ID: Type: **Poster**

# of U.S. NDC Performance Metrics Through Large Scale Analysis of System Log Files With Hadoop Distributed File System Based Tools

Optimizing the performance of the United States National Data Center (U.S. NDC) geophysical data processing system is critical for efficiently monitoring international compliance to nuclear test ban treaties. The U.S. NDC software stores system performance information for each data processing interval in a collection of semi-structured alphanumeric log files. On average, the system generates 140,000 log files per day which are stored in different directories. Currently, acquisition of process specific performance information or isolation of error messages must be parsed from each log-file individually. This manual parsing process is time consuming and often leads to incomplete collections of system performance information. The U.S. NDC system has been modified to output log files in JavaScript Object Notation (JSON), which is a highly structured data format that can be easily parsed. Here, we show how U.S. NDC system performance information can be parsed from JSON files and analyzed using a collection of Hadoop Distributed File System (HDFS) based tools such as Hive, Zeppelin, and PySpark. The HDFS system architecture is designed to store and process large alphanumeric datasets, which can be easily scaled to accommodate our continuously growing collection of performance information extracted from U.S. NDC log files.

Primary author: JUNEK, W. N. (United States National Data Center)

Presenter: JUNEK, W. N. (United States National Data Center)

Track Classification: 3. Advances in sensors, networks and processing