Michael J. Pedersen : (908) 283-0318 :
datacyclist@gmail.com : My Site : LinkedIn

Summary

Executive Summary

Web Summary

With a 28 year career, I have filled a variety of technical roles. As a manager, I have spent 5 years of my career guiding teams through challenging conditions, ranging from technical challenges through COVID-19 related challenges. I have kept my team together, and we have thrived on the challenges of taking care of the data for a $200 million company. This includes managing my team through the COVID-19 pandemic, without losing a single member of my team to another company.

I am a Google Cloud Certified Professional Data Engineer. I have spent 9 years working with streaming challenges, ETL challenges, and data latency challenges as we keep our overall latency to 5 hours or less. We have done this using Hadoop, Python, Kafka, and Alluxio. We have worked with different data storage formats, different compute engines for accessing the data, and different ways to coordinate our ETL pipeline to avoid having jobs crash into each other.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Furthermore, as a DevOps Engineer, I have 23 years of experience managing Linux, UNIX, Windows, and OSX/macOS systems. This means that I look at the whole picture, not just System Administration or Software Development. Shepherding a system through the creation and deployment process, and seeing the customer's happiness at having things work the way they need it to, is a particular joy of mine. Making people's lives better is the point of technology, after all.

Finally, as a Software Engineer, I have spent 6 years of my career focused on delivering high quality software to my company's customers, with their focus and needs being on sorting through large numbers of documents in a timely fashion. This has meant understanding ingestion, storage, and display of arbitrary data. It has included custom data visualizations. This was primarily done with Python and Ubuntu Linux, but has also included work with Perl and PHP.

I am comfortable in a wide range of working conditions. 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).

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Programming languages have included Python, PHP, Perl, and Java.

Data Engineer Summary

I am a Google Cloud Certified Professional Data Engineer. I have 9 years of experience doing data engineering. I've scaled up a company from receiving 4T of new data/day through to its current 40T/day of new data. We have grown from about 50 servers to 450 servers for our big data architecture, as well as holding over 2P of data we are storing and actively using.

DevOps Engineer Summary

DevOps Engineer with 28 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).

Michael J. Pedersen : (908) 283-0318 :
datacyclist@gmail.com : My Site : LinkedIn

Manager Summary

Manager with 5 years of experience (28 years overall) guiding teams through challenging conditions, ranging from technical challenges through COVID-19 related challenges. I have kept my team together, and we have thrived on the challenges of taking care of the data for a $200 million company. Most importantly, all of this has been done while keeping an eye on our budget, making sure to keep our costs down to keep the company profitable right through COVID-19.

Software Engineer Summary

As a Software Engineer, I have spent 6 years of my career focused on delivering high quality software to my company's customers, with their focus and needs being on sorting through large numbers of documents in a timely fashion. This has meant understanding ingestion, storage, and display of arbitrary data. It has included custom data visualizations. This was primarily done with Python and Ubuntu Linux, but has also included work with Perl and PHP.

Michael J. Pedersen : (908) 283-0318 :
datacyclist@gmail.com : My Site : LinkedIn

Job History

Pulsepoint - Data Engineer and Director of Infrastructure for Data

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

Pulsepoint is an internet healthcare marketing company with a focus on activating health care providers. Pulsepoint was acquired by WebMD in June 2021.

My role evolved over time from dealing with individual data jobs to overseeing the entire ETL pipeline to leading the entire department.

- 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 pipelines to new data center.
- Split data management team into data platform and data product development.
- Guided the team through splitting our ETL pipelines into multiple repositories.
- Organized the migration of ETL pipelines from Python 2 to Python 3.
- Insituted and formalized processes and

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

procedures for the team.
- Planned capacity to ensure we could handle incoming data throughout the year.
- Replaced Vertica with Trino.
- Acted as scrum master for the team.
- Reported on system wide data latency using ElasticSearch, Kibana, and Grafana.
- Conducted interviews for my team and for teams that work closely with my team.
- Automated distribution of incident reports to all affected parties.
- Changed hardware profiles for Hadoop to remove storage and compute colocation.
- Onboarded new team members, helping them to fully integrate into the team.
- Held weekly 1 on 1 meetings with team members.
- Participated in on-call rotation.
- Developed new stories (including estimates) for our Jira board.
- Prioritized tickets for our Jira board.
- Passed annual HIPAA training for data protection.
- Upgraded Kafka with zero downtime for producers and consumers.
- Deployed and configured Alluxio for caching and data orchestration.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

