Harsha Reddy

Big Data Lead

Big Data Lead with 8+ years of experience architecting petabyte-scale, **real-time data platforms and AI-native infrastructures** . Expert in building low-latency systems and Agentic AI workflows that drive autonomous decision-making in production environments . Proven track record of scaling platforms from scratch to support 10B+ monthly interactions, optimizing for high-throughput anomaly detection, and leading cross-functional teams within the open-source ecosystem (Kafka, Spark, Druid, Trino, Kubernetes..etc)

mailme.sha97@gmail.com

+918847839187

Bangalore, India

linkedin.com/in/i-am-harsha-reddy

SKILLS

Algorithms

Data Structures

Spark | Presto

Trino | Druid

Elastic Search

Qdrant | Airflow

Change Data Capture

Superset | Datahub

Debezium | Mongo

SQL Server | Grafana

Tableau | Kubernetes

AWS | Apache Atlas

Redshift | Azure

Prometheus | Psql

File Formats (Avro, Parquet, ORC, Delta)

Metabase

Agentic AI

Langchain / Langgraph

EDUCATION

B.Tech
IIT Bhubaneswar

07/2014 - 06/2018 *82%*

WORK EXPERIENCE

Big Data Lead
Zzazz

05/2024 - Present *Bangalore, India*

Architecting and leading a globally distributed real-time analytics platform *that captures 10B+ monthly user interactions, serving as the foundational layer for Agentic AI workflows to process 500k+ articles daily across 100+ countries.*

Projects

- **Led the engineering of an enterprise-grade web analytics platorm** utilizing a 100% open-source stack to disrupt market-leading proprietary solutions like Google Analytics .
- Differentiated the product by launching a **conversational AI interface** which empowered users to ask open-ended questions and receive real-time, multi-modal responses containing direct answers, data insights, and custom graphs.
- **Engineered an ultra-low-latency Feature Store** for a high-throughput ad exchange, meeting mission critical SLAs for real-time bidding (RTB) and WTP models .
- **Engineered a comprehensive self-serve data platform** to democratize analytics, enabling cross-functional teams to seamlessly author, deploy, and monitor custom data pipelines.
- **Spearheaded the C4X platform as Project Director** , leading the end-to-end design, architecture, and engineering of a centralized data hub that acts as the absolute source of truth for founders, investors, and all internal stakeholders.
- **Architected a proprietary wordpress automation engine** that generated and continuously published to a fleet of 20k+ sites, serving as a large-scale, real-world sandbox for rigorous A/B testing of ad revenue, user engagement, and dynamic pricing strategies.
- **Owned complete DevOps responsibilities** for the data engineering unit, independently managing the end-to-end setup, configuration, and performance tuning of the entire data infrastructure.

Tech Lead, Big Data
Gameskraft

02/2023 - 05/2024 *Bangalore, India*

Led the design and development of a ***high-performance Lakehouse platform*** *to power company-wide analytics, enabling scalable, self-serve access to structured and unstructured data for diverse use cases across product, growth, and business teams.*

Projects

- Built and scaled a centralized Lakehouse platform to support cross-functional analytics use cases company-wide, enabling structured access to raw and processed data.
- Built a real-time data quality and SLA monitoring module to ensure accuracy across data pipelines ingesting from multiple microservices.

LANGUAGES

Java | Python

Bash | SQL

WORK EXPERIENCE

Senior Software Development Engineer
OYO Rooms

09/2021 - 02/2023 *Bangalore, India*

Built and maintained end-to-end big data pipelines for both batch and real-time processing, while mentoring individual contributors and providing architectural guidance across multiple data-driven projects.

Projects

- Achieved a 90% reduction in platform costs by optimizing each tech stack layer during COVID-induced constraints, without compromising analytics capabilities.
- Developed a real-time and batch anomaly detection system for service-level metrics, improving monitoring coverage across OYO's data services.
- Integrated Apache Druid to replace latency-heavy Hive/Presto workflows, enabling sub-second query performance on billions of events.
- Deployed highly available Kafka clusters using Terraform to support both transactional and analytical workloads, implemented active-passive disaster recovery with MirrorMaker 2.0 for zero-downtime failover, and standardized cluster versions and naming conventions across environments.
- Successfully migrated the entire data platform stack from AWS to Azure with minimal downtime, ensuring continuity for analytics and batch jobs through robust cutover planning and validation.

Senior Data Engineer
Trip Advisor

03/2021 - 08/2021 *Gurugram, India*

Led the migration of data infrastructure from on-premise systems to the cloud, driving a strategic shift from enterprise solutions to open-source and in-house tools for improved scalability, cost-efficiency, and ownership.

Projects

- Migrated legacy SQL Server and Oozie workflows to Snowflake and a proprietary in-house tool (WhamPipe), modernizing the orchestration layer and reducing operational burden.

Data Engineer
Dailyhunt

01/2021 - 03/2021 *Bangalore, India*

Built a highly scalable big data architecture that parses, enriches, and aggregates 9–10 billion events and over 3 TB of data daily, ensuring low-latency processing and high data quality for downstream analytics.

Software Development Engineer - II
OYO Rooms

07/2018 - 12/2020 *Gurugram, India*

Developed end-to-end big data pipelines for both batch and real-time processing, while actively mentoring individual contributors and providing technical leadership across cross-functional data initiatives.

Projects

- Implemented an early form of a transactional data lake before Delta Lake existed, enabling consistent Spark aggregations by generating snapshot versions of dependent tables at runtime.
- Ensured data consistency and reduced sync-related failures by decoupling aggregations from live tables, laying the foundation for versioned and atomic data processing workflows.
- Assessed Snowplow's feasibility to replace Google Analytics for event tracking and enrichment, focusing on data ownership and pipeline customizability.
- Implemented automated alerting for data freshness, completeness, and schema violations to improve trust in analytics.
- Built a real-time data quality and SLA monitoring module to ensure accuracy across data pipelines ingesting from multiple microservices.
- Ingested data from third-party sources (GMB, Snapchat, MoEngage, Singular) into the data lake for unified analysis with transactional datasets.
- Created persistent views for real-time Presto queries to reduce query time and offload compute from live dashboards.