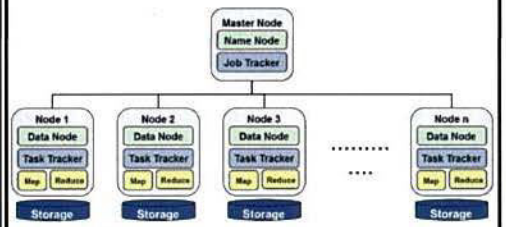
Development of U.S. NDC Performance Metrics Through Large Scale Analysis of System Log Files With Hadoop Distributed File System Based Tools

W. N. Junek, C.A. Houchin, J. A. Wehlen III, A. Poffenberger, and R. C. Kemerait
Air Force Technical Applications Center, Patrick Air Force Base, Florida

Abstract

Optimizing the performance of the United States National Data Center (U.S. NDC) geophysical data processing pipeline 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 based tools such as HIVE, Zeppelin, and PySpark. The Hadoop 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.

Hadoop Distributed File System Architecture Overview

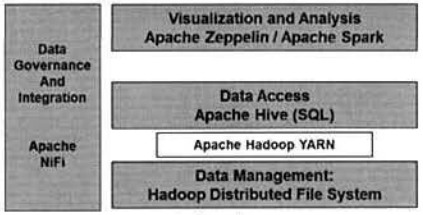


- **Hadoop Distributed File System (HDFS)**: Java based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers.
- **Map/Reduce**: Framework for processing parallelizable problems across a large number of nodes. Takes advantage of data locality, processing it on or near the storage asset to reduce the distance it must be transmitted.

U.S. NDC HDFS System Specifications

- **Nodes**: 12 Dell PowerEdge Servers, 6c3Ghz Cores, 8x4TB SAS Drives, 128 GB Ram
- **Storage**: 148 TB of usable storage

Log File Analysis System



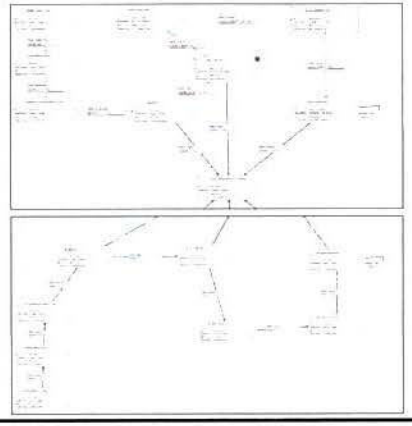
- US NDC System generates over 140,000 log files per day
- Manual parsing of log files is time consuming, error prone, and leads to incomplete data sets
- HDFS systems are ideal for ingesting, storing, and analyzing large scale alpha numeric datasets

JSON Data File Format

- U.S. NDC log files store system performance information in numerous unstructured text files
- The U.S. NDC software now outputs system log files in JavaScript Object Notation (JSON)
- JSON files are a highly structured information storage and exchange mechanism that is easy to parse

NiFi JSON Data Parsing Processor

- Custom NiFi pipeline was designed to move JSON based log files to custom Hive database
- Top portion transfers data processing (e.g., DFX, GA) log files
- Bottom portion transfers event correlator log files

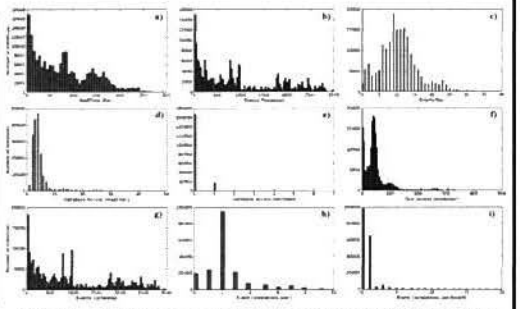


U.S. NDC Event Correlator Performance Metrics

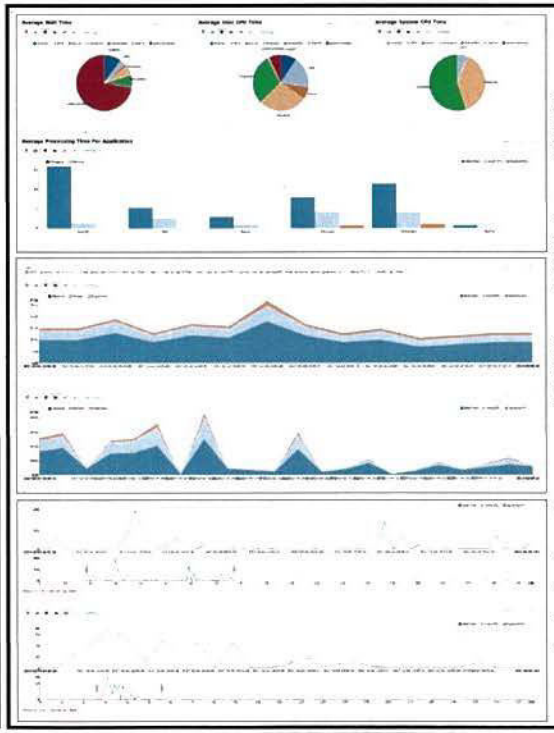
- Apache Hive is a data warehouse infrastructure built on top of HDFS for providing data summarization and querying capabilities
- Hive uses a Structured Query Language (SQL) like interface known as HQL
- Allows SQL like interaction with distributed data sets for uses that are familiar with this

Hive Schema for Event Correlator Metrics

Data record for parsed event correlator JSON file log entry



Automated event correlator performance statistics over a 365 day period. a) Wall Time in seconds illustrating the total elapsed time each correlation process required to reach completion. b) Event processed by each correlation process. c) Number of events processed per second. d) Number of database reads per second. e) Database access in seconds per read f) Disc access in reads per second g) Number of events correlated. h) Event correlations per second. i) Event correlations in seconds per event



Web Browser Based Performance Metrics Dashboard

Apache Zeppelin

- Web-based notebook for interactive analytics
- Multi-purpose toolbox for data ingestion, analytics, visualization
- Supports multiple languages including Apache Spark and Python
- Renders numerous chart styles to create interactive output
- Share notebooks among collaborators

Apache Spark

- Engine for large-scale data processing
- Runs on Hadoop and can access HDFS
- Fast in-memory computing platform
- Supports SQL, streaming and machine learning libraries
- Application development in Java, Scala, Python, and R

Apache NiFi

- Flexible data collection, transformation, and routing platform
- Supports numerous formats, schemas, and protocols
- Web-based user interface for design, control, feedback and monitoring
- Data Provenance for real-time visual chain of custody and metadata
- Real-time streaming data: send, ignore, or store received data

Summary

- The U.S. NDC has acquired a HDFS cluster to study the large scale data acquisition and analysis
- Modified US NDC system to output log files in JSON format
- Developed HIVE based data model for archiving log file data
- Analyzed log file data using Spark and Zeppelin based tools
- Results of this work will influence U.S. NDC future data processing and hardware architecture design decisions for processing, storing, and analyzing large datasets