File Transfer Log Analysis: A meteorological data center case study

Ivo Koga^{a1}, Eugenio S. Almeida^a

^aNational Institute for Space Research, Cachoeira Paulista, SP, Brazil

Abstract

The Center for Weather Forecast and Climate Studies (CPTEC) from the Brazilian National Institute for Space research (INPE) provides valuable meteorological information through file transfer services. Logs usually provide information related to services usage, furthermore they can show meaningful insights from data, user's location and predict data infrastructure growth. Due to the high demand of these services, it is very difficult to keep track of the information stored in the file transfer service logs. In this work, we captured file transfer logs and used the open source Elasticsearch, Logstash and Kibana (ELK) stack for data analysis. Logstash serves for data collection and transport pipeline, Elasticsearch allows data index and Kibana provides a user-friendly interface for queries, results, and visualizations. Our preliminary results with File Transfer log data from INPE/CPTEC show interesting details from data: high rate of access due to a misconfigured script, activity of real and anonymous users and more. We would like to emphasize that this work can be extended and used with other kinds of data, providing useful insights to researchers from many other scientific areas.

Keywords: File transfer service, Meteorological data, Elasticsearch, Data Analysis.

1. Introduction

Data centers usually hosts many computational resources (computing systems, networks and storage devices) and provide different kinds of services. Its management is a non-trivial task. The management team need to be aware of the current use of services and at the same time provide information about processing problems, account management, problems detected and solved, etc [1].

Once a problem happens, the data center management team need to analyze different sources of information in order to discover its causes. Another concern rises when treating log data format heterogeneity, since they

¹E-mail Corresponding Author: ivo.koga@inpe.br

38 come from many different sources and services. They also need to deal with
39 rapid growth of data logs size and the increasing need for real-time analysis.

40 The complexity in acquiring and storing log files rises proportionally to
41 the increase of computational resources and services. Getting, analyzing and
42 making all the information useful requires a coordinated effort, especially in
43 an overwhelming data center management context. The proper information
44 management is key for problem solving.

45 A meteorological data center usually hosts a supercomputer to execute
46 atmospheric numerical models and additional computational resources to
47 pre/post-process meteorological data. It requires a file transfer service to
48 bridge meteorological data from different sources, and to everyone who needs
49 access to meteorological products.

50 The CERN Data Centres production IT monitoring infrastructure uses
51 Elasticsearch as one of its components [2]. Moh [3] has used the Bayes Net
52 model (a supervised machine learning method) with the ELK stack to detect
53 SQL injection attacks that cause severe security problems on any web-based
54 application hosted on the Internet or on the cloud.

55 Awad et. al [4] used the ELK (Elasticsearch, Logstash, Kibana) ecosys-
56 tem to parse Apache Tomcat access logs and to infer the repeatable user be-
57 havior, which is transformed into Customer Behavior Model Graph (CBMG).

58 Kononenko [5, 6] has created a tool called DASH that provides informa-
59 tion about patch review process for Mozilla developers. DASH used Elastic-
60 search instead of other traditional storing tools, such as traditional relational
61 databases. This approach improved the query performance, allowing the de-
62 velopment review process to occur in real-time.

63 This paper aims to evaluate the ELK stack in a context of a meteoro-
64 logical data center, which led to the development of a framework to ingest,
65 process and analyze meteorological data center logs. We used file transfer
66 logs as a proof of concept to get insights. These files are especially important
67 since they show the input and output data flow from the FTP data transfer
68 service. The log files were collected at the Center for Weather Forecast and
69 Climate Research (CPTEC) from the Brazilian National Space Research
70 Institute (INPE).

71 This paper is organized as follows: Section 2 provides an overview of
72 indexing and search technologies and section 3 details the framework archi-
73 tecture to process meteorological data center logs. Section 4 provides the
74 results and analysis from the FTP logs data and section 5 concludes the
75 paper.

2. Indexing and Search platforms

Search engines such as Apache Lucene [7] provides index and query functionalities. However, the need of new features, such as language independency, scalability, User Interface (UI) and others, promoted the development of search platforms. Two of the most widely used open source platforms based on Apache Lucene are Solr [8] and Elasticsearch [9].

