Exploratory study of the ELK stack for meteorological observation system data analysis

Eugênio S. Almeida[a 1], Ivo Koga [a], Márcio A. A. Santana[a], Patrícia L. O. Guimarães[a], Luciana M. Sugawara[a] and Tero Eklin[b]

[a]National Institute for Space Research, Cachoeira Paulista, SP, Brazil
[b]Finnish Environment Institute (SYKE), Finland

Abstract

Different kinds of sensors compose a meteorological observation system that measures meteorological variables. Sensors can collect data for a long period of time in a high sampling frequency. Some meteorological parameters can be determined by making measurements that ranges from a few seconds to annual measurements which depends on the kind of equipment and application needs. In this scenario, data management is not a trivial task due to heterogeneity, large amount of data and also to the usage of proprietary software for data gathering and handling. We used a data acquisition system (datalogger) to collect and store data from a thermo-baro-hygrometer, and a pyranometer, which were calibrated previously in the laboratory. This paper aimed to analyze the open source Elasticsearch, Logstash and Kibana (ELK) stack to capture, transform, enrich, store, index, select relevant time slots and generate graphs that were integrated in a dashboard for combined visualization and analysis. Additionally, we explored its capacity to embed metadata from sensors and correct data based on a calibration certificate, also showing some relevant graphics. In this weather application, we observed that this set of computational tools are well suited to manage the daily difficulties in handling meteorological data and metadata.

Keywords: Meteorological observation system, sensors, data analysis, metrological metadata.

1. Introduction

An environmental monitoring system may be composed of different sensors. Each sensor produces information, as electrical or optical signals, about various environmental variables (surface wind speed, wind direction, air temperature, humidity relative, barometric pressure, precipitation, etc.) from the area where it is installed [1].

[1]E-mail Corresponding Author: eugenio.almeida@cptec.inpe.br

The sensor measurement frequency and data collection period are quite variable. Each meteorological variable requires different data gathering frequencies and periods, which depends on the equipments used and the application.

The World Meteorological Organization (WMO) [2] documents suggest some parameters. Others are defined depending on the need or according to the manufacturer's specification. The limitation given by the instrumentation must follow some standardization. Often it is the instrument that reads and collects data that limits the measurements and not the sensor.

The data accessibility depends on the availability of computational resources and owner's desire. After overcoming these problems, the available period depends on the application, generally ranging from years to decades.

Governmental institutions often make the meteorological information available quickly using the Global Telecommunication System (GTS) of the WMO while research projects only after their use. The data are usually available offline as files, encoded in ASCII format. Some institutions provide graphics with values of selected environmental variables of the past few days. Data analysis often occurs offline, requiring the use of spreadsheets or specific software (Figure1).

Figure 1 - Common collection and analysis mode.

The requirement of having observational networks acquiring and recording regular observations is insufficient to provide accurate data that supports climate monitoring and service provision. The data must be homogeneous and accompanied by good supporting information (metadata), which can be sensor information and calibration information. Additionally it must also be properly quality-controlled, archived and easily accessible [3].

Numerous factors affect the field measurements and laboratory tests [4]. Some of them are considered systematic errors and can be used to correct the measurement by a correction factor (C). Others are considered random, named uncertainty contributions (U). The Equation 1 presents the measurement result (MR), defined by the measurement (M) corrected by the correction factor (C) and the range of uncertainty (U).

$$MR = M + C \pm U \tag{1}$$

The following example explains the variable value presentation with its uncertainty: MR = 23.1 °C ± 0.2 °C, where M = 23.1 °C is the measurement corrected by a C factor with an uncertainty of 0.2 °C.

Error correction, calibration and other kinds of metadata is fundamental to a proper environmental study. Most of the data generated by environmental observation systems provide just measurement data captured by sensors, without sharing information about the data itself (metadata). The information that describes data is essential on its treatment and usually is unavailable for quick and easy access.

