**Aravind Ravella**

**Sr. AWS Data Engineer | Sr. Snowflake Data Engineer | Sr. ETL Developer**

**Contact no: 214-814-5028**

[**ravellaaravind14@gmail.com**](mailto:ravellaaravind14@gmail.com)

[**www.linkedin.com/in/aravind-ravella**](http://www.linkedin.com/in/aravind-ravella)

**PROFESSIONAL SUMMARY:**

* Extensive experience migrating on-premises data infrastructure to AWS Cloud and Snowflake, achieving significant improvements in data accessibility, query performance, and cost savings
* Expertise in designing and implementing petabyte-scale data lakes on AWS using services like S3, Glue, and Athena, and integrating with Snowflake for enhanced analytics and reporting capabilities
* Proficient in utilizing AWS services such as Lambda, Step Functions, and Glue for serverless data integration and application development
* Skilled in developing and optimizing complex SQL queries, Scala scripts, and Python ETL pipelines for efficient data manipulation, optimization, and performance tuning.
* Proficient in NoSQL (MongoDB, Cassandra, DynamoDB, Couchbase), SCD (Type 1, 2 & 3), CDC, and OLAP for efficient data storage, historical tracking, real-time updates, and advanced analytics across healthcare, insurance, e-commerce, and financial services.
* Experience in implementing real-time data streaming solutions using Amazon Kinesis, Kafka, and Snowpipe(Streaming API ) for processing high-volume data streams and delivering timely insights
* Expertise in Snowflake's cloud data platform, including data loading, transformation, and querying using Snowflake SQL.
* Familiarity with Snowflake's features and best practices such as virtual warehouses, data sharing, and data governance capabilities.
* Proficient in configuring Snowpipe for near real-time data ingestion, including setting up stages, file formats, and auto-ingest.
* Experience in performance tuning and scaling Snowflake warehouses to handle high-volume real-time data workloads efficiently
* Automating data workflows using Airflow, creating reusable DAGs for scheduling and monitoring data pipelines
* Proven ability to optimize data storage and processing using partitioning, compression, and vacuum strategies on Redshift and AWS EMR, resulting in significant cost savings and improved query performance
* Proficient in handling semi-structured data using Snowflake's VARIANT data type and implementing data processing techniques such as scheduling tasks, creating conjugate tasks, and monitoring table changes using streams
* Implemented SQL best practices for data warehousing in Snowflake, leveraging materialized views, stored procedures, and user-defined functions to streamline data processing and enhance query performance
* Expertise in leveraging AWS Glue for complex data transformations, utilizing PySpark to develop efficient ETL scripts that handle data cleansing, normalization, and aggregation, to reduce data transformation runtime.
* Skilled in designing and implementing data pipelines on AWS using services like S3, EMR, and Glue in conjunction with PySpark, ensuring data quality and integrity throughout the pipeline
* Experience in migrating data processing workflows from traditional frameworks to PySpark on AWS, leveraging the scalability and performance benefits of Spark within the AWS ecosystem
* Expertise in optimizing PySpark performance on AWS EMR through in-depth analysis, tuning cluster configurations, and implementing best practices for data partitioning and caching
* Proficient in designing and implementing scalable data architectures on Unix/Linux servers, leveraging Bash scripting to automate data ingestion, transformation, and loading processes

Experience in developing and maintaining robust data pipelines using Scala and Spark, processing over 10 TB of data daily, and implementing DevOps practices with Git and Jenkins for version control and CI/CD

* Proven ability to collaborate with cross-functional teams using agile framework, including data scientists, business analysts, and stakeholders, to gather requirements and deliver high-quality data solutions that meet business requirements
* Proven leadership skills in mentoring and guiding junior data engineers, fostering a culture of continuous learning and development within the team

**KEY SKILLS:**

**Data Engineering**: Architecting and managing end-to-end data solutions, including robust data pipelines, efficient data warehousing, and scalable data lakes

**ETL and ELT**: Proficient in both ETL and ELT methodologies, with a strong background in data modeling to support complex analytics

**Cloud Platforms**: Deep knowledge of AWS cloud services (S3, EC2, DynamoDB, SNS, SQS, CloudWatch, RDS, Lambda, Redshift, EMR, Glue, Kinesis, Athena, Quick sight) and Snowflake(Time travel, SnowSQL, SnowPipe, SnowPark, Zero Copy Cloning, Virtual Warehouses, Tables and Views, Stages, Snowpipe, Tasks, Streams) for cloud-based data warehousing

**Big Data Technologies**: Extensive experience with Apache Spark, Hadoop, Hive, Kafka, and Airflow for large-scale data processing, storage, and real-time data streaming

