SURAVI KC

suravik257@gmail.com +1 347-241-6108 https://www.linkedin.com/in/suravi-kc/ New York, NY

PROFESSIONAL SUMMARY

Seasoned Data Engineer with a proven track record in designing and optimizing data solutions across diverse industries. Skilled in orchestrating Apache Spark integration, mastering AWS and Azure services, and leading Kafka streaming pipelines. Demonstrated expertise in Python, SQL, Hadoop ecosystem, and real-time data processing. Adept at enhancing data warehouses, collaborating cross-functionally, and applying AI/ML technologies to drive data insights and decision-making. Seeking a challenging role as a Data Engineer to leverage expertise in big data technologies, cloud services, and data analytics. Aim to contribute to innovative data solutions, drive efficiency in data processing, and foster a data-driven culture in a dynamic organization.

PROFESSIONAL EXPERIENCE

Data Engineer - Capital One, New York NY **Sep 2022 - Present**

- Orchestrated the integration of Apache Spark with diverse data sources, optimizing HDFS, Hive, and Spark SQL for efficient data processing and analytics.
- Demonstrated mastery in AWS and Azure ecosystems, crafting scalable cloud solutions with services including EC2, S3, RDS, Lambda, Glue, and Redshift and achieved a 40% improvement in data transfer speeds by implementing data compression techniques and optimizing S3 bucket configurations.
- Developed and optimized Kafka streaming pipelines, focusing on topic and partition distribution, and implementing stringent security protocols.
- Initiated and developed real-time data processing solutions using Apache Kafka, Spark Streaming, and PySpark, enabling instant data insights and improving customer experience through real-time analytics.
- Mastered the Hadoop ecosystem for robust data storage and processing, optimizing HDFS interactions to enhance data retrieval and analysis efficiency.
- Collaborated with teams to troubleshoot and fine-tune Hadoop cluster performance, ensuring high availability and seamless data operations.
- Designed and managed ETL workflows using Informatica PowerCenter for seamless data integration.
- Enhanced Snowflake and Redshift data warehouse performance and reliability by implementing index optimization strategies, resulting in a 40% improvement in query performance.
- Showcased proficiency in Python coding with an emphasis on maintainability, adept in crafting complex SQL queries for data manipulation and analysis.
- Designed, managed, and optimized Snowflake and Redshift data warehouses, ensuring SQL query efficiency and data integrity.
- Engineered real-time data processing solutions on Azure using Delta Table Streaming, Cosmos DB, MongoDB, and Databricks, enhancing data insights and analytics.
- Managed Kafka-based ETL processes, focusing on data transformation for downstream compatibility and compliance with data standards.
- Collaborated cross-functionally to enhance Databricks and Spark capabilities for advanced data processing and analytics.
- Documented ETL processes and database designs, ensuring clarity and collaboration within the team.
- Applied AI and ML technologies, particularly in Azure ML, to drive enhanced data insights and decision-making.

Environment: Apache Spark, Hadoop, Informatica, Hive, AWS (EC2, S3, RDS, Lambda, Glue, Redshift), Azure Services, Apache Kafka, SQL databases (MySQL, PostgreSQL, SQL Server), Python, Delta Table Streaming, Cosmos DB, MongoDB, Databricks, AI/ML technologies.

Data Engineer – Mastercard, Purchase NY **Dec 2019 - Aug 2022**

- Engineered robust data pipelines leveraging Python, SQL, and Shell Scripting to facilitate seamless ETL processes, ensuring efficient data transformation and loading.
- Innovated with Scala, Java, and R to develop customized data solutions, enhancing system performance for dynamic data manipulation and insightful analysis.
- Ensured data quality and compliance using Informatica data governance tools.
- Spearheaded the integration of Hadoop and Spark for advanced analytics, amplifying data insights to support strategic decision-making through optimized processing.
- Designed and managed HBase and Cassandra databases, elevating data storage and retrieval capabilities, and bolstering system performance and availability.
- Implemented data warehousing solutions using Amazon Redshift for large-scale analytics.

