AVINASH MUVVA

## Sr. Data Engineer

## G-Mail: avinash.muvva01@gmail.com| Contact: 972-275-9579

## PROFESSIONAL SUMMARY

* Sr. Data Engineer with overall 10+ years of IT experience in data management and proficient in the areas of Database development, ETL development, data modelling and experience on various technologies, tools and databases like Big Data, Azure, AWS, S3, Snowflake, Hadoop, Hive, Spark, python, java, scala, CDL(Cassandra), SQL Server, Oracle, MySQL, PLSQL, PostgreSQL, Matillion, SSIS .
* Proficient in AWS Glue and Apache Spark ETL process design, development, and optimization for effective data extraction, transformation, and loading.
* Proficient in data governance tools like Collibra and analytics automation platforms such as Alteryx.
* Experience with ETL workflow Management tools like Apache Airflow and have significant experience in writing the python scripts to implement the workflow.
* Experience in building of SPARK applications using Spark tools such as Spark core, Spark MLlib, Spark Streaming, and Spark SQL.
* Strong experience and knowledge of real time data analytics using Spark Streaming, Kafka and Flume.
* Good working knowledge on Snowflake and Teradata databases.
* Experience with Amazon Web Services (AWS) cloud and its services, including Snowflake, EC2, S3, RDS, EMR, VPC, IAM, Elastic Load Balancing, Lambda, RedShift, Elastic Cache, Auto Scaling, Cloud Front, Cloud Watch, Data Pipeline, DMS, Aurora, ETL, and other AWS Services.
* Proficient in analytics and large data processing using Amazon EMR (Elastic MapReduce).
* Experience with EC2 in setting up of instances, working with infrastructure teams to resolve complex issues, and the establishment of security groups.
* Experience in the integration of data from multiple source systems, including the import of nested JSON formatted data into a Snowflake table, using the AWS Snowflake cloud data warehouse and AWS S3 bucket.
* Experience in working with Azure Blob Storage, Azure Data Lake, Azure Data Factory, Azure SQL, Azure SQL Data warehouse, Azure Analytics, PolyBase, Azure HDInsight, Azure Databricks.
* Experience at the development level in Microsoft Azure, where data migration and scheduling capabilities are provided to cloud-based technologies, including Azure Blob Storage and Azure SQL Database.
* Expert in the Python programming language and its various libraries, including NumPy, Matplotlib, Pandas, and others, for data analysis and visualization.
* Proficient in data development and aggregation via scripting and coding using Python, PySpark, and Spark-SQL; proficient in XML, JSON, CSV, Avro, Parquet, and ORC, among other formats.
* Experience in the creation and implementation of relational database schemas, which include tables, indexes, constraints, and relationships, as well as the ability to create effective SQL queries for data retrieval, manipulation, and analysis.
* Familiar with containerization technologies such as Docker and container orchestration systems such as Kubernetes for the purpose of deploying and managing applications that require a significant amount of data.
* Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using R Studio, Hadoop, MongoDB, Tableau, Power BI and Alteryx.
* Experience in developing Conceptual, logical models and physical database design for Online Transactional processing (OLTP) and Online Analytical Processing (OLAP) systems.
* Good knowledge in the core concepts of programming such as algorithms, data structures, collections.
* Experience in the use of version control systems such as Git for collaborative development, as well as the ability to write clear, maintainable code in accordance with best practices and coding standards.
* Strong experience in Software Development Life Cycle (**SDLC**) including Requirements Analysis, Design Specification, application development, application migration and maintenance and testing as per Cycle in both Waterfall and Agile methodologies.
* Proficient in the implementation of Agile methodologies and project management tools such as JIRA to facilitate efficient collaboration, task monitoring, and iterative development cycles.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data / Hadoop technologies** | HDFS, HBase, Hive, Kafka, MapReduce, Spark, SparkSQL, PySpark, Sqoop, Pig, Oozie, ,Collibra, Airflow, Zookeeper |
| **Scripting Languages** | Python, Java, Scala, Shell Scripting, HTML, CSS |
| **Databases** | Microsoft SQL Server, MySQL, DB2, Oracle, Teradata, PostgreSQL, PLSQL |
| **Cloud technologies** | AWS (Amazon Web Services), Microsoft AZURE |
| **Reporting tools** | Power BI, Tableau, MS Office, SSIS, SSRS |
| **ETL tools** | Informatica, Pentaho,DBT, Alteryx, Matillion |
| **Version control** | Git, SVN, Jira, Postman |
| **Operating Systems** | Unix, Linux, Windows |
| **API Integration** |  REST, SOAP XML, OAuth, IAM User, Security Tokens |