**Programming Languages**: Advanced proficiency in Python for data processing and automation, Advanced SQL for database querying and manipulation, and Scala for big data processing

**Databases**: Skilled in managing and optimizing various databases, including PostgreSQL, MySQL, Microsoft SQL Server, DynamoDB, and Snowflake

**Data Visualization**: Proficient in Tableau and AWS Quicksight for creating interactive dashboards and utilizing Snowflake's data-sharing capabilities

**DevOps and Version Control**: Applied DevOps practices and tools like Git, Github, Bitbucket, and Jenkins to enhance continuous integration and deployment (CI/CD) of data pipelines

**Automation**: Utilized Python and PySpark for advanced ETL processes and AWS Glue for serverless ETL, optimizing data operations, DBT for building data models

**Methodologies:** Actively engaged in Agile Methodologies (Scrum, Kanban) and Waterfall Methodologies, as well as the Software Development Life Cycle (SDLC), to consistently deliver high-quality products.

**PROFESSIONAL EXPERIENCE:**

**Icario Health   
Minneapolis, MN November 2023 – Present**

**Senior Data Engineer**

* **Enhanced Data Processing Efficiency**: Managed intricate ETL systems utilizing Python, SQL, and AWS services such as Step Functions, EMR, Lambda, S3, and Redshift, achieving a 30% increase in data processing efficiency.
* **Rapid Issue Resolution**: Engineered comprehensive platform issue resolutions using MongoDB, AWS Redshift, and Sumo Logic; ensuring stable system performance slashed downtime by 40%, and fortified system dependability.
* **Executed Scalable Data Warehousing Solution**: Engineered a high-performance data warehousing solution using AWS Glue and Step Functions, enhancing data storage and retrieval processes resulting in a 25% decrease in data processing time.
* **Managed Complex Data Migration Projects**: Led the successful execution of data migration projects, ensuring smooth integration of new data sources with minimal disruption; expanded data analytics capabilities by 60%, enabling enhanced insights.
* **Utilized MongoDB and implemented SCD Type 2** **in Snowflake** for personalized healthcare communications and historical tracking and **leveraged CDC and transformed OLTP data into OLAP format** for real-time updates and efficient data analysis
* **Implemented Airflow for Workflow Orchestration:** Utilized Airflow to orchestrate complex data workflows, ensuring efficient execution and monitoring of data pipelines. Developed custom Airflow operators using Python to handle specific data processing tasks.
* **Developed and optimized complex SQL queries and Scala scripts** for data transformation and analysis, collaborating with data scientists and business stakeholders in an Agile environment to deliver actionable insights and drive data-driven strategies.
* **Implemented CI/CD pipelines using GitHub and Jenkins**, automating the build, test, and deployment processes for data pipeline code, and ensuring code quality and reliability through rigorous testing and code reviews.
* **Utilized Time Travel in Snowflake for accessing historical data**, enhancing query performance and data analysis capabilities, and applying performance optimization techniques in Snowflake, such as clustering.
* **Drove Personalized Healthcare Communications**: Led a cross-functional effort with Data Scientists and Business Intelligence Engineers to develop data-driven strategies for personalized healthcare communications and achieved a 30% boost in actionable data insights and targeted outreach.
* **Automated ETL Processes**: Implemented advanced SQL, Python, and Scala programming skills to automate ETL processes, reducing data processing time by 30% enabled the delivery of timely insights for healthcare engagement strategies.
* **Architected and implemented scalable data integration frameworks** using Python and AWS services, connecting multiple data sources, including REST APIs, to create unified data platforms for clients, enabling efficient data processing and analysis across various business domains