- Performance tuned Kafka.
- Enabled integration with Active Directory for Hadoop systems.
- Built tool to graphically show the ETL pipelines.
- 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.
- Created ELT jobs to ingest third party data to make it available internally.
- Installed and configured multiple Hadoop clusters.
- Developed new ETL jobs to aggregate data from Pulsepoint’s RTB exchange.
- Optimized Hadoop jobs.
- Maintained Vertica cluster, including

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

troubleshooting.
- Tested Cassandra as a potential reporting database.
- Converted Sqoop jobs to use FreeBCP instead.

Weight Watchers - Systems Engineering Lead

New York City, NY - Nov 2014 - Feb 2015

Weight Watchers is a Fortune 500 company focused on helping customers manage their weight and reduce health problems caused by it.

My role was focused on providing internal support within the company to enable other groups to support the customer base.

- Developed lightweight monitoring tool for use within my group.
- Configured Vormetric products to ensure HIPAA compliance for customer data.
- Worked to transfer from Rackspace Cloud to Openstack based private cloud.

OrcaTec, LLC - Developer

Atlanta, GA (Telecommute) - Jun 2012 - Oct 2014

OrcaTec is in the litigation support industry (they help their clients reduce the costs of being sued). OrcaTec is primarily a software-as-a-service company, allowing OrcaTec to host customer data.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

While working here, my focus has been on improving the GUI. This has involved refactoring code heavily, adding new features, and adding new tests to cover existing and new code.

The team structure at OrcaTec is geographically very diverse. In addition to my own telecommuting, I have teammates in many states. We all work remotely, and we all work together to make the product the best that it can be.

- Mentored other developers in the use of TurboGears, SQLAlchemy, Python, and JavaScript.
- Organized weekly meetings for members of the frontend (OTGUI) team, providing a chance to discuss (in depth) the issues the team was facing.
- 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

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

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.
- Wrote Python framework to manage long running background jobs.
- Reduced multi-hour SQLAlchemy bulk database jobs to minutes.
- Spearheaded conversion from YUI 2 to jQuery and jQueryUI.
- Documented internal server API, wrote a Python class to standardize it's use.
- Added tag cloud (using awesomecloud plugin for jQuery).
- Added support for allowing customers to login using OpenID.
- Developed advanced search tool using Python, TurboGears, and jQuery.
- Created new document production framework from scratch.
- Installed and configured WSO2 Identity Server for our OpenID implementation
- Created a tool to allow copying settings between instances.
- Added user preferences to the frontend.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

- Resolved intermittent issue with drag/drop events that had been unsolvable by the existing team.
- Implemented login idle timeout functionality.
- Refactored Python and JavaScript code on a regular basis to reduce code repetition and increase legibility.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Relevant Technical Skills

- **Programming Skills:** Docker, Jenkins, Jira, Intellij IDEA, Object-Oriented Design, Object-Oriented Programming, Refactoring
- **Database Skills:** PostgreSQL Database Administration, Relational Schema Design, Structured Query Language (SQL)
- **Big Data:** Google Big Query, HDFS, Hive, YARN, Alluxio, Impala, Trino, Kafka, Kubernetes, Dagster
- **Programming and Scripting Languages:** Bash, C/C++, Java, Javascript, Perl, PHP, Python
- **Software Configuration Management Tools:** Git, GitHub, GitHub Actions, Mercurial, Subversion
- **Database Servers:** MySQL, PostgreSQL, Microsoft SQL Server
- **Operating Systems Administered:** Linux (Debian, RedHat, Suse, Ubuntu), Microsoft Windows (10/2008/7/Vista/2003/XP/NT/98/95), UNIX (Solaris, AIX, HP-UX)
- **Markup Languages:** CSS, HTML, Markdown, XML
- **Applications:** Ipswitch What's Up, Nagios, OpenStack, Slack, VirtualBox, VMware,