## PROFESSIONAL EXPERIENCE

## Truist – Charlotte, NC | Sr. Data Engineer Feb 2023 – Present

## Responsibilities:

* Experience in the design and development of many data pipelines, including the end-to-end ETL and ELT processes for data ingestion and transformation.
* Worked extensively with AWS services like EC2, S3, VPC, ELB, Auto Scaling Groups, Route 53, IAM, CloudTrail, CloudWatch, CloudFormation, CloudFront, SNS, and RDS.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, Parquet/Text Files into AWS Redshift, integrating Immuta for enhanced data security and compliance.
* Designed and implemented data models using Kimball methodology, including star schema creation, to optimize data storage and retrieval.
* Developed and automated ETL processes using DBT, Matillion and Alteryx to streamline data transformations and loading into Snowflake.
* Developed python scripts to parse XML, Json files and load the data in AWS Snowflake Data warehouse, with data governance managed by Immuta.
* Developed Python scripts for automating data extraction, transformation, and loading (ETL) processes, significantly reducing manual intervention and operational costs by 30%.
* Designed and implemented PLSQL procedures to handle complex business logic and data transformations, optimizing data flow across various Oracle databases.
* Created advanced PLSQL scripts for data migration and integration tasks, ensuring seamless data consistency and reliability across multiple environments.
* Strong background in Data Warehousing, Business Intelligence and ETL process (Informatica, AWS Glue) and expertise on working on large data sets and analysis.
* Apache Kafka, Nifi and some other Big Data tools are used to ingest data in batch mode for the data warehouse.
* Designed and implemented ETL pipelines from various Relational Data Bases to the Data Warehouse using Apache Airflow, incorporating Starburst for optimized querying across data sources and used SSIS to automate data extraction.
* Integrated Matillion for ETL processes to manage data pipelines, ensuring smooth data ingestion and transformation workflows.
* Created, configured, and subscribed to REST and SOAP APIs to streamline data ingestion and integration processes, enhancing data availability and reliability for analytics teams.
* Integrated OAuth, IAM User, and Security Tokens for secure API access and data governance, ensuring compliance with organizational security policies.
* Integrated Collibra for data governance, ensuring data quality, compliance, and efficient data stewardship across the organization.
* Working knowledge of Spark RDD, Data frame API, Data set API, Data Source API, Spark SQL, and Spark Streaming.
* Worked on Data Extraction, aggregations and consolidation of Adobe data within AWS Glue using PySpark.
* Used Spark Data Frame API to process Structured and Semi Structured files and load them into S3 Bucket.
* Used Spark Data Frames Operations to perform required validations in the data and to perform analytics.
* Deployed Lambda and other dependencies into AWS to automate EMR Spin for Data Lake jobs.
* Worked on analyzing the Hadoop cluster using different big data analytic tools including Pig, Hive and Map Reduce, and utilized Starburst for fast, interactive analytics across distributed data sources.
* Worked with ETL tools to migrate data from various OLAP and OLTP databases to the AWS Cloud.
* Leveraged AWS data storage services such as Amazon S3 for building data lakes and Amazon Redshift for data warehousing, facilitating centralized storage and analysis of data.
* Worked on AWS research stories for developing new Infra using various AWS services like SQS, SNS, S3, EC2, Lambda, API Gateway, Kinesis.
