PAVEL YAKOVLEV

yvpavel@gmail.com
github.com/paulyakovlev

EXPERIENCE

General Motors April 2025 - Present
Data Engineer, Connected Vehicle Data Engineering *Austin, TX*

- Implement connected vehicle telemetry collection by configuring data element definitions and event-based triggers, deploying to production fleets via automated tools, and validating data accuracy using SQL, Kafka, and Databricks.
- Resolved production data collection failures for rental fleet partner (100,000+ vehicles) by implementing direct read method to bypass queue bottleneck, ensuring accurate fuel level and odometer telemetry for billing systems.
- Developed Python ETL pipeline parsing and combining CSV/XML files from Git repository into Databricks tables, coordinated cross-functional deployment and permissions, enabling PowerBI analytics on production configurations.

General Motors February 2021 - April 2025
Software Engineer, Big Data Engineering & Infrastructure *Austin, TX*

- Led migration from Cloudera Data Analytics Studio to Tez UI, eliminating closed-source dependencies and reducing licensing costs.
- Responded to Log4Shell zero-day vulnerability by developing automated remediation script that scanned and patched production Hadoop clusters, then integrated solution into standard cluster startup process to ensure ongoing protection.
- Contributed to Apache Ambari and Apache Bigtop open-source projects by implementing Livy service integration and service management (Ambari PR#3790, Bigtop PR#1282).
- Integrated open-source Hadoop components (Sqoop, Livy, Tez UI) into GM's custom Hadoop platform, collaborating with platform engineers to migrate from Cloudera Hadoop Platform and reduce annual costs by $1.2M.
- Developed Python-based ETL pipeline converting cluster usage audit logs to Parquet format and ingesting into Hive tables with Oozie automation, enabling PowerBI analytics to support data cleanup decisions during cloud migration planning.
- Enhanced Docker-based Hadoop development environment with local RPM repos and auto-build + deployment of Ambari, significantly improving setup time for developers.

PROJECTS

Long Marine Lab Stranding Map https://github.com/lmlstrandingnetwork/lml-stranding-map

- Built full-stack geospatial web application with Node.js/Express REST API backend and React frontend, enabling scientists to digitize marine mammal stranding records with location tracking, photo uploads, and filterable search
- Mentored next cohort of engineering students on codebase, development practices, and technical implementation.
- Integrated Mapbox for interactive mapping, Algolia for search, and Firebase backend; deployed to Heroku

EDUCATION

University of California, Santa Cruz 2020
Bachelor's Degree in Computer Science

SKILLS

Languages: Python, Java, Bash, SQL, JavaScript

Backend & Integration: REST APIs, ETL Pipelines, Kafka, Databricks, Node.js, Express, Docker

Data & Tools: SQL, PowerBI, Hive, Apache Hadoop, Apache Ambari, Oracle

DevOps: Git, CI/CD, Azure DevOps, Maven, Gradle, Jenkins, RPM