HEMANT SETHI

Staff Infrastructure Engineer

+1-401-636-6527 • sethi.hemant@gmail.com • https://www.linkedin.com/in/hemantsethi/ • San Francisco Bay Area, CA

Summary

Staff Infrastructure Engineer with 12+ years building high-performance data platforms and AI-optimized storage systems for mission-critical ML workloads. Specialized in S3-compatible object storage with intelligent caching for GPU-intensive training (Crusoe), high-throughput fully-managed data ingestion and delivery platform processing billions of events daily (AWS Firehose), and large-scale event-driven infrastructure (WeWork). Expert in distributed systems architecture, performance optimization for AI workloads, and cross-functional technical leadership. Aspiring technical author and industry speaker on AI infrastructure and storage optimization.

Experience

Crusoe
San Francisco, United States

Staff Infrastructure Engineer
07/2025 - Present

- Leading the architecture and development of **Crusoe's AI-optimized object storage platform**, enabling high-performance data access for GPU-intensive machine learning workloads at scale.
- Supporting **dual-protocol** storage interface supporting both **NFS** and **S3** access patterns over unified datasets, enabling seamless integration with legacy HPC workflows and modern cloud-native AI workloads.
- Designing and implementing intelligent **node-local(Tier-0)** caching and pre-fetching strategies to minimize data access latency for distributed AI training workloads, reducing data bottleneck time resulting in higher GPU Utilization reducing cost-per-GPU-hour(TCO).
- Building **multi-cluster**, **multi-vendor** storage infrastructure enabling rapid deployment of new storage clusters reducing operational overhead.
- Conducted technical research on storage access patterns for transformer models and large language model training and inference.
- **Benchmarking** and profiling data loading patterns for PyTorch/TensorFlow training pipelines, identifying and resolving storage bottlenecks.

Amazon
Palo Alto, United States

Software Development Engineer
07/2020 - Present

- Led the architectural strategy and execution of migrating **Firehose's Dynamic Partitioning** to microservices-based **Gen2** architecture, driving long-term scalability and operational efficiency while managing a team of 5 engineers.
- Redesigned the **partitioning** and **scaling strategy** to improve the scalability of **Dynamic Partitioning Streams,** increasing per-partition throughput by **400%** and partition count by **1000%** using **Akka's PartitionHub/MapAsync,** eliminating manual operator intervention for scaling and related customer tickets.
- Defined and drove the cross-team architectural design (across a team of 15 engineers) of **data Processing framework** enabling efficient workload offloading and streamlining data transformations using **SQS**, **DynamoDB/Redis Pub-Sub**.
- Implemented a generic Firehose **lease management library** to support various apps during **Gen2 migration** across **Firehose's backend services**.
- Led root cause analysis and performance optimizations, profiling **Java services (Native Memory, Heap, & Thread Dumps)** to diagnose and resolve **high GC/CPU** utilization issues due to **Native Memory leaks,** improving system stability under extreme load conditions.
- Improved **operational posture** by defining **SOPs, dashboards, & alarming** to on-call response workflows reducing **MTTR** for high-sev incidents.
- Spearheaded Firehose's **Load Balancer (LB) migration**, replacing hardware-based VIPs with Software **Elastic Load Balancers (ELBs)**, reducing operational costs, enabling **on-demand scaling** for traffic bursts, and cutting **VIP expansion** time from **months to days**.
- Led the Firehose region expansion (with 3 engineers) in **KIX** and introduced automations to reduce the region launch times by **67%**.
- Mentored engineers across teams, leading **tech deep dives** and **onboarding sessions** to accelerate ramp-up, fostering engineering excellence.

WeWork
San Francisco, USA

Software Engineer
10/2018 - 07/2020

- Worked in a team of 5 engineers to migrate WeWork's legacy PHP-based space service to a Go/gRPC/Protocol Buffers/Kubernetes-based micro-services architecture, reducing latency by 90% and achieving P99 latency of 100ms. This service served as the source of truth for all spatial data at WeWork, supporting core business use cases such as inventory management and occupancy map visualization.
- Integrated Kafka-based event-driven system to synchronize spatial data with downstream services, ensuring real-time data updates for critical business functions like Inventory Management. Used Apache Avro for data serialization and schema registry for schema enforcements for events sent to Kafka. Implemented Observability (Logs, Metrics, Traces, and Dashboards) using Jaeger, Prometheus, and Grafana.

Experience

Tata Consultancy Services
Woonsocket, United States & Gurgaon, India

Software Engineer
06/2012 - 10/2018

- Worked on a Large Scale Batch Processing System(**ReadyFill**) for **CVS's** Pharmacy application and across various functional releases.
- Worked as a lead developer for a consumer facing web application, **VaccineClinicScheduler**. Responsible for designing data models and implementing Secure Login API (Security standards around username/password functionality for the application) and application workflows. Used spring boot, Hibernate, Apache HTTP Server, tomcat, Mariadb. Wrote Chef Scripts for infra-automation (httpd, and tomcat setup/configuration) and host service's environments on Openstack. Implemented CI/CD Jenkins pipeline for applications within CVS that resulted in faster deploy and feedback cycles across applications that helped save 30% of Dev/QA's effort in deployment/test cycles.

Education

NIT Jalandhar
Jalandhar, India

Bachelor of Technology in Computer Science
07/2008 - 06/2012

- CGPA : 7.41/10
- Undergraduate Subjects: Operating Systems, Databases, Algorithms, Object Oriented Programming, Distributed Systems, Computer Networks etc.

Skills

ECS Fargate · Cloudwatch · StepFunctions · DynamoDB · Lambda · SQS · Redis · S3 · CloudFormation · IAM · STS · Java · Go · Python · Docker · gRPC · Kafka · Kubernetes(K8s) · Mysql · Postgres · Grafana · Prometheus · REST · Gradle · Jaegar · GIT · Akka · Kinesis · MongoDB · ElasticSearch