The open source Elasticsearch, Logstash and Kibana (ELK) stack has great potential to analyze historical and near real-time data: geo-identify the website users [5], real-time analytics on Electric Vehicle (EV) mobility and behavior [6], collection and indexing of Tweets with a geographical focus [7], log management [8].

This work aimed at the analysis of the ELK stack to capture, process, store, index and display data from a meteorological observation system. Additionally we explored its capacity to embed metadata from sensors and correct data based on a calibration certificate, also showing some relevant result graphics.

This paper is organized as follows. Section 2 describes the meteorological observation system used to collect data in this paper. Section 3 presents the infrastructure used to handle the problems of capturing, analyzing and providing data. Section 4 presents the results and discussions. Finally, we conclude the paper in Section 5.

2. The weather observation system

In order to provide information for our analysis, we used the weather observation system deployed in the city of Cachoeira Paulista, SP, Brazil and located at geographical coordinates 22°41'21.12" S and 45°0'22.96" W. The Figure 2 shows equipments used to collect the environmental data. It consists of a thermo-baro-hygrometer Vaisala PTU303[2], a pyranometer Kipp & Zonen CM22[3] and a datalogger Campbell CR1000[4].

[2]http://www.vaisala.com/en/products/pressure/Pages/PTU300.aspx
[3]http://www.kippzonen.com/Product/15/CMP22-Pyranometer#.V3wJEVfj80o
[4]https://www.campbellsci.com/cr1000

Figure 2 - The weather observation system.

The thermo-baro-hygrometer has sensors for measuring the atmospheric pressure (hPa), the air temperature (°C) and the air relative humidity (%). The pyranometer measures the global solar irradiance in watts per square meter (W/m^2) on a plane surface and a spectral range from 300 to 3000 nm. Additional sensors information (metadata) are:

- **Air temperature sensor**: Vaisala brand, PTU303 model, S/N F2910008, calibration certificate LIM 078-14, valid until 01/15/2016;
- **Air relative humidity sensor**: Vaisala brand, PTU303 model, S/N F2910008, calibration certificate LIM 079-14, valid until 01/15/2016;
- **Atmospheric pressure sensor**: Vaisala brand, PTU303 model, S/N F2910008, calibration certificate LIM 029-14, valid until 11/13/2015;
- **Solar irradiance sensor**: Kipp & Zonen brand, modelo CM22 model, S/N 040102, calibration certificate PMOD/WRC-2014-C-57, valid until 08/25/2016;
- **Datalogger**: Campbell brand, CR1000 model, S/N 13220, calibration certificate LIT06-CPT-CC-10011, valid until 05/20/2017.

The thermo-baro-hygrometer sensors are connected on the digital ports of the datalogger and the pyranometer on the analog port (output in millivolts). The datalogger connects to a computer via a serial RS232 interface. Figure 3 briefly shows how the equipment are connected.

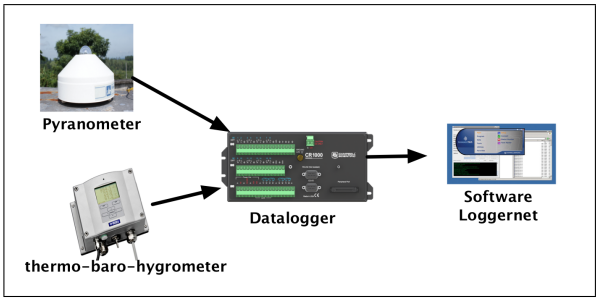

Figure 3 - Components of a weather observation system.

Sensors are installed in the observation area to measure the environmental variables. Next, the datalogger collects sensor data, making it available. The last step is communication and data retrieval, controlled by a code developed using Loggernet[5], proprietary of Campbell Scientific. This allows the usage of data by users on local or remote computers.

