Swathi Eerala
Swathieerala31@gmail.com
(479) 265-3228

www.linkedin.com/in/
Senior Data Engineer

**Summary:**

* Over 10+ years of experience as a Data Engineer, Data Analyst, Data Integrating, Big Data, Data Modelling - Logical and Physical, and Implementation of Business Applications using the Oracle Relational Database Management System RDBMS.
* Development and Implementation of various applications including Web and Client/Server Application development using Microsoft .NET Framework, C#.NET, ASP.NET, AJAX.net.
* Experience as Analysis Services Cube Developer (OLAP) and Report Developer (SSRS) Excellent knowledge of creating Matrix Reports, List Reports, Chart Reports and Drill-down Reports.  Experience in SSAS all phases of data warehouse development lifecycle, from gathering requirements to testing, implementation, and support.
* Strong experience working with Oracle, SQL, SQL Loader and open Interface to analyze, design, develop, test, and implement database applications using Client / Server applications.
* Hands on experience in Data Modelling, Data Mapping, Data Profiling, Data Analytics, Data Mining, Data Integration, Data Cleansing, Data Loading, Data Transformation, Data Migration and Data Validation.
* Extensive experience in designing the Data models for OLTP & OLAP database system. Strong Data Modeling experience using ER diagram, Dimensional data modeling, Star Schema modeling, Snow-flake modeling using tools like Erwin, EMBARCADERO translating business requirements into technical designs and development of the physical aspects of a specified design.
* Knowledge in database conversion from Oracle and SQL Server to PostgreSQL and MySQL.
* Worked on projects that involved Client/Server Technology, customer implementation involving GUI Design, Relational Database Management Systems RDBMS, and Rapid Application Development Methodology. Practical knowledge in PL/SQL for creating stored procedures, clusters, packages, database triggers, exception handlers, cursors, cursor variables.
* Worked on ASP.NET AJAX Extensions and Toolkit Controls.
* Strong knowledge and experience on object oriented programming (OOPS). Strong programming experience with web technologies like HTML JavaScript, ASP.NET, Web services
* Experience in using Snowflake Clone and Time Travel.
* Experience in various data ingestion patterns to hadoop.
* Participates in the development improvement and maintenance of snowflake database applications
* Experience in various methodologies like Waterfall and Agile.
* Understanding and analysis of Monitoring/Auditing tools to gain in depth knowledge in AWS such as Cloud Watch and Cloud Trail.
* In-depth familiarity with AWS DNS Services via Route53. Understanding the several sorts of routes: simple, weighted, latency, failover, and geo location.
* Effective Algorithm expertise with Hadoop ecosystem components such as Hadoop Map-Reduce, HDFS, HBase, Hive, Sqoop, Pig, Zookeeper, Hortonworks, and Flume, as well as installing, configuring, monitoring, and using them.
* Amazon EMR, Spark, Kinesis, S3, Boto3, Bean Stalk, ECS, Cloud Watch, Lambda, ELB, VPC, Elastic Cache, Dynamo DB, Redshift, RDS, Aena, Zeppelin, and Airflow professionals.
* Handling, organizing, and operating databases such as MySQL and NoSQL databases such as MongoDB and Cassandra.
* Sound knowledge of AWS cloud building templates and how to transmit information using the SQS service via Java API.
* AWS **AEM Developer** experience generating separate virtual data warehouses with differently sized classes
* Worked on Teiid and Spark Data Virtualization, RDF graph Data, Solr Search, and Fuzzy Algorithm.
* Thorough understanding of MPP databases, wherein data is partitioned across multiple servers or nodes, with each server/node having memory and processors to interpret data locally.
* Data simulation, database development, and OLTP (Star Schema, Snowflake Schema, Data Warehouse, Data Marts, Multi-Dimensional Modelling, and Cube design) for OLTP, OLAP (Star Schema, Snowflake Schema, Data Warehouse, Data Marts, Multi-Dimensional Modelling, and Cube design), Business Intelligence, and data mining.
* For data analysis and pattern classification, I used SQL, NumPy, Pandas, Scikit-learn, Spark, and Hive extensively.
* Established and sustained several Existing BI dashboards, reports, and content packs.
* Customized POWER BI Visualizations and Dashboards in line with the client's needs
* Working expertise of Amazon Web Services databases such as RDS (Aurora), Redshift, Dynamo DB, and Elastic Cache (Memcached & Redis)
* Productizing and constructing a Data Lake employing Hadoop and its ecosystem components.
* Long Working hours on real-time data streaming solutions using Apache Spark/Spark Streaming and Kafka, as well as developing Spark Data Frames in OLTP.
* Shifting Cultivation an API to manage servers and run code in AWS using Amazon Lambda.
* Have rich experience programming python scripts to implement the workflow and have experience with ETL workflow management technologies such as Apache Airflow.
* Adequate knowledge of databases such as MongoDB, MySQL, and Cassandra.
* For performance tuning and database optimization, a working grasp of SQL Trace, TK-Prof, Explain Plan, and SQL Loader is needed.
* Provide asynchronous replication, including Amazon EC2 and RDS, for regional MySQL database deployments and fault-tolerant servers (with solutions tailored for managing RDS).
* Extensive knowledge of Dynamic SQL, Records, Arrays, and Exception Handling, as well as data sharing, data caching, and data pipelines. Nested Arrays and Collections are being used to do complex processing.
* Hands on experience in integrating databases such as MongoDB and MySQL with webpages such as HTML, PHP, and CSS to update, insert, delete, and retrieve data using simple ad-hoc queries.
* Expertise in Extraction, Transformation, and Loading (ETL) processes, including UNIX shell scripting, SQL, PL/ SQL, and SQL Loader
* For Distributed Processing, I created both Spark RDD and Spark Data Frame APIs.

