Sruthi Erra Hareram
Ph: +1 (403)399-6483
E-mail: sruthihari.sh16@gmail.com
LinkedIn: www.linkedin.com/in/sruthi-hareram

Professional Summary:

Data Engineer having 6+ years of experience with strong background in end-to-end enterprise data warehousing and big data projects. Proficient in big data tools like Hive and Spark and relational data warehouse tool Teradata etc. Excellent hands-on business requirement analysis, designing, developing, testing and maintaining the complete data management & processing systems, process documentation and ETL technical and design documents. Responsible for data engineering functions including, but not limited to data extract, transformation, loading, integration in support of enterprise data infrastructures - data warehouse, operational data stores and master data management. Expertise in resolving production issues, hands-on experience in handling all phases of the software development Life cycle. Adept at multitasking, working independently and as part of a team as required. Very flexible at adapting to changing client needs and deadlines. Possessing strong problem solving and communication skills.

A solid experience and understanding of designing and operationalization of large-scale data and analytics solutions on Snowflake Data Warehouse. Developing ETL pipelines in and out of data warehouses using a combination of Python and Snowsql. Experience in Google Cloud components, Google container builders and GCP client libraries and cloud SDK's. Substantial experience in Spark 3.0 integration with Kafka 2.4. Experience in setting up monitoring infrastructure for Hadoop clusters using Nagios and Ganglia. Sustaining the BigQuery, PySpark and Hive code by fixing the bugs and providing the enhancements required by the Business User. Working with AWS/GCP cloud using GCP Cloud storage, Data-Proc, Data Flow, Big- Query, EMR, S3, Glacier and EC2 Instance with EMR cluster. Proficient in Statistical Methodologies including Hypothetical Testing, ANOVA, Time Series, Principal Component Analysis, Factor Analysis, Cluster Analysis, Discriminant Analysis. Expertise in transforming business resources and requirements into manageable data formats and analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data. Worked with various text analytics libraries like Word2Vec, Glove, LDA and experienced with Hyper Parameter Tuning techniques like Grid Search, Random Search, model performance tuning using Ensembles and Deep Learning. Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS specific features. Knowledge of working with Proof of Concepts (PoC's) and gap analysis and gathered necessary data for analysis from different sources, prepared data for data exploration using data munging and Teradata. Experience in developing customized UDF's in Python to extend Hive and Pig Latin functionality.

Work Experience:

J&M Group Inc., Calgary, Canada — Oct 2024 - Present

Sr. ITA/AST

Nova Chemicals

- Built highly performant scalable services by designing microservices architecture that handled 1M+ daily API requests with 99.9% uptime.
- Applied outstanding technical problem-solving and debugging skills to resolve bottlenecks, improving system efficiency by 25%.
- Showcased a passion for quality by implementing automated unit and integration tests, achieving 90% test coverage.
- Optimized data ingestion pipelines by using Sqoop to migrate terabytes of data between RDBMS and HDFS, reducing data transfer time by 30%.
- Streamlined data exports to Teradata via Sqoop, improving query performance by 25%.
- Developed a route optimization system for a logistics project, designing a graph-based model that reduced delivery time by 20%.
- Enhanced SQL query performance by refactoring scripts with PySpark SQL, achieving a 40% faster execution time.
- Implemented Kerberos authentication, improving cluster security and access control across 500+ nodes.

