**NAVEEN KUMAR SURAPANENI
Data Engineer**

**Contact: +1 469-970-5043**

**Email:** **Kumarsurapaneni1008@gmail.com**

**Professional Summary:**

* 10+ years of professional experience in information technology as Software engineer
* Experienced in analyzing data using Hadoop Ecosystems including HDFS, Spark, Hive, Sqoop, Flume, Snowflake, Spark, HBase, and Scala.
* Experienced in Machine learning and Generative AI to gain insights from Data.
* Worked on cloud platforms AWS, Azure, and GCP data ecosystems.
* Experience in building Data pipelines In Azure integrating Azure DevOps platform with Databricks and Data Factory.
* Experienced in Databricks and Azure Cognitive services to utilize AI-based analytics.
* Experienced with AWS services Glue, EMR, Redshift, Athena, Kinesis
* Worked on the GCP cloud platform to utilize Big Query and Dataflow
* Proficient in SSIS and SSRS(MSBI) development and creating dash boards
* Familiarity with code versioning tools, like Git, SVN, and TFS.
* Knowledge of building and deploying data analytics and machine learning solutions with Azure services.
* Creating LLM applications with Lang chain and connecting them through APIs
* Ability to work well with business unit SMEs, Business Analysts, vendors, and end users in cross-functional teams.
* Strong experience in time series forecasting, anomaly detection, reinforcement learning, deploying models in production, data governance, data quality, and data security.
* Proficient in using visualization tools like Tableau Power BI and MSBI.
* Used DBT cloud to aggregate and transform Data and migrate to Snowflake Warehouse
* Worked with relational databases like SQL, MySQL, and Postgres for advanced analytics using stored queries.
* Worked with Apache Beam to connect with GCP BIG Query and Dataflow.
* Experienced in Integrating Both Streaming and Batch data to ADLS (Azure Data Lake) using Kafka and Spark Streaming.
* Experienced in creating jobs using Airflow from Databricks to HDFS
* Additionally, experienced in building data models, creating UDFs, retrieving and manipulating data, and designing and developing ETL processes.
* Assessed the impact of using new data sources on data quality and comprehensiveness of graph.
* Utilized Predictive modelling using generalized linear models, random forests, and ensemble learning to deliver robust and accurate results.
* Optimized algorithms to improve performance and ensure efficient processing.
* Experience in developing and implementing machine learning models using techniques such as supervised and unsupervised learning.
* Experienced in working with Python, TensorFlow, and Keras.
* Adept in data analytics and machine learning using GCP services such as Cloud ML Engine and BigQuery ML.
* Good understanding and knowledge of NoSQL databases like MongoDB, PostgreSQL, HBase, and Cassandra.

**Education:**

* Bachelor Of Mechanical Engineering– JNTUK, India

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming languages** | Python, Java, Scala, PySpark, Shell Scripting, SQL, PL/SQL and UNIX Bash |
| **Operating Systems** | UNIX, LINUX, Solaris, Mainframes, Cloudera |
| **ETL tools** | Alteryx, Informatica, Azure Data Factory, Talend, SSIS |
| **Cloud and Databases** | Databricks, Microsoft Azure, Azure Data Factory, Domino Data Lab, PostgreSQL, MSSQL, Oracle, Data Robot, SPARQL, MongoDB, : Snowflake, DBT, Teradata, Qlik |
| **Libraries:** | Keras, Pandas, Matplotlib, NumPy, SciPy, Streamlit, Scikit-learn, TA-Lib, nltk, SQ Lalchemy, pyspak, apache beam. |
| **Visualization** | Tableau, Power BI, RDF, Star Dog, MSBI |
| **Cloud Technologies** | **Aws:** CDK, Lambda functions, Glue, Athena, EMR**Azure**: ADF, Synapse analytics, Azure DevOps, Azure Databricks, Azure CLI**GCP**: Dataflow, Pub-Sub, Apache Beam, Big Query |
| **IDE Tools** | IntelliJ, Eclipse, Teradata SQL Assistant, PyCharm, Jupyter, Sublime Text, Visual Studio, SQL Developer. |
| **Data ecosystem**  | **Streaming**: Kafka, Pub Sub, Azure Event Hub, Flink**Orchestration**: Apache airflow, ADF, NiFi, oozie**Enterprise Software:** Bitbucket, VMware/Virtual box/Hyper-V, Docker, Miro, Jenkins, Office, Linux and PowerShell, Hyper-V, Virtual box, VMware.**Collaboration tools:** Jira, Miro, Confluence.**Methodologies:** Agile/Scrum, Auto ML, ML Ops (ML Flow. Data version control (DVC)), CI/CD, design, reviews, Waterfall. |
| **Machine Learning** | TensorFlow, Keras Regression, K-Means, SVM, NLP, Neural Networks, like Decision Tree, Random Forest, LLM, Langchain |