Solr search features include faceted search, clustering, database data import handler, rich document handling (PDF, Word, etc), and geospatial search. Written in Java, it has a highly scalable and fault tolerant architecture, providing distributed search and index replication. It provides REST-like HTTP/XML and JSON APIs, allowing its usage with any programming language that supports these interfaces [8].

Elasticsearch is part of the ELK stack and differs from Solr in its unity of information, which is JSON based. It allows full-text search, indexing, replicas, automatic nodes discovery, recovery among other functionalities [10, 11]. Logstash collects and parses data from many different kinds of sources. Kibana is used for GUI display and dashboard visualization.

There is a lot of commercial search engine based solutions, many of them are built on top of open source projects. For instance, Cloudera Search [12] provides search data inside their Hadoop and HBase products using open source Apache Solr.

Among the open source search engine platforms we chose Elasticsearch, which provides a full-stack development environment. It has many plugins that allows the ingestion of different kinds of data, from environmental sensor data to database logs and a wide variety of visualizations, that can be integrated in a dashboard. Finally, it also has a strong community support which reflects in their continuously evolving infrastructure.

3. ELK applied to INPE/CPTEC FTP log data analysis

ProFTPD (short for Pro FTP daemon) is a free and open source software used as an FTP server. The transfer log file (xferlog) from ProFTPD contains interesting logging information. It shows useful information about service usage such as transferred filename, filesize, date and time. The xferlog default format of this file has the following information:

- **Current-time:** current local time in the format “DDD MMM dd hh:mm:ss YYYY”, where DDD is the day of the week, MMM is the month, dd is the day of the month, hh is hours, mm is minutes, ss is seconds, and YYYY is year in Gregorian calendar.

- 116 • **Transfer-time:** total time in seconds for the transfer
- 117 • **Remote-host:** remote host name
- 118 • **File-size:** size of transferred file in bytes
- 119 • **File-name:** name of the transferred file
- 120 • **Transfer-type:** a = ascii; b = binary
- 121 • **Special-action-flag:** C = compressed; U = uncompressed; T =
122 tar'ed; _ = no action was taken
- 123 • **Direction:** direction of the transfer (o = outgoing; i = incoming; d
124 = deleted)
- 125 • **Access-mode:** method by which the user is logged in (a = anony-
126 mous; r = real) username
- 127 • **Service-name:** ftp
- 128 • **Authentication-method:** 0 = none; 1 = RFC931 Authentication
- 129 • **Authenticated-user-id:** user id or '*'
- 130 • **Completion-status:** c = complete; i = incomplete

131 We have analyzed data from the ProFTPD service of CPTEC from
132 28/Feb/2016 to 02/Mar/2016. A four core server @2.27GHz with 4 GB
133 of memory and 5.4 TB of storage was installed exclusively to store log data.
134 We have used these data to provide the analytics presented in this paper.

135 To tackle the problems of storing and analysing data we have used the
136 open source ELK stack to parse (Logstash 2.3.2), index (Elasticsearch 2.2.1)
137 and analyse (Kibana 4.4.2) FTP log data. We have programmed Logstash to
138 read the xferlog file, which comes in ASCII format. Then we have converted
139 the date to decimal date format, and the transfer time and the file size
140 to number. Finally, we forward the transformed data to Elasticsearch for
141 storage and indexing. In Kibana, we selected a data subset for analysis,
142 presenting the results in format of graphics.

143 144 **4. Results and discussions**

145 After ingesting and indexing the FTP log data in Elasticsearch, we have
146 generated different types of graphics in Kibana to understand users' behavior

147 on remote transferring files using FTP services. Next, we have included
148 all the graphics in a Dashboard to provide an overview of the analyzed
149 information.

150 Figure 1 shows a pie chart with an aggregated view of FTP service usage.
151 Starting from the inner circle (1), it shows that all the 2 million documents
152 collected are from the file analysed: xferlog (100%).

153 The second and third circle represents the service name (ftp) and special-
154 action-flag (_ symbol means that no action was taken), respectively, meaning
155 that all the documents are from FTP and has no special flag set. The fourth
156 circle shows that the majority of data access (92.58%) are from authenticated
157 real (r) users and a few (7.42%) from anonymous (a) users.