The collected data, stored in files encoded in ASCII format, has a header and the measurements. The gathering date and time is embedded into every data collected, separated by commas. We used data starting from 09/01/2015 until 09/30/2015, collected every minute. The Figure 4 shows part of the file content, where the columns used in this paper are highlighted:

- **TIMESTAMP**: date (yyyy:dd:mm) and time (hh:mm:ss);
- **PresAtm**: Atmospheric pressure in hPa;
- **TptAr**: Air temperature in °C;
- **UmidRel**: Air relative humidity in %;
- **Pir(1)**: Voltage in mV (millivolts).

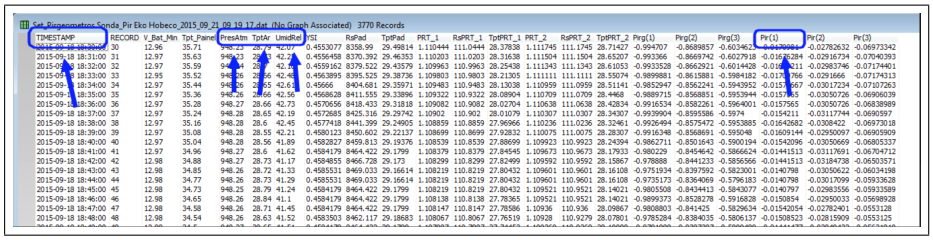

Figure 4 - Collected data in ASCII format.

[5]https://www.campbellsci.com/loggernet

The pyranometer provides outputs in millivolts (mV). We applied the correction factor from Equation 2 to provide the irradiance in W/m^2, where the value 0.1168 is the default calibration factor in $W.m^{-2}/\mu V$ informed in the pyranometer calibration certificate and E the solar irradiance:

$$E = Pir(1) * 1000 * 0.1168 \tag{2}$$

3. Capturing, storing, indexing and visualizing data with the ELK stack

Indexing and querying are operations provided by search engines. Some of its core functionalities include: scalability, user interface (UI), administration and filtering. Solr [9] and Elasticsearch [10] are widely used open source search platforms. Both use the Apache Lucene project [11] to provide their functionalities.

Solr search features include faceted search, clustering, database import handler, rich document handling (PDF, Word, etc), and geospatial search. It is highly scalable and fault tolerant, providing distributed search and index replication. Written in Java, it uses the Lucene Java Search library as a core full-text indexing and search. It provides REST-like HTTP/XML and JSON APIs, allowing the use in any programming language that supports these interfaces [9].

Elasticsearch is part of the ELK ecosystem and differs from Solr in the format the entries are stored, its distributed architecture and automatic nodes discovery, among other functionalities [12, 13]. Elasticsearch allows full-text search, indexing, replicas and recovery. Logstash is a tool used to collect and parse data, which provides means to read data from many different kinds of sources. Kibana is used for GUI display and dashboard visualization.

For this study we opted for the ELK, which provides a full-stack development environment. The ELK fulfills our needs in terms of availability of plugins to read the different kinds of data from environmental sensors and to visualize them in a dashboard for analysis.

Figure 5 shows data and metadata integration inside the ELK infrastructure. Initially, we used Logstash for data and metadata ingestion. First we input data from different kinds of sources (step 1) and then the metadata associated with these data (step 2). In the meteorological observations context this step is fundamental since there are different kinds of sensors, collecting data at different frequencies and formats, susceptible to different kinds of errors.

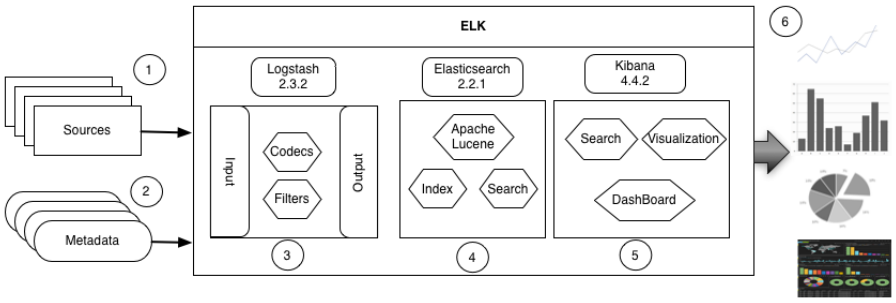

Figure 5 - Sensor data and metadata integration using the ELK ecosystem.

