

Aviral Srivastava

• [Github](#) • [LinkedIn](#) • [Website](#) • +1-617-283-3811
• [Email](#) • [Blog](#) • [Google Scholar](#)

EDUCATION

Boston University, Boston, MA. MS, Computer Science. CGPA : 3.77/4 Sep 2019 - Dec 2020
Vellore Institute of Technology, Chennai, B.Tech, CS and Engg. CGPA : 3.56/4. Jul 2014 - May 2018

SKILLS

- **Languages:** Rust, Python, Scala, C++, Go, Java, PHP, SQL
- **Data & Streaming:** Apache Spark, Apache Kafka, Apache Flink, Apache Iceberg, Debezium, Airflow, Argo, Hive, PrestoDB
- **Databases:** Aurora (MySQL), RocksDB, KyotoDB, Cloud Spanner, AlloyDB, TiDB, Aerospike, Key-value DBs
- **Infrastructure:** Kubernetes, Docker, Vagrant, Mesos, Pulumi, Terraform, Helm
- **Cloud & Platforms:** AWS, GCP (Cloud Run, Spanner, AlloyDB), Azure, On-Premise Data Centers
- **Observability:** Datadog, Prometheus, Grafana

COMPENSATION

- **Current Total Compensation (Affirm, 2025):** \$200,000 base salary + \$7,500 annual cash stipend + \$150,000–\$200,000 in publicly traded Affirm stock (RSUs). Total annual compensation: ~\$357,000–\$407,000.
- This places compensation well above the 90th percentile for software engineers in the United States, reflecting the specialized and senior nature of the role in data infrastructure and distributed systems.

WORK EXPERIENCE

Affirm, Senior Software Engineer (L6), Lakehouse Infrastructure, Data Storage Services, NYC
Mar 2025 - Present

- Designed and built an end-to-end streaming **CDC pipeline** (Debezium → Kafka → Flink → Iceberg) for MySQL-to-Iceberg replication, replacing unreliable DMS-based ingestion with a provably correct, auditable data capture system targeting **2000 tables, 20M rows, <10 min latency SLO**.
- Authored the architectural decision record (ADR) for **out-of-band validation** — a key design insight that decouples capture completeness verification from compaction correctness, enabling independent assurance of per-transaction data integrity.
- Built comprehensive **failure mode and load testing frameworks** (65+ tests across Go and Python with CI integration) to validate pipeline resilience under production-scale conditions. Authored 57+ pull requests in the first year.
- Owned and led resolution of multiple **production incidents** across CDC and data replication infrastructure, driving root cause analysis and actionable postmortem follow-through across tightly-coupled financial pipelines (Ledger, Funding). Authored 7 proposed improvements to the DMS restart runbook, of which 6 were adopted.
- Supported migration of OLAP workloads off MySQL onto a governed **Lakehouse platform** (Apache Iceberg), addressing ~17 onboarding requests in 6 months from teams building ad-hoc pipelines. Pipeline architecture underpins Affirm's long-term **database strategy** (Aurora replacement), as both TiDB and Vitess candidates rely on the same Debezium-based binlog extraction path.
- Partnered cross-functionally with NewDB (TiDB/Vitess evaluation), Ledger Engineering, and Analytics teams to ensure pipeline compatibility across database engine candidates and downstream consumers.
- Mentored a teammate on CDC pipeline internals and lake infrastructure patterns, contributing to their technical growth and independent ownership of workstreams.
- Contributed a major feature to the open source **Debezium** project (industry-standard CDC platform used by thousands of organizations worldwide) in the context of building Affirm's streaming infrastructure.
- Implemented HA KafkaConnect (3 replicas, multi-AZ), Avro serialization with Schema Registry, Terraform IaC for Kafka (service accounts, ACLs, secrets), and multi-cluster architecture support.
- Stack: **Go, Python, Debezium, Kafka, Flink, Apache Iceberg, Terraform, Kubernetes**.

Pinecone, Senior Software Engineer, Platform, NYC

Apr 2023 - Oct 2024

- Single handedly expanded **Pinecone's serverless offering to GCP from zero to GA**. Redesigned core services for cross-cloud compatibility, including a custom serverless solution with auto-scaling for AlloyDB. This was the first cross-cloud expansion of Pinecone's serverless product, a critical business milestone.
- Engineered robust **failover detection** and zero-downtime migration strategies for **distributed database systems**. Enhanced comprehensive benchmarking tools and identified performance bottlenecks across cloud providers.
- **Built Global Control Plane**, enabling product's transition to serverless architecture with zero-downtime migration. Reduced request time from 10 seconds to around 1 second, and decreased operational costs by 10x. Implemented **Observability** for real-time metrics and alerting for proactive issue resolution in distributed environments.
- Early joiner at a high-growth startup: owned broadly across the Database tech stack including infrastructure-wide oncall, Cost-Engineering, System Stability, and CI/CD.
- Served as on-call engineer for both Pinecone Serverless (v4) and Pod-based (v3) distributed systems, gaining comprehensive knowledge and troubleshooting experience across the full product surface.
- Stack: **Rust, Google Cloud Run, Cloud Spanner, AlloyDB, Pulumi, SeaORM, Datadog, Kubernetes**.