158 The fifth circle shows that all anonymous users are only allowed to down-
159 load files. Real (authenticated) users had 48.84% of download (outgoing [o]),
160 49.64% of upload (incoming [i]) and 1.52% of removals (deleted [d]).

161 From the sixth circle, we verified that anonymous users download 98.11%
162 of the files in binary format and 1.89% in ASCII format. The real users
163 downloads only binary data (100%). The uploads were 8.63% binary and
164 91.37% ASCII and they deleted 54.91% of the binary and 45.09% ASCII
165 files.

166 Through our analysis, we detected that the majority of the uploads were
167 binary data transfer but the data was stored in ASCII format. The majority
168 of downloads from anonymous users were also in binary format. This infor-
169 mation leads us to infer that the ASCII data transfer mode is not usually
170 used, i.e. users are uploading and downloading data using mostly binary
171 mode. This may be due to the default configuration of FTP clients.

172 The number of accesses performed by real users (blue) is higher than by
173 anonymous (yellow) users as shown in Figure 2. It indicates that users from
174 institutions with formal partnership with INPE/CPTEC transfer more data
175 with higher frequency. Figure 2 also shows a peak in data access, which
176 should indicate an anomalous access in 29/Feb/2016 around 14:00 from real
177 users.

178 The Figure 3 graph shows the number of files downloaded (o), uploaded
179 (i), and deleted (d) obtained from the direction field of xferlog. In this
180 particular case, the amount of file uploads (blue) and downloads (red) are
181 quite similar.

182 In Figure 4 we explore more details of the information provided by Fig-
183 ures 2 and 3, focusing on the analysis for the date 29/Feb/2016. For each
184 bar, we plotted the ten hosts with more access. In most cases, two hosts A
185 (green) and B (blue) have the highest number of file transfers. The excep-
186 tion is the peak between 14:00 and 15:00, where we have a different host,

Figure 1: Overview information of FTP log.

187 represented in orange, performing more than 66% of overall access.

188 We splitted the direction field in three graphs (Figure 5): deleted (deleted
 189 files, in the top), incoming (uploaded files, in the middle) and outgoing
 190 (downloaded, in the bottom). Most of the downloads and uploads belong
 191 to two hosts located in different institutions represented by the blue (down-
 192 loads) and red (upload) bars, shown in the middle and bottom bar charts.

193 Analysing these charts, we observed an inadequate behavior. Getting
 194 into the details using the data table visualization feature of Kibana, we had
 195 a detailed visualization, in tabular format, of the results. We identified that
 196 uploads come from an external institution, which is transferring duplicated
 197 data to the FTP server. The other host, which downloads data, is inside of
 198 the meteorological data center and downloads the duplicated data.

199 The Figure 6 plots the top ten hosts that access the FTP server for the
 200 analyzed period, acknowledging the predominance of access by two hosts,
 201 represented by 1 and 2, which we observed that one is an internal and other
 202 external. This shows that they exchange about 350,000 files between the
 203 servers. We observed that some of the files were repeatedly transferred

images_used/fig2.png

Figure 2: Number of access by real (blue) and anonymous (yellow) users.

204 between them which indicates the mistake found previously on the usage of
205 the file transfer service.

206 All these graphs were integrated in a Dashboard (Figure 7), presenting
207 an overview of the FTP log service status. Kibana provides mechanisms
208 to visualize historical or near real-time, i.e. it is possible to configure the
209 graphs to be recreated in a constant period of time in order to provide the
210 current happenings. Thus, FTP server events can be presented and analyzed
211 properly. It is also possible to customize the dashboard to show past events
212 of interest.

images_used/fig3.png

Figure 3: Number of file downloaded (red), uploaded (blue) and deleted (green).

213 Conducting data analysis using the ELK over FTP log data allow impor-
214 tant informations to guide storage capacity plan, identification of threats,
215 promote data transfer better usage and support the correct use of technolo-
216 gies for data transfer.

217

218 **5. Conclusion**

219 Every day the INPE/CPTEC FTP service has an enormous amount
220 of access from internal/external users, which download, upload and delete

images_used/fig4.png

Figure 4: Host access from 29/Feb/2016.