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Zenoss

- **Networking and Security:** Checkpoint VPN, Cisco, Firewall Design, TCP/IP

Education

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

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Project History

Migrate To New Data Center

Period	2022-2023
Company	Pulsepoint
Tools	Alluxio, Hadoop, Kafka, Python
Platform	CentOS, Kubernetes

Pulsepoint is in the process of migrating between data centers. A significant portion of the existing hardware has gone past its end of life, so we chose to build a new data center, with new hardware. At the same time, we used the latest versions of all relevant software that we could (Hadoop, Kubernetes, etc).

This provided us with an opportunity to fix some design flaws in the original big data clusters, and we used this chance to make things better for us overall.

The work remaining at this point comes down to verifying that the new versions of the ETL jobs function as expected, producing valid output. The process is expected to complete in 2025.

- Created new clusters, with new versions of relevant software, in the new data center.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

- Updated ETL jobs as needed so that they would run exclusively in the new data center.
- Configured those ETL jobs to output copies of their data to the original data center.
- Removed those ETL jobs from the original data center, configuring the original to use the output from the new data center.

Migrate From Python 2 to Python 3

Period	2022-2023
Company	Pulsepoint
Tools	Python
Platform	CentOS, Kubernetes

Pulsepoint built the entire ETL pipeline using Python 2. On January 1, 2020, Python 2 reached its end of life. In order for the ETL pipeline to continue to grow, we needed to migrate to Python 3.

The path we chose was to extract the code that was common to the pipeline, and turn that code into a library. We then began the normal route of making backwards incompatible changes. Because of the scope of this work (nearly 200K lines in Python files), and the work being done during a data center migration, the project is still ongoing. However, over 50K lines have been successfully completed so far.

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

- Established a library cutoff version, after which the library would no longer support Python 2.
- Began regular release cycles for the library
- Ensured that developers outside of the library maintenance team could use the library to easily migrate ETL jobs.

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

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

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.

Cassandra for User Reporting

Period	2015
Company	Pulsepoint
Tools	Cassandra
Platform	CentOS Linux

Pulsepoint has a fairly significant Microsoft SQL Server installation, and we were asked if we could use Cassandra as a replacement for it. We set up a small cluster, and began trying to run various reports against it.

The actual performance was impressive, but we ran into a significant roadblock: Cassandra is, in significant ways, a disk based key/value store. In order to use this as a reporting database, and avoid triggering table scans for the user reporting, we would have had to load up many copies of the

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

same data into different tables with different primary keys.

In the end, this was deemed non-feasible for the number of combinations we would have had to provide, along with the amount of maintenance as new reports could be brought online.

- Deployed a Cassandra cluster.
- Produced data sets into that cluster.
- Confirmed queries ran, and ran well.
- Ultimately recommended against because of the issues with table scans and primary keys.

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

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

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.

Sqoop to FreeBCP(FreeTDS) Conversion

Period	2016
Company	Pulsepoint
Tools	Sqoop, FreeTDS
Platform	Hadoop, Microsoft SQL Server

Apache Sqoop has long been deprecated, with its eventual complete retirement in June 2021. As part of Pulsepoint's platform, we needed a replacement for Sqoop before it was fully retired. We settled on FreeBCP, which is part of the FreeTDS project. Using this tool, we were able to migrate our processes for transferring data from Hadoop to MS SQL Server.

- Developed migration strategy to transition from Sqoop to FreeBCP.
- Tested FreeBCP as a substitue for Sqoop.
- Updated our ETL pipelines to use FreeBCP in

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

place of Sqoop.

Vertica Decommissioning

Period	2018
Company	Pulsepoint
Tools	Vertica, Trino
Platform	CentOS Linux

Pulsepoint had used Vertica, but we were outgrowing it in 2017. In 2018, when we came up for the most recent support renewal, we had fully outgrown it and needed to replace it with something else. After trying out several other options (including Clickhouse, Trino, [MariaDB] [MARIODB], and others), we settled on Trino as the option that provided us with the best capabilities while being nearest to the performance that Vertica provided.