- Resolved production issues in a 10+ node cluster, reducing downtime by 35% through proactive monitoring and troubleshooting.
- Designed and deployed an Apache NiFi data flow, facilitating seamless data movement between middleware and the EBI team, reducing manual intervention by 50%.
- Automated infrastructure deployment with Ansible and Chef, ensuring 99.9% system availability across multiple environments.
- Built real-time text analytics pipelines using Apache Spark (Scala), reducing data processing time from 5 hours to 1 hour.
- Transformed raw data from Data Lake into structured datasets in Databricks, leveraging Azure Databricks and ADF, leading to a 30% improvement in ETL efficiency.
- Provisioned and maintained Kubernetes clusters on AWS using Docker, Ansible, and Terraform, increasing deployment efficiency by 40%.
- Automated Datadog Dashboards using Terraform scripts, enabling real-time monitoring and alerting for critical applications.
- Developed and fine-tuned Spark applications, optimizing parallelism and memory usage, leading to a 25% improvement in processing time.
- Exported and analyzed data from Hadoop to relational databases using Sqoop, providing valuable insights for business intelligence and reporting teams.
- Leveraged Spark's in-memory computing with Scala, performing text analytics and complex data processing, reducing execution time by 45%.
- Implemented Google BigQuery jobs, automating data ingestion from Google Cloud Storage (GCS) every 15 minutes, ensuring real-time analytics availability.
- Converted SAS code to Python/Spark-based jobs on Google Cloud Dataproc and BigQuery, improving performance and reducing cloud costs by 20%.
- Built data pipelines using Apache Airflow in GCP Composer, automating ETL workflows with branching logic and custom operators, improving data pipeline efficiency by 35%.
- Developed store-level metrics and data pipelines using Google Cloud's big data stack, enhancing business decision-making with real-time analytics.
- Implemented infrastructure-as-code (IaC) solutions in Google Cloud Platform (GCP) using Terraform, ensuring scalable and maintainable cloud environments.
- Developed robust data validation frameworks using Spark, reducing data discrepancies by 95%.
- Created Ansible playbooks to automate infrastructure setup, cutting deployment time by 40%.
- Implemented advanced data modeling techniques for CosmoDB, improving query response times by 50%.

Pyramid Consulting, Calgary, Canada — May 2025 - Present

Data Engineer

Telus International Inc.

- Designed and deployed real-time telemetry ingestion pipelines using Azure IoT Hub, Event Hubs, and Databricks for 500+ industrial assets.
- Built PySpark Structured Streaming pipelines with Delta Lake, achieving sub-second latency and high-throughput analytics.
- Integrated Azure Digital Twins to create contextual models of physical assets for accurate fault detection.
- Developed ML-based predictive maintenance models using MLflow, improving failure prediction accuracy by 35%.
- Orchestrated large-scale data workflows across Azure Data Factory, Blob Storage, and Synapse Analytics.
- Automated ML model retraining pipelines with feedback loops (human-in-the-loop) to enhance long-term model performance.
- Reduced unplanned equipment downtime by 25%, improving operational continuity and plant safety.
- Increased asset utilization by 18% through data-driven diagnostics and performance monitoring.
- Implemented observability dashboards using Grafana, Azure Monitor, and Log Analytics for real-time diagnostics.
- Improved data processing throughput by 40% via optimized PySpark and Delta Lake transformations.
- Reduced manual root cause analysis by 60% through automation and anomaly detection pipelines.
- Achieved 99.7% ingestion reliability across IoT edge devices streaming millions of data points per hour.
- Applied Terraform and Kubernetes to automate infrastructure provisioning and scale analytics workloads.

- Enabled full data lineage and monitoring with Great Expectations to ensure data quality and integrity.
- Managed CI/CD pipelines with Azure DevOps for seamless deployment of streaming and batch workloads.
- Built time-windowed aggregations and stateful alerts to detect performance degradations in real time.
- Supported Nova Chemicals' Reliability-Centered Maintenance (RCM) strategy via advanced analytics.
- Partnered with plant engineers to validate anomaly alerts and drive continuous model refinement cycles.