**TECHNICAL SKILLS:**

**Hadoop Technologies:** HDFS, Map Reduce, YARN, Hive, Pig, HBase, Impala, Zookeeper, Sqoop, OOZIE,

Apache Cassandra, Flume, Spark, AWS, EC2

**Java Technologies:** J2EE, JSP, JSTL, EJB, JDBC, JMS, JNDI, JAXB, JAX-WS, JAX-RPC, SOAP,
WSDL

**Web Technologies:** HTML, CSS, JavaScript, AJAX, Servlets, JSP, DOM, XML, XSLT.

**Languages:** C, Java, SQL, PL/SQL, Scala, Shell Scripts

**Operating Systems:** Linux, UNIX, Windows

**Databases:** NoSQL, Oracle, DB2, MySQL, SQL Server, MS Access, HBase.

**Application Servers:** WebLogic, WebSphere, Apache Tomcat, JBoss

**IDEs:**  Eclipse, NetBeans JDeveloper, IntelliJ IDEA.

**Version Control:** CVS, SVN, Git

**Reporting Tools:**  Jasper soft, Qlik Sense, Tableau, JUnit

**PROFESSIONAL EXPERIENCE**

**Change Healthcare, Nashville, TN Apr 2021 to Present**

**Sr. Data Engineer**

**Responsibilities:**

* Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
* Managed both Jira, and Confluence installations, including upgrades and adding user features.
* Optimized current infrastructure and implemented changes to increase uptime and service
* efficiency.
* Implemented SOA architecture with Windows Services, web services using SOAP, MSMQ, WSDL, UDDI, and XML. OLTP Protocol.
* Developed Data Access Layer( ADO.NET) in C# 3.5 to do all the database related operations like apply changes of the strongly typed dataset, fill strongly typed dataset.
* Used data controls like Data Grid, Telerik AJAX, Data List, Data caching,Data Binding and Repeater controls throughout the application to display data in a customized format in the ASP.NET 3.5 web pages.
* Implemented Auto-scaling and created a process to auto-scale front end infrastructure to adjust on
* traffic increases.
* CI/CD tools and configuration management tools like CHEF
* Identified and informed development teams to remove single points of failure within applications
* Strong experience in developing jobs using different stages in DataStage like link collector, join, merge, lookup, remove duplicates, XML Stage, filter, dataset, transformer, aggregator
* Proficient with all major PostgreSQL procedural languages (PL/PgSQL, PL/Perl, PL/PgPython, PL/Tcl) as well as some Oracle PL/SQL and SQL - Server T-SQL.  Excellent SQL skills including query optimization, complex nested and co-related queries, common table expressions, window functions, and business analytical reporting
* Created new DataStage Jobs and sequences for data validation to detect inconsistency on the data loaded by existence ETL processusing different stages in DataStage like link collector, join, merge, lookup, remove duplicates, filter, dataset, transformer, aggregator
* Working knowledge in Core Java and JavaEE platform with Servlets, JSP, JDBC, Multithreading, Hibernate, Spring MVC, Spring Boot
* Ensured that Business Requirements can be translated into Data Requirements. Created Business Requirement documents (BRD’s), such as SRS & FRS and integrated the requirements and underlying platform functionality.
* Partnered with ETL developers to ensure that data is well cleaned and the data warehouse is up-to-date for reporting purpose by Pig.
* Working knowledge in Core Java and JavaEE platform with Servlets, JSP, JDBC, Multithreading, Hibernate, Spring MVC, Spring Boot
* Selected and generated data into CSV files and stored them into AWS S3 by using AWS EC2 and then structured and stored in AWS Redshift.
* Good experience in writing Spark applications using Python and Scala.
* Clear knowledge of Python and Scala for creating Spark applications.
* Processed some simple statistical analysis of data profiling like cancel rate, var, skew, kurt of trades, and runs of each stock every day group by 1 min, 5 min, and 15 min.
* Used Pyspark and Pandas to calculate the moving average and RSI score of the stocks and generated them into data warehouse.
* Oracle can be used to integrate data from multiple sources and make it available for analysis and reporting.
* Exploring with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, PostgreSQL, Data Frame, Open Shift, Talend, pair RDD's
* Involved in integration of Hadoop cluster with spark engine to perform BATCH and GRAPHX operations.
* Performed data preprocessing and feature engineering for further predictive analytics using Python Pandas.
* Developed and validated machine learning models including Ridge and Lasso regression for predicting total amount of trade.
* Working experience with Functional programming languages like Scala, and Java.
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Developed highly complex Python and Scala code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
* Hands-on use of Spark and Scala APIs to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Developing architecture to move the project from Abinitio to Pyspark and Scala spark.
* Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
* Experience in designing and developing POCs in Spark using Scala to compare the performance of Spark with Hive and SQL/Oracle.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala
* Developed predictive analytics using Apache Spark Scala APIs.
* Utilized Agile and Scrum methodology for team and project management.
* Used Git for version control with colleagues.