We used Logstash version 2.3.2 in step 3 and the plugins available. In addition to input functionality, Logstash can filter or change the data format using filters and codecs. The data output can be provided to many different formats and systems such as databases, CSV, mongoDB and Elasticsearch.

The Elasticsearch (step 4) used in this work is version 2.2.1, which has Apache Lucene for its main functionalities (index and search). Elasticsearch creates the indices from the data and metadata ingested. Kibana (step 5) version 4.4.2 has visualization and dashboard functionalities, and is the ELK platform for visualization and analysis.

Other possible functionalities of this infrastructure includes different kinds of graphics and also a dashboard to join many graphics together (step 6). The integration of these components provides insights through the data.

In our particular case study, we started by reading the file containing the data collected. For each line read we separated the data into fields, according to the file header, and filtered only the selected variables fields used in this paper. Data regarding the values collected were transformed into float variable type.

From the pyranometer collected value, we assigned a conversion factor to modify the solar irradiance value from mV to W/m^2 unit, using Equation 1. Next we added the information from each sensor and its calibration certificate (metadata). For each sensor data collected we applied the corresponding calibration curve, generated by the sensor calibration specialists and included the calibration certificate of each sensor. The Figure 6 shows an extract of the results of the data processed by Logstash.

```
{
"message" => "\"2015-09-18 18:34:00\",34,12.97,35.44,948.26,28.65,42.61,0.45666,8404.681,29.35971
,1.109483,110.9483,28.13038,1.110959,111.0959,28.51141,-0.9852947,-0.8562241,-0.5943952,-0.015756
67,-0.03017234,-0.07107263",
          "@version" => "1",
          "@timestamp" => "2015-09-18T21:34:00.000Z",
          "path" => "/Users/usuario/dados/LIM/Set_Pirgeometros_3lines",
          "host" => "xxxx.xxxx.xxxx.br",
          "type" => "Pirgeometros",
          "date_time" => "2015-09-18 18:34:00",
:
:
"UmidRelSensor" => {
"Relative humidity" => 68.03,
"Brand" => "Vaisala",
"Model" => "PTU303",
"SN" => "F2910008",
"CalibCert" => "079-14",
       "CalibDate" => "Jan-15-2014",
          "CalibCurve" => "0.9689 * X - 0.5892",
"Calibrated Relative humidity" => 65.33
},
:
:
"latitud" => -22.689339,
"longitud" => -45.006373
}
```


Figure 6 - Data processed with Logstash.

After ingesting data, Elasticsearch receives data and generate indexes from them to provide high performance and accurate searches. Finally, Kibana enables querying and visualization. It is the ELK element that allows data analysis, and insights. We selected a data subset from 09/18/2015 until 09/21/2015, generating graphics of each meteorological variable with the original and corrected values. We created visualizations using Kibana data histogram with the measured variables: air temperature, air relative humidity, atmospheric pressure, and solar irradiance. Aiming a unified visualization of the variables graphics, they were all integrated into a Dashboard.

4. Results and discussions

Through this work, we collect some interesting information regarding the data collected from sensors. Figure 7 shows a bar chart that represents the data subset, collected 60 times per hour (one sample per minute) and without gaps in the collection.

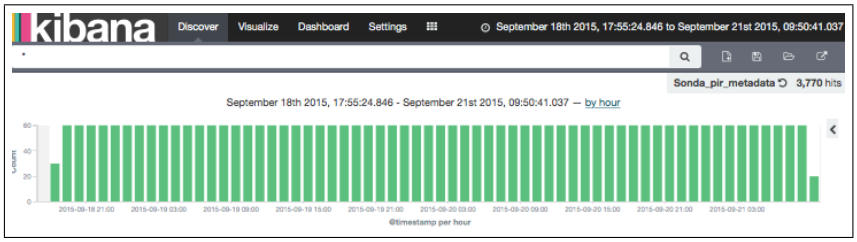

Figure 7 - Histogram of collected data.

