Saptashwa Mitra

saptashwamitra7790@gmail.com • (216) 600-6382
• GitHub • Google Scholar • LinkedIn • Fort Collins, CO 80521

Data Scientist Profile

Skillful and highly analytical professional with substantial experience in research and development. Demonstrated strong expertise and innovation in large-scale, distributed data analytics, interactive visualization, deep learning, project management, and software development. Well-versed in several programming languages and frameworks. Excel at researching, developing, and deploying innovative in-memory storage and analytics frameworks to alleviate excessive/redundant disk and network I/O and hotspots. Implemented deep learning to increase interactivity of data exploration and other in-memory distributed frameworks to collect, explore, and extract insights from structured and unstructured data. Competent at performing large scale experimentation; and crafting, prototyping, and delivering advanced algorithmic solutions. Ability to adapt to change and willingness to be flexible in a global team environment. Recognized for innovation with Best Paper Award (IEEE Cluster, 2019), Graduate Research Fellowship (2020) and the Wim Bohm & Partners Award (2022).

Technical Proficiencies

Big Data Ecosystem:	Apache Hadoop, Spark, Storm, GraphX, Kafka, Ignite, Elasticsearch.
AI/Machine Learning:	PyTorch, PyTorch Lightning, Tensorflow, Spark MLlib, Horovod, Scikit-Learn.
Containerization:	Docker, Kubernetes.
Database:	MongoDB, IBM DB2, Oracle 9i/10g, PostgreSQL, Microsoft SQL Server.
Languages & Technologies:	Java, Scala, Python, Matlab, HTML, PHP, C, C++, C#, Hibernate, Spring MVC, Kibana, Unity, Tableau.
Web Frameworks:	Flask, Django, Tomcat, JBOSS, WAS.
Operating Systems	Linux, Windows.

Work Experience

Intel Corporation | Chandler, AZ **May 2022-Present**
Machine Learning Graduate Intern

- Implemented automated deployment of client applications over a Kubernetes cluster.
- Designed and implemented a *distributed, fault-tolerant online AI Inference Engine* for large inference jobs for OpenVINO deep-learning Image Recognition models (trained using Intel's Sonoma Creek).
- Developed a robust server with *multi-priority queues* with parallel evaluation and load-balancing of *concurrent inference jobs* of varying importance. Implemented checkpointing for error-recovery from failure scenarios and to reduce computational overloads.
- Designed lightweight, web-based front-end, using Flask-RESTful, for users to interact with the inference framework.

Department of Computer Science, Colorado State University | Fort Collins, COJan' 2018-May 2022, Summer 2016 & 2017
Graduate Research Assistant

- Utilized Kubernetes containers to coordinate efficient parallel model-fitting over a distributed cluster, leading to reduction of workloads via identification of dependent data regions (through clustering) for transfer learning.
- Facilitated interactivity by generating requested high-resolution satellite data through a progressive deep GAN network (GLANCE) to bypass communication with server and achieving *~297x-6627x* speedup.
- Supported in-memory data storage and analytics with dynamic auto-scaling capability on top of Apache Ignite (STRETCH).
- Implemented *fast spatiotemporal integration (order of seconds)* of voluminous, heterogeneous datasets through real-time, distributed query relaxation over a framework (CONFLUENCE) with a lightweight spatiotemporal boundary-based data indexing.
- Monitored and identified critical risk factors for childhood obesity through predictive modeling over biometric and non-biometric attributes over diverse health datasets combined through geospatial data integration.
- Slashed resource requirements and enhanced throughput by *27-42x* by deploying distributed framework (RELAY) for multi-user parallel deep learning and utilizing spatial clustering.
- Enhanced interactive visualization on clients with speedups of *~3.3x* compared to other cache-driven distributed analytics framework by designing server-side in-memory distributed cache (STASH) for large scale spatiotemporal data.
- Drove quick data ingestion and interactive visualization by designing a 2-tier hashgrid-based index (RADIX+) to support high-throughput geo-referencing of voluminous sensor data streams and in-memory feature aggregation for fast query evaluations.
- Reviewed multiple technical papers submitted for publication in different international data science conference proceedings.
- Mentored for *SWiFT: Education Outreach for Female High School Students*, a STEM summer-camp for aspiring women engineers.
- Served as Technical Program Chair at Graduate Symposium at Department of Computer Science, CSU, 2019.

