**Name: Vinay Chandra Email: mahesh@masitlabs.com**

**Phone: +1 (469) 598-1624 Linkedln:**

**Senior Data Engineer**

**PROFESSIONAL SUMMARY**

* Experienced Data Engineer with 10+ years of expertise in SQL, Python, and Data Engineering, specializing in advanced SQL query optimization, stored procedures, and ETL pipeline development.
* Data Engineer & Analyst with expertise in SQL, Splunk, and Tableau. Skilled in data extraction, reporting, and visualization for business insights.
* Expert in Python, PySpark, SQL, and Apache Airflow for ELT pipeline development, workflow orchestration, and DataOps. Proven track record in transforming raw data into actionable insights, improving decision-making efficiency by 40%.
* Well-versed with developing data-intensive applications using GCP services such as Cloud Storage, Dataflow, BigQuery , Cloud Functions, Cloud Run, Pub/Sub, Dataproc, and Cloud Composer
* Optimized ETL pipelines across Azure Synapse, Snowflake, and AWS Redshift, reducing processing time by up to 40% and improving transformation efficiency by 30%.
* Optimized SQL & cloud-based queries in AWS Athena & Azure SQL, reducing execution time by 50% and cutting costs by 30%.
* Automated cloud monitoring using AWS CloudWatch, Azure Monitor, and Splunk, reducing system downtime by 40%.
* Developed distributed ETL pipelines using PySpark & Spark SQL, improving data transformation efficiency by 50%.
* Expert in SQL performance tuning, reducing query execution time by 40-50% across various platforms like Snowflake, Redshift, and Azure SQL.
* Designed and optimized Spark-Scala & PySpark jobs to process terabyte-scale datasets in Azure Databricks & AWS EMR.
* Designed OLAP cubes and materialized views for faster query execution in HBase and Snowflake.
* Designed HIPAA-compliant data pipelines, ensuring end-to-end encryption and secure patient data transmission across AWS and Snowflake.
* Deployed and managed AWS Parallel Cluster for high-performance computing (HPC) workloads in seismic workflows.
* Designed ETL pipelines for financial reporting and risk analytics using Matillion, Snowflake, and AWS Glue. Optimized finance data workflows, reducing processing time by 40% and ensuring compliance with IFRS 17 & SOX standards.
* Refactored Snowflake SQL queries into optimized Spark SQL for Databricks, reducing execution time by 40%.
* Proficient in Apache Airflow, automating data pipelines to improve workflow efficiency by 50%, reducing manual intervention.
* Worked with supply chain teams to analyze procurement trends and optimize supplier performance using data-driven insights.
* Designed ETL workflows using Matillion, SSIS, and Informatica for financial data processing. Automated data transformations, reducing manual effort by 50%.
* Developed and optimized Alteryx workflows for data integration, ETL transformations, and data blending for business insights. Used Alteryx SDK for advanced workflow automation.
* Built Power BI dashboards to track inventory levels, supplier performance, and demand forecasting, improving logistics efficiency by 30%. Developed UI components using Angular for interactive data visualization dashboards.
* Skilled in developing ETL processes, data pipelines, data models, security protocols, and Terraform configurations for GCP environments. Proven track record of reducing operating costs, increasing storage capabilities, decreasing latency, and improving system performance.
* Developed Spark applications in Java for distributed data processing, reducing execution time by 30%.
* Managed Windows and Linux VDIs, ensuring performance optimization and security compliance.
* Developed optimized ELT pipelines in Databricks, migrating large-scale healthcare claims datasets (100M+ records) from MS SQL to cloud, improving data availability by 40%.
* Optimized Alteryx workflows, resolving bottlenecks and improving execution speed by 30%. Designed Alteryx macros to automate complex ETL tasks, improving data transformation efficiency.
* Implemented and managed Amazon DCV Connection Gateway for secure external access to cloud-based VDIs.
* Developed Unix shell scripts for ETL job monitoring, automated alerts, and system maintenance, reducing downtime by 35%.
* Developed AI-driven data quality framework using Python & PySpark, improving data accuracy by 25%.
* Led cross-functional teams in Azure data migration projects, aligning business goals with technical implementation, improving analytics capabilities by 35%.
* Implemented ML models using Decision Trees, Random Forest, K-Means Clustering, and Gradient Boosting for predictive analytics and anomaly detection.
* Led enterprise-wide cloud data modernization strategy, migrating legacy data warehouses from Teradata & SQL Server to Snowflake & GCP BigQuery, reducing infrastructure costs by 35%.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Big Data Frameworks | HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala, MongoDB |
| Programming languages | Python, PySpark, Scala, SQL, PL/SQL, C++, Spark, Rest APIs, Java |
| Hadoop Distribution | Cloudera CDH, Horton Works HDP, Workday Data Processing, Fabric Pipelines Apache |
| Machine Learning Classification Algorithms | Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Principal Component Analysis, Fraud Detection, Data Governance, Grafana, OCI, TensorFlow C++ API, OpenCV, ONNX, Eigen, Scikit-Learn. ensorFlow |
| Version Control | GITHUB, Jenkins, Bitbucket, CI/CD, Android NDK, iOS Core ML |
| IDE & Tools, Design | Eclipse, Visual Studio, Net Beans, MySQL, PowerBI, Tableau, Splunk, Qlik Sense, Microsoft Fabric, MySQL Workbench, Monte Carlo, VaR, Time Series |
| Databases | Oracle, SQL Server, MySQL, DynamoDB, Cassandra, Teradata, PostgreSQL, MS Access, Workday Data Warehouse Snowflake, NoSQL Database (HBase, MongoDB). |
| Cloud Technologies | | MS Azure, Amazon Web Services (AWS), Google cloud, Microsoft OneLake |