**Professional Experience:**

**WarnerMedia-discovery, Los Angeles, USA Mar 2022 - present Senior Data Engineer**

**Responsibilities:**

* Designing and develop Power BI graphical and visualization solutions with Python 3.12 business requirement documents.
* Spearheaded the migration of on-premises data infrastructure to AWS cloud, leveraging services such as Amazon S3, AWS Glue, and Amazon Redshift, resulting in a 30% reduction in operational costs.
* Designed and implemented scalable, fault-tolerant data pipelines using AWS Glue and Apache Spark, processing terabytes of data daily, enhancing data availability and reliability.
* Extensively worked on ETL process to load the data from/into Oracle, Teradata 17.2, DB2, Flat files, etc.
* Experience in OLTP/OLAP system study, Analysis, and modelling, developing data warehouse schemas like Star and Snowflake
* Developed ETL workflows utilizing AWS Glue to extract data from diverse sources, transform it to meet business requirements, and load it into Amazon Redshift for analytics and reporting purposes.
* Implementing Agile Methodology for building an internal application Data Manipulation and Aggregation from different source
* Optimized query performance and reduced costs by 20% through fine-tuning Amazon Redshift clusters, implementing appropriate distribution keys, and optimizing schema design.
* Collaborated with cross-functional teams to gather requirements, design solutions, and deliver high-quality data products, fostering a culture of collaboration and innovation.
* Expert using Business Intelligence tools - Microsoft SSRS 16.0, SSIS 16.0 and SSAS 16.0, Visual Studio, Informatics Power.
* Developing MS SQL Server 2022 and Oracle 12c Database development experience using tables, triggers, views, packages and
* Experienced working in the Agile methodology and data viewers and performance tuning of ETL data flows involved in creating
* Leveraged Google Cloud Platform (GCP) services such as BigQuery, Dataflow, and Cloud Storage to build data pipelines and analytics solutions.
* VB.Net Script for Data Flow and Error Handling using Script component in SSIS 16.0.
* A highly immersive Data Science program involving Data Manipulation &amp; Visualization, Web Scraping, Machine Learning,
* Working with Tabular reports, Matrix reports, Gauges &amp; Chart reports, Parameterized reports, Sub reports, Ad-hoc reports,
* Reverse engineered backend database to change the T-SQL scripts, create Views, Stored Procedures, Triggers and Functions to
* Implemented automated monitoring and alerting systems using Stackdriver to ensure the reliability and availability of data services.
* Participating in all phases of data mining; data collection, data cleaning, developing models, validation, visualization
* Developing Voice Bot using AI (IVR), improving the interaction between Human and the Virtual Assistant on different data
* Creating various database objects (tables, indexes, views, stored procedures and triggers) and implement referential integrity
* Performing T-SQL tuning and optimization of queries that take longer execution time using MS SQL Profiler and Database

**Technologies Used:**  NumPy 1.26, SSRS 16.0, SSIS 16.0, SSAS 16.0, MDM, GIT, Unix, Python 3.12, SAS, Hadoop 3.3, Teradata 17.2,NoSQL, OLTP, OLAP, HDFS, ODS, NLTK, SVM, JSON, XML. ER Studio 9.7, Tableau, AWS, Snowflake 7, Oracle 12c, MS SQLServer 2022

