Divide & Recombine for Big Data Analysis for Cybersecurity

Application on DNS Blacklist Query Study

Ashrith Barthur [1], William S. Cleveland [1], and John Gerth [2]

[1] CERIAS, Purdue University

[2] Department of Computer Science, Stanford University

Abstract

D&R is a statistical approach to big data that provides comprehensive, detailed analysis. This is achieved because almost any analytic method from machine learning, statistics, and visualization can be applied to the data at their finest level of granularity. D&R also enables feasible, practical computation because the computations are largely embarrassingly parallel. Our work has two core threads.

- Tailor the D&R environment to analyse big data in cybersecurity.
- Apply this tailored environment the Spamhaus traffic at the Stanford University mirror.

D&R Statistics

In D&R the data is divided into subsets by a statistical division method. Each of a collection of analytic methods used in the analysis is applied to the subsets independently without communication among the computations. A statistical recombination method is applied across the outputs of the application of each analytic method to form the final D&R result. Work in D&R statistical theory and methods seeks division and recombination methods that “optimize” the statistical accuracy of the final D&R research.

The D&R Computational Environment

D&R can readily exploit powerful distributed parallel computational environments running on a cluster. The front-end of our D&R environment is R. The back-end is Hadoop on a Linux cluster. We are building layers in between that together with front and back are the D&R computational environment. Hadoop does the computations for division into subsets, application of analytic methods to subsets, and some of the recombinations.

RHIPE, the R and Hadoop Integrated Programming Environment, enables the analyst to program wholly from within R. The analyst writes R code for division, application, and recombination, and passes it to RHIPE R routines that manage communication with Hadoop.

RHIPE protects the analyst from many details of Hadoop. Our datadr R package is a layer between R and RHIPE that provides more protection. The analyst does not have to manage key-value pairs, Map, and Reduce, and thinks solely in terms of division methods and recombination methods developed as part of our statistics research.

Application of The D&R Environment to Spamhaus Big Data

We have been collecting data at Stanford for 6 weeks. So far, there been 2.05 billion queries. We capture just response packets which have needed query packet data. We use 5 response-packet fields.

- Timestamp
- Source IP address
- Response code
- IP Address or domain name queried
- Spamhaus response

From these fields we form 17 variables that are analyzed:

- Timestamp
- Source IP address
- IP Address or domain name queried
- Query type (IP address or domain request, query error)
- Response code type
- State of the query (blacklist, whitelist, or no information)
- IP address spam flag
- IP address CSS spam flag
- IP address exploit flag
- IP address ISP policy violation flag
- IP address Spamhaus policy violation flag
- IP address individual mail whitelist flag
- IP address transaction mail whitelist flag
- Domain address spam flag
- Domain address spam redirection flag
- Domain address individual mail whitelist flag
- Domain address transaction mail whitelist flag

We create R dataframe objects for our analysis with 17 columns. Each dataframe object holds 50,000 queries to optimize Hadoop performance.

Collection Summary Statistics

Here are summary statistics for our first 6 weeks.

6- Week Collection Statistics	Number of Occurrences
All queries	2,051,268,078
IP address queries	1,679,915,476
Domain address queries	370,958,042
Queries with incorrect format	122,896
Queries unanswered by Spamhaus	1,657,054,290
Queries for text information	124,465,836
IP addresses blacklisted for exploits	183,497,081
IP addresses blacklisted for violating ISP policy	24,326,177
IP addresses blacklisted for violating Spamhaus policy	101,221,765
IP addresses blacklisting spamming	11,728,943
IP addresses blacklisted for spamming by CSS detection	10,420,020
Domain addresses blacklisted for spamming	12,503,879
Domain addresses blacklisted for spam redirection	589,814
Blacklisted domain address queries	251,694,934
Whitelisted IP and domain address queries	15
IP addresses whitelisted for sending individual emails	15

- 81% of the queries that come to Spamhaus are for IP address verification.
- 80% of queries are unanswered by Spamhaus. This prompts us towards the idea that a huge part of the malicious internet might still be invisible.
- 9% of the queries indicate hosts that are responsible for transferring malicious programs.
- 0.005% of the queries are incorrectly formatted. This might be an indication of misuse.