The Power of Pentaho and Hadoop in Action

Demonstrating MapReduce Performance at Scale

Introduction

Over the last few years, Big Data has gone from a tech buzzword to a value generator for many organizations. At the forefront of this trend has been Hadoop, a common framework for distributed data storage and processing that is well-suited to handle very large quantities of unstructured and semi-structured data from sources like web logs, social media, and industrial sensors. Forrester expects that two-thirds of enterprises will have deployed Hadoop or a related solution for at least one use case by the end of 2016.[1]

As mainstream enterprise interest grows for a particular technology, so does the demand for reliable scalability and performance with respect to that technology. Large, complex organizations require that a system not only meets today's requirements, but also meets future needs. For Big Data integration and analytics solutions like Pentaho, this means demonstrating an ability to perform efficiently even as data volumes rise very rapidly. This report discusses a recent scalability test conducted using Pentaho Data Integration to execute and orchestrate MapReduce jobs in Hadoop with the purpose of demonstrating sustained performance at scale.

Pentaho and Hadoop

Pentaho provides an end-to-end data integration and business analytics platform that allows users to blend, orchestrate, and analyze data from a wide array of sources including relational databases, Hadoop distributions, NoSQL stores, enterprise applications, and many others. A part of the broader platform, Pentaho Data Integration (PDI) is comprised of Pentaho's data integration engine and associated server and workstation tools. It provides a flexible, user-friendly interface for creating visual data flows for transforming and integrating data.

Pentaho began investment in the development of data integration solutions for Hadoop in early 2009 and delivered its first commercial Hadoop integration capabilities in 2010. For reference, Cloudera - the first company formed to commercialize Hadoop - was founded in 2008. Today, Pentaho offers extensive native integration capabilities for Hadoop that facilitate data ingestion, execution of complex data transformations in-cluster via MapReduce, and management of clusters through the Hadoop YARN infrastructure.

The scalability tests were focused on Pentaho Visual MapReduce – PDI's native ability to visually design Hadoop MapReduce jobs and run them in-cluster, leveraging Hadoop's distributed cache. This capability allows technical teams to accelerate time to value by avoiding complex and time-consuming hand coding of MapReduce jobs using Java. For example, in one use case a major financial institution was able to create Hadoop MapReduce jobs in Pentaho approximately 15 times faster than they could using a hand-coded approach.

[1]From Forrester research report – "Hadoop Ecosystem Overview, Q4 2014," Brian Hopkins; Published 21 November 2014

The Scalability Tests

According to IDC, all data created and copied worldwide is expected to grow by 40% per year through 2020.[2] As such, enterprises want to be sure they can accommodate major growth in data volume as they evaluate their current environment and new Big Data technology options. In particular, a key challenge for parallel processing systems, like Hadoop, is to remain ecient as work demands increase. With this in mind, Pentaho wanted to demonstrate that running Pentaho Visual MapReduce delivers consistently high performance when scaling to billions of rows and multiple Terabytes (TB) of data processed across large Hadoop clusters.

The data set used in these tests represented a common pattern of customer data, including orders and detailed line items with date, time, and geographic attributes. It is similar to the type of data an enterprise might expect to offload from a large data warehouse into a Hadoop cluster to optimize its overall data infrastructure, storage costs, and processing performance. The tests were conducted using Pentaho Data Integration and a 129-node Cloudera Hadoop cluster, deployed on Amazon Web Services cloud-based EC2 machines.

Test Case Architecture Diagram

The tested Pentaho Visual MapReduce job included operations to filter, score, and rank the data. This overall job included 2 component MapReduce jobs, each of which was comprised of a Mapper and Reducer transformation. A final transformation at the end of the job prepared a correlated top 10 output set. All jobs and transformations were create in PDI's visual drag-and-drop interface.

The data processed by PDI Visual MapReduce in the Hadoop cluster was increased by a factor of 8 over the course of the tests, with each trial approximately doubling the previous data volume processed - as indicated in the following chart:

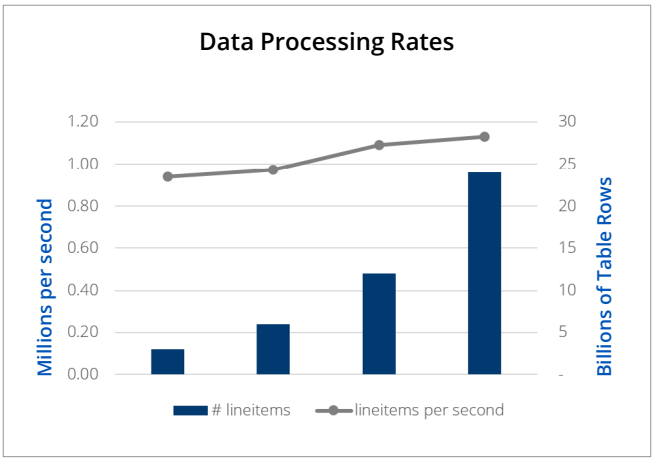
Test Data Characteristics by Trial

Trial Number	Data Volumes Processed			
	Gigabytes (GB)	Table Rows	Orders	Customers
1	471	3 Billion	750 Million	75 Million
2	948	6 Billion	1.5 Billion	150 Million
3	1,886	12 Billion	3 Billion	300 Million
4	3,773	24 Billion	6 Billion	600 Million

[2]From IDC Whitepaper – "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," Published April 2014

Results

Pentaho Data Integration delivered consistent and efficient data processing rates of approximately 1 million table rows per second over the course of the 4 tests. In fact, while data processed increased by a factor of approximately 8 times between trials 1 and 4, Pentaho's processing rate actually became 20% faster over that interval. While we would not necessarily expect improved performance at higher and higher data volumes, the overall results demonstrate that Pentaho Visual MapReduce delivers sustained performance at scale in a large Hadoop cluster with several Terabytes of data processed.



Conclusion

The rapidly evolving Big Data ecosystem presents challenges regarding how to create data processing solutions that can adapt and scale efficiently. Pentaho delivers a time-tested, cost-effective data and analytics platform that meets these challenges in a way that is efficient for organizations to implement and maintain.

Pentaho Data Integration has shown its ability in this test case to compliment large Hadoop clusters by sustaining a high rate of processing over increased dataset sizes of multiple terabytes. Overall, PDI provides the enterprise with the ability to accelerate IT productivity in bringing Big Data projects into production, while ensuring that their environment is future-proofed against large and unexpected increases in data volume.