* Developed serverless architectures leveraging AWS Lambda and API Gateway to build event-driven, microservices-based systems, reducing infrastructure costs and improving deployment speed.
* Integrated Immuta for advanced data governance and policy enforcement, ensuring secure access and compliance across data pipelines and analytics.
* Utilized Alteryx for data preparation, blending, and analytics, enabling more efficient data-driven decision-making and reporting.
* Incorporated Tableau for data visualization and reporting, enhancing the ability to create insightful dashboards and interactive reports for better business intelligence and decision-making.
* Provide ongoing support and maintenance for existing data pipelines, ensuring seamless integration and operation.
* Assist customers with issues on a dedicated Slack channel, leveraging technical expertise to provide timely solutions.
* Execute complex data queries and manage databases using SQL and Snowflake.
* Extensive experience with upstream and downstream data processes, ensuring data consistency and accuracy across the pipeline.
* Collaborate with cross-functional teams and optionally use Amazon QuickSight and Tableau for data visualization and reporting.
* Designed and implemented CloudFormation templates for various AWS services including EC2 instances, S3 buckets, IAM roles, Lambda functions, and VPC configurations.
* Developed and maintained RESTful APIs and integrated third-party services using AWS API Gateway, enabling seamless communication between front-end applications and backend services.
* Utilized Apache Airflow to orchestrate data pipelines, ensuring reliable and timely data flow across various platforms and systems.
* Hive queries were optimized to retrieve the customer information from HDFS.
* Integrated Dynatrace for monitoring and optimizing application performance, ensuring high availability and reliability of data systems.
* Utilized SonarQube to maintain code quality and ensure compliance with coding standards across all data engineering projects.
* Conducted data analysis using Hive, which included partitioning and bucketing the data, and generated a variety of metrics for reporting purposes.
* Implemented advanced procedures like text analytics and processing using the in-memory computing capabilities like Apache Spark written in Scala.
* Automated routine data reporting processes using Tableau and Python scripts, reducing manual effort and minimizing errors.
* Developed infrastructure as code (IaC) solutions using Terraform, enabling automated and consistent deployment of AWS resources.
* Developed customized data solutions for specific business requirements using Python, SQL, and Tableau, ensuring that all data needs were met efficiently and effectively.
* Implemented deployment automation using Octopus Deploy, streamlining the release process and reducing deployment time for data pipelines and applications.
* Used GIT Repositories for code check-ins.
* Utilized Postman for API testing and documentation, improving the efficiency of API development and integration cycles.
* Used SDLC (System Development Life Cycle) methodologies like Agile and Scrum methodology.
* Interacted with the infrastructure, network, database, application, and BA teams to ensure data quality and availability.
* Actively participated in code and design document reviews and provided constructive feedback to peers, contributing to improved project outcomes.
* Implemented performance optimization strategies, such as database query tuning, code refactoring, and cache systems, to improve the responsiveness and scalability of the application.

