Intel IT Business Review

One big data strategy, three Massively Parallel Processing platforms (MPP)

Let's face it: big data is still very new. And that means the infrastructure platforms on which big data analytics are performed are also relatively immature.

What it doesn't mean, however, is that organizations can't or shouldn't pursue big data success. They simply need to be resourceful and calculated as they step into the big data waters and evolve their strategies and platforms over time.

Intel IT, for example, previously employed a "one size fits all" approach to business intelligence (BI) and analytics using a centralized enterprise data warehouse (EDW).

However, with data volumes growing—much of it unstructured data from proxy, machine, server, and access logs—and more business groups wanting to mine value from those datasets, there was an emergent need for new analytical capabilities and less expensive platforms.

"Due to its cost and the type of data it contains, our EDW is largely reserved for structured enterprise data that has horizontal use across the entire company," says Chandhu Yalla, BI Engineering Manager for Intel IT. "To be honest, we don't want to pollute the EDW with unstructured data and vertically focused, one off analyses."

"By providing an array of platforms, we are helping Intel mine a broader range of data—faster, deeper, and more cost-effectively."

Chandhu Yalla,
BI Engineering Manager for Intel IT

One big data strategy, three Massively Parallel Processing platforms (MPP)

Intel IT's overarching BI and big data goal remains constant: provide the right data to the right people at the right time. The methods for achieving this goal are evolving to accommodate an increasingly wide variety of business use cases, analytics, and data types.

To support the big data needs of Intel business groups, Intel IT is now employing a multiple platform strategy. Doing so has allowed the organization to move from an expensive, "one size fits all" approach to a more cost effective, multi-tiered data warehouse architecture that better matches the requirements, opportunities, and types of data available to business groups.

"By providing an array of platforms," says Yalla, "we are helping Intel mine a broader range of data—faster, deeper, and more cost-effectively."

Three primary platforms are being utilized:

- EDW platform—for analysis of enterprise-wide, structured data;
- Apache Hadoop™ platform—for analysis of raw, unstructured data
- Low cost Massively Parallel Processing Platform (MPP)™ – for analysis of structured and semi-structured data.

Different platforms for different purposes

According to Anil Varhadkar, Enterprise Architect for Intel IT, each platform has different attributes and capabilities. Therefore, each is used for different purposes and use cases.

"Each platform has unique strengths, which allows us to marry project requirements with the platform best suited to meet them," he says.

"We have a process and a tool for evaluating each use case and for determining the most appropriate platform to employ, considering the purpose of the project, the type of data being examined, the analytics needed, data integration requirements, and who will consume the data."

The EDW platform is used for enterprise-wide analyses involving relatively small quantities of structured data. For example, examining the company's sales performance—with structured data encompassing products, regions, customers, and salespeople—would be an optimal job for the EDW.

The low-cost MPP platform is used to analyze both structured and semi-structured data for more vertical use cases and line of business (LOB) groups. IT security teams, for example, have used the MPP platform for server access log analyses to detect and prevent fraud.

"The MPP platform can support lots of users and analyses at a much lower cost than the EDW," Varhadkar explains. "And we can open it up to business groups without impacting consumption of enterprise data in the EDW."

Designed for batch processing, the Apache Hadoop™ platform is used to analyze huge amounts of unstructured data. It also acts as an inexpensive storage repository. Intel marketing teams, for example, are using the Apache Hadoop™ platform to analyze more than 6 gigabyte of unstructured log data daily.

Doing so allows Intel's marketing professional to better understand and act upon Intel.com web traffic and customers' online behaviors.

While each platform has unique capabilities and purposes, multiple platforms are sometimes utilized for certain use cases. For example, it is often advantageous to use the Apache Hadoop™ platform to process and normalize massive amounts of raw data, and then send the refined data to the MPP platform for analyses. In fact, both are being employed for the aforementioned customer insights and web analytics use case. Intel's marketing team processes 6 gigabyte of unstructured log data daily using the Hadoop platform, and then sends 2 gigabyte of refined data to the MPP platform for further analysis.

"We need both batch and real-time processing," says Yalla. "The combination of Hadoop™ and MPP platforms meets the vast majority of our big data use cases today. And as technology landscape evolves, we will continue to evaluate, evolve, and improve our big data strategies and platforms."

Brought to you by IT@Intel.