Systech Solutions, Glendale, CA, USA Apr 2021 - March 2024

Data Engineer

FoxData Products

- Architected and optimized Hadoop clusters, translating functional and technical requirements into scalable and high-performance solutions, improving data processing efficiency by 35%.
- Analyzed and compared results between traditional systems and the Hadoop environment, identifying discrepancies and reducing data inconsistencies by 40%.
- Developed a high-performance data processing engine using Hortonworks distribution, enhancing ETL speed by 50%.
- Designed, developed, and tested ETL processes in AWS Glue, migrating campaign data from S3, ORC, Parquet, and text files into AWS Redshift, reducing ETL runtime by 30%.
- Deployed and scaled HBase clusters in AWS, ensuring seamless scalability based on business growth and reducing storage costs by 25%.
- Implemented AWS IAM policies, improving access control and security across 100+ applications.
- Automated deployments with AWS Lambda and API Gateway, reducing manual effort by 60%.
- Developed and optimized ETL pipelines using Hive, improving data transformation speed by 45%.
- Built scalable Spark applications in PySpark, processing millions of records from CSV files into Hive ORC tables, reducing data ingestion time by 40%.
- Streamlined PDF document ingestion from Microsoft SharePoint to HDFS using Apache NiFi, enhancing metadata extraction and indexing efficiency by 30%.
- Led the implementation of an end-to-end Azure data solution, integrating Azure Databricks, ADF, Data Warehouse, and Power BI, accelerating report generation by 50%.
- Applied Spark MLlib for advanced analytics, optimizing classification, regression, and clustering tasks, improving model accuracy by 20%.
- Commissioned and decommissioned Hadoop nodes, scaling infrastructure based on demand, reducing downtime by 35%.
- Developed automation scripts for ETL validation across Oracle, SQL Server, Hive, and MongoDB, reducing manual effort by 70%.
- Optimized SQL queries across MySQL, PostgreSQL, Redshift, SQL Server, and Oracle, enhancing query performance by 30%.
- Collaborated with cross-functional teams to migrate complex customer data between clusters, ensuring minimal downtime and zero data loss.
- Benchmarked Spark vs. Hive and SQL performance, leveraging Spark SQL and Scala for 20% faster data frame manipulations.
- Designed and implemented historical and incremental data loads in Databricks, optimizing batch processing by 35%.
- Built robust data ingestion pipelines using Flume and NiFi, reducing latency in log file processing by 50%.
- Developed scalable NoSQL data solutions using HBase and Cassandra, improving query efficiency by 25%.
- Orchestrated data pipelines from Data Lake to Databricks and from Databricks to Azure SQL DB, ensuring real-time data availability.
- Automated data loads using shell scripts, improving HDFS ingestion efficiency by 40%.
- Implemented Airflow and Oozie for workflow orchestration, reducing ETL failures by 30%.
- Worked with Cloudera and Hortonworks distributions, optimizing configurations for better cluster performance.
- Reduced migration time by 30% by optimizing T-SQL queries and data transformation processes, ensuring seamless data integrity.
- Developed high-performance SSIS/DTS packages, reducing ETL execution time by 25%.
- Designed and implemented microservices architecture, improving scalability by 40%.
- Resolved critical production bottlenecks, reducing system latency by 25%.
- Achieved 90%+ test coverage by implementing automated unit and integration tests, improving code quality and

reliability.

- Effectively collaborated with cross-functional teams, aligning on goals and ensuring on-time delivery of high-quality solutions.
- Built and optimized data pipelines that extract, classify, merge, and deliver actionable insights, improving data processing efficiency by 40%.
- Automated and scheduled workflows using Python and Shell scripting on Azure, reducing manual effort by 60%.
- Designed and implemented modern data solutions using Azure PaaS services, enhancing data visualization and reporting accuracy by 35%.
- Developed ETL pipelines for structured and unstructured data ingestion into Azure Data Lake, Azure Storage, Azure SQL, and Azure DW, reducing data latency by 30%.
- Optimized performance troubleshooting and tuning for Azure services like HDInsight clusters, ADF, Databricks, and networking, reducing job execution time by 25%.
- Designed and executed Ad Hoc data pulls using Azure Data Factory, automating data migration from on-prem to SQL DW, improving query performance by 40%.
- Developed Spark jobs using PySpark and Spark-SQL, efficiently processing large-scale data and reducing transformation time by 35%.
- Installed and configured Hive, Sqoop, Flume, and Oozie, ensuring seamless integration and workflow automation.
- Managed and optimized 70-node Hadoop clusters, reducing downtime and enhancing resource allocation efficiency by 30%.
- Proven expertise in Microsoft Azure ecosystem, leveraging Terraform, Databricks, Azure Data Factory, and Azure Function Apps to design and deploy cloud-native data solutions.
- Developed and automated real-time data ingestion pipelines using Google Cloud Pub/Sub and BigQuery, reducing data latency from 30 minutes to 5 minutes, enabling near-instant analytics for decision-making.
- Led the design and implementation of ETL workflows on Google Cloud Dataflow, processing over 10TB of data daily and improving data processing efficiency by 25%.
- Optimized BigQuery performance, reducing query costs by 40% by implementing partitioned tables and materialized views, which resulted in faster data retrieval and reduced operational costs.
- Automated 15-minute interval data ingestion jobs from Google Cloud Storage to BigQuery using Cloud Functions, improving data freshness for real-time reporting and analysis.
- Collaborated with a cross-functional team of 8 data engineers and analysts to develop a unified data warehouse on BigQuery, supporting real-time business intelligence tools like Google Data Studio and Looker.
- Implemented robust security measures, ensuring compliance with industry regulations by integrating IAM roles and DLP policies in GCP, reducing security risks by 30%.
- Led the integration of third-party data sources into BigQuery using Cloud Functions, enabling seamless analytics from multiple data streams and reducing manual data extraction time by 50%.
- Managed and optimized data pipelines on Google Cloud Composer (Apache Airflow), ensuring 99.9% uptime for critical data flows and improving overall pipeline reliability by 20%.

