Sudhir singh

New Delhi, India · de.crepantherx@gmail.com · +91 8587001379
LinkedIn · Portfolio Website · GitHub

Work Experience

Tiger Analytics — Remote
Data Engineer — Jan 2023 - Present

- **Built a single enterprise-wide unified data repository** by unifying **10 Petabytes** of data requirements from numerous products, re-developing to standardize the development process using an on-prem Python-based framework, by individually contributing and leading a team of 7 Data Engineers **as a Data Engineering Lead**.
- **Reduced Databricks compute resources cost by 50%**, by optimizing a Spark job to process 7.5 TB of data in a single run, through detailed analysis of data transformations, Spark submit configurations, data spills, partition sizes, and worker utilization.
- **Achieved 40% reduction in job computes cost of 35 daily Spark jobs** responsible for ingesting incremental loads by reviewing SLA & re-developing to remove skewness in data by rewriting the partitioner & optimizing Spark configurations, through detailed understanding of data.
- **Reduced re-development needs by 90% and minimized operational costs**, by developing a metadata-driven, object-oriented PySpark API-based Data Ingestion framework and pairing it with Azure DevOps CI/CD pipelines to automate and keep metadata always in control of client.

Tech Mahindra — Sydney, Australia
Data Engineer — Jan 2021 - Jan 2023

- **Achieved a 95% improvement in estimating Estimated Time of Arrival (ETA)** by delivering a decision support system to technicians that enabled relocation based on forecasted ticket numbers, through the deployment, and prediction of a Machine Learning model on Kubeflow.
- Ingested billions of customer transactions by leveraging **Cloudera Hortonworks Spark Architecture** to process and manage large-scale data efficiently.
- **Led the establishment of 3+ ML CI/CD pipelines on AWS using Kubernetes**, Jenkins, and GitHub, resulting in improved deployment processes and increased productivity for the team.
- **Conducted** comprehensive training sessions for 60 employees on SQL Analytics, data modeling techniques, and building scalable data pipelines using PySpark and Apache Airflow.

Skills

- **Transactional DBMS:** Microsoft SQL Server, MySQL, PostgreSQL
- **Distributed Storage System:** ADLS, Amazon S3
- **Big Data Processing & Streaming:** Apache Spark, Apache Kafka
- **Data Orchestration:** Apache Airflow, Azure Data Factory
- **Data Processing:** Python, PySpark API, SparkSQL API, Pandas, SpaCy, NLP
- **Data Warehousing:** Snowflake, BigQuery, Redshift
- **Machine Learning:** Scikit-learn, TensorFlow, Keras, Seaborn, Matplotlib, NumPy, SciPy
- **DevOps:** Azure DevOps, Splunk, Kubernetes, Kubeflow, Jenkins, JFrog, Docker, GIT

Education

- **Liverpool John Moores University** — Liverpool, UK
  Master of Science in Machine Learning & AI — Graduation Date: Sep 2025
- **Gautam Buddha University** — Greater Noida, India
  Bachelor of Technology in Information Technology — Graduation Date: Jul 2021

Awards

- **Tech Mahindra** / Bravo Award - Project Delivery — Jun 2021 & Mar 2022
- **Tech Mahindra** / Pat On The Back Award - Trained 50+ employees, in Apache Spark, SQL — Dec 2021

Certifications

- **Microsoft** / Azure Data Scientist Associate, DP-100 Exam — Jun 2022
- **Microsoft** / Azure Data Engineer Associate, DP-203 Exam — Jun 2022

Project Experience

Tiger Analytics, Remote

- **Common Data Foundation** Jan 2024 - Present
  Tech Stack: Azure Databricks, Python IDE-based Development, API, Large Data Volume
  - **Developed & migrated unified foundation layer** coming from various retail source systems
  - **Lead for delivery** of data sources migration from Databricks to on-prem frameworks.
  - **Developed** Python frameworks for **data quality checks, PII encryption, and data transformation** from raw to trusted and feature layers.
- **Launchpad** Feb 2023 - Dec 2023
  Tech Stack: Azure DataFactory, Azure DevOps CI/CD, Databricks, Logic Apps, Log Analytics
  - **Lead the end-to-end design and development** of an object-oriented PySpark-based Data Ingestion framework, increasing data processing speed by 30%.
  - Built Azure DevOps CI/CD pipeline, automating 95% of deployments and accelerating timelines.

Tech Mahindra, Sydney, Australia

- **DQR-DAC Ingestion System** June 2022 - Jan 2023
  Tech Stack: Hadoop, Scala, Hive, Delta Tables, Linux, Zeppelin, Ambari, Jira, Confluence
  - **Implemented** a workflow to determine customers eligibility for migration to new strategic plans.
  - **Developed & maintained** data ingestion layer, audit & configuration layer, Hadoop layer for seamless operational needs as per GDPR compliance.
  - Transformed 25+ data sources using Scala as per business & operational requirements.
  - Provided operational support in detecting any discrepancy in 25+ data sources.
- **Pharma Insights Data Extraction & Analytics** Apr 2022 - May 2022
  Tech Stack: SQL, Python, PySpark, Azure Data Factory, Data Bricks, Data Lake, Key Vault
  - **Created ADF pipeline to fetch data from S3 to ADLS**, using ELTL approach.
  - Written Pyspark script in Azure Databricks to extract text from images.
  - Categorized data retrieved into JSON objects, loaded into ADLS.
  - Provided analytics insight report using Azure Databricks & presented it to clients.
- **Decision Support For Technician Allocation** Aug 2021 - Mar 2021
  Tech Stack: Python, SQL, ML, CI/CD, AWS, Kubernetes, JFrog, Jenkins, Kubeflow
  - **Improved end-to-end service processes & cut down manual deployment** by 75% by predicting proper utilization of technicians and minimizing customer wait time for the best customer experience by generating 3000 recommendations per day, in near real-time, to mitigate the risk of the day.
  - **Deployed** 5+ container-based applications orchestrated with Kubernetes.
  - Established 2+ CI/CD pipelines of ML models on AWS.
  - **Written 40+ highly optimized SQL stored procedures** to transform/aggregate raw data, API integrations to fetch live data.
- **Control Tower Intrusion Detection Analytics** Jan 2021 - July 2021
  Tech Stack: Azure Databricks, Azure Datafactory, CI/CD, Spark SQL, PySpark, Data Lake, Delta Tables
  - **Worked closely with stakeholders, BA** to gather the business requirements.
  - Extracted complex Fields from different types of Log files using **Regular Expressions** in Databricks.
  - **Exploratory data analysis to analyze security anomalies** from server's generated data.
  - Wrote and optimized simple and complex SQL queries to load information extracted into Splunk datasets as per business needs.