**BNSF Railways, Fort Worth, TX Feb 2021 – Feb 2022
Senior Data Engineer**

**Responsibilities:**

* Databricks integration and deployment of connected engines algorithmic framework for IOT data.
* Developed and maintained data pipelines on AWS to ingest, transform, and load large volumes of freight movement data, enabling real-time analytics and decision-making processes.
* Data Extraction, aggregations, and consolidation of Adobe data within AWS Glue using PySpark.
* Experience in Migrating production infrastructure into Amazon Web Services cloud utilizing AWS Server Migration Service (SMS), AWS Database Migration Service, Elastic Bean Stalk, Cloud Formation, Code Deploy, Code Commit, EBS, and Ops Works.
* Developed and maintained customer relationship management (CRM) systems on AWS, enabling BNSF to personalize services, track customer preferences, and enhance overall customer satisfaction.
* Wrote parsers and data quality check scripts for DAGs deployed on Airflow.
* Implemented geospatial databases and visualization tools on AWS to monitor train movements, track conditions, and infrastructure layout across BNSF's extensive network.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline.
* Expertise in Data Masking and Row level security over data
* Conducted performance tuning and troubleshooting of AWS infrastructure to enhance reliability and scalability of data pipelines.
* Managed databases and storage systems on AWS to store and analyze maintenance records, enabling predictive maintenance algorithms to reduce downtime and optimize asset utilization.
* Built data pipelines and analytics solutions on AWS to process financial transactions and generate reports, supporting financial analysis and decision-making processes within BNSF.
* Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
* Extensively used AWS Athena to ingest structured data from S3 into various systems such as RedShift or to generate reports.
* Collaborated with cross-functional teams to understand business requirements and translate them into technical solutions on AWS platform.
* Developed Python scripts to automate the ETL process using Apache Airflow and CRON scripts in the UNIX operating system as well.
* vision, and speech recognition capabilities into machine learning models.
* Skilled in working with large datasets and distributed computing frameworks, such as Apache Spark, on Azure HDInsight clusters.

**Technologies Used:** Spark, DBT cloud, Azure EventHub, Dat Lake, PowerBI, Scala, Python, Databricks, ADF, Snowflake, Synapse Analytics, GitHub Flow, Airflow.

**SiriusXM, Irving, TX Dec 2018 – Jan 2021
Data Engineer**

**Responsibilities:**

* Participated in Designing databases (schemas) to ensure that the relationship between data is guided by tightly bound Key constraints.
* Writing PL/SQL stored procedures, function, packages, triggers, view to implement business rules into the Application level.
* Extensive experience in Data Definition, Data Manipulation, Data Query and Transaction Control Language.
* Good experience in developing web applications implementing Model View Control architecture using Django, Flask, and spring web application frameworks.
* Understanding the requirements by interacting with business users and mapping them to design and implement it following the AGILE Development methodology.
* Experience in Installing, Upgrading and Configuring Microsoft SQL Server and Migrating data from SQL Server 2008 to SQL Server 2012.
* Developed and tested many features for dashboard using Flask, CSS and JavaScript.
* Experience in designing and creating Tables, Views, User Created Data Types, Indexes, Stored Procedures, Cursors, Triggers and Transactions.
* Developed warmup programs to load recently logged in user profile information into MSSQL DB.
* Performed manual testing and used logging tools like Splunk, putty to read the application logs for elastic search.
* Configured the data mapping between Oracle and SQL Server and tested performance accuracy related queries under SQL Server.
* Created connections to database using Hibernate Session Factory, used Hibernate APIs to retrieve and store data with Hibernate transaction control.
* Experienced in writing JUnit test cases for testing.
* Helped in creating Splunk dashboard to monitor MDB modified in the project.

**Technologies Used:** SQL Developer, Hibernate, Restful Web Services, Splunk, Agile Methodology, UNIX, Oracle, TOMCAT, Django, Flask, Eclipse, Jenkins, CVS, JSON, Oracle PL/SQL.

