

Zhuocheng (Zoe) Shang

Software Engineer, Distributed Query Execution and Data Infrastructure | PhD Candidate, Computer Science (Dec 2026)

Riverside, CA | 636-206-9665 | zshan011@ucr.edu

[linkedin.com/in/zhuocheng-shang](https://www.linkedin.com/in/zhuocheng-shang) | github.com/ZhuochengShang | cs.ucr.edu/~zshan011

PROFILE

- Designed Spark-based query engines that carry multi-terabyte geospatial and scientific workloads
- Managed partitioning, operator pipelining, cost-based optimization, and shuffle reduction on distributed systems
- Builds agentic systems that autonomously generate, execute, and repair code against specialized, undocumented codebases
- Author of 13 peer-reviewed publications at top systems and data venues (VLDB, SIGSPATIAL, SSTD)
- Open-source contributor to Apache Sedona

TECHNICAL SKILLS

- **Distributed Query Execution:** Apache Spark, Spark SQL, Apache DataFusion, Apache Arrow, task/stage scheduling, partitioning, operator pipelining, execution reuse, materialization, shuffle reduction, spill control, compression
- **Query Optimization:** query planning, domain-specific optimization, spatial joins, indexing
- **Data and Metadata:** columnar formats (Parquet, Arrow), unstructured satellite imagery, scientific data (GeoTIFF, NetCDF), schema and metadata design, PostgreSQL, PostGIS
- **Agentic AI and Verification:** LangGraph, MCP servers, retrieval-augmented generation (RAG), microbenchmarks
- **Languages and Tools:** Scala, Java, Rust, Python, SQL, Git, CI/CD

EXPERIENCE

University of California, Riverside

Sep 2021 – Present

PhD Researcher, Distributed Systems and Agentic AI

Riverside, CA

- Lead the architecture, implementation, and evaluation of the four production-grade systems described below
- Work with Spark-based query engines and agentic code-generation tools
- Translate research challenges in remote sensing and scientific fields into distributed operators on terabyte-scale datasets
- Publish research and release open-source code for all four systems, maintaining continuous integration, regression tests

Wherobots

Jun 2025 – Sep 2025

Software Engineering Intern, Apache Sedona and Spark SQL

Bellevue, WA

- Ported 5,000+ lines of Google's S2Geography from C++ to Java, delivering a production Geography type integrated into Apache Sedona and Spark SQL's query planner and DataFrame API.
- Implemented core geometry algorithms (S2 closest-edge distance, chord-angle methods, topological predicates, WKT/WKB parsing) with Kryo encode/decode support for distributed execution.
- Authored Java microbenchmarks and CI regression suites that caught correctness and performance regressions before release, hardening a library now used in production spatial SQL queries.

LenderPrice

May 2021 – Aug 2021

Software Engineering Intern

Pasadena, CA

- Built backend REST API features over relational databases and implemented unit tests for front-end components
- Diagnosed and resolved database-related performance bottlenecks with the application team, optimized user experience.

SELECTED SYSTEMS AND PROJECTS

RDPro and FieldSAT: Distributed Engine for Multi-Terabyte Raster/Vector Workloads (VLDB 2024, SIGSPATIAL 2025)

github.com/ZhuochengShang/RDPro-Big-Raster-Data-Processing

- Designed the Maplet data model on Spark (Scala) to process imagery datasets ranging from 700 MB to 6.7 TB on a 12-node cluster.
- Reduced ingestion overhead to near zero through lazy evaluation and metadata-only alignment of heterogeneous datasets.
- Reduced network shuffle by 50% for complex window operations; reduced post-shuffle data volume by 60% through compression.
- Achieved 3–6× speedups over GeoTrellis, which ran out of memory on datasets larger than 150 GB in these benchmarks.
- Extended RDPro into FieldSAT to accelerate real-world precision agriculture workflows produce time-series based analyzed results.

GRAIL and AIDEAL: AgenticCode Genera on Against a Specialized Codebase (VLDB 2026 Demo)

github.com/ZhuochengShang/GRAIL

- Built GRAIL, a LangGraph coding agent that converts natural-language specifications or Python scripts into runnable Scala code.
- Designed a context layer that provides API definitions, structured documentation, execution constraints, and error feedback
- Improved end-to-end task success from 0/5 to 5/5 across raster and raster-vector workflows.
- Developed AIDEAL to assess how coding agents use unfamiliar SDKs through standardized evaluation metrics and recommendations
- Designed controlled experiments comparing original, generated, and repaired documentation, isolating the effects of improvements
- Evaluated 532 APIs across three open-source libraries, increasing aggregate first-attempt pass rates from 63.2% to 92.7% with generated and repaired documentation, without code-repair retries.