**Environment**: Python, AWS Glue, RDS, DynamoDB, Redshift, EC2, S3, lambda, EMR, Snowflake, Alteryx, Collibra, Informatica, Spark, Kafka, Hadoop, Scala, Hive, Terraform, SonarQube, Dynatrace, Octopus, Immuta, Starburst, Tableau, GitHub, Linux.

## Equifax – Atlanta, GA | Azure Data Engineer Jul 2021 – Jan 2023

## Responsibilities:

* Analyze, create and build modern data visualizations solutions using Azure Paas. Examine current Production application and how new implementation would affect business operations.
* Expert in using Databricks with Azure Data Factory (ADF) to compute large volumes of data.
* Used Azure Data Factory, T-SQL, Spark SQL, and U-SQL, Azure Data Lake Analytics to extract, transform and load data from source Systems to Azure Data Storage.
* Integrated Collibra to ensure data governance and quality throughout the ETL processes.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Created Spark applications utilizing PySpark and Spark-SQL to extract, process and aggregate data from various file formats. Leveraged Alteryx for data preparation and analytics to provide valuable insights into client usage patterns.
* Proven hands-on experience as a software engineer with a focus on one or more technologies inclusive of NodeJS, TypeScript, JavaScript, ExpressJS.
* Employed Python to automate data aggregation and reporting tasks, leveraging libraries such as Pandas and Matplotlib for dynamic visualizations and dashboards.
* Wrote complex PLSQL scripts to extract and transform data from various sources, optimizing the ETL processes and improving overall data accuracy and performance.
* Experience in Mocha, CouchDB, Redis, Kafka, Docker and Kubernetes in Azure platform.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Work closely with IT and the Business group to understand business reporting requirements and analyses logical model and develop subject matter expertise in a short time.
* Helped other teams such as Ops team to configure Data Lake for their data processing.
* Implemented the Spark best practices to efficiently process data to meet ETAs by utilizing features like partitioning, resource tuning, memory management and check pointing features.
* Migration of on-premises data (SQL Server / MongoDB) to Azure Data Lake Store (ADLS) using Azure Data Factory (ADF).
* Implement data security measures including encryption, data masking, and access control using Azure IAM, Azure Key Vault, and Azure Policy.
* Used Azure Key Vault for managing and protecting sensitive data such as encryption keys and secrets.
* Used Pig as ETL tool to do transformations, joins and some pre-aggregations before storing the data into HDFS.
* Used python and shell scripts to automate Teradata ELT and Admin activities.
* Developed Databricks ETL pipelines using notebooks, Spark DataFrames, Spark SQL and python scripting.
* Developed complex SQL queries using stored procedures, common table expressions (CTEs), temporary tables to support Power BI and Tableau reports.
* Created Power BI and Tableau visualizations and dashboards as per requirements, providing comprehensive and interactive reporting solutions.
* Integrated Jenkins with Azure services such as Azure DevOps, Azure Data Factory, and Azure Databricks to streamline and automate data engineering workflows.
* Working closely with DevOps teams to integrate data engineering solutions into CI/CD pipelines, ensuring smooth deployment and operation of data services.
* Leveraged Collibra for data governance and quality management and used Alteryx for advanced data preparation, blending, and reporting, ensuring robust data handling and insightful analytics.
* Provided assistance in the implementation of deployment operations and ensure the successful migration of all necessary components into the production environment using Microsoft TFS, a project management tool. The migration components consist of a windows service, configuration loaders, azure blob objects, data lake, web tasks, and azure data warehouse components.
* Actively participating in Agile ceremonies such as daily stand-ups, sprint planning, sprint reviews, and retrospectives to ensure alignment with team goals and continuous improvement.
* Incorporated Tableau to enhance data visualization and reporting, creating dynamic dashboards and reports that provided actionable insights and improved decision-making.
* Used Agile practices to track sprint progress, manage workload, and report on task completion and project status.

**Environment**: Python, Azure Data Factory, Azure Data lake, Azure Blob, Power BI, Collibra, Alteryx, Tableau, Shell script, Spark, Databricks.

