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

Executive Summary

I'm a data engineer with 9 years of experience. 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.

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.
- Guided the team through splitting our ETL monolithic repository.
- Organized the migration of the ETL pipeline from Python 2 to Python 3.
- Built tool to graphically show the flow of data through the system.
- Replaced Vertica with Trino.
- Ingested third party data to make it available internally.
- 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.
- Installed and configured multiple Hadoop clusters.
- Changed hardware profiles for Hadoop to remove storage and compute colocation.
- Implemented data duplication between two Hadoop clusters.
- Upgraded Hadoop clusters with minimal downtime.
- Passed annual HIPAA training for data protection.
- Tested Cassandra as a potential reporting database.
- Transitioned ETL pipeline from crontabs to Mesos and then into Kubernetes.
- Reported on system wide data latency using ElasticSearch, Kibana, and Grafana.

OrcaTec, LLC - Developer

Atlanta, GA (Telecommute) - 2012-2014

- Reduced multi-hour SQLAlchemy bulk database jobs to minutes.
- 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.
- Wrote Python framework to manage long running background jobs.
- Debugged and resolved memory issues that were causing systems to shut down.

Relevant Technical Skills

Big Data

	Time Used	Last Used	**Proficiency**
HDFS	9 years	2023	Very Good
Hive	9 years	2023	Good
YARN	9 years	2023	Good
Alluxio	3 years	2023	Good
Impala	9 years	2023	Fair
Trino	3 years	2023	Good
Kafka	9 years	2023	Very Good
Kubernetes	4 years	2023	Good

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

Database Skills

	Time Used	Last Used	**Proficiency**
PostgreSQL Database Administration	1 year	2011	Fair
Relational Schema Design	14 years	2023	Very Good
Structured Query Language (SQL)	14 years	2023	Very Good

Software Configuration Management Tools

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

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.
- 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.

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][MARIOADB], 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.

Michael J. Pedersen datacyclist@gmail.com 908-283-0318