**PROFESSIONAL EXPERIENCE**

**Client: Citigroup, Irving, TX June 2023 to Present**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Developed and optimized SQL-based ETL pipelines in Snowflake, Redshift, and AWS Glue, improving data ingestion efficiency by 40% and reducing query execution time by 50%.
* Involved in loading and transforming large sets of the structured, semi structured dataset and analyzed them by running Hive queries.
* Implemented data ingestion and transformation workflows in Palantir Foundry, integrating real-time healthcare claims processing data. Optimized Foundry Query Pipelines & Data Streams, reducing latency by 50%.
* Developed finance data models using Kimball methodology in Snowflake & Redshift. Optimized ETL processes, reducing query execution time by 50%.
* Developed real-time data streaming pipelines using Kafka & Spark Streaming, reducing event processing latency by 45%.
* Automated AWS CloudWatch monitoring for NaviNet® AllPayer’s infrastructure, reducing downtime and improving real-time error detection.
* Developed Apache Spark applications using Java for large-scale data processing, reducing ETL execution time by 30%.
* Designed finance data models in Snowflake & Redshift using Kimball methodology, optimizing ETL processes for financial reconciliation, ledger consolidation, and SAP finance data integration. Ensured compliance with Basel II, SOX, and IFRS 17, improving data accuracy by 40%.
* Implemented serverless ETL workflows using AWS Lambda & DynamoDB, improving real-time data processing efficiency by 40%.
* Designed and optimized Tableau & Power BI dashboards for payer and provider networks, reducing report load times by 60% and improving real-time insights for claim approvals.
* Migrated 20TB+ patient claims and financial data from SQL Server to AWS Redshift & S3 using AWS DMS, reducing data transfer time by 40%.
* Implemented Delta Lake for efficient data versioning and optimized table storage, reducing query execution time by 35%
* Designed controlled experiments (A/B testing, hypothesis testing) to measure business impact and optimize decision-making.
* Developed serverless ETL pipelines using AWS Lambda & Glue, automating data transformation tasks.
* Documented cloud architecture and best practices to streamline onboarding for providers using NaviNet® AllPayer.
* Implemented data anonymization & pseudonymization techniques for handling PHI (Protected Health Information) in compliance with HIPAA & HITECH.
* Optimized AWS HPC networking using Elastic Fabric Adapter (EFA) for low-latency seismic computing.
* Designed interactive Tableau dashboards to visualize KPIs and track business performance. Created LOD expressions, custom calculations, and data blending for advanced reporting.
* Presented data-driven insights to stakeholders, helping engineering teams optimize performance and reduce costs by 30%.
* Developed and automated Apache Airflow DAGs for Snowflake data pipelines, reducing manual intervention by 50%
* Built automation scripts in Python & Perl to optimize claim data ingestion, reducing processing time by 25%.
* Led the migration of high-volume transactional data from Oracle to MongoDB, optimizing schema design, indexing, and document storage for better performance. Integrated Kafka-based real-time data ingestion into MongoDB, ensuring low-latency access to critical business data.
* Worked on audience targeting, data privacy (GDPR, CCPA), and cross-device graphs for digital advertising analytics, improving campaign performance by 30%
* Developed dynamic data visualization dashboards using Angular 10+ and integrated with AWS API Gateway for real-time analytics.
* Actively participated in Agile project management for NaviNet® AllPayer’s enhancements, leading sprint planning and retrospective discussions.
* Designed ELT workflows in Snowflake & Redshift for financial reconciliation, ensuring compliance with IFRS 17 & SOX.
* Developed DBT models and Matillion workflows to automate ETL for healthcare claims processing, reducing manual intervention by 40%.
* Worked with clinical research datasets, pharma regulatory data (FDA, CFR Part 11), and patient health records to ensure compliance with HIPAA & GDPR.
* Led stakeholder workshops to define SAP data cleansing rules, improving migration efficiency by 30%
* Applied time series forecasting (ARIMA, Prophet) to predict business trends and optimize decision-making.
* Analyzed Salesforce CRM data to track customer behavior and retention metrics. Integrated customer interaction data from Salesforce, Snowflake, and external sources.
* Partnered with finance & procurement teams to analyze supplier cost trends, leading to $2M in annual savings on contract negotiations.
* Developed interactive BI dashboards in Tableau and Looker, ensuring intuitive user experience for non-technical stakeholders.
* Developed ETL solutions for SAP finance data integration using Matillion & SSIS. Improved data ingestion efficiency by 35%.