## Change Healthcare – Bridgeton, MO| Azure Data Engineer Oct 2019 – Jun 2021

## Responsibilities:

* Worked with data transfer from on-premises SQL servers to cloud database (Azure).
* Created CI-CD Pipelines using Azure DevOps.
* Created Azure data factory (ADF pipelines) using Azure PolyBase and Azure blob.
* Performed ETL using Azure Data Bricks. Migrated on-premises Oracle ETL process to Azure Synapse Analytics.
* Designed and managed data storage solutions like Azure SQL database and Azure Data Lake Storage.
* Worked directly with Big Data Architecture Team, which created the foundation of this Enterprise Analytics initiative in a Hadoop-based Data Lake.
* Utilize tools such as Python, Jupyter notebooks, Azure Data Lake, Azure Databricks, Azure SQL database, Azure SQL Data Warehouse.
* Developed and maintained PLSQL scripts for extracting, transforming, and loading data into Azure SQL databases, supporting data warehousing and reporting requirements.
* Designed and implemented end-to-end data pipelines using Azure Data Factory to facilitate efficient data ingestion, transformation, and loading (ETL) from diverse data sources into Snowflake data warehouse.
* Orchestrated robust data processing workflows utilizing Azure Databricks and Apache Spark for seamless large-scale data transformations and advanced analytics improving data processing speed by 14%.
* Developed real-time data streaming capabilities into Snowflake by seamlessly integrating Azure Event Hubs and Azure Functions, enabling prompt and reliable data ingestion.
* Deployed Azure Data Lake Storage as a reliable and scalable data lake solution, implementing efficient data 0partitioning and retention strategies to store and manage both raw and processed data effectively.
* Employed Azure Blob Storage for optimized data file storage and retrieval, implementing advanced techniques like compression and encryption to bolster data security and streamline storage costs.
* Enforced data governance and comprehensive data quality checks using Azure Data Factory and Snowflake, guaranteeing the highest standards of data accuracy and consistency.
* Implemented robust data replication and synchronization strategies between Snowflake and other data platforms leveraging Azure Data Factory and Change Data Capture techniques, ensuring data integrity and consistency with a 98% reduction in data inconsistencies.
* Developed custom monitoring and alerting solutions using Azure Monitor and Snowflake Query Performance Monitoring (QPM), providing proactive identification and resolution of performance bottlenecks.
* Integrated Snowflake seamlessly with Power BI and Azure Analysis Services to deliver interactive dashboards and reports, empowering business users with self-service analytics capabilities.
* Optimized data pipelines and Spark jobs in Azure Databricks through advanced techniques like Spark configuration tuning, data caching, and data partitioning, resulting in superior performance and efficiency.
* Used Azure Databricks to do ETL (Extract, Transform, Load).
* Responsible for building scalable distributed data solutions using Hadoop.
* Monitored and optimized data pipelines and Azure services for performance and cost efficiency.
* Conducted performance tuning and query optimization for Azure SQL databases and data warehouses.
* Developed business intelligence solutions using SQL Server data tools and load data to SQL & Azure cloud databases.
* Stayed updated with Azure data technologies, best practices, and industry trends to drive continuous improvement and innovation.

**Environment**: Python, Azure Data Factory, Azure Synapse Analytics, PolyBase, Blob, SQL Server, Jupyter.

## ERT - Boston, MA | Bigdata Engineer Jan 2017 – Sep 2019

## Responsibilities:

* Designed and implemented a scalable ETL framework using Sqoop, Pig, and Hive to efficiently extract, transform, and load data from various sources, ensuring seamless data availability for consumption.
* Processed data stored in Hadoop Distributed File System (HDFS), leveraging Hive to create external tables and developing reusable scripts for efficient table ingestion and repair across the project.
* Developed robust ETL jobs using Spark and Scala to migrate data from Oracle to new MySQL tables, ensuring smooth data transfer and maintaining data integrity.
* Leveraged the powerful capabilities of Spark, including RDDs, Data Frames, and Spark SQL, along with Spark-Cassandra Connector APIs, for diverse data tasks such as data migration and generating comprehensive business reports.
* Developed common Flink module for serializing and deserializing AVRO data by applying schema.
* Engineered a high-performance Spark Streaming application for real-time sales analytics, enabling timely insights and decision-making.
* Conducted comprehensive analysis of source data, effectively handled data type modifications, and utilized Excel sheets, flat files, and CSV files to generate on-demand Power BI reports.
* Analyzed SQL scripts and devised optimal solutions using PySpark, ensuring efficient data processing and transformation.
* Setup and build AWS infrastructure various resources, VPC EC2, S3, IAM, EBS, Security Group, Auto Scaling and RDS in Cloud Formation JSON templates.
* Created python scripts for migration of data from Oracle database to Postgres database.
* Created pl/pg sql functions to return data through ref cursor in Postgres database.
* Overloading of functions is performed in Postgres DB based on different number of input parameters.
* Leveraged Sqoop to efficiently extract data from multiple data sources into HDFS, facilitating seamless data integration.
* Orchestrated data imports from various sources, executed transformations using Hive and MapReduce, and loaded processed data into HDFS.
* Successfully extracted data from MySQL databases into HDFS using Sqoop, enabling seamless data transfer and integration.
* Implemented streamlined automation for deployments using YAML scripts, resulting in accelerated and efficient build and release processes.
* Expertly utilized Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka, and Sqoop, leveraging their capabilities to optimize data processing and management.
* Developed data classification algorithms using MapReduce design patterns, enhancing data processing efficiency and accuracy.
* Employed advanced techniques including combiners, partitioning, and distributed cache to optimize the performance of MapReduce jobs.
* Effectively utilized Git and GitHub repositories for comprehensive source code management and version control, fostering efficient collaboration and ensuring traceability of code changes.

