An Oracle White Paper
September 2012

Big Data and Natural Language: Extracting Insight From Text

Table of Contents

Executive Overview

There is a wealth of information to be extracted from natural language, but that extraction is challenging. The volume of human language we generate constitutes a natural Big Data problem, while its complexity and nuance requires a particular expertise to model and mine. In this paper we illustrate the impressive combination of Oracle Big Data Appliance and Digital Reasoning Synthesys software. The combination of Synthesys and Big Data Appliance makes it possible to analyze tens of millions of documents in a matter of hours. Moreover, this powerful combination achieves four times greater throughput than conducting the equivalent analysis on a much larger cloud-deployed Hadoop cluster.

Introduction

Extracting useful information from any piece of human language is a challenging problem. Sentences must be parsed, entities identified, and relationships must be modeled. Consolidating large amount of human language into a body of information that can be leveraged to derive business value is an even more challenging and complex problem. As the volume of language to analyze increases, so too does the complexity of that analysis and the need for scalable, parallel processing of that language. Indeed, extracting actionable knowledge from large volumes of unstructured human language is a quintessential Big Data problem.

In this paper, we explore the power of applying Digital Reasoning Synthesys semantic analysis software on Oracle Big Data Appliance to extract and resolve billions of entities, facts and relationships from millions of pieces of human language. The paper describes the following topics:

- An introduction to Oracle Big Data Appliance, the only engineered system optimized for enterprise Big Data

- An introduction to Digital Reasoning Synthesys, which provides advanced semantic analysis and leverages the Hadoop ecosystem to scale over hundreds of millions of documents and billions of entities, facts and relationships.

- The motivating problem of using human resources to flag documents for audit and an explanation of the testing conditions used in our tests

- Testing results for the Big Data Appliance and Synthesys on document sets of 1, 11, 50, and 233 million documents

- The dramatic improvements to cost and efficiency by applying this joint solution in the financial market – where over $400M is spent annually to prepare unstructured documents for back office applications[1].

Oracle Big Data Appliance

Oracle Big Data Appliance is an enterprise-class engineered system to provide an optimized and complete solution for Big Data workloads. The appliance contains 18 Sun servers with a raw storage capacity of 648TB. Each server contains two 6-core Intel© Xeon© CPUs and 48GB of memory. Oracle Big Data Appliance runs Cloudera's Distribution including Apache Hadoop (CDH). CDH provides the #1 Hadoop-based distribution in commercial and non-commercial environments. CDH provides a comprehensive Hadoop-based environment on which to develop Big Data solutions.

Networks which simultaneously enable rapid data acquisition and massive analytics operations are necessary to keep up with the unprecedented volume and velocity presented by many Big Data problems. Oracle Big Data Appliance features high-speed networking for both fast internode data streams and for data ingestion. Oracle Big Data Appliance provides a 10GbE client network for data ingestion. This allows new data to be consumed as quickly as it arrives at the datacenter. To ensure that analytics operations on Oracle Big Data Appliance do not hinder data acquisition, internal communication in the appliance utilizes an Infiniband fabric capable of 40Gb/s. As internal networking is important for HBase-backed applications, the Infiniband fabric provides the best possible communication backbone for Synthesys' workload. Moreover, the Infiniband fabric provides the fastest possible communication with other Oracle Engineered Systems, ensuring rapid consolidation of structured and unstructured data sources.

[1] Source: Shore Communications Research, August 2012

Synthesys

Synthesys is a software platform for automatically making sense of Big Data. By focusing on entities in a broad set of documents rather than individual documents, Synthesys avoids the need for analysts to "read to understand," and takes them directly to the information they seek. Synthesys ingests and resolves both structured and unstructured data, but its unique benefit is its ability to automatically distill critical information in context (e.g., people, places, events, relationships) from massive amounts of unstructured data. Integrated with Cloudera's Distribution including Apache Hadoop (CDH), Synthesys uses patented algorithms and machine learning methods in a three-phase process called "Read-Resolve-Reason." Synthesys analytics scale horizontally to virtually any corpus size.

Via a combination of model-building and unsupervised learning, Synthesys discovers the information as a human would – in context and without the need for a pre-defined ontology. Synthesys understands related terms and associations (e.g., synonyms and acronyms) that lead to more accurate entity recognition, improved ability to deal with "coded" language, and a contextual understanding of concepts across large sets of text.. Extracted information is stored in a knowledge graph which continuously processes data and more deeply refines it. The knowledge graph exposes a RESTful API, allowing easy integration with enterprise applications and workflows.