**Nielsen Corporation - Tampa, FL Sep 2016 - Nov 2018
Data Engineer**

**Responsibilities:**

* Involved in phases of software development life cycle from requirement gathering, development, testing and deployment.
* Data is then cleaned and processed according to the needs of data science team so that a model can be run and generate results helps to analyze the future requirements of products and customer requirements.
* Responsible for moving huge loads of data from sap/Hana database using Databricks to AWS S3 which is the primary staging location.
* Worked on end-to-end pipelines from gathering the data, performing operations according to requirement, storing it in s3/producing the data to streaming data platform.
* Used JDBC for database connectivity to invoke tables and stored procedures from database.
* Developed PySpark code to move the data loads from raw to enhance and generated timeseries data from enhanced data which is used for ML purposes.
* Developed code using PySpark to restore the failed data sets if a daily load gets failed.
* Used Pandas, NumPy, and Spark in Python for developing Data Pipelines.
* Perform Data Cleaning, features scaling, features engineering using Pandas and NumPy packages in Python.
* Written Terraform code to spin up the AWS infrastructure EC2, EMR, S3, RDS with required security protocols.
* Used spark SQL to create views on dataframes, added new columns to the existing views and stored the results in AWS s3 in parquet format and partitioned by date.
* Developed data ingestion modules (both real time and batch data load) to data into various layers in S3, Redshift and Snowflake using AWS Kinesis, AWS Glue, AWS Lambda and AWS Step Functions.
* Handled large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations during ingestion process from staging to enhance.
* Implemented Spark using Scala and SparkSQL for faster testing and processing of data.
* Data Processing: Processed data using Map Reduce and Yarn. Worked on Kafka as a proof of concept for log processing.
* Used Snowflake to migrate data from other databases.
* Monitoring the Hive Meta store and the cluster nodes with the help of Hue.
* Involved in auditing the daily data loads to check the accuracy of data.
* Written an utility job that checks the record counts of staging location and enhance location.
* Developed Jenkins pipeline scripts to deploy Jenkins job to validate daily Databricks jobs are executed at scheduled timing, notification email will be sent if any failure occurs.
* Converted old freestyle Jenkins jobs to pipeline scripts for better understanding of jobs purpose and scheduled them using CRON expressions.
* Implemented liquibase an opensource version control for databases, deployed database scripts contains creation of tables, functions and procedures to AWS RDS.
* Agile/scrum was used as the project management methodology and rally, confluence is used to keep things in check.
* Used Git for version control.

**Technologies used**: AWS, S3, EC2, EMR, S3, RDS, Kinesis, Glue, Lambda, Python, Pandas, PySpark, SQL, JDBC, Scala, Hive, Snowflake, Agile/Scrum, Git

**Anthem - Mason, Ohio Nov 2013 – Aug 2016
Big Data Engineer**

**Responsibilities:**

* Responsible for the design, implementation, and architecture of very large-scale data intelligence solutions around big data platforms.
* Analyzed large and critical datasets using HDFS, HBase, Hive, HQL, Pig, Sqoop and Zookeeper.
* Developed multiple POC’s using Spark, Scala and deployed on the Yarn Cluster, compared the performance of Spark, with Hive and SQL.
* Use Amazon Elastic Cloud Compute (EC2) infrastructure for computational tasks and Simple Storage Service (S3) as storage mechanism.
* Capable of using AWS utilities such as EMR, S3 and Cloud Watch to run and monitor Hadoop and Spark jobs on AWS.
* Worked on SQL queries in dimensional data warehouses and relational data warehouses. Performed Data Analysis and Data Profiling using Complex SQL queries on various systems.
* Troubleshoot and resolve data processing issues and proactively engaged in data modelling discussions.
* Worked on RDD Architecture and implementing spark operations on RDD and optimizing transformations and actions in Spark.
* Written programs in Spark using Python (PySpark) packages for performance tuning, optimization, and data quality validations.
* Worked on developing Kafka Producers and Kafka Consumers for streaming millions of events per second on streaming data.
* Implemented a distributing messaging queue to integrate with Cassandra using Apache Kafka.
* Hands on experience on fetching the live stream data from UDB into HBase table using PySpark streaming and Apache Kafka.
* Evaluate Snowflake Design considerations for any change in the application.

**Technologies used**: HDFS, Python, SQL, Spark, Scala, Kafka, Hive, Yarn, Sqoop, Snowflake, Tableau, AWS Cloud, GitHub, Shell Scripting.

**Hydes’s tech, Bangalore, India Sep 2012- Sep 2013
Big data Engineer**

**Responsibilities:**

* Importing and exporting data into HDFS using Sqoop. Created hive tables which is a one-time activity.
* Writing complex SQL queries by using the joins, Functions, and sub-queries to generate reports on the user or client requirement.
* Managing the various database logs as available in MySQL, Analyzing & fine-tuning the slow query logs
* Design and develop new Power BI solutions and migrate reports from SSRS.
* Developed and executed a migration strategy using Scala applications to move Data Warehouse from Greenplum to the Oracle platform.
* Load data into Amazon Redshift and use AWS Cloud Watch to collect and monitor AWS RDS instances within Confidential.
* Used Talend connectors integrated into Redshift - BI Development for multiple technical projects running in parallel.

**Technologies Used:** HDFS, Hive, Sqoop, Scala, Python, Databricks**,** Talend, Kafka, NiFi,Informatica Power Center 10.x/9.x, IDQ, AWS Redshift, Snowflake, S3, Postgres, Google Cloud Platform (GCP), MS SQL Server, Big query, Salesforce Sql, Python, Postman, Tableau, Unix Shell Scripting, EMR, GitHub.