**Environment:** Sqoop, Pig, HDFS, Power BI, GitHub, Cassandra, Apache Spark, Scala, Hive, Hadoop, Cloudera, HBase, MySQL, YAML, JIRA, Git, GitHub

## Yana Software Private Limited - Hyderabad, INDIA | ETL Developer Oct 2013 – Dec 2016

## Responsibilities:

* Data was collected from a variety of sources like databases, APIs and flat files and it was stored in a consistent format to ensure that it was clear and reliable.
* Databases are created, maintained, and optimized in order to store vast amounts of data in an effective manner.
* Built and managed ETL (Extract, Transform, Load) processes to move data between systems, ensuring data quality and consistency.
* Conducted comprehensive requirement analysis to identify data extraction needs from various source systems, including Netezza, DB2, Oracle, and flat files, for seamless integration into the Salesforce application.
* Designed and developed robust ETL processes using Informatica Power Center to efficiently extract data from diverse sources and load it into the target data warehouse.
* Implemented advanced performance tuning techniques to optimize data mappings and address bottlenecks in the data transfer process, resulting in improved efficiency and faster data processing.
* Utilized Informatica Power Center Tools, such as Designer, Workflow Manager, Workflow Monitor, and Repository Manager, to streamline development, monitoring, and management of ETL workflows, ensuring smooth execution and enhanced productivity.
* Created intricate data mappings from scratch, leveraging a wide range of Informatica Designer Tools, including Source Qualifier, Aggregate, Lookup, Expression, Normalizer, Filter, Router, Rank, Sequence Generator, Update Strategy, and Joiner transformations, to ensure accurate data transformation and seamless integration.
* Implemented efficient Incremental Loading mappings using Mapping Variables and Parameter Files, enabling incremental data transfer, and optimizing the overall ETL process for efficient data synchronization.
* Developed reusable Transformations and Mapplets to promote code reusability, reduce development efforts, and enhance the maintainability of the ETL workflows.
* Identified and resolved performance bottlenecks by leveraging the capabilities of the Netezza Database, optimizing Index Cache and Data Cache, and utilizing Rank, Lookup, Joiner, and Aggregator transformations for efficient data processing.
* Created and executed Netezza SQL scripts to ensure accurate table loading, and developed SQL scripts for validating row counts and verifying data integrity, ensuring data accuracy and reliability.
* Designed and implemented data pipelines to automate the flow of data from source systems to data storage solutions.
* Ensured that datasets were of high quality by performing data cleaning duties to eliminate inaccuracies and inconsistencies.
* Worked closely with data scientists, analysts, and other stakeholders to understand data requirements and provide the necessary data infrastructure.
* Query structures, data pipelines, and SQL queries have been tuned for the best performance and effectiveness.
* Documentation of data processes, data flows, and data structures was kept in a clear and thorough manner in order to provide transparency and ease of maintenance.
* Basic data analysis was carried out, and findings were reported, in order to provide support for decision-making processes.
* Continuously learned new data technologies, tools, and best practices to keep up with the evolving data landscape.

**Environment**: Python, Shell scripting (bash), Apache sqoop, Apache hive, MySQL, Informatica, ETL, Oracle, Microsoft Excel.