Motivation

In the financial services industry, audits are necessary to mitigate risk. Traditionally, teams of people are needed to read and "tag" documents in order to make the details in these documents accessible for search tools. The staff must carefully read each document, identifying the people, places and concepts contained within. This information must be collected into a set of facts. The set of facts must then be analyzed for signs of risk. This labor-intensive process constitutes a cycle of: reading documents, gathering facts, and assessing risk. Recent research shows that over $400M is spent by enterprises in the financial services market to support these teams and the tools that they use – a significant and growing cost.

Suppose a major financial firm employs 300 people to conduct this work. If one assumes the average analyst reads at a rate of 200 words per minute and the average document contains 10 kilobytes of unstructured text (i.e. 2.5 pages), a team of 300 people can effectively analyze

4 to 5 million documents per year. Not only is this process time consuming and expensive, but it scarcely addresses exponentially growing amount of data many firms currently produce. .

As an alternative, we examine the application of Synthesys on Oracle Big Data Appliance to the process of reading, tagging and analyzing documents. Our test set consists of three data sets of different sizes: 11 million documents, 50 million documents, and 233 million documents. Each of these far exceeds the capability of our theoretical 300-person team, but are conceivable set sizes for a major financial firm.

Our test environment is an Oracle Big Data Appliance running Cloudera's Distribution including Apache Hadoop version 3u4 (CDH3u4). Digital Reasoning's Synthesys application is deployed on the Big Data Appliance and uses MapReduce and HBase services to process document sets stored in HDFS. For each document set, we examine the total time needed to ingest, analyze, and write facts into a searchable knowledge base. Additionally, we compare throughput against a 30-node Hadoop cluster in Amazon's Elastic Compute Cloud (EC2) using High Memory Quadruple Extra Large instances.

Results

For each of our test sets, we find that Synthesys deployed on the Big Data Appliance dramatically reduces time necessary for ingesting and analyzing documents when compared to a deployment on Amazon Elastic Compute Cloud cluster with higher parallelism. As shown in Figure 1, using Synthesys our hypothetical financial services company can process a year's worth of documents before lunch.

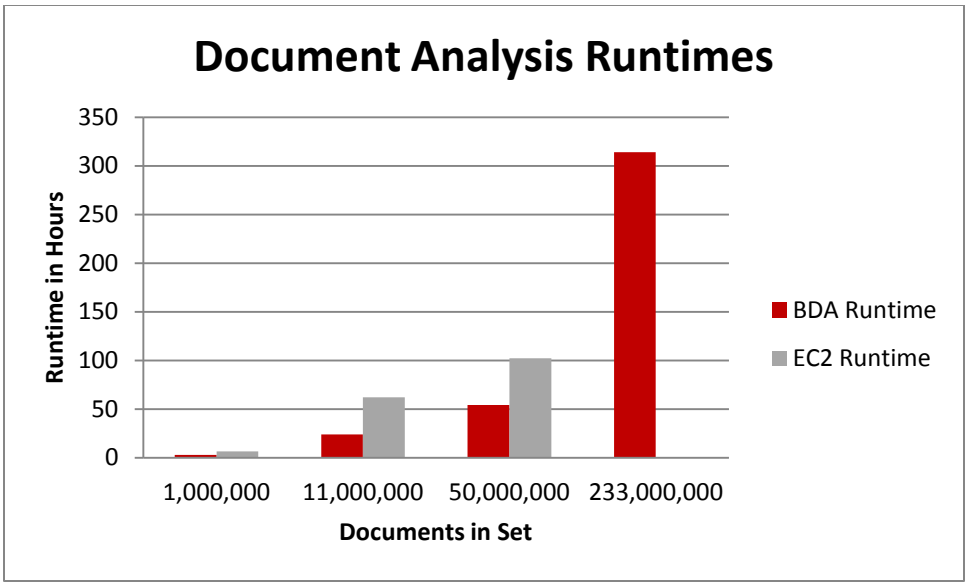

Figure 1: Compared runtimes for document analysis on the Big Data Appliance and an Amazon EC2 cluster. Not only does the Big Data Appliance dramatically improve runtime, but Synthesys scales at less than a factor of 1 as set-size increases. Lower runtimes equate to more documents analyzed per hour, except the the case of 233M documents, which did not run on EC2.

Document Processing Runtimes

As the size of the document set increases, so too does the time necessary to ingest and analyze it. However, leveraging the high-performance infrastructure of the Big Data Appliance, massive document sets are processed in a matter of hours, not days. Note that Synthesys processes the 11 million document set in approximately 24 hours, yet processing time scales at only a fraction of the document set size. In just over twice the time (54 hours) Synthesys can process five times (50 million) the number of documents. During our test Synthesys running on the Big Data Appliance was used to process 233 million documents; far more than Digital Reasoning had ever before attempted to process in a single test.

