

Michael J. Pedersen
42 Henry St
Oxford, NJ 07863
datacyclist@gmail.com
(908) 283-0318

Executive Summary

DevOps Engineer with 23 years of experience managing Linux, UNIX, [Windows](#), and OSX/macOS systems. Programming languages have included Python, PHP, Perl, and Java. Work environments have been heterogeneous (several flavors of Linux, [Windows](#), and OSX/macOS), small to medium sized (from 10 to 1200 servers, 20 to 300 workstations), and mixed locations (all local to all remote teams).

Relevant Job History

Pulsepoint - Data Engineer (2015-2018), Director of Infrastructure for Data (2018-2023)

New York City, NY & Newark, NJ (Telecommute) - 2015-2023

- Architected data streaming that manages 40T of data/day.
- Established new data centers in Europe and in Virginia.
- Migrated data center, moving processing of data flows to new data center.
- Upgraded [Kafka](#) with zero downtime for users of [Kafka](#).
- Deployed and configured [Alluxio](#) for caching and data orchestration.
- Performance tuned [Kafka](#).
- Enabled integration with [Active Directory](#) for [Hadoop](#) systems.
- Replaced [Vertica](#) with [Trino](#).
- Built tool to graphically show the flow of data through the system.
- Transitioned ETL pipeline from crontabs to [Mesos](#) and then into [Kubernetes](#).
- Troubleshooting of issues with [Hadoop](#), [Kafka](#), [SQL Server](#), and [Kubernetes](#).
- Production maintenance of data pipelines, including after hours support.
- Tested new tools for suitability, including [MariaDB](#), [Clickhouse](#), and [Kudu](#).
- Switched build server from [TeamCity](#) to [Jenkins](#), recreating all build jobs.
- Implemented data duplication between two [Hadoop](#) clusters.
- Upgraded [Hadoop](#) clusters with minimal downtime.
- Participated in on-call rotation.

OrcaTec, LLC - Developer

Atlanta, GA (Telecommute) - 2012-2014

- Found major security hole (remote code execution) and closed it.
- Debugged and resolved memory issues that were causing systems to shut down.
- Incorporated memcached into our stack to handle sessions and cached data.
- Switched web server from [Paster](#) to [Apache](#) with [mod_wsgi](#).
- Corrected Unicode handling errors in the code.
- Added holds and matters framework, allowing customers to state that documents belong to specific cases and should not be deleted while the cases are ongoing.
- Identified weaknesses in the database model, and added code to prevent those weaknesses from being hit.

For more history going back to 1995, please visit my website at <https://www.icelus.org/>
Michael J. Pedersen datacyclist@gmail.com 908-283-0318

Relevant Technical Skills

Programming and Scripting Languages

	Time Used	Last Used	Proficiency
Bash	10 years	2014	Good
C/C++	12 years	2009	Good
Java	2 years	2021	Fair
Javascript	3 years	2021	Good
Perl	6 years	2012	Fair
PHP	2 years	2012	Fair
Python	15 years	2023	Excellent

Software Configuration Management Tools

	Time Used	Last Used	Proficiency
Git	11 years	2023	Good
Mercurial	4 years	2014	Fair
Subversion	2 years	2010	Fair

Database Servers

	Time Used	Last Used	Proficiency
MySQL	3 years	2021	Fair
PostgreSQL	5 years	2011	Fair
Microsoft SQL Server	3 years	2023	Fair

Operating Systems Administered

	Time Used	Last Used	Proficiency
Linux (Debian , RedHat , Suse , Ubuntu)	21 years	2023	Excellent
Microsoft Windows (2008/7/Vista/2003/X P/NT/98/95)	12 years	2011	Very Good
UNIX (Solaris , AIX , HP-UX)	5 years	2011	Very Good

Markup Languages

	Time Used	Last Used	Proficiency
CSS	2 years	2023	Fair
HTML	5 years	2023	Very Good
Markdown	3 years	2023	Good
XML	2 years	2023	Fair

Education

Bachelor of Science in Computer Science, 2000
East Stroudsburg University, East Stroudsburg, Pennsylvania

For more history going back to 1995, please visit my website at <https://www.icelus.org/>
Michael J. Pedersen datacyclist@gmail.com 908-283-0318

Project History

Dataflow Explorer

Period	2015
Company	Pulsepoint
Tools	Python , Graphviz Dot , Luigi
Platform	Mesos , CentOS , NGINX

At Pulsepoint, we have a large number of data aggregation jobs that are coordinated with each other via Spotify's [Luigi](#) tool. [Luigi](#) has the user create a [Python](#) codebase that resolves which order to do jobs similar to how [GNU Make](#) actually works. A negative side effect of this is difficulty for humans to understand the order of jobs that will be run when the number gets to any significant size.

The Dataflow Explorer would walk the [Python](#) code that represented all of the jobs, and extract the attributes that would allow construction of a dependency tree. It would then pass that tree to the [Graphviz DOT](#) tool, which would run dot to produce an SVG file showing the graph of all the jobs. Finally, it would publish that output onto [Mesos](#) using [NGINX](#), allowing people to browse, zoom, and search the resulting graph.

- Wrote code to walk a [Python](#) code base and extract specific attributes
- Produced syntactically valid [Dot](#) files.
- Automatically published updated versions of the graph for myself and others to use.

California Hadoop Cluster

Period	2015
Company	Pulsepoint
Tools	Hadoop
Platform	CentOS Linux

Pulsepoint needed to establish a disaster recovery site, and had chosen an existing data center to do so. In the process, establishing a [Hadoop](#) cluster was required for business continuity. My task was to get everything configured to the point that the same data jobs running in the primary cluster ran in the backup cluster and provided equivalent data, even though everything was running independently.

- Installed [Cloudera Distribution of Hadoop](#) across the cluster.
- Ensured that [HDFS](#), [Hive](#) and [Impala](#) were functioning properly.
- Ensured that the same data jobs running in the primary cluster were running in the secondary cluster.
- Ensured that equivalent output was happening in both data centers.