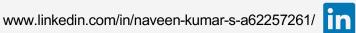
# NAVEENKUMAR S



\_\_\_\_\_ naveenkumarsai10@gmail.com

Data Engineer





# SUMMARY

I'm a Data Engineer with 3 years of experience designing, developing, and maintaining data pipelines, specializing in big data technologies like PySpark, SparkSQL, Python, SQL, and Databricks. I have developed pipelines for ingesting data from API sources and storage services (ADLS) and automated 90% of ETL workflows using tools like Azure Data Factory. Additionally, I have experience in Agile project management, including sprint planning, daily stand-ups, and review meetings, ensuring efficient delivery and continuous improvement

### SKILLS

- Data Modelling
- Data Cleaning
- · Stream Processing
- · Batch Processing
- · Data Warehousing
- ETL/ELT
- · Pipeline Monitoring

### **TECH STACK**

#### **Microsoft Azure**

- Azure Data Factory (ADF)
- Azure Databricks
- · Azure SQL Database
- Azure Data Lake Storage (ADLS)
- · Azure Logic Apps
- · Azure Function Apps

## **Amazon Web Services (AWS)**

- Amazon EMR (Elastic MapReduce)
- Amazon S3

## **Big Data Technologies**

- PySpark
- Hive

### **Programming Languages**

- Python
- SQL

#### CI/CD

· Azure DevOps

## **CERTIFICATION**

#### **Microsoft Certified**

• DP - 900

## **Hackerrank Certified**

• SQL

## **Udemy Certified**

Airflow

# PROFESSIONAL EXPERIENCE

MathCo - Bangalore (2022 - 2025 April)

#### **Azure Data Engineer - Product Development**

- Project Overview: Collaborated with the development team to collect employee data from APIs
  and built a project using ADF and Azure Databricks to onboard the data to Data Warehouse.
- Data Ingestion: Build ADF pipelines to transfer data from APIs to ADLS.
- Architecture Design: Implemented Medallion Architecture to onboard the data.
- **Data Quality Management:** Implemented schema checks and data validation processes in the refined layer, resulting in a 40% improvement in data accuracy.
- **Data Transformation:** Transformed the data in the consumption layer according to the specific requirements provided by the development team.
- Orchestration and Monitoring: Apache Airflow was used to orchestrate and monitor the pipeline.
- CI/CD Integration: Implemented CI/CD pipelines using Azure DevOps for automated build, test, and deployment processes.

## Azure Data Engineer - Retail Industry - Marketing Analytics

- **Project Overview:** Collaborated with the insights team to transfer the data from source database to target database.
- Data Modelling: Collected essential data from the source database based on the specified requirements and constructed a new database utilizing a star schema.
- Data Ingestion: Used ADF to transfer data from source database to target database.
- **Marketing Mix Modeling (MMX)**: Developed MMX models using multiple linear regression and log-log regression to assess marketing impact on revenue.
- Campaign Performance Analysis: Performed ROI calculations and incremental sales analysis to measure campaign effectiveness.

### Python / Pyspark developer - Pharma Industry

- **Project Overview**: Collaborated with the business team to onboard data from a customized FTP location into Data warehouse.
- Architecture Design: Implemented Medallion Architecture to onboard the data into different layers(raw,refined,consumption).
- Framework Developement: Built a customized data onboarding framework that automated 60% of coding tasks across Raw and Refined layers.
- Business Logic Development: Created optimized code to apply the business team's transformation logic and load to the data.
- CI/CD Integration: Utilized CI/CD processes to deploy code from Azure DevOps to S3.

# Code optimization:

- Reduced time complexity by 30% using cache and persist methods to optimize DataFrame operations
- Applied partitions and bucketing techniques for large dataset, cutting execution time by 40%
- Tuned PySpark configurations to enhance job performance.
- Implemented checkpoints to ensure fault tolerance and recover intermediate states during long-running jobs.

# PROFESSIONAL EXPERIENCE

Loyalytics Consulting (March 2025 - present):

Senior Associate Data Engineer - Azure Databricks

Project: Retail Client Performance Management Data Warehouse.

- Built a scalable data warehouse to onboard and process real-time streaming data from Kafka on retail client transactions and performance metrics.
- Designed and developed streaming ETL pipelines using Delta live tables to transform and load data into the warehouse for analytics, reporting, and dashboarding.
- Delivered a customized product enabling client performance tracking, acknowledgment workflows, and actionable insights for business management.
- · Ensured high data accuracy and completeness to support near real-time decision-making and reporting.

## **PROJECTS**

#### Formula 1 Data Warehouse & Analytics

Tech Stack: Azure Databricks, Spark SQL, Delta Lake, Databricks Workflows, Unity Catalog

- Designed and implemented a Formula 1 Data warehouse using Medallion Architecture (Raw, Refined, and Consumption layers) to onboard and process historical F1 race data.
- Developed schema validation checks to ensure data consistency and quality before ingestion into Delta Lake storage.
- · Utilized Databricks Workflows for automated data ingestion, transformation, and aggregation.
- · Structured the data warehouse using a Star Schema, optimizing it for high-performance analytical queries.
- Performed analytical insights on race results, driver statistics, and team performances using Spark SQL.
- Tracked end-to-end data lineage, providing complete visibility into data transformations.

#### **Covid 19 Reporting & Analytics**

Tech Stack: ADF, Azure Databricks, ADLS Gen2, Delta Tables, Azure SQL DB, CI/CD, Azure DevOps

- Used Azure Data Factory (ADF) to ingest COVID-19 data from various sources into Azure Data Lake Storage Gen2 (ADLS Gen2).
- Designed and implemented a Medallion Architecture using Azure Databricks to process and transform the data.
- · Stored raw data in the bronze layer.
- · Cleaned and refined data in the silver layer.
- · Optimized and structured data in the gold layer using Delta Tables for better performance.
- Loaded the processed data into Azure SQL Database for further analysis and reporting.
- · Used SQL queries to generate analytical insights, such as daily infection rates, trend analysis, and data patterns.
- Fully orchestrated the end-to-end pipeline using CI/CD in Azure DevOps, ensuring automation, monitoring, and seamless deployment.