Additionally, Kibana provides mechanisms to verify the documents stored in Elasticsearch. We observed the presence of the collected data and the sensor metadata, according with the initial specification.

For each sensor data subset we generated a graph of the collected variable, with the curves of the original and corrected data. We produced graphics of air temperature, air relative humidity, atmospheric pressure and solar irradiance.

The Figure 8 shows the air relative humidity behavior for the selected period. The blue curve corresponds to the original sensor data and the red curve represents the data corrected using the calibration curve informed on calibration certificate LIM 079-14.

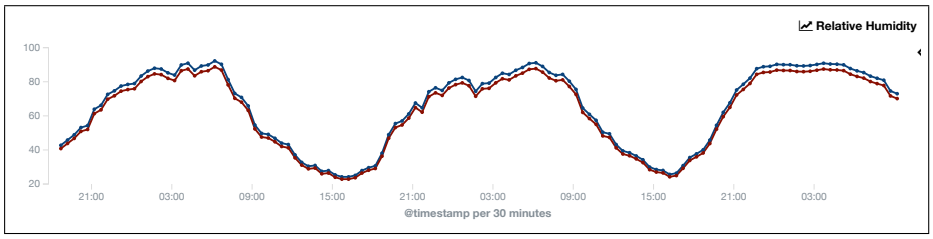

Figure 8 - Air relative humidity graphic.

After generating the environmental variables graphics, they were integrated into Kibana's Dashboard for a unified visualization and analysis. Figure 9 shows a Dashboard with the integration of various graphics related to sensors data used in this paper. Although the graphics present historical data, the ELK stack is ideal to treat near real-time data. The data query and visualization to be monitored or analyzed can be updated automatically. This is useful when dealing with real-time events.

5. Conclusions

The activity of collecting, storing and providing information from weather observation systems are generally automatic. The data are usually available in ASCII format and most of the time users employ spreadsheets or other tools for data analysis without combining different kinds of data sources and metadata.

This paper presents a new methodology to collect, treat, enrich, store, index, visualize and analyze meteorological data, using the numerous facilities we have identified in the ELK stack