LearnDB: A Distributed Query Execution Engine in Rust (Personal Project)

github.com/ZhuochengShang/LearnDB

- Building a distributed query execution engine in Rust using Apache DataFusion to study the full query lifecycle, from SQL parsing and planning to distributed execution.
- Implementing task and stage scheduling, partition-aware execution, and a reference-counted in-memory store to explore asynchronous execution and memory management.
- Developing a visual dashboard to show how query plans map to stages, tasks, and partitions distributed across workers.
- Comparing broadcast and shuffle hash joins using DataFusion's physical plans, examining stage boundaries at shuffle and merge operators, and implementing merge and eviction logic to validate execution behavior.

GS-QA2: Benchmark for Question Answering over Raster and Vector Data (SIGSPATIAL 2026)

github.com/ZhuochengShang/QARV

- Designed and built GS-QA2, integrating OpenStreetMap features with a 265,950 tile, 17.4 billion cell U.S. elevation model across 25 templates spanning local, focal, zonal, and global map-algebra operations.
- Generated 500 question-answer pairs with executable PostGIS ground truth and zero manual annotation, scored with output-type-specific metrics (token F1, geodesic error, directional error, relative error).
- Benchmarked Text2SQL, RAG, CHESS, and a GIS coding agent, showing the dominant failure mode is semantic rather than syntactic: the best system reaches 0.234 accuracy on raster queries and near zero on zonal/global operations despite emitting syntactically valid SQL.

Viper and DynoViz: Low Latency Retrieval over Unstructured Satellite Imagery (SSTD 2023)

github.com/ZhuochengShang/DynoViz-Visualization

- Built a query-time index over imagery spanning hundreds of billions of pixels, answering zonal statistics up to 100x faster than baselines at 90%+ accuracy via SuperPixel approximation.
- Served interactive queries across 10,000+ files in under 0.5 seconds using 60x less disk than full materialization, via a frequency-driven materialization strategy.
- Deployed both engines inside FutureFarmNow, a live application covering the 640,000 sq km Colorado River Basin.

PUBLICATIONS

- RDPro: Distributed Processing of Big Raster Data. VLDB 2024.
- GRAIL: AI Translation for Scientists' Application Workflow on Satellite Data. VLDB 2026 Demo.
- FieldSAT: A Scalable Query Workflow for Precision Agriculture. SIGSPATIAL 2025 (Best Application Paper).
- GS-QA2: A Benchmark for Question Answering over Raster and Vector Data. SIGSPATIAL 2026.
- Geo-ReAct: An Agentic AI Framework with Concept Grounded Tools for Geospatial Question Answering. SIGSPATIAL 2026.
- Scalable Raster Processing: Models, Systems, Algorithms, and Open Challenges. SIGSPATIAL 2025 Tutorial.
- Viper: Interactive Exploration of Large Satellite Data. SSTD 2023.
- End-to-End Raw Scalable Raster Processing with Customized Pipeline and Foundation Model Integration. SIGSPATIAL 2025.
- FutureFarmNow: Interactive Spatial Data Exploration for Precision Agriculture. SIGSPATIAL 2025 Demo.
- Additional: AgriParcel (GeoAI Workshop 2025) | DynoViz (Big Geospatial Data 2024) | Object Delineation in Satellite Images (SpatialGems 2023) | Can SAM Recognize Crops? (NeurIPS 2023 Workshop).

AWARDS, TEACHING, AND SERVICE

- Best Application Paper, SIGSPATIAL 2025 | First Place, Student Research Competition, SIGSPATIAL 2025
- Teaching assistant for CS 167 Big Data and CS 226 Big Data Management. Reviewer and tutorial presenter at SIGSPATIAL.

EDUCATION

University of California, Riverside | PhD, Computer Science – Expected Dec 2026

University of California, Riverside | MS, Computer Science – 2019 to 2021

Lindenwood University | BS, Computer Science, Magna Cum Laude – 2015 to 2019