**State Farm   
Dallas, TX July 2022 - October 2023**

**Senior Data Engineer**

* **Led Comprehensive Data Infrastructure Migration:** Migrated on-premises data infrastructure to AWS Cloud and 10 TB on-premises data warehouse to Snowflake, achieving a 40% improvement in data accessibility and a 30% increase in query performance.
* **Architected a Petabyte-Scale Data Lake on AWS:** Designed and implemented a petabyte-scale data lake using AWS S3, Glue, and Athena, reducing data processing time by 60%, and integrated Snowflake to enhance data analytics and reporting capabilities.
* **Designed and implemented secure, scalable storage solutions using Amazon S3,** optimizing data retrieval and reducing costs by implementing lifecycle policies and intelligent tiering.
* **Managed petabyte-scale data warehousing using Amazon Redshift**, achieving faster query performance through columnar storage and data compression techniques.
* **Developed serverless applications using AWS Lambda**, significantly reducing infrastructure management overhead, and achieving high availability.
* **Utilized AWS Glue for serverless data integration**, creating and automating ETL jobs that prepare and combine data for analytics and **utilized dbt (Data Build Tool) to develop and maintain data transformation models** within the Snowflake
* **Streamlined Data Ingestion Pipelines**: Transitioned from batch loading to continuous, real-time data ingestion with Snowpipe, improving data freshness and accessibility for business analysis
* **Implemented real-time data streaming using Snowpipe's Streaming API**, enhancing data ingestion and reducing latency by directly streaming data into Snowflake, eliminating intermediate storage.
* **Worked closely with architects to design scalable and fault-tolerant data streaming architectures using Snowpipe,** ensuring high data availability and consistency across Snowflake clusters.
* **Configured Snowpipe Streaming with exactly once delivery and dead-letter queue (DLQ)** support to maintain high data integrity and reliability in streaming processes
* **Integrated Kafka with Snowflake using Snowflake's Kafka Connector and Snowpipe Streaming**, streamlining data ingestion and reducing latency.
* **Configured Snowpipe error notifications with cloud messaging services like AWS SN**S to promptly address any ingestion failures and maintain data pipeline integrity
* **Handled semi-structured data using Snowflake's** VARIANT data type, enabling flexible data processing and analysis, and Implemented data processing techniques in Snowflake, including scheduling tasks, creating conjugate tasks, and monitoring table changes using streams.
* **POC on Snowflake’s Snowpark to develop** a PySpark-based ETL pipeline for optimizing State Farm's claims data processing and **Organized data using Snowflake's tables** (permanent, transient, temp, and external), object hierarchy, including databases, schemas, enhancing data structure and accessibility.
* **Implemented secure data sharing with other accounts** and non-Snowflake customers using reader accounts, ensuring data privacy and compliance
* **Designed NoSQL solution (DynamoDB) and implemented SCD Type 1 and 2 Utilized CDC (AWS DMS) and created OLAP cubes in Snowflake**
* **Collaborated with cross-functional teams to migrate existing data streaming processes to Snowpipe**, ensuring a smooth transition and minimal disruption to business operations.
* **Optimized Snowpipe performance by implementing best practices, such as data partitioning**, file sizing, and error handling, to ensure reliable and efficient data ingestion into Snowflake.
* **Implemented SQL Best Practices for Data Warehousing**: Designed and implemented SQL best practices for the Snowflake data warehouse, SQL features such as materialized views, stored procedures, and user-defined functions to streamline data processing and enhance query performance such as query optimization, database indexing, and query parallelization.
* **Developed and maintained Snowpipe data ingestion pipelines** to stream data from AWS S3 buckets into Snowflake tables in near real-time, ensuring data availability for critical business analytics and reporting.
* **Implemented Monitoring and Automation**: Implemented and managed agents for monitoring, logging, and automation within AWS environments. Utilized tools such as CloudWatch and Lambda to monitor data pipelines, set up alerts, and automate routine tasks, enhancing the reliability and efficiency of the data platform.
* **Implemented Python ETL Pipelines with:** Developed Python-based ETL pipelines leveraging Airflow for orchestration, enabling efficient data processing and workflow management. Utilized Python libraries such as pandas and NumPy for data manipulation and transformation tasks.
* **Developed and maintained data pipelines to integrate data from various sources, including REST APIs**, databases, and flat files, using Python, Spark, and AWS services like Glue and Kinesis, ensuring seamless data integration for multiple clients.