Cognizant Technology Solutions (CTS), Chennai, India — Jan 2017 - Jul 2019

Program Analyst

DNB Compliance Check

- Led the design and implementation of a complex billing module as part of an agile team, delivering the module 20% ahead of schedule and improving system efficiency by 30%.
- Developed and optimized data pipelines using Apache Spark to ingest, transform, and load large datasets into Hive and HBase, reducing data processing time by 40%.
- Built and deployed map-reduce programs in Java, leveraging distributed cache for map-side joins, improving data join performance by 25% in large-scale datasets.
- Enhanced Hive query performance by implementing partitioning, dynamic partitions, and bucketing, reducing query execution times by up to 50% for critical reports.
- Designed and developed real-time data processing solutions using Spark Structured Streaming and Kafka, reducing data processing latency from 5 minutes to near real-time (sub-second).
- Ensured data quality and access security by validating transactional and profile data from RDBMS and implementing robust access control measures, improving data integrity by 30%.
- Worked on end-to-end data integration from source systems to Data Lake using Hadoop, streamlining data flows and cutting data load times by 20%.
- Led the selection of Hadoop components such as Hive, Pig, and Sqoop, ensuring high-performance data processing while improving scalability by 25%.
- Managed scalable data ingestion solutions using Sqoop and Flume, processing over 50TB of data monthly, and

optimizing throughput by 30%.

- Streamlined web log data ingestion using Flume, capturing and aggregating real-time data from web servers, mobile apps, and IoT devices, improving log processing efficiency by 40%.
- Collaborated on designing and implementing Hive and Pig use cases, improving query performance for complex data sets, cutting processing time by 35%.
- Automated job scheduling and orchestration using Airflow, reducing job failure rates by 20% and ensuring timely data availability for downstream systems.
- Built a CI/CD pipeline using Jenkins and Ansible, reducing deployment times by 60% and enhancing deployment frequency, contributing to faster releases.
- Developed custom UDFs in Python for Hive and Pig, extending functionality and enabling efficient data transformations, reducing ETL job execution time by 15%.
- Enhanced data loading performance with Teradata utilities (Fast Export, Multi Load), improving ETL throughput by 50% for large-scale datasets.
- Designed and implemented OLAP models using Kimball methodology, creating a robust data model for analytics that reduced data query times by 40%.
- Improved application uptime and fault tolerance by leveraging Kubernetes on Google Cloud, ensuring 99.9% availability for mission-critical applications.
- Automated data movement to cloud storage using Sqoop, Oozie scripts, and GCP services, streamlining data migration to Google Cloud Storage and reducing manual intervention by 50%.
- Configured and deployed big data solutions on GCP (Cloud Dataproc, BigQuery, Cloud Storage), enabling scalable and efficient cloud data processing for over 100TB of data monthly.
- Built and deployed RESTful APIs for seamless integration with real-time data pipelines, enabling data exchange with third-party platforms and improving system interoperability by 30%.
- Integrated external data sources via APIs into real-time data pipelines, enabling IoT devices and SaaS platforms to stream data directly into data lakes for analytics, reducing data ingestion time by 20%.
- Developed microservices using Spring Boot, enabling scalable, high-performance systems that reduced processing time for key workflows by 25%.
- Integrated REST APIs with Spark and Kafka, enabling micro-batch and stream processing, which resulted in improved real-time analytics capabilities and reduced latency by 40%.
- Optimized data storage for analytics using Delta Lake, MS SQL Server, and DB2, improving retrieval times by 50% for large-scale datasets.

EDUCATION:

ME in Electrical and Computer Engineering
University of Calgary, Calgary, AB, Canada — Aug 2019 – Apr 2021

BE in Electrical and Electronics Engineering
Sathyabama University, Chennai, TN, India — Aug 2012 – Apr 2016

Certifications:

Databricks Certified Data Engineer Associate.
Issued by: Databricks — Validity: March 16, 2026