More importantly, running Synthesys on the Big Data Appliance far outperforms a much larger cluster on Amazon's EC2 platform. Despite the higher parallelism the Amazon EC2 30-node cluster with 300 map slots, required approximately 62 hours to process 11 million document and 102 hours to process 50 million. The single-rack Big Data Appliance used in this test provides 180 map slots, delivering much higher throughput and a shorter runtime to finish the test.

Analysis Throughput

The improved runtime performance the Big Data Appliance show compared to a larger cluster in Amazon EC2 shows that Synthesys achieves greater throughput on the Big Data Appliance. Figure 2 illustrates this difference using a normalized throughput for both clusters. Throughput is normalized as the number of bytes processed per second per map slot. This allows the reader to do an apples-to-apples comparison between the two systems.

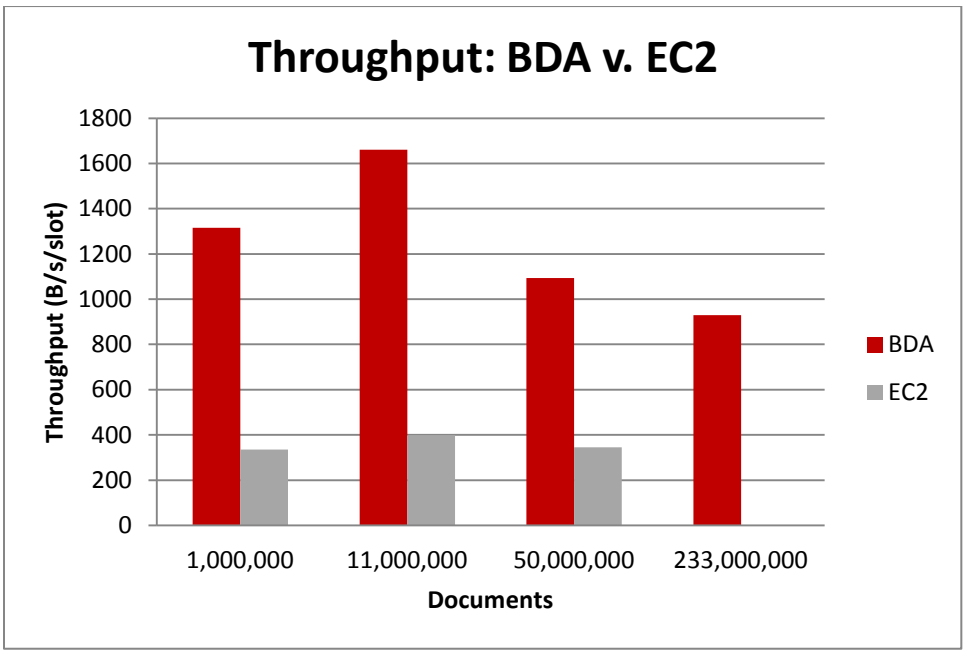

Figure 2: Compared per-slot throughput for the Big Data Appliance and a comparable cloud-based cluster. The Big Data Appliance achieves approximately 4 times the throughput. Higher throughput numbers equate to more documents processed per hour.

Using this normalization, we find that the Big Data Appliance achieves per-slot throughputs 3 to 4 times faster than the Amazon© EC2© cluster. At this per-slot throughput, we project a Big Data Appliance installation with a comparable number of nodes to the EC2 cluster could be expected to perform the 50 million document analysis in just over 1 day. Furthermore in this test, the combination of Oracle Big Data Appliance and Synthesys demonstrated the capability

of persisting more that 4 Billion entities, facts and relationship[2], a staggering amount of knowledge that far exceeds the capacity of even the most capable human analyst.

Conclusion

The tests find that Synthesys on Oracle Big Data Appliance is capable of reading and analyzing in a few hours the same number of documents that a team of 300 people could analyze in a year. There is tremendous value which results from freeing human resources to spend more of their time reasoning about risk and making valuable decisions. Moreover, Synthesys on Oracle Big Data Appliance has 4 times greater per-slot throughput and dramatically reduces runtimes compared to a larger, more costly deployment on Amazon EC2. The combination of the two technologies provides more documents processed per hour for fewer dollars spent than any other available solution for natural language processing on tens of millions of documents. Analysis teams can potentially save millions of dollars a year and better devote human resources to the more nuanced business of reasoning about risk exposure.

[2] This quantity only reflects the results of this test and not a limitation. Larger tests and customer installations demonstrate significantly larger numbers of entities, facts and relationships.

Big Data and Natural Language: Extracting Insight From Text
September 2012
Author: Dan McClary, Ph.D.
Contributing Authors: Dave Danielson, Digital Reasoning Systems

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together