**State of Iowa Department of Health and Human Services(DHHS) February 2021 - June 2022**

**DesMoines, IA**

**Senior Data Engineer**

* **Migrated a legacy on-premises data warehouse to AWS Redshift**, reducing infrastructure costs by $200K annually involved extensive planning, data migration strategies, and the implementation of AWS services for a seamless transition to the cloud.
* **Developed AWS Lambda functions** for serverless ETL processing of streaming data from Kinesis Firehose, to leverage AWS for efficient, scalable data processing solutions.
* **Implemented data governance policies using AWS Glue Data Catalog and Athena** for secure data access, demonstrating my expertise in managing data access and security within the AWS ecosystem. Automated Data Workflows with Airflow: Implemented Airflow to automate and manage data workflows, creating DAGs for scheduling and monitoring data pipelines. Utilized Python operators and hooks to interact with AWS services and Snowflake within Airflow workflows.
* **Deployed Snowflake infrastructure** for Iowa DHHS, demonstrating 5x faster query performance on Medicaid claims and recipient data, enabling efficient fraud detection and improved public health reporting.
* **Spearheaded the integration of Snowflake within AWS environments**, leveraging Snowflake's cloud data platform to enhance data analytics and reporting capabilities including setting up Snowflake environments, migrating data, and optimizing Snowflake configurations for performance and cost-efficiency.
* **Created and managed standard and materialized views to improve query performance** and facilitate data analysis within the Snowflake environment.
* **Developed ETL pipelines to load and transform large datasets into Redshift**, utilizing its massively parallel processing (MPP) architecture for efficient data analysis.
* **Developed Lambda functions** in response to AWS events, such as S3 uploads and DynamoDB updates, for real-time data processing and workflow automation.
* **Managed AWS Glue Data Catalog**, cataloging data assets and enabling a unified metadata repository across various data stores.
* **Employed NoSQL (Cassandra) and implemented SCD Type 2 Utilized AWS DMS for CDC and designed OLAP cubes in Redshift**
* **Leveraged DBT (Data Build Tool) to manage data transformations and build reusable data models.**
* **Configured Kinesis Data Firehose** for data delivery to S3 and Redshift, enabling continuous capture, transformation, and loading of streaming data.
* **Integrated Step Functions with Lambda** and other AWS services to build resilient, stateful applications that execute in response to events and conditions.
* **Developed PySpark Solutions on AWS**: Utilized PySpark and AWS services to develop scalable and efficient data processing solutions. Leveraged PySpark's distributed computing capabilities to handle large-scale data workloads and optimize performance on AWS infrastructure.
* **Managed Data Pipelines on AWS:** Designed, implemented, and managed data pipelines using PySpark and AWS services such as S3, EMR, and Glue. Ensured efficient data ingestion, transformation, and loading processes while maintaining data integrity and optimizing performance. Monitored and troubleshot pipeline issues to minimize data processing disruptions.
* **Optimized SQL Queries for Performance**: Analyzed and optimized SQL queries to improve data retrieval and processing performance. Implemented query optimization techniques such as indexing, partitioning, and query restructuring, resulting in a 40% reduction in query execution time and enhanced system efficiency.
* **Collaborated on the design and implementation of data pipelines on AWS** using S3, Glue, and Redshift, integrated with Snowflake for a 360-degree customer view across multiple sources improved data availability and insights, supporting strategic decision-making.
* **Implemented AWS solutions architecture and directly managed** 2 junior Data Engineers, sharing knowledge on AWS best practices and Snowflake integration, fostering a culture of learning and innovation within the team.
* **Designed and implemented a scalable data architecture on Unix/Linux servers,** leveraging Bash scripting to automate data ingestion, transformation, and loading processes, resulting in a 30% improvement in data processing efficiency.
* **Utilized Scala to build a real-time data streaming pipeline**, processing millions of events per second, and employed DevOps practices with Docker and Kubernetes for containerization and orchestration, ensuring high availability and fault tolerance.

**Change Health Care**

**Nashville, TN April 2018 - January 2021**

**Data Engineer**

* **Architected Fault-Tolerant Data Pipeline Architecture on AWS**: Designed and implemented a highly available and fault-tolerant data pipeline architecture on AWS, leveraging services like S3, Kinesis, Lambda, and Glue, infrastructure ingested, processed, and stored over 1 TB of daily data with 99.9% to create reliable and scalable data solutions.
* **Optimized Big Data Processing:** Developed scalable data pipelines using Apache Spark, Kafka, and Hadoop to process over 5 TB+ of daily data from 100+ sources, improving efficiency by 40% and enhancing data processing capabilities.
* **Optimized PySpark Performance on AWS EMR**: Conducted in-depth performance analysis and optimization of PySpark jobs running on AWS EMR clusters. Tuned cluster configurations, optimized resource allocation, and implemented best practices for data partitioning and caching, resulting in a 40% reduction in job execution times and a 30% cost savings on EMR usage.
* **Integrated S3 with other AWS services for log storage**, website hosting, and data lake solutions, leveraging S3's versioning and event notifications for real-time processing.
* **Tuned Redshift clusters for optimal performance** by analyzing query execution plans and adjusting distribution keys and sort keys to minimize query runtimes.
* **Optimized Lambda functions for performance** by fine-tuning memory allocation and execution timeouts, resulting in lower execution costs and improved response times
* **Enhanced data quality and reduced data preparation time by using Glue's data transformation** capabilities and built-in connectors for various data sources.
* **Monitored and tuned Kinesis streams** for throughput and performance, ensuring smooth data flow and efficient processing using Kinesis Data Analytics.
* **Visualized and debugged application workflows using Step Functions**' graphical console, streamlining development and troubleshooting.
* **Built complex data models using star and snowflake schemas**, and developed ETL workflows in Python and SQL to establish a centralized data warehouse, enabling self-service analytics for business stakeholders and facilitating a unified view of data for better decision-making
* **Designed and implemented a real-time monitoring** and alerting system using Grafana and Prometheus on Unix/Linux servers, enabling proactive identification and resolution of data pipeline issues, and reducing downtime by 50%.
* **Utilized Scala and Apache Flink to build a stream processing** application for fraud detection, leveraging machine learning models and DevOps practices with AWS CloudFormation for infrastructure as code, enabling rapid deployment and scalability.
* **Built Real-time Data Processing with PySpark Streaming**: Implemented real-time data processing pipelines using PySpark Streaming and Kafka to ingest and analyze healthcare data streams. Developed stateful stream processing logic to identify anomalies, and trigger alerts, and enable real-time decision-making, improving patient care and operational efficiency.
* **Using system and custom roles within the Snowflake environment**, ensuring proper data governance and security.
* **Automated Data Pipelines with Airflow**: Implemented Airflow for orchestrating and scheduling data pipelines, creating DAGs to manage complex data workflows. Utilized Python operators and hooks to integrate with big data technologies like Spark and Hadoop within Airflow workflows.
* **Collaborated with cross-functional teams to design and implement data integration solutions, leveraging REST APIs** and diverse data sources, to meet client requirements and enable comprehensive data analytics across different projects.