**Environment:** Spark (Pyspark, SparkSQL, Spark MLLib), Python (Scikit-learn, NumPy, Pandas), Tableau, GitHub, AWS EMR/EC2/S3/Redshift, Pig, and Oracle.

**Global Atlantic Financial Group, Indianapolis, IN Jul 2018 to Mar 2021**

**Big Data/Hadoop Developer**

**Responsibilities:**

* Developed an ETL pipeline to source these datasets and transmit calculated ratio data from AWS to DataMart (SQL Server) and Credit Edge.
* Team leader with large-scale, widely distributed database systems, including relational (Oracle, SQL server) and NoSQL (MongoDB, Cassandra) databases.
* Designed and implemented in all scenarios through configuring Topics in a new Kafka cluster.
* Developing and maintaining best practices and standards for data pipelining and Snowflake data warehouse integration.
* Extensively used Metadata & Data Dictionary Management; Data Profiling; Data Mapping. Applied Data Governance rules (primary qualifier, class words and valid abbreviation in Table name and Column names)
* Maintained Data Model and synchronized it with the changes to the database. Involved with all the phases of Software Development Life Cycle (SDLC) methodologies throughout the project life cycle
* Oracle can be used to process and analyze large volumes of data from a variety of sources, such as Hadoop, NoSQL databases, and cloud data sources.
* Streamlined the speed of both External and Managed HIVE tables.
* Worked mostly on requirements and technical phase of the Streaming Lambda Architecture, that uses Spark and Kafka to provide real-time streaming.
* Created and developed a system that uses Kafka to collect data from multiple portals and then processes it using Spark.
* Oracle can be used to perform real-time analytics on streaming data, such as IoT data, social media data, and financial data.
* Responsible for building scalable distributed data solutions using EMR cluster environment with Amazon EMR.
* Developed Spark scripts by writing custom RDDs in Scala for data transformations and perform actions on RDDs.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala.
* Designed and developed Data Access layer, Business layer and Presentation Layer (user interface) for the n-tier architecture web application using C#, ADO. NET and ASP. NET.
* Loaded information from the data warehouse and other systems such as SQL Server and DB2 using ETL tools such as SQL loader and external tables.
* Using a REST API, implemented Composite server for data isolation and generated multiple views for restricted data access.
* Developed Spark code using Scala and Spark-SQL for faster testing and data processing.
* Arranged Visual Studio to integrate with AWS, making it easier to write code in a pleasant setting.
* Employed Python's pandas and NumPy libraries to clean data, scale features, and engineer features, as well as Predictive Analytics to create models.
* Applied Apache Airflow and CRON scripts in the UNIX operating system to develop Python scripts to automate the ETL process.
* Building distributed data scalable using Hadoop.
* Use Python, Scala programming on a daily basis to perform transformations for applying business logic.
* Converting data load pipeline algorithms written in python and SQL to Scala spark and pyspark.
* Managed Amazon Web Services like EC2, S3, ELB, Auto-Scaling, Dynamo DB, and Elastic Search
* Using Hadoop stack technologies SQOOP and HIVE/HQL, implemented Data Lake to consolidate data from multiple source databases such as Exadata and Teradata.
* AWS SQS was used to transmit the processed data to the next working teams for processing.
* Design and install secure data pipelines into a Snowflake data warehouse from on premise and cloud data sources using the Redshift Database (ETL data pipelines from AWS Aurora - MySQL Engine to Redshift).
* Used Data Transformation Services to convert a SQL server database to MySQL.
* Developed complicated SQL queries that included joins, sub-inquiries, and nested queries.
* Loaded data into different schema tables using SQL loader and control files.
* Created Pyspark and SparkSQL code to process data in Apache Spark on Amazon EMR and conduct the required transformations depending on the STMs developed.
* The jars and input datasets were stored in S3 Bucket, and the processed output from the input data set was stored in Dynamo DB.
* Used the AWS data pipeline for data extraction, transformation, and loading from homogeneous and heterogeneous data sources, as well as the Python matplot toolkit to try unique graphs for corporate decision-making.
* Participate in the design and architecture of Master Data Management (MDM) and Data Lakes. Cloudera Hadoop is used to create Data Lake.
* Mainframes are known for their reliability, scalability, and performance, making them well-suited for transaction processing applications, such as banking and financial systems.
* Data intake is handled via Apache Kafka.
* On HDFS, Hive tables were built to store the Parquet-formatted data processed by Apache Spark on the Cloudera Hadoop Cluster.
* For modest data sets, AWS services like EC2 and S3 were used.
* Involved in Design, Development and Testing of Windows Forms, Web Forms and Web Services using IIS, ASP.NET, C#, XML Dataset and .NET Framework.
* Participated in Data Integration by defining information needs within and across functional domains of an enterprise database update and scripting/data migration using SQL Server Export Utility.
* Implemented and integrated several NOSQL databases such as HBase and Cassandra.
* Clear knowledge of Python and Scala for creating Spark applications.
* Knowledge of cloud versioning systems such as GitHub.
* Using Python, formulated and constructed automation test scripts.
* Good experience in writing Spark applications using Python and Scala.

