Computers & Geosciences 76 (2015) 96–111

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

The data quality analyzer: A quality control program for seismic data

A.T. Ringler [a,*], M.T. Hagerty [b,1], J. Holland [a], A. Gonzales [c], L.S. Gee [a], J.D. Edwards [d], D. Wilson [a], A.M. Baker [a]

[a] U.S. Geological Survey, Albuquerque Seismological Laboratory, P.O. Box 82010, Albuquerque, NM 87198, USA
[b] Boston College, 140 Commonwealth Ave., Chestnut Hill, MA 02467, USA
[c] Honeywell Technology Solutions Incorporated, Albuquerque Seismological Laboratory, P.O. Box 82010, Albuquerque, NM 87198, USA
[d] Ed-Craft Software Solutions, 5309 John Thomas Dr. NE, Albuquerque, NM 87111, USA

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form
11 December 2014
Accepted 22 December 2014
Available online 24 December 2014

Keywords:
Data quality
Data metrics
Seismic Network Performance

ABSTRACT

The U.S. Geological Survey's Albuquerque Seismological Laboratory (ASL) has several initiatives underway to enhance and track the quality of data produced from ASL seismic stations and to improve communication about data problems to the user community. The Data Quality Analyzer (DQA) is one such development and is designed to characterize seismic station data quality in a quantitative and automated manner.

The DQA consists of a metric calculator, a PostgreSQL database, and a Web interface: The metric calculator, SEEDscan, is a Java application that reads and processes miniSEED data and generates metrics based on a configuration file. SEEDscan compares hashes of metadata and data to detect changes in either and performs subsequent recalculations as needed. This ensures that the metric values are up to date and accurate. SEEDscan can be run as a scheduled task or on demand. The PostgreSQL database acts as a central hub where metric values and limited station descriptions are stored at the channel level with one-day granularity. The Web interface dynamically loads station data from the database and allows the user to make requests for time periods of interest, review specific networks and stations, plot metrics as a function of time, and adjust the contribution of various metrics to the overall quality grade of the station.

The quantification of data quality is based on the evaluation of various metrics (e.g., timing quality, daily noise levels relative to long-term noise models, and comparisons between broadband data and event synthetics). Users may select which metrics contribute to the assessment and those metrics are aggregated into a "grade" for each station. The DQA is being actively used for station diagnostics and evaluation based on the completed metrics (availability, gap count, timing quality, deviation from a global noise model, deviation from a station noise model, coherence between co-located sensors, and comparison between broadband data and synthetics for earthquakes) on stations in the Global Seismographic Network and Advanced National Seismic System.

Published by Elsevier Ltd.

1. Introduction

The Albuquerque Seismological Laboratory (ASL) operates nearly 200 seismic stations as part of the Global Seismographic Network (GSN) and the Advanced National Seismic System (ANSS). The data produced from these stations are fundamental to research studies of earthquake sources and earth structure and underpin the operations of the National Earthquake Information Center (NEIC) to provide accurate and timely earthquake data to produce products such as alerts, Web pages, ShakeMaps, and Prompt Assessment of Global Earthquakes for Response (PAGER) impact estimates (Earle et al., 2009). In order to insure the usability of the data, the ASL staff members perform data quality analysis. Traditionally, this has been conducted by waveform review, both through a daily and weekly "run" through the stations, supplemented by automated notifications about problems with availability, timing quality and other data integrity issues, evaluation of power-spectral density, and use of tidal synthetics to catch large-scale problems in polarity and gain. These techniques generally work well for verifying state of health of a station but are not well suited to capturing subtle problems or issues that develop gradually over time, such as the case of degradation of STS-1 responses resulting from humidity in the feedback electronics boxes

* Corresponding author.
E-mail address: aringler@usgs.gov (A.T. Ringler).
[1] Now at Instrumental Software Technologies, Inc., P.O. Box 963, New Paltz, NY 12561, USA.

http://dx.doi.org/10.1016/j.cageo.2014.12.006
0098-3004/Published by Elsevier Ltd.