**Environment:** AWS (S3, Redshift, Glue, EC2, IAM, StepFunctions, CloudWatch, SES), Snowflake, Airflow, AWS Glue, PySpark, Tableau, Power BI, Python (Boto3, NumPy, Fabric), SQL, Workday Data Warehouse, Chef, CloudFormation, SAS, Agile

**Client: Target, Minneapolis, MN Feb 2021 – May 2023**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Configured and implemented the Azure Data Factory Triggers and scheduled the Pipelines; monitored the scheduled Azure Data Factory pipelines and configured alerts to detect failures in urban infrastructure data workflows.
* Developed ETL pipelines to extract data from city planning systems, transforming it into standardized formats, facilitating seamless integration with SQL databases for housing, zoning, and transportation analysis.
* Collaborated with development teams to design cloud-native applications for smart city initiatives, ensuring seamless integration with existing urban management services.
* Designed and optimized data models in Snowflake & Azure Synapse, reducing transformation time by 30%.
* Developed Python-based predictive analytics models to detect irregularities in city utility usage, increasing detection accuracy by 30%.
* Developed end-to-end finance process maps integrating SAP & Oracle ERP, aligning financial reporting and compliance tracking with business goals.
* Built and optimized real-time streaming pipelines using Kafka and Spark Streaming, reducing event processing latency by 45%
* Developed high-performance ETL pipelines in Databricks & Azure Synapse, integrating structured & unstructured data.
* Designed optimized data models in Snowflake & Azure Synapse, reducing query execution time by 40% and enhancing business reporting efficiency.
* Built entity resolution pipelines using TigerGraph, enabling a 50% improvement in fraud detection accuracy.
* Enhanced the performance of ETL processes for urban infrastructure data by optimizing SSIS package configurations, utilizing parallel processing, and implementing best practices for data flow management.
* Automated dashboards refresh cycles in Looker and Power BI using dbt and Airflow DAGs, reducing manual interventions by 70%.
* Worked closely with city planners, data architects, and analysts to ensure that the data infrastructure supports urban expansion, housing, and transportation development.
* Developed and optimized data pipelines using PySpark and Scala for large-scale data processing in Azure Synapse.
* Implemented comprehensive monitoring and alerting solutions using tools like Splunk and New Relic, ensuring proactive detection and resolution of performance issues in smart city systems.
* Designed and optimized event-driven data pipelines using Kafka and Azure Functions for real-time city traffic and infrastructure monitoring.
* Implemented security protocols and access controls to protect sensitive zoning, housing, and citizen data, ensuring compliance with government data privacy regulations.
* Integrated SAP transactional data from city infrastructure management systems into SQL databases, developing robust ETL pipelines for data extraction, transformation, and loading.
* Automated Azure Synapse deployments and infrastructure provisioning using Terraform and Azure DevOps Pipelines to improve scalability for city planning applications.
* Designed and optimized search queries using Elastic Search and OpenSearch, implementing full-text search indexing for zoning laws, real estate records, and permit data, improving data retrieval speed by 30%.
* Developed real-time log analytics pipelines, integrating Elastic Search with Kafka and Spark Streaming to process high-volume city management and public utility data.
* Developed scalable ETL workflows in Azure Synapse Analytics using PySpark, reducing data processing time by 30% for urban growth projections.
* Analyzed financial data processing workflows, identifying inefficiencies and implementing SQL-based optimizations, reducing reconciliation time by 30%.
* Developed ML-based predictive modeling in Databricks using PySpark & SQL, improving demand forecasting accuracy by 30% for public transportation and housing development.
* Implemented a CI/CD pipeline using Jenkins, Airflow for Containers from Docker and Kubernetes, automating the deployment of smart city analytics platforms.
* Developed governance workflows using Apache Airflow and Azure Logic Apps for tracking zoning changes, infrastructure updates, and smart city metadata.