**Environment:** Kafka, Spark, Hive, Scala, HBase, Snowflake, Pig, AWS, CI/CD, API, Data stage, SQS, Git, Oracle Database 11g, PowerBI, Oracle HTTP Server 11g, PostgreSQL, Windows 2007 Enterprise, RDBMS, Data Pipelining, NoSQL, MongoDB, Dynamo DB, Python, ETL, SDLC, Waterfall, Agile methodologies, SOX Compliance, Mainframes.

**Sams Club, Bentonville, AR Oct 2015 to Jun 2018**

**Azure Data Engineer**

**SQL Server**

**Responsibilities:**

* Created and executed Hadoop Ecosystem installation and document configuration scripts on Google Cloud Platform(GCP).
* Transformed batch data from several tables containing tens of thousands of records from SQL Server, MySQL, PostgreSQL, and CSV file datasets into data frames using Pyspark.
* Researched and downloaded jars for Spark-Avro programming.
* Hands on Experience in MS AZURE (Azure Data Factory, Azure Data Lake, Data Bricks)
* Experienced in designing and implementing Service Oriented Architecture underlined with Ingress and Egress using Azure Data Lake Store & Azure Data Factory by adding blobs to lakes for analytic results and so pull data from Azure Data Lake to the Blobs.
* Developed a Pyspark program that writes data frames to HDFS as Avro files.
* Utilized Spark's parallel processing capabilities to ingest data.
* Created and executed HQL scripts that create external tables in a raw layer database in Hive.
* Developed a Script that copies Avro formatted data from HDFS to External tables in raw layer.
* Created Pyspark code that uses Spark SQL to generate Data Frame from Avro formatted raw layer and writes them to data service layer internal tables as orc format.
* In charge of Pyspark code, creating data frames from tables in data service layer and writing them to a Hive data warehouse.
* ETL: Employ tools like Azure SQL Server, Azure Data Factory, and Data Bricks to create end-to-end data pipelines for collecting, cleansing, and processing client data.
* Mainframes are known for their ability to process large volumes of data in a batch mode, making them well-suited for applications such as payroll processing and billing systems.
* Experience in developing pipelines in Azure Data Factory using SQL Azure.
* Designed and developed a horizontally scalable APIs using Python Flask.
* Worked on migrating Map Reduce programs into Spark transformations using Scala.
* Used SCALA to store streaming data to HDFS and to implement Spark for faster processing of data.
* Installed Airflow and created a database in PostgreSQL to store metadata from Airflow.
* Configured documents which allow Airflow to communicate to its PostgreSQL database.
* Developed Airflow DAGs in AWS by importing the Airflow libraries.
* Experienced in working with spark ecosystem using Spark SQL and Scala queries on different formats like text file, CSV file.
* Oracle can be used to perform real-time analytics on streaming data, such as IoT data, social media data, and financial data.