- Performance tested existing Vertica queries.
- Stood up several competitors and compared their performance using the same queries.
- Compared maintenance of these environments to Vertica.
- Finally chose Trino, implemented it, and fully decommissioned Vertica.

Data Management Team Split

Period	2021

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Company	Pulsepoint
Tools	Git
Platform	Jira, GitHub

As part of the growth of Pulsepoint, the Data Management team reached a point wherein the team was no longer able to do everything that was required: New data products were needed, and the data platform itself needed both maintenance and new features as well. I made the decision to split the team in two, creating a Data Platform team and a Data Product Development team. Each team would be focused on exactly one role, instead of trying to split the focus between two distinct functions.

- Divided the team into two distinct functional teams.
- Divided the code between the two teams to reflect their individual functions.
- Divided the Jira board between the two teams.
- Established new teams on GitHub, with each team getting only the portion of the code belonging to them.

Data Management Code Split

Period	2021
Company	Pulsepoint
Tools	Git

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Platform	GitHub

Pulsepoint needed to split the Data Management team into a Data Platform team and a Data Product Development team. This also meant splitting the code, since the entirety of the ETL pipeline was in one monolithic repository. The team had to develop a means of crossing repository boundaries to establish the pipeline steps (e.g.: Job A in repository 1 is dependent on Job B in repository 2). We also had to come to agreements on how to determine which team got which pieces of code.

- Developed cross-repository dependency system for ETL jobs.
- Agreed on terms to decide which team got which piece of code from the original repository.
- Created new repositories to get that code.
- Created new teams on GitHub to assign ownership over the newly divided code.

Advanced Search Tool

Period	2014
Company	OrcaTec, LLC
Tools	Python, jQuery, jQueryUI
Platform	Server: TurboGears,

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

Browser (Cross Browser)

At OrcaTec, the primary tool we provided to our customers was the ability to search collections of documents quickly. In addition to having simple search tools, we also had a helper tool in the “Advanced Search”.

This tool allowed the user to search based on a dozen different fields, but was still limited and fragile. It was unable to help the user build queries which combined different fields in a single clause. In addition, it had issues with encoding <> in email addresses, and did not support drag and drop on all of our supported browsers.

When this project was completed, this tool had transformed noticeably. It now is its own miniature investigative tool, allowing customers to easily search through collections of documents. One customer reported narrowing their searches from 80,000 possible documents down to under 2,000 within an hour through use of this tool. Due to extensive test coverage when the code was published, even the problems that were found were quickly fixable. All of this was accomplished while reducing the total code for it by 50%.

- Debugged issues with drag/drop on mobile

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

browsers.
- Designed new interface for maximum flexibility, and to allow easy refinement of queries as they are being built.
- Incorporated user feedback to improve that design.

Paster to Apache/mod_wsgi Conversion

Period	2013
Company	OrcaTec, LLC
Tools	Python, Apache, mod_wsgi, Paster
Platform	Ubuntu Linux

Paster is meant to be used in a development environment, allowing the developer to use a (single threaded) lightweight, easily managed webserver while writing code before it goes to production. At OrcaTec, we were using Paster both in development and in production. Due to the demands being placed on Paster (in many instances, loading up documents that were over 100M), the entire system could appear (to one user) to freeze up due to it responding to a request from another user.

After analysis, we were able to determine that Paster was no longer suitable for our needs. Since Apache, with mod_wsgi, provides an at least

Michael J. Pedersen : (908) 283-0318 : datacyclist@gmail.com : My Site : LinkedIn

adequate performance web server (in comparison to others like Nginx), and the Apache configuration was already known to the team, we chose to switch from Paster to Apache. This allowed us to have Apache itself serve up static files (like images, css files, and javascript files), leaving the dynamic pages to the Python code.

- Debugged threading/locking/memory usage issues with Paster.
- Recompiled and repackaged Python 2.6.8, Apache, and mod_wsgi for use with Ubuntu 10.04.
- Developed automatic Apache configuration for use within our local stack.