- Orchestrated automated data pipelines using Apache Airflow, ensuring consistent and monitored data flow for enhanced reliability in the data ecosystem.
- Collaborated on data modeling using Erwin, translating complex insights into efficient databases to facilitate visualization and informed database design.
- Proficiently managed AWS services, including Glue, EC2, S3, and Redshift, dynamically allocating resources to scale capabilities in response to evolving data needs.
- Pioneered the integration of Kafka and Apache Impala, enabling real-time data streaming and interactive analytics for rapid decision-making.
- Consolidated data from multiple sources into a central data warehouse for unified analysis.
- Ensured data security and compliance using IAM, maintaining governance and integrity across the organization's data ecosystem.
- Drove the adoption of Docker and Kubernetes for application efficiency, containerizing, and orchestrating data-centric apps for streamlined deployment and scalability.

Environment: Python, SQL, Shell Scripting; Informatica, Scala, Java, R, Spark, Hadoop, HBase, Cassandra; Apache Airflow; AWS (Glue, EC2, S3, Redshift); Kafka, Apache Impala, IAM, Docker, Kubernetes.

Data Analyst/Engineer – Comcast, New York NY **Mar 2017 - Nov 2019**

- Collaborated with various teams to extract insights using SQL, Pandas, and NumPy, supporting informed business decisions through statistical analysis.
- Developed data pipelines to clean and transform diverse data sources, ensuring accuracy and integrity for reliable analysis with SQL and Python libraries.
- Designed data models to maintain integrity and adhere to best practices, contributing to optimized insights through comprehensive data analysis.
- Identified trends and anomalies through data analysis, driving decision-making with actionable insights using SQL and Python proficiency.
- Visualized data findings using tools like Tableau, Power BI, and Matplotlib, translating complex data into clear visual narratives for stakeholders.
- Worked closely with Data Engineers to enhance ETL processes, providing data quality feedback to refine pipelines for robust data handling.
- Supported machine learning initiatives by preparing and refining datasets for Data Scientists, aiding in model development and improvement.
- Utilized Kafka and Spark Streaming for real-time data insights, extracting valuable information from dynamic data sources.
- Ensured data security and compliance with access controls and regulations, maintaining data integrity and aligning with privacy standards.
- Enabled complex analytical processing and business intelligence through OLAP-optimized data warehousing.
- Drove innovation by exploring new analysis techniques and staying updated with emerging trends and technologies in data analysis.
- Documented processes, models, and analyses to create comprehensive records, enhancing transparency and future decision-making.

Environment: SQL, Python (Pandas, NumPy), Data Visualization (Tableau, Power BI), Statistical Analysis, Data Cleansing and Transformation, Data Modeling, Machine Learning Foundations, Real-Time Data Processing, Data Security, Documentation Practices.

TECHNICAL SKILLS

Programming Languages:	Java, Python 3. x, Scala 2.12 (for Spark), SQL, Unix-Shell scripting, React 17.x, API/Redux, NodeJS 14.x
Data Processing & Analytics:	PySpark 3.x, TensorFlow 2.x, Pandas 1.x, NumPy 1.x, NLTK 3.x, Scikit-Learn 0.24.x, OpenCV 4.x, Matplotlib 3.x, Seaborn 0.11.x
Databases & Data Warehousing:	Oracle 12c, MySQL 8.x, MongoDB 4.x, Hive 3.x, Redshift, Azure SQL, Snowflake
Big Data Frameworks and Tools:	Hadoop 3.x, HDFS, MapReduce, Kafka 2.8.x, Hive, Zookeeper, Apache Spark 3.x
Cloud Platforms:	AWS (Lambda, S3, Glue, EC2, EMR, Redshift), Azure (Data Lake Gen2, Data Factory, Azure SQL), Snowflake (Stage, Snow pipe, Snowpark), Databricks
DevOps & CI/CD:	Docker 20.x, Kubernetes 1.21.x, Jenkins 2.x, GitLab CI, Azure DevOps
Version Control:	Git 2.3x, GitHub, Bitbucket, GitLab
Business Intelligence:	Tableau 2021.x, Power BI, Looker

EDUCATION

New York City College of Technology, New York NY || BTech in Computer Science.