Cognizant Technology Solutions, India **June 2012 – July 2015**
Associate Analyst

Fostered sending, receiving, and processing of information between various GDS and users for travel booking services for Travelport.

- Promoted to associate analyst within 2.5 years via demonstration of impeccable performance.

IBM, India **2011 – 2012**
Intern – Remote Mentoring Program

Enabled users to visualize spending habits over an interface through IBM Cognos over Smart Banking. Met all targets and delivered the desired banking analytics software.

Department of Computer Science, Colorado State University | Fort Collins, CO **Aug' 2015 – Dec' 2017**
Graduate Teaching Assistant

- Assisted in teaching the following courses: Introduction to Big Data (CS435), Big Data (CS535), Advanced Networking (CS 557), Introduction to Unix & C (CS 155/156/157), Introduction to Java (CS 164)
- Conducted recitation sessions and guest-lectured on big data systems and in-memory analytics frameworks.

Academic Projects

Evaluating Topic Virality on Twitter Based on User Rankings (Term Project for Distributed Systems, CS455)

- Used Spark GraphX to combine virality of diffusion tree along with the rank of the author to evaluate the popularity of tweets.

Distributed Deep Learning for Large-Scale Image Processing (Term Project for Machine Learning, CS545)

- Utilized Deep Convolutional Neural Network (CNN) in SparkNet to categorize CIFAR-100/ImageNet dataset

File Sharing Software Using Named Data Network (NDN) (Advanced Networks, CS557)

- Leveraged DeterLab testbed for deployment and ndn-cxx C++ library for coding to develop an NDN-based file-sharing network.
- Supported many-to-many file transfer for multiple files in the network along with system performance feedback.

Training Flappy Bird using Reinforcement Learning (Term Project) (Introduction to Machine Learning, CS480A3)

- Utilized Ω500 JavaScript game engine to train Flappy bird game using Q-Learning.
- Trained bot on parameters, including distance from the 2 closest upcoming pipes at any time -> bird never died

FreeBits, a file sharing network loosely resembling BitTorrent (Advanced Networks, CS557)

- Developed file sharing network, like Bit-Torrent, via effective utilization of concepts of peer-to-peer networking in the language C.

Quantifying Virality of Hashtags from Live Twitter Streams using Apache Storm (Term Project for Big Data, CS535)

- Designed and deployed real-time streaming data analytics system by leveraging Apache Storm using Lossy Count Algorithm.

Estimating PageRank Values of Wikipedia Articles using Hadoop MapReduce (Big Data, CS535)

- Facilitated distributed computation of PageRank values with(out) taxation for 50GB of Wikipedia Articles using Hadoop.

Education

Ph.D. in Computer Science | Aug'2018 – Dec'2022 (Expected) | Colorado State University, Fort Collins, CO
M.S. in Computer Science | Aug'2015 – June'2018 | Colorado State University, Fort Collins, CO
B.E. in Computer Science | 2008 – 2012 | Indian Institute of Engineering. Science and Technology, Kolkata, India

Honors and Awards

- Best Paper at IEEE International Conference on Cluster Computing (CLUSTER), 2019.
- Graduate Research Fellowship, 2020, at Dept. of Computer Science, Colorado State University.
- Wim Bohm & Partners Ph.D. Award, 2022, at Dept. of Computer Science, Colorado State University.

