A blue hexagon with white text

Description automatically generated A blue and white logo

Description automatically generated

**Sumith Raj**

Email: [sumithrj1@gmail.com](mailto:sumithrj1@gmail.com)

Phone: +1 206-333-8304

[www.linkedin.com/in/sumith-jala](https://d.docs.live.net/c6c97381d8f5ec8d/Documents/www.linkedin.com/in/sumith-jala)

PROFESSIONAL SUMMARY:

* Around 9 years of IT experience in the design and development of big data using Hadoop, the design and development of web applications using Python, and data warehousing development using My SQL, Oracle, and Informatica. Solid work experience in big data technologies with hands-on experience in installing, configuring, and using ecosystem components like Hadoop Map Reduce, HDFS, HBase, Zookeeper, Hive, Flume, Snowflake, Kafka, and Spark.
* Experienced in Agile methodologies, Scrum stories, and sprints.
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing and transforming the data to uncover insights into customer usage patterns.
* Well-versed with big data on AWS cloud services, i.e., EC2, S3, EMR, Glue, Lambda, DynamoDB, and RedShift.
* Proficient in implementing and configuring DBT for data transformation workflows, ensuring efficient and scalable data pipelines.
* Prepared scripts to automate the ingestion process using Pyspark and Scala as needed through various means, such as API, AWS S3, Teradata, and Redshift.
* Developed complex SQL queries to extract, transform, and analyze large datasets from various sources, including relational databases, data warehouses, and data lakes.
* Experience in building the orchestration on Azure Data Factory for scheduling purposes.
* Responsible for estimating the cluster size, monitoring, and troubleshooting the Spark data brick cluster.
* Developing Spark programs with Python and applying principles of functional programming to process complex structured data sets.
* Complete monitoring of AWS Athena query performance and resource utilization using AWS CloudWatch metrics and logs. Capable of identifying and troubleshooting performance bottlenecks to ensure smooth data processing operations.
* Proficient in utilizing Azure DevOps for managing end-to-end software delivery pipelines, including source code management, continuous integration, and continuous deployment.
* Implemented automated build, test, and deployment pipelines using Azure Pipelines to accelerate software delivery and ensure consistent quality.
* Experience integrating DBT with version control systems like Git, enabling collaborative development and code management practices.
* Expertise in writing different RDD (resilient distributed datasets) transformations and actions in Scala, Python, and Java.
* Hands-on experience in handling database issues and connections with SQL and NoSQL databases like MongoDB, Cassandra, Redis, CouchDB (Azure), and DynamoDB (Amazon) by installing and configuring various packages in Python.
* Good experience in using data modeling techniques to find the results based on SQL and PL/SQL queries.Proficient in SQLite, MySQL, and SQL databases with Python.
* Acute knowledge of Spark streaming and Spark machine learning libraries.
* Worked on the ETL process of data loading from different sources and the data validation process from the staging area to the Actavis data warehouse
* Designed and implemented dimensional data models following the Kimball Methodology, including star schemas and snowflake schemas, to support efficient querying and analysis of business data.
* Optimized SQL queries and database performance by analyzing query execution plans, indexing strategies, and database configurations, resulting in significant improvements in query response times and overall system performance.
* Proficient in writing infrastructure code using Azure Resource Manager (ARM) templates and Azure CLI to provision and manage Azure resources programmatically.
* Implemented Apache Airflow for authoring, scheduling, and monitoring data pipelines.
* Experience in using Cloudera Manager for installation and management of single-node and multi-node Hadoop clusters (CDH4&CDH5).
* Familiar with data architecture, including data ingestion pipeline design, Hadoop architecture, data modeling, data mining, and advanced data processing. Experience optimizing ETL workflows.
* Experience in Data Extraction, Transformation, and Loading of data from multiple data sources into target databases, using Azure Databricks, Azure SQL, PostgreSQL, SQL Server, and Oracle
* Good Exposure to Data Quality, Data Mapping, and Data Filtration using Data Warehouse ETL tools like Talend, Informatica, and DataStage.
* Experience in writing distributed Scala code for efficient big data processing.
* Experience working with text, sequence files, XML, parquet, JSON, ORC, AVRO file formats, and Click Stream files.
* Created Tableau reports with complex calculations and worked on ad hoc reporting using Power BI.
* Experience with Azure Big Data technologies like Azure Data Lake Analytics, Azure Data Lake Store, Azure Data Factory, and Azure Synapse, and created POC in moving the data from flat files and SQL Server using U-SQL jobs.
* Strong hands-on programming, design, understanding, and deep functional experience in creating, deploying, and consuming XML Web Services, WCF Services, REST Services, and Web APIs.
* Experience in Designing and Building the Dimensions and Cubes with Star Schema using SQL Server Analysis Services (SSAS).
* Demonstrated experience in deploying AWS Athena in production environments for real-world use cases such as data analytics, business intelligence, log analysis, and ad-hoc querying, delivering actionable insights to stakeholders.
* Experience in tracking tools like JIRA and ServiceNow.
* Well-versed in building CI/CD pipelines with Jenkins, used tech stacks like Gitlab, Jenkins, Helm, and Kubernetes.
* Experience in working with different SDLC techniques like waterfall, agile scrum, and TDD for creating and conveying applications.
* Worked with different record organizations like CSV, JSON, XML, Avro, and Parquet document designs.
* Carried out many data quality checks on all data assets, including ETL scripts, jobs, and data pipeline workflows.

