Data Engineer
Name: Karthik Sai
Phone: 203-804-8091
Email: karthiksai197@gmail.com

PROFESSIONAL SUMMARY:

- **4** years of experience as a Data Engineer with hands-on experience on Big Data technologies in **Scala, Spark, Hadoop, Pig, Kafka, MapReduce, HBase, Zookeeper, Sqoop, Apache NiFi, Oozie, Impala, Flume, Yarn,** and **HDFS.**
- Extensive experience with (AWS) concepts, including **S3, Redshift, Glue, EMR, Lambda** CloudWatch, ELB, Autoscaling, Elastic Cache, DynamoDB, and Athena, Kinesis.
- Expertise in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Azure Data Factory, Azure Analysis Services**, Azure Synapse, Data Lake, Data Factory, Databricks, Cosmos DB.**
- Extensive experience working with GCP Cloud services like **GCP cloud dataflow, GCP Big Query, GCP Cloud Storage, GCP Composer.**
- Hands on experience with different languages such as **Java, Python, Scala, SQL, R, SAS, Linux, and UNIX Shell.**
- Proficient in ETL/ELT processes, utilizing tools like **Apache Airflow, dbt (data build tool), Talend, Informatica** for ingesting and processing data from disparate sources**.**
- Good experience in designing and developing logical and physical data models that utilize concepts like **Star Schema, Snowflake Schema,** and Slowly Changing Dimensions.
- Experience in **fact/dimension data warehouse design models**, including star and snowflake schemas.
- Extensive experience working with distributed computing systems such as **Hadoop, HDFS, Spark Core, Hive, and Kafka** for large-scale data processing and storage.
- Experienced in data governance and compliance, ensuring data quality and adherence to regulatory requirements like **GDPR, HIPAA, and CCPA.**
- Adept at implementing CI/CD pipelines using **GitHub Actions, Docker, Kubernetes, Jenkins and Terraform**, ensuring reliable and automated data workflows.
- Proficient in project management and documentation using **JIRA**, Prometheus, **Grafana**, and **ELK Stack** and **JS docs**.
- Expertise in working with both **SQL and NoSQL databases like MongoDB, Cassandra, HBase, and SQL Server.**

TECHNICAL SKILLS:

Languages	Python, Scala, Java, C, C++, Shell Script, Perl Script, SQL
Big Data Technologies	Hadoop, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Oozie, Flume, NiFi, Yarn, Airflow, Apache Spark, DBT, Databricks
Cloud Technologies	Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), IAM
Python Libraries	Pandas, NumPy, SciPy, Matplotlib, Beautiful Soup, Scikit-Learn
Data Modelling Tools	Snowflake, Star Schemas
Versioning Tools	SVN, Git, GitHub
Project Management Tools	Jira
ETL Tools	Apache NiFi, Informatica and SSIS.
Reporting Tools	Tableau, Excel, SSRS, Power BI
Databases	Oracle, MySQL, DB2, SQL Server, PostgreSQL, MongoDB, Cassandra, DynamoDB

CERTIFICATIONS

- AWS Certified Data Engineer – Associate
- Databricks Certified Apache Spark Developer

PROFESSIONAL EXPERIENCE:

Client: CVS, NYC **Nov 2023-Till Date**
Role: Data Engineer
Responsibilities:

- Architected and deployed cloud-based data pipelines using Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, and NoSQL DB to handle 2TB+ of daily data, while integrating IAM security best practices.

- Optimized big data processing by extensively working with **HDFS, MapReduce, and Spark**, ensuring high scalability and fault tolerance in distributed computing environment.
- Developed scalable AWS data solutions by integrating **AWS EMR, EC2, Redshift, S3, Glue, DynamoDB, and Kinesis with** Big Data frameworks like Zookeeper, Yarn, Spark, Scala, and NiFi.
- Used Data Frame API and Scala API, with Python and Spark SQL, to ingest data to SQL ad NoSQL databases in Azure like **Azure SQL DB, PostgreSQL, MySQL, Cassandra, Cosmos DB.**
- Experienced in **Dimensional** Data Modelling using tools like ER/Studio, Erwin, Sybase Power Designer, and proficiency in Star Join Schema/Snowflake modelling.
- Developed **Airflow DAGs** to automate workflow scheduling, integrating SLA monitoring, dependency tracking, and error handling for enterprise-wide job orchestration.
- Implemented CI/CD pipelines with **Git, Jenkins, and Terraform**, improving code deployment automation and release management.
- Developed microservice on boarding tools leveraging Python and Jenkins allowing for easy creation and maintenance of build jobs and **Kubernetes** deploy and services.
- Proficient in ETL processes, utilizing tools like **Apache Airflow, dbt, Talend, Informatica** for ingesting and processing data from disparate sources.
- Advanced knowledge of programming languages such **as Python, SQL, Java, and Scala is** crucial for developing robust data solutions.

Environment: Azure (Data Factory, Data Lake, Synapse, SQL, Storage), Snowflake, AWS (EC2, S3, Glue, Athena, Redshift, EMR, DynamoDB, Lambda, Route53, CloudFormation), Big Data (Hadoop, Hive, Spark, Kafka), Jenkins, Git, Airflow, DBT, Terraform.

Client: Dimension Data India Pvt. Ltd, India **Jan 2020-July 2022**
Role: Data Analyst/Engineer
Responsibilities:

- Proficient in **Python**, **SQL**, and **Shell scripting**.
- Created various Spark applications using **Py-Spark** to perform a series of enrichments of clickstream data combined with enterprise user data.
- Developed data transition programs from **DynamoDB** to **Snowflake (ETL Process)** using AWS Lambda by creating functions in Python for certain events based on use cases.
- Involved in Developing a Restful service using **Python Flask** framework.
- Interact with AWS services programmatically using **Boto3's Python SDK**.
- Designed **ETL** Process using **Informatica** to load data from Flat Files, and Excel Files to target Oracle Data Warehouse database.
- Strong hands-on experience with modern **streaming data architectures** using **Apache Kafka**, **AWS Kinesis**, **Azure EventHub**, **Google Pub/Sub**, and **Apache Flink** for real-time analytics.
- Implemented data cataloguing and lineage with **Apache Atlas** and **AWS Glue Data CatLog** to support GDPR and audit requirements.
- Developed Shell scripts to transfer data from Hadoop to Google Cloud Storage (**GCS**) and from GCS to Big Query.
- Worked on both supervised and un-supervised Machine Learning models, Deep Learning etc.
- Extract data from various sources, including databases, files, and APIs, using Qlik's data connectors or scripting capabilities
- Utilized **Databricks** on **GCP** for advanced machine learning model development, automating model training and deployment processes using **MLflow**.
- Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including **Pig, Hive, SQOOP.**
- Experienced working with **Agile** Methodologies and **SCRUM** Process.

Environment: Python, Django, MySQL, AWS, Linux, Informatica Power Centre, HTML, XHTML, CSS, AJAX, JavaScript, ETL, Oracle, NumPy, Pandas, Unix, SDLC, Jira.

Education:

- Master of Science in Data Science, University of New Haven, May 2024 | CGPA 3.55
- Bachelor of Engineering, Sathyabama University, 2021 | CGPA 8.54