221 data. Users' behavior are stored in logs, which has many information of this
222 important service usage.

223 The traditional FTP log analysis restricts the insights related to the use
224 of the FTP server and the data, since it is performed manually. Additionally,
225 automatic software operations generate the majority of the download and
226 upload requests. A misconfigured program can lead to huge problems.

227 Data transfer is the users' objective, which generally are not conscious
228 about FTP services misuse. Monitoring this service is key to prevent prob-
229 lems and to the better maintenance of the service.

images_used/fig5.png

Figure 5: Downloads, uploads and deletions from 29/Feb/2016.

230 In this paper we analyzed the potential of the ELK stack to provide
231 different insights related to the FTP server usage. Its functionalities includes
232 getting log from the FTP service, building a log collection, storing and
233 indexing, and visualizing different information from FTP log data.

234 We started by configuring Logstash to collect, transform and send FTP
235 log data to Elasticsearch for storage and indexing. We used Kibana to build
236 different graphics combined in a Dashboard for a complete overview of FTP
237 service data usage.

238 This study used data from 28/Feb/2016 to 02/Mar/2016 period. Results

images_used/fig6.png

Figure 6: Host access (whole period).

239 showed different kinds of insights such as common problems on repetitive
240 file transfers between internal/external servers, files stored without dele-
241 tions that leads to an ever increasing storage, and the predominance of real
242 (authenticated) users access over anonymous users.

243 This study revealed important behavior of the CPTEC meteorological
244 datacenter FTP server usage. We intend to continue this work searching for
245 new insights and adding more data sources.

images_used/image07.png

Figure 7: Dashboard from 29/Feb/2016.

References

- 247 [1] C. Gough, I. Steiner, and W. Saunders. Data center management. In
248 *Energy Efficient Servers*, pages 307–318. Springer, 2015.
- 249 [2] P. Andrade, B. Fiorini, S. Murphy, L. Pigueiras, and M. Santos. Mon-
250 itoring evolution at cern. In *Journal of Physics: Conference Series*,
251 volume 664, page 052002. IOP Publishing, 2015.
- 252 [3] M. Moh, S. Pininti, S. Doddapaneni, and T. Moh. Detecting web at-

- 253 tacks using multi-stage log analysis. In *2016 IEEE 6th International*
254 *Conference on Advanced Computing (IACC)*, pages 733–738. IEEE,
255 2016.
- 256 [4] M. Awad and D. Menascé. Automatic workload characterization using
257 system log analysis. In *Computer Measurement Group*, pages 733–738,
258 2015.
- 259 [5] O. Kononenko, O. Baysal, R. Holmes, and M. W. Godfrey. Mining mod-
260 ern repositories with elasticsearch. In *Proceedings of the 11th Working*
261 *Conference on Mining Software Repositories*, MSR 2014, pages 328–
262 331, New York, NY, USA, 2014. ACM.
- 263 [6] O. Kononenko, O. Baysal, R. Holmes, and M. W. Godfrey. Dashboards:
264 Enhancing developer situational awareness. In *Companion Proceedings*
265 *of the 36th International Conference on Software Engineering*, ICSE
266 Companion 2014, pages 552–555, New York, NY, USA, 2014. ACM.
- 267 [7] M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action,*
268 *Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co.,
269 Greenwich, CT, USA, 2 edition, 2010.
- 270 [8] D. Smiley and E. Pugh. *Apache Solr Enterprise Search Server*. Packt
271 Publishing, 3 edition, 2015.
- 272 [9] H. Akdogan. *Elasticsearch Indexing*. Packt Publishing, 2015.
- 273 [10] S. Bagnasco, D. Berzano, A. Guarise, S. Lusso, M. Masera, and
274 S. Vallero. Towards monitoring-as-a-service for scientific computing
275 cloud applications using the elasticsearch ecosystem. In *Journal of*
276 *Physics: Conference Series*, volume 664, page 022040. IOP Publishing,
277 2015.
- 278 [11] S. Chhajed. *Learning ELK Stack*. Packtpub Co., 2015.
- 279 [12] Cloudera search documentation. [http://www.cloudera.com/](http://www.cloudera.com/documentation.html)
280 [documentation.html](http://www.cloudera.com/documentation.html), 2016.