**Environment:** Erwin, SQL, PL/SQL, Kafka, AWS, API's, Agile, ETL, HDFS, OLAP, HDFS, T-SQL, SSIS, Teradata, Hive, SSRS, Sqoop, Tableau, Map Reduce, XML, Azure, Oracle, Mainframes.

**Huda Infotech Private Limited, Hyderabad, India Jan 2014 to May 2015**

**Data Engineer**

**Responsibilities:**

* Reports based on SQL queries were created using Business Objects. Executive dashboard reports provide the most recent financial data from the company, broken down by business unit and product.
* Conducted data analysis and mapping, as well as database normalization, performance tuning, query optimization, data extraction, transfer, and loading ETL, and clean up.
* Developed reports, interactive drill charts, balanced scorecards, and dynamic Dashboards using Teradata RDBMS analysis with Business Objects.
* Gathering requirements, status reporting, developing various KPIs, and project deliverables are all responsibilities.
* In charge of maintaining a high-availability, high-performance, and scalability MongoDB environment.
* Created a NoSQL database in MongoDB using CRUD, Indexing, Replication, and Sharing.
* Assisting with the migration of the warehouse database from Oracle 9i to Oracle 10g.
* Worked on assessing and implementing new Oracle 10g features in existing Oracle 9i applications, such as DBMS SHEDULER create directory, data pump, and CONNECT BY ROOT.
* Improved report performance by rewriting SQL statements and utilizing Oracle's new built-in functions.
* Used Erwin extensively for data modelling and ERWIN's Dimensional Data Modeling.
* Tuning SQL queries with EXPLAIN PLAN and TKPROF.
* Created BO full client reports, Web intelligence reports in 6.5 and XI R2, and universes with context and loops in 6.5 and XI R2.
* Worked on Informatica, Oracle Database, PL/SQL, Python, and Shell Scripts as an ETL tool.
* Built HBase tables to load enormous amounts of structured, semi-structured, and unstructured data from UNIX, NoSQL, and several portfolios.

**Environment:** Quality Center, Quick Test Professional 8.2, SQL Server, J2EE, UNIX, .Net, Python, NoSQL, MS Project, Oracle, Web Logic, Shell script, JavaScript, HTML, Microsoft Office Suite 2010, Excel

**Yana Software Private Limited, Hyderabad, India Oct 2012 to Dec 2013**

**SQL Data Engineer**

**Azu**

**Responsibilities:**

* Migrating data from FS to Snowflake within the organization
* Imported Legacy data from SQL Server and Teradata into Amazon S3.
* Exported Data into Snowflake by creating Staging Tables to load Data of different files from Amazon S3.
* Compare the data in a leaf level process from various databases when data transformation or data loading takes place. I need to analyze and look into the data quality when these types of loads are done (To look for any data loss, data corruption).
* As a part of Data Migration, wrote many SQL Scripts for Mismatch of data and worked on loading the history data
* from Teradata SQL to snowflake.
* Developed SQL scripts to Upload, Retrieve, Manipulate and handle sensitive data (National Provider Identifier Data I.e. Name, Address, SSN, Phone No) in Teradata, SQL Server Management Studio and Snowflake Databases for the Project
* Worked on to retrieve the data from FS to S3 using spark commands
* Built S3 buckets and managed policies for S3 buckets and used S3 bucket and Glacier for storage and backup on AWS.
* Created Metric tables, End user views in Snowflake to feed data for Tableau refresh.
* Generated Custom SQL to verify the dependency for the daily, Weekly, Monthly jobs.
* Using Nebula Metadata, registered Business and Technical Datasets for corresponding SQL scripts
* Experienced in working with spark ecosystem using Spark SQL and Scala queries on different formats like text file, CSV file.
* Developed spark code and spark-SQL/streaming for faster testing and processing of data.
* Strong proficiency in C# and the .NET framework, including LINQ, Entity Framework, and ASP.NET.

**Environment:** Snowflake, AWS S3, GitHub, Service Now, HP Service Manager, EMR, Nebula, Teradata, SQL Server, Apache Spark, Sqoop. Azure Synapse AnalyticsPo