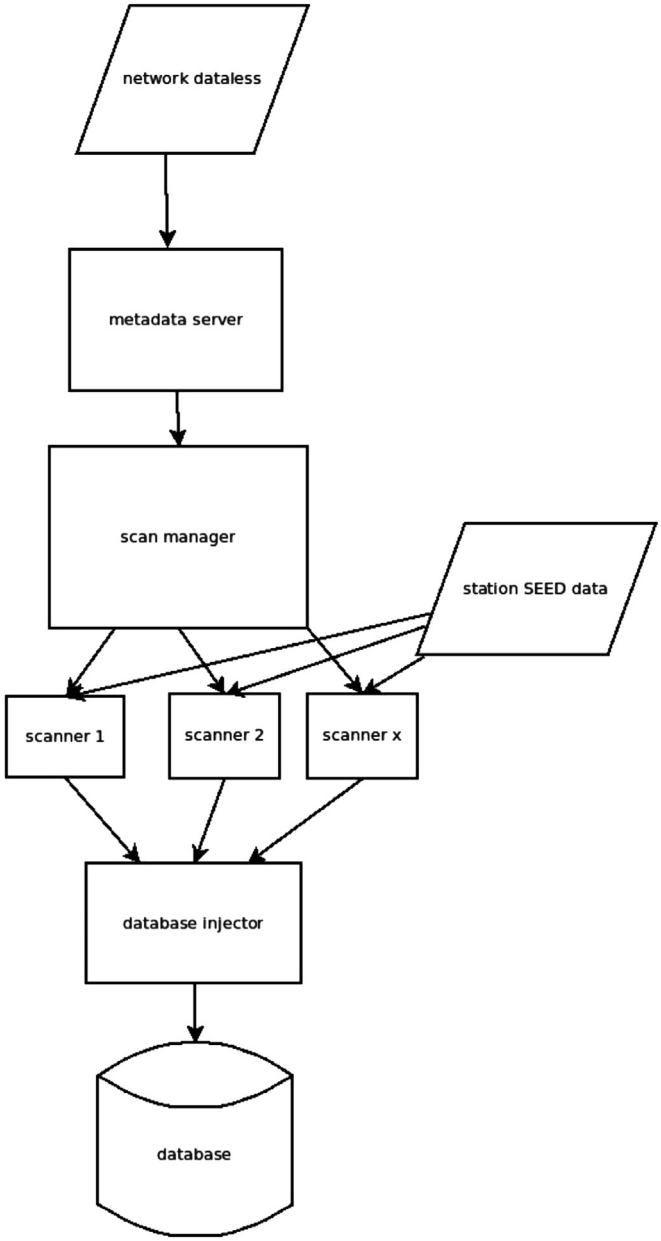

Fig. 1. Logical flow of the SEEDscan process. Using the network dataless SEED volumes provided by the metadata server (top), the Scan Manager produces scanners which then scan the data for a given day and station, check the hash with the database (bottom), and compute the metric if necessary. The metrics are then injected into the database through the database injector.

(Hutt and Ringler, 2011). As a result, the ASL recently has developed and implemented a number of tools to monitor station performance in situ, such as using PQLX (PASSCAL Quick Look eXtended; McNamara and Buland, 2004) and synthetic seismograms to identify changes in gain at GSN stations (Ringler et al., 2010, 2012a) as well as implementing an annual calibration process (Ringler et al., 2012b).

In order to facilitate the use of multiple metrics to identify problems and to enable the quantification of data quality, we developed a framework, called the Data Quality Analyzer (DQA) to compute data metrics routinely and display the results in an easy-to-use interface. The DQA consists of components for scanning miniSEED (Ahern et al., 2009) data and computing the metrics (SEEDscan), storing them in a database, and displaying the results on a Web interface. The system is configurable to deal with future developments or changes, and we are able to add and modify metrics through an Extensible Markup Language (XML) configuration file. The code may be run as a scheduled task (e.g., nightly) or on command to ensure the latest metrics are available. The DQA makes extensive use of hash signatures to ensure that changes in either metadata or data trigger a rescan to update the metrics.

In this paper we discuss the overall DQA structure including the flow of SEEDscan, the database, and the Web interface as well as describe the currently implemented metrics. Using these metrics, we illustrate a number of common data problems, including some subtle problems not obvious from simple inspection of time series or power spectra. Finally, we discuss future development plans.

2. The code

The DQA naturally breaks into three distinct pieces: the SEEDscan metric calculator, the database, and the interface. In addition, there is auxiliary code that supports the DQA process.

2.1. SEEDscan

The SEEDscan program is written in Java and compiled using an Apache Ant (Apache Software Foundation, 2011) build file. Upon successful compilation, SEEDscan reads an XML configuration file to identify the location and structure of the data files, the location of the metadata, the metrics to be computed, and the time interval for which these metrics will be computed (Fig. 1). Once SEEDscan has read its configuration file, it configures the metadata server that reads in the network metadata and serves up station metadata for specific deployment profiles (epochs) upon request from the scanner module (see below). The metadata server can be setup as a local or remote running instance, in case the user wants to run multiple instances of SEEDscan using a single metadata server while running a remote SEEDscan process. By default, the metadata server reads dataless SEED format metadata that does not contain any time series data, but may be configured to use other formats. Once the metadata have been successfully read, the scan manager creates a thread pool of scanners. A scanner is created for each station and is executed in its own thread. The scanner requests the data for its assigned station day. For each metric, SEEDscan compares the hash of the data and metadata used to compute the metric with values contained in the database; if they differ, the metric is recomputed and injected into the database, otherwise SEEDscan moves on without performing the computation.

We have implemented SEEDscan using a nightly build routine. This makes it possible to incorporate new metrics as well as modify metrics without having to make changes to the infrastructure of the code (e.g., we may add a metric by simply including it in the code and adding a few lines to the configuration file). The managing of nightly builds is discussed more in the auxiliary code section.

2.2. Database

The PostgreSQL database acts as the central hub where metric values and limited station descriptions are stored. The database is populated by the SEEDscan program, with data stored as double precision floats at the channel level with one-day granularity. Certain metrics are multivalued and are stored as separate metrics in each of a number of selected frequency bands. The database also contains several stored procedures that are called depending on what data are needed. Most stored procedures return comma-delimited responses that the Web interface parses. Each metric has a stored computation type for determining how best to summarize data over a time period; metrics such as the gap count are summed, while most metrics are averaged.