Twitter, Software Engineer → Software Engineer II, Graph Storage (Platform/Real Time Storage), NYC/Seattle

Apr 2021 - Feb 2023

- Promoted from Software Engineer to Software Engineer II within 1 year 3 months (Apr 2021 → Jul 2022).
- Replaced the decade old **Flock** (Graph Storage service) with a new service (Flock v1.5) that scales the write pipeline, improves data consistency, and reduces human ops. Led setting up v1.5 as a **multi-data center setup** — a critical infrastructure initiative for Twitter's graph storage layer.
- Proposed, designed, and developed **Backup Verification** in **Scala**: a service to verify **300 TB** of backups as they are used in v1.5 to serve live traffic. This was an original system design to ensure data integrity at massive scale.
- Performed **oncall operations** such as rebalancing shards, restoring servers during heatwave, modifying mirrorsets to cater fast product iterations, etc.
- Led the team for building a **managed graph database offering** for analytics purposes, graph machine learning, and applications. Reduced onboarding time from **1 week to minutes** for data teams at Twitter.
- Supervised request-based authorization for Flock data. **Mentored two interns** through the full development lifecycle.
- Stack: **Scala, Java, Mesos, Manhattan (KV store), MySQL, multi-datacenter replication**.

Atlan (formerly SocialCops), Software Engineer (Data), New Delhi

May 2018 - Jul 2019

- Atlan (formerly SocialCops) is the world's leading data-for-good company, recognized as a **New York Times Global Visionary** and Yourstory Best Impact Startup in India. By 2018, the company's platforms had impacted the lives of **over a billion people**.
- Ensured the development and on-premise deployments of dashboards including: (1) **United Nations SDGs monitoring** (global partner for the UN for countries including India, Sri Lanka, and Papua New Guinea), (2) **Ujjwala Dashboard** (gas subsidy for 50M Indian women below the poverty line — the first government program in India to surpass 22% of its annual target), and (3) the **National Data Platform (DISHA)** used by the Prime Minister of India and every member of parliament. Stack: **Python3, Kubernetes, Docker and AWS S3**.
- Integrated serverless JupyterHub: a multi-user hub that spawns, manages, and proxies multiple instances of the single-user Jupyter notebook server. Stack: **papermill, Python3, Docker, Helm, Kubernetes**. Used by companies like **Unilever, Mindshare, McKinsey & Co**.
- Built and deployed **Data Repository**: a version control system for tabular data, including features like 'git diff', rollback, at a scale of over **billion rows in one cycle**. Stack: **Python3, S3, RocksDB, KyotoDB, Hive metastore and PrestoDB**. This scaled the production releases by **300%**: 1 project release to 3 project releases per quarter.
- Co-built an in-house project, **Pallet-Core**: Airflow's abstraction for all dashboards (Ujjwala, UNSDG, Disha, etc.).
- Built **Collect's** pricing unit (Woodward) and introduced event streaming using Kafka in the product.

Red Hat, Software Engineering (Data) Intern, AICoE / Internal Data-Hub team, Boston May 2020 - Dec 2020

- Worked in the AI Center of Excellence (AICoE) team to set up the internal data platform, making data access efficient while removing hidden technical debts to empower AI and ML teams across the organization.
- Re-partitioned the schema of **SOS reports (10 Trillion+ rows, 100+ TeraBytes data)** to ensure high data availability and low latency.
- Stack: **Argo, Spark, JupyterHub, Kubernetes, OpenShift, OpenStack, Ansible, Python.**

PUBLICATIONS

- Srivastava, A., Patel, F., and Sivagami, M. *A Software Requirement Engineering Technique Using OODA-RE And CSC For IOT Based Healthcare Applications*. International Journal of Software Engineering and Applications (IJSEA), 2018. Available on [Google Scholar](#).

OPEN SOURCE

- **Debezium — Custom Binlog Position Signal**: Implemented a new signal action for the Debezium MySQL connector enabling CDC streaming from a custom binlog file and position ([DBZ-3829](#)). Added SetBinlogPositionSignal with validation, integration tests (365 lines), and unit tests (193 lines). 778 additions across 7 files. **Merged**. Debezium is the industry-standard open source CDC platform used by thousands of organizations worldwide.
- **Zcash Service Status Dashboard**: Built a health check dashboard for monitoring Zcash (ZEC), one of the leading privacy-focused cryptocurrencies. Received **funding from the Zcash Foundation** through their community grants program to implement this project. Includes a companion [Python library](#). **Python**.
- **CPython — end_lineno for pylbr**: Added end_lineno attribute to pylbr Function and Class objects in the Python standard library ([bpo-38307](#)), enabling tools to determine the scope of classes and functions in a module. Shipped in **Python 3.10**. **Merged**.
- **GRUML (Generates Rectangular UML)**: Automatically generates UML diagrams from source code in a novel rectangular format (RUML) rendered as spreadsheets. Shows class dependencies, inheritance hierarchies, and control flow. Research project with Dr. Eric Braude at Boston University. Includes a [web application](#). **Python**.
- **Other**: [E-Commerce Framework](#) (PHP/MySQL, 8 stars, 7 forks), [Cloud CI/CD Pipeline](#), and forks/contributions to Apache Kafka, Apache Flink, pandas, PyTorch-BigGraph, raft-rs, and Zulip. 62 total repositories on GitHub.