Professional Certification

Oracle Certified Java SE6 Programmer; EI Systems Certified L1- Level Android Application Developer

Publications & Technical Reports

[1] Caleb Carlson, Menuka Marushavithana, **Saptashwa Mitra**, Kassidy Barram, Sudipto Ghosh, Jay Breidt, Sangmi Lee Pallickara, Shrideep Pallickara, *Resource Efficient Profiling of Spatial Variability in Performance of Regression Models,* IEEE International Conference on Big Data, 2022 *[Under Review]*

[2] **Saptashwa Mitra**, Menuka Warushavithana, Mazdak Arabi, Jay Breidt, Sangmi Pallickara, Shrideep Pallickara, *Alleviating Resource Requirements for Spatial Deep Learning Workloads*, IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid) 2022. [Accepted]

[3] Menuka Warushavithana, **Saptashwa Mitra**, Mazdak Arabi, Jay Breidt, Sangmi Lee Pallickara, Shrideep Pallickara, *Containerization of Model Fitting Workloads over Spatial Datasets*, IEEE Big Spatial Data (BSD) 2021. [Link]

[4] **Saptashwa Mitra**, Daniel Rammer, Shrideep Pallickara, Sangmi Lee Pallickara, *Glance: A Generative Approach to Interactive Visualization of Voluminous Satellite Imagery*, IEEE International Conference on Big Data, 2021. [Link]

[5] M. Warushavithana, Caleb Carlson, **Saptashwa Mitra**, D. Rammer, M. Arabi, Jay Breidt, Sangmi Lee Pallickara, Shrideep Pallickara, *Distributed Orchestration of Regression Models Over Administrative Boundaries*, IEEE/ACM BDCAT, 2021. [Link]

[6] **Saptashwa Mitra**, Maxwell Roselius, Pedro Andrade-Sanchez, John K. McKay, Sangmi Lee Pallickara, *RADIX+: High-Throughput Georeferencing and Data Ingestion over Voluminous and Fast-Evolving Phenotyping Sensor Data*, Concurrency and Computation: Practice and Experience (Journal). [Accepted]

[7] **Saptashwa Mitra**, Daniel Rammer, Shrideep Pallickara, Sangmi Lee Pallickara, *A Generative Approach to Visualizing Satellite Data*, IEEE International Conference on Cluster Computing (CLUSTER), 2021. [Link]
[8] Menuka Warushavithana, **Saptashwa Mitra**, Mazdak Arabi, Jay Breidt, Sangmi Lee Pallickara, Shrideep Pallickara, *A Transfer Learning Scheme for Time Series Forecasting Using Facebook Prophet*, IEEE International Conference on Cluster Computing (CLUSTER), 2021. [Link]
[9] **Saptashwa Mitra**, Paahuni Khandelwal, Shrideep Pallickara, and Sangmi Lee Pallickara, *STASH: Fast Hierarchical Aggregation Queries for Effective Visual Spatiotemporal Explorations*, IEEE International Conference on Cluster Computing (CLUSTER), 2019. [Link] **Best Paper
[10] Bibek Shrestha, **Saptashwa Mitra**, and Sangmi Lee Pallickara, *STRETCH: In-memory Storage with Autoscaling for Cluster Computing*, IEEE International Conference on Cloud Computing (IEEE CLOUD), 2019. [Link]
[11] **Saptashwa Mitra**, and Sangmi Lee Pallickara, *CONFLUENCE: Adaptive Spatiotemporal Data Integration Using Distributed Query Relaxation Over Heterogeneous Observational Datasets*, IEEE/ACM Conference on Utility and Cloud Computing (UCC), 2018. [Link]
[12] **Saptashwa Mitra**, Yu Qiu, H. Moss, K. Li, and Sangmi Lee Pallickara, *Effective Integration of Geotagged, Ancillary Longitudinal Survey Datasets to Improve Adulthood Obesity Predictive Models*, IEEE Big Data Science and Engineering (IEEE BigDataSE), 2018. [Link]
[13] Kevin Bruhwiler, Philip Sharpp, Nick Czarnecki, Jim Xu, Fawad Ahmed, **Saptashwa Mitra**, Sangmi Lee Pallickara, *Immersive Analytics for Traffic Analysis using Machine Learning Techniques*, CURC, Colorado State University, April 2018.