggregate their data. Then the multiple nodes' output gets merged to generate the final result data. In a separate post, I'll provide a more detailed and precise explanation of MapReduce, but this high-level explanation will do for now.

But Big Data's not all about MapReduce. There's another computational approach to distributed query processing, called Massively Parallel Processing, or MPP. MPP has a lot in common with MapReduce. In MPP, as in MapReduce, processing of data is distributed across a bank of compute nodes, these separate nodes process their data in parallel and the node-level output sets are assembled together to produce a final result set. MapReduce and MPP are relatives. They might be siblings, parent-and-child or maybe just kissing cousins.

products started out as offerings from pure-play companies, but there's been a lot of recent M&A activity that has taken MPP mainstream. MPP products like Teradata (http://www.teradata.com/) and ParAccel (http://paraccel.com/) are independent to this day. But other MPP appliance products have been assimilated into the mega-vendor world. Netezza (http://www.netezza.com/) was acquired by IBM; Vertica (http://www.vertica.com/) by HP, Greenplum (http://www.greenplum.com/) by EMC and Microsoft's acquisition of DATAllegro resulted in an MPP version of SQL Server, called Parallel Data Warehouse Edition (http://www.microsoft.com/sqlserver/en/us/solutions-technologies/data-warehousing/pdw.aspx) (SQL PDW, for short).

MPP gets used on expensive, specialized hardware tuned for CPU, storage and network performance. MapReduce and Hadoop find themselves deployed to clusters of commodity servers that in turn use commodity disks. The commodity nature of typical Hadoop hardware (and the free nature of Hadoop software) means that clusters can grow as data volumes do, whereas MPP products are bound by the cost of, and finite hardware in, the appliance and the relative high cost of the software.

MPP and MapReduce are separated by more than just hardware. MapReduce's native control mechanism is Java code (to implement the Map and Reduce logic), whereas MPP products are queried with SQL (Structured Query Language). "Hive (http://hive.apache.org/)," a subproject of the overall Apache Hadoop project, essentially provides a SQL abstraction over MapReduce. Nonetheless, Hadoop is natively controlled through imperative code while MPP appliances are queried though declarative query. In a great many cases, SQL is easier and more productive than is writing MapReduce jobs, and database professionals with the SQL skill set are more plentiful and less costly than Hadoop specialists.

But there's no reason that SQL + MPP couldn't be implemented on commodity hardware and, for that matter, no reason why MapReduce couldn't be used in data warehouse appliance environments. MPP and MapReduce are both Big Data technologies. They're also products of different communities and cultures, but that doesn't justify their continued separate evolution.

The MPP and Hadoop/MapReduce worlds are destined for unification. Perhaps that's why Teradata's Aster Data nCluster (http://www.asterdata.com/resources/assets/ds_Aster_Data_nCluster_4.6.pdf) mashes up SQL, MPP and MapReduce. Or why Teradata and Hortonworks (http://hortonworks.com/) (an offshoot of Yahoo's

Hadoop on Windows Azure (https://www.hadooponazure.com/) (Microsoft's cloud computing platform) and Windows Server, but also to integrate it with SQL Server business intelligence products and technologies.

Big Data is data, and it's big, whether in a hulking data warehouse or a sprawling Hadoop cluster. Data warehouse and Hadoop practitioners have more in common than they might care to admit. Sure, one group has been more corporate and the other more academic- or research-oriented. But those delineations are subsiding and the technology delineations should subside as well.

For now, expect to see lots of permutations of Hadoop and its ecosystem components with data warehouse, business intelligence, predictive analytics and data visualization technologies. In the future, be prepared to see these specialty areas more unified, rationalized and seamlessly combined. The companies that get there first will have real competitive advantage. Companies that continue to just jam these things together will have a tougher time.

RELATED TOPICS: ENTERPRISE SOFTWARE | DIGITAL TRANSFORMATION | ROBOTICS | INTERNET OF THINGS | INNOVATION | CXO

By Andrew Brust for Big on Data | March 2, 2012 | Topic: Big Data Analytics

SHOW COMMENTS

MORE RESOURCES

Hiring Kit: Network Engineer

Research from TechRepublic Premium

DOWNLOAD NOW

Hiring Kit: Video Game Writer

Research from TechRepublic Premium

DOWNLOAD NOW

2022 IT Budget Research Report: COVID-19 prompts organizations to tighten budgets

Research from TechRepublic Premium

DOWNLOAD NOW