Sandia National Laboratories

SAND2019-8269D

Data Normalization

Thomas Byrd, Mississippi State University
Noah Pittenger, Brigham Young University
Thanh Nguyen, 8765
Troy Stevens, 6617

Introduction – Data normalization is the process of transforming data from its original format into a standardized format. In cyber security, normalizing sensor/log data is of increasing importance, given the exponential growth of the internet and the sheer volume of data that each system generates. In our study, we normalize five different types of logs, each generated in their own unique format that capture different interactions occurring on computer networks.

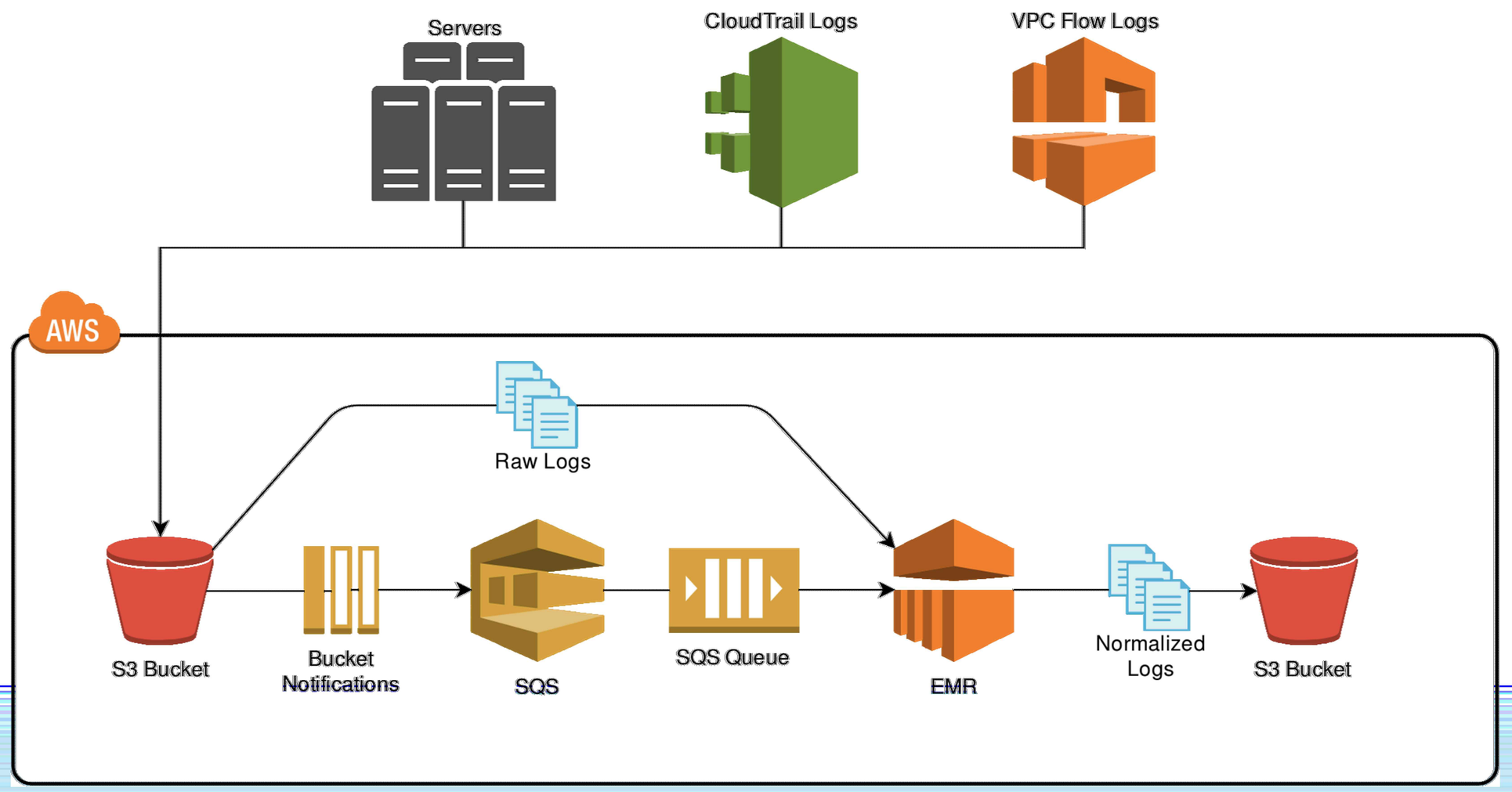
Challenges

- Configuring the platform's infrastructure in the most cost-effective and efficient way
- Ensuring cross-region compatibility for each solution
- Optimizing throughput capable of handling terabytes of data over a given time period
- Ensuring all logs could be accurately and timely normalized

Learning Outcomes

- Understand Amazon Web Services and their ability to provide services for normalization
- Sharpen skills in cloud and systems engineering to provide high-throughput solutions
- Integrate custom code templates and cloud computing for data normalization
- Architect secure cloud-based solutions

Problem – Log data sources can be generated in many unique formats and require normalization before they are useful to an analyst. In our case, log data could be delivered as a compressed JSON file, a compressed text file, or even a comma-separated spreadsheet. The contents of each log covers different areas of interest to network defenders such as network flow data, API call logs, and server logs. In addition to the breadth of information, we observed data in multiple formats. IP Addresses given in their integer equivalent along with their string counterpart, timestamps expressed in seconds (UNIX EPOCH) as well as a GMT based date/time strings, and differing field names were detected despite representing the same network data. The above complications compound when the volume of data grows to include various cloud and local infrastructures.



Results – Using AWS, we developed several solutions to automate log processing based upon specific needs. Displayed above is an example of one such solution. Agents running on Servers as well as log information collected from AWS Components deliver their logs for processing into an S3 bucket. The bucket is configured to send notifications to SQS which then notifies an EMR cluster. The cluster then normalizes the logs using a custom Python script to format the logs according to the desired output. The now normalized logs are transferred to an S3 bucket where they can be further enhanced or queried.

U.S. DEPARTMENT OF ENERGY
National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. _____

Sandia National Laboratories