**Environment:** Azure, PySpark, SSIS, SAP Data Services, Python, Databricks, Snowflake, Teradata, SQL Server, Hadoop, ETL operations, Data Warehousing, Data Modeling, Urban Development Analytics, Cassandra, Workday ETL Pipelines, Synapse Pipelines, Advanced SQL methods, Workday Data Processing, NiFi, Linux, Apache Spark, Scala, Spark-SQL, HBase.

**Client: Johnson & Johnson, New Brunswick, NJ March 2018 – Dec 2020**

**Role : Data Engineer**

**Responsibilities:**

* Responsible for analyzing large data sets to develop multiple custom models and algorithms to drive innovative business solutions.
* Involved in designing data warehouse and data lakes on regular (oracle, SQL Server) high performance on big data (Hadoop-Hive and HBASE) databases. Data modeling, Design, implement , and deploy high-performance, custom applications at scale on Hadoop / Spark.
* Experience setting up instances behind Elastic Load Balancer in AWS for high availability and cloud integration with AWS using ELASTIC MapReduce (EMR).
* Experience working with Android/iOS data collection and processing pipelines. Worked with Android NDK and iOS Core ML for deploying ML models on mobile devices.
* Automated data ingestion and ETL processes on AWS EMR, reducing processing time by 30% and ensuring data availability for analytics in near real-time.
* Experience in working in Hadoop eco-system integrated into the Cloud platform provided by AWS with several services like Amazon EC2 instances, S3 bucket and RedShift.
* Implementation of new tools such as Kubernetes with Docker to assist with auto-scaling and continuous integration (CI) and upload a Docker image to the registry so the service is deployable through Kubernetes. Use the Kubernetes dashboard to monitor and manage the services.
* Built API-driven ETL workflows to extract metadata from Oracle, Snowflake, and AWS for Collibra integration.
* Developed ETL pipelines for financial reconciliation and risk analytics using AWS Glue & Snowflake, ensuring SOX & IFRS 17 compliance.
* Developed ETL pipelines for financial reconciliation & risk analytics in Snowflake & AWS Glue, ensuring compliance with SOX & IFRS 17.
* Contributed to a multi-year enterprise-wide risk analytics program by designing ETL pipelines that ensured compliance with AML/OFAC regulatory frameworks.
* Written script from scratch to create AWS infrastructure using languages such as BASH and Python, created Lambda functions to upload code and to check changes in S3, DynamoDB table.
* Improvised a python module that de-normalizes data from RDBMS to JSONs and saved 35 hours as part of the migration.
* Created program in python to handle PL/SQL functions like cursors and loops which are not supported by snowflake.
* Utilized Rapid Deployment Solution, BPDMs and customized Jobs to migrate data with SAP Data Services (BODS) for following items: Purchase Orders, DMS, Inventory, PIRs, Misc. PRTs, Batch Master, Work Centers, EWM Stock Transfer, Product Master, Storage Bins, and Inspection Rules.
* Orchestrated and migrated CI/CD processes using Cloud Formation and Terraform, packer Templates and Containerized the infrastructure using Docker, which was setup in OpenShift, AWS and VPCs.
* Create dashboards on snowflake cost model, usage in QlikView.
* Worked on Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, and Spark Yarn.
* Created Power BI dashboards to track sales performance, leading to a 15% increase in revenue by optimizing pricing strategies.
* Partnered with finance and marketing teams to analyze customer behavior and optimize marketing campaigns.