**Macy’s   
New York, NY February 2016 - March 2018**

**Data Warehouse Developer**

* **Optimized BI Data Model**: Designed and developed a Business Intelligence (BI) Physical Data Model, optimizing data structures for efficient querying and reporting streamlined data access, and improved the performance of reporting tools, enhancing the overall analytics capabilities of the organization.
* **Advanced SQL Query Development**: Orchestrated the creation of efficient ETL processes through the development of intricate SQL queries and stored procedures in Oracle SQL and PL/SQL, resulting in a 25% reduction in data processing errors.
* **Performance Optimization**: Optimized BI report performance by 40% through strategic utilization of Oracle Data Integrator (ODI), accelerating report generation times and enhancing user experience for 50+ business analysts and stakeholders.
* **Developed Python Scripts for Data Processing**: Created Python scripts to automate data processing tasks, improving efficiency, and reducing manual effort. Leveraged Python libraries such as pandas and NumPy for data manipulation and transformation.
* **Stakeholder Collaboration**: Collaborated with stakeholders to ensure deliverables met or exceeded expectations, providing technical support, and troubleshooting as necessary to deliver high-quality results.
* **Developed and maintained ETL scripts using Python and Shell scripting** on Unix/Linux servers, automating data extraction, transformation, and loading processes, and collaborating with data engineers and analysts in an Agile environment to ensure data quality and consistency.
* **Utilized NoSQL (MongoDB) and implemented SCD Type 2** in Oracle data warehouse for e-commerce analytics and sales data and **leveraged Oracle Golden Gate for CDC and designed OLAP cubes** in Oracle OLAP for real-time data analysis and marketing analytics.
* **Implemented a CI/CD pipeline**, automating the build, test, and deployment processes for data pipeline code, and leveraging Docker containers for consistent and reproducible deployments across development, staging, and production environments.

**Experian   
CostaMesa, CA February 2014 - January 2016**

**Data Warehouse Developer**

* **Streamlined Data ETL Processes**: Spearheaded the design and implementation of SSIS packages for efficient data extraction, transformation, and loading processes significantly enhanced the data workflow efficiency, showcasing my ability to optimize data operations for large-scale systems.
* **Enhanced Data Warehouse Integrity**: Designed and maintained robust data warehouse models, ensuring high data integrity and performance data modeling and maintenance supporting critical business analytics and decision-making processes.
* **Automated ETL Processes with Python**: Developed Python scripts to automate ETL processes, streamlining data extraction, transformation, and loading tasks. Utilized Python's extensive libraries and frameworks to enhance data processing capabilities.
* **Advanced Data Reporting:** Developed complex reports using SQL Server Reporting Services (SSRS), leveraging SSRS functionalities for effective data visualization and presentation, enabling stakeholders to gain insightful and actionable data insights, and enhancing the organization’s data-driven decision-making capabilities.
* **Optimized Database Performance**: Directed the strategic utilization of stored procedures, triggers, views, and functions to enhance database performance, leading to a 25% increase in system efficiency and reliability, supporting the organization’s operational and analytical requirements.

**EDUCATION DETAILS:**

Masters in Computer Science University of Dayton 2013

Bachelors in Computer Science Veltech University India 2011

**CERTIFICATIONS:**

AWS Certified Data Engineer - Associate

SnowPro® Core Certification