INVITED TALKS

- **Versionator: Data Repository, integrated with Airflow** — Invited speaker at **Google DevFest '18**, organized by **GDG New Delhi** (Google Developer Group), Nov 2018. Presented original work on a scalable data versioning system (version control for tabular data at the scale of billions of rows per cycle), covering design tradeoffs vs. Pachyderm and git-lfs, and a diff-based patch pipeline feeding Elasticsearch-backed dashboards that power United Nations SDG monitoring and India's Ujjwala Yojana programs. DevFest is Google's flagship global developer conference series, hosted annually by GDG chapters worldwide. Speaker profile: [commudle.com/users/user1023](#). Talk listing: [commudle.com/speaker-resources/12](#).

ENTREPRENEURSHIP

- Centify Technologies**, Co-founder & CTO, India Mar 2018 - Sep 2018
 - Built audience engagement tools enabling digital media platforms to integrate social reactions with video or textual content.
 - **Selected for Y Combinator Startup School '18, mentorship track**. Shelved the project after iterating on initial customer feedback revealed distribution challenges.

- BetaApple**, CEO, Chennai Apr 2015 - Nov 2016
 - Created a blogging service for Apple consumers, bridging trusted local stores with customers needing quality device services at affordable prices.
 - Total revenue: 120k INR. Customers bridged in Lucknow: **1,000+**.

- Find Signal Studio**, Engineering Advisor, New Delhi Sep 2019 - Present
 - Technical advisor consulting on blockchain infrastructure and databases. Long-standing advisory role spanning 6+ years.

RESEARCH

- Boston University**, Research Associate, Greater Boston Area Sep 2019 - May 2020
 - Worked with **Dr. Eric Braude** on making UML diagrams work at scale in large software codebases. This research produced the GRUML/RUML open source project.
 - **Led the team at MIT-ICORPS** for the commercialization of the research: discovered Value Propositions and Customer Segments for the research projects. MIT I-Corps is a highly selective NSF-funded program for translating research into commercial impact.
 - Conducted interviews with Senior Software Engineers and Engineering Managers from different organizations and companies to pivot the research work and make it more objective.

- Developed a **Neural Mass Computational Framework** to study synaptic mechanisms underlying alpha and theta rhythms. Successfully built a neural-mass computational model which can generate rhythms for both the thalamus and hypothalamus regions of the brain.
- This research directly produced the [py-eeg](#) open source project (20+ stars), a Python library for simulating EEG readings using the Kinetic LGN model. Fills a gap in computational neuroscience where no existing software could run this model. Targets brain-computer interface (BCI) and medical research.

University of Liverpool, Data Science Research Intern, Liverpool, UK May 2017 - Aug 2017

- Implemented **Mapper's Algorithm** in C++ with Dr. Vitaliy. The existing Python version lacked visualization. Improved visualization by generating intermediate steps of processing. Cloud generation for all algorithms (Gauss Density, kNN, etc.) showed significant improvement in understandability.
- **Led a team of 3 interns** in project management including development, testing, version control, and deployment of the software.

Polkadot-IPFS (PIPFS), Independent Research Project Jan 2020 - Jun 2020

- Built a CLI tool bridging the Polkadot/Substrate blockchain with IPFS (InterPlanetary File System). Upload files to IPFS and record ownership on a Substrate chain via a custom pallet. Includes a companion [custom Substrate node](#). **Rust**.

INTERNSHIPS

Wingify, Software Development Intern, New Delhi Dec 2017 - Apr 2018

- Built **Segmenter Services**: a microservice for PushCrew and VWO (Visual Website Optimizer) that parses and executes database queries for A/B testing. Applied Shunting Yard algorithm for rightful parsing of mathematical equations generated by the platforms.
- Coordinated with QA team to sort out all permutations and combinations across different user stories and sequences in VWO and PushCrew.
- Stack: **PHP, PHPUnit, MySQL**.

Commutatus, Full Stack Developer Intern Jun 2016 - Jul 2016

- Developed e-commerce websites and business models using HTML, CSS, JavaScript, Angular2.js, Python/Django and PHP.

LeagueSX, Python/Django Backend Developer Intern Dec 2015 - Jan 2016

- Built parsing code and Private League functionality for a fantasy football platform.