**Environment**: Hadoop, ETL operations, Data Warehousing, SAP Data Services, Data Modelling, Cassandra, AWS Cloud computing architecture, EC2, S3, Advanced SQL methods, NiFi, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase, AWS Redshift

**Client: All State, Northfield, IL July 2016 to Feb 2018**

**Role: Data Engineer**

**Responsibilities:**

* Created and maintained detailed documentation of GCP infrastructure, configurations, and processes to support ongoing operations and future scaling.
* Created Schema objects like Indexes, Views, and Sequences, triggers, grants, roles, Snapshots.
* Developed end-to-end ETL pipelines using Snowflake's Snow pipe and other ETL tools to extract, transform, and load data from various sources into Snowflake.
* Managed multiple Life & Annuity data projects concurrently, ensuring on-time delivery while maintaining data integrity.
* Used advanced Excel formulas and functions like Pivot Tables, Lookup, If with and/index, match for data cleaning.
* Performed SQL validation to verify the data extracts integrity and record counts in the database tables.
* Built enterprise-wide ETL solutions for SAP-based Life & Annuity insurance data integration using Matillion & SSIS. Automated policy data ingestion, claims reconciliation, and premium processing, improving data ingestion efficiency by 35% while ensuring compliance with NAIC regulatory reporting standards.
* Effectively used data blending feature in Tableau to connect different databases like Oracle, MS SQL Server.
* Transferred data with SAS/Access from the databases MS Access, Oracle into SAS data sets on Windows and UNIX.
* Utilized Google Cloud AI and Machine Learning services, such as AI Platform, for developing and deploying ML models.
* Optimized Google BigQuery transformations, reducing compute costs by 30% and improving query execution speeds by 50%.
* Developed fraud detection models using PySpark and Kafka, identifying anomalous financial transactions in real time.
* Performed data cleaning and preprocessing in SQL, Python, and Power BI, ensuring high data accuracy and consistency.
* Wrote complex SQL queries for validating the data against different kinds of reports generated by Business Objects.
* Designed microservices architectures using Google Kubernetes Engine (GKE) for container orchestration.
* Used ad-hoc queries for querying and analyzing the data.

**Environment:** Google cloud, cloud functions, SQL, PL/SQL, Oracle9i, SAS, Business Objects, Tableau, Crystal Reports, T-SQL, SAS, UNIX, MS Access2010.

**Client: ADP, Roseland, NJ. Sep 2014 – June 2016**

**Role: DATA ANALYST**

**Responsibilities:**

* Collaborated with database administrators and application developers to optimize and fine-tune Oracle PL/SQL stored procedures, functions, and packages for improved performance and efficiency.
* Played a key role in the design phase of database schemas, ensuring that data relationships were established through tightly bound key constraints, thus enhancing data integrity and consistency.
* Actively engaged with users and application developers to understand their business requirements, translating them into effective database solutions that met their needs.
* Developed rest API's using python with flask framework and done the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Worked on Python Open stack API's and used Python scripts to update content in the database and manipulate files.
* Demonstrated proficiency in creating various database objects, including tables, indexes, views, and constraints, aligning them with business requirements and best practices.
* Utilized a wide range of Data Definition Language (DDL), Data Manipulation Language (DML), Data Query Language (DQL), and Transaction Control Language (TCL) statements to manipulate and manage data within the Oracle database.
* Enforced data integrity by implementing primary keys and foreign keys, ensuring data consistency and accuracy across the database.
* Involved in migration of databases from UAT servers and prod deployment
* Monitoring Python scripts run as daemons in the UNIX/Linux system background to collect trigger and feed arrival information. Created Python/MySQL backend for data entry from Flash
* Enhanced the performance of existing PL/SQL programs by identifying bottlenecks and applying optimization techniques, resulting in improved query response times.
* Developed complex SQL queries and integrated them into Oracle Reports, facilitating the generation of comprehensive and informative reports for end-users.
* Implemented data validations using database triggers to enforce data quality standards and business rules.
* Leveraged import/export utilities, such as UTL\_FILE, to facilitate seamless data transfer between Oracle database tables and flat files.
* Conducted SQL tuning by analyzing execution plans (Explain Plan) to identify and resolve performance issues, optimizing query execution.
* Provided valuable support during the project's implementation phase, ensuring the successful deployment of database solutions.
* Worked with Oracle's built-in standard packages like DBMS\_SQL, DBMS\_JOBS, and DBMS\_OUTPUT to streamline database operations and automate tasks.
* Designed and implemented report modules within the database, integrating them with client systems using Oracle Reports, aligning with business requirements and reporting needs.
* Responsible for scheduling and deployment of data pipelines using corn jobs
* Developed Job Scheduler scripts for data migration and built automation jobs using UNIX Shell scripting.

# Environment: python, Oracle 9i, Oracle Reports, SQL, PL/SQL, SQL\*Plus, SQL\*Loader, Unix/Linux, Windows XP.