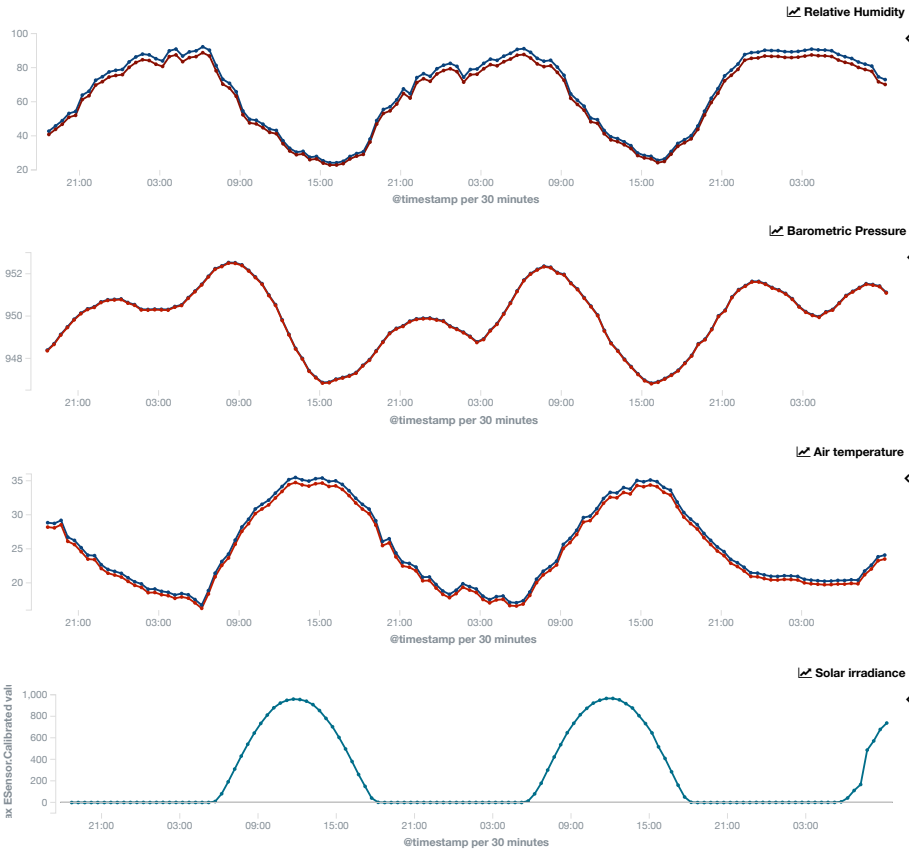

Figure 9 - Integrated view of data.

The data collected by the weather observation systems is indexed with the metadata, which has great value to ensure the measurement reliability. In addition to sensor information, the metadata provides information from the sensor calibration certificate.

In this paper, we used the ELK stack in a meteorological observation context to index sensor data and metadata. Besides allowing the application of the sensor calibration curve to correct the measured data, the usage of such infrastructure also enabled the analysis of data.

Many extensions are possible for this work. The ingestion of other kinds of metadata and also provenance data might be interesting to track data quality. Other kinds of visualization plugins to present data, such as mea-

surement uncertainties, should be investigated as well.

References

[1] Ritsche, M. Surface meteorological observation system (smos) handbook. Tech. Rep., DOE Office of Science Atmospheric Radiation Measurement (ARM) Program (United States) (2008).

[2] World Meteorological Organization. Guide to Meteorological Instruments and Methods of Observation. Part I - Measurement of Meteorological Variables. (2008). URL `http://library.wmo.int/pmb_ged/wmo_8_en-2012.pdf`.

[3] Wright, W. Observing the Climate – Challenges for the 21st Century (2008).

[4] Santana, M.A.A. and Guimarães, P.L.O. and Almeida, E.S. and Eklin, T. The importance of metrological metadata in the environmental monitoring. In *Journal of Physics: Conference Series*, vol. 733, 012033 (2016). URL `http://stacks.iop.org/1742-6596/733/i=1/a=012033`.

[5] Prakash, T., Kakkar, M. & Patel, K. Geo-identification of web users through logs using elk stack. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 606–610 (2016).

[6] Moore, J. *et al.* Devops for the urban iot. In *Proceedings of the Second International Conference on IoT in Urban Space*, Urb-IoT '16, 78–81 (ACM, New York, NY, USA, 2016). URL `http://doi.acm.org/10.1145/2962735.2962747`.

[7] Barbaresi, A. Collection and indexation of tweets with a geographical focus. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 24–27 (2016).

[8] Nguyen, N. & Cuong, T. V. An efficient log management framework. *VNU Journal of Science: Computer Science and Communication Engineering* **32** (2016).

[9] Smiley, D. & Pugh, E. *Apache Solr Enterprise Search Server* (Packt Publishing, 2015), 3 edn.

[10] Akdogan, H. *Elasticsearch Indexing* (Packt Publishing, 2015).

[11] McCandless, M., Hatcher, E. & Gospodnetic, O. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0* (Manning Publications Co., Greenwich, CT, USA, 2010), 2 edn.

[12] Bagnasco, S. *et al.* Towards monitoring-as-a-service for scientific computing cloud applications using the elasticsearch ecosystem. In *Journal of Physics: Conference Series*, vol. 664, 022040 (IOP Publishing, 2015).

[13] Chhajed, S. *Learning ELK Stack* (Packtpub Co., 2015).