**Education:**

Bachelor’s in computer engineering from Osmania University, India

Masters in MSc Analytics & Systems, University of Bridgeport, US.

TECHNICAL SKILLS

|  |  |
| --- | --- |
| **Big Data Technologies** | HDFS, Hive, MapReduce, Hadoop distribution, HBase, Spark, Kafka Streaming, Scala, ETL |
| **Programming languages** | R, Scala, Terraform, Python, Java |
| **Databases** | MySQL, MS-SQL Server 20012/16/17/19, Oracle 10g/11g/12c, Teradata No SQL Mongo DB, Cassandra, HBase |
| **Scripting/Web Languages** | HTML5, XML, SQL, Shell/Unix, Python. |
| **Operating Systems** | Linux, Windows XP/7/8/10, Mac. |
| **Software Life Cycle** | SDLC, Waterfall, and Agile models. |
| **Utilities/Tools** | Eclipse, Tomcat, NetBeans, JUnit, SQL, SVN, Log4j, SOAP UI, ANT, Maven, Alteryx, Visio, Jenkins, Jira, IntelliJ, bamboo. |
| **Data Visualization Tools** | Tableau, PowerBI, Informatica, Talend, SSIS, and SSRS |
| **Cloud Services** | AWS (EC2, S3, EMR, RDS, Lambda, CloudWatch, Auto scaling, Redshift, Cloud Formation, Glue, etc.), Azure (Databricks, Azure Data Lake, Azure HDInsight) |
| **Version Control & Containerization tools** | Jenkins, Git, and SVN |

# 

**Professional Experience:**

**Client: Barclays- New York City Mar 2023 - Present**

**Role: Senior Data Engineer**

**Responsibilities:**

* As a data engineer for Constellation, I was responsible for building and architecting multiple data pipelines and performing end-to-end ETL processes for data ingestion and transformation. Utilize tools like Informatica Power Center, AWS, Spark RDD, and Apache Spark to process large datasets.
* Skilled in setting up automated testing frameworks within DBT, ensuring data quality and reliability across development, staging, and production environments.
* Designed and set up Enterprise Data Lake to support various use cases, including analytics, processing, storing, and reporting voluminous, rapidly changing data, and developed a security framework to provide fine-grained access to objects in AWS S3 using AWS Lambda and Dynamo DB.
* Set up and worked on Kerberos authentication principles to establish secure network communication on cluster and testing of HDFS, Hive, Pig, and Map Reduce to access cluster for new users.
* Performed end-to-end architecture and implementation assessment of various AWS services like Amazon EMR, Redshift, and S3.
* Built Extract, Transform, Load (ETL) processes to populate dimensional models from source systems, adhering to Kimball's principles of conformed dimensions and slowly changing dimensions (SCDs).
* Built and maintained ETL (Extract, Transform, Load) processes using SQL scripts and stored procedures to transform raw data into standardized formats suitable for analytics and reporting.
* Experience automating AWS Athena workflows using AWS Step Functions, AWS Lambda, and AWS Glue for scalability and reliability. Proficient in designing serverless architectures to handle varying workloads efficiently.
* Experienced in configuring and customizing Azure DevOps services such as Azure Repos, Azure Pipelines, Azure Boards, and Azure Artifacts to meet project requirements.
* Implemented the machine learning algorithms using Python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and S3 data lake.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Simple Storage Service (Amazon S3) and Amazon DynamoDB.
* Used Spark SQL for Scala, a Python interface that automatically converts RDD case classes to schema RDD.
* Worked with the ETL team involved in loading data from the staging area to the data warehouse. Provided all business rules for the database for loading data.
* Familiarity with CI/CD pipelines and integrating DBT into automated deployment processes enable rapid and reliable deployment of data transformations.
* Integrated Azure Pipelines with version control systems like Git to automate code deployment to target environments, including development, testing, and production.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Creating Lambda functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost of EC2 resources.
* Designed and implemented data warehouse architectures based on Kimball's bus architecture or hybrid approaches, ensuring scalability, flexibility, and ease of maintenance for evolving business needs.
* Implemented data quality checks and validation rules within SQL scripts to identify and remediate data quality issues, ensuring accuracy, completeness, and consistency of data for analysis and decision-making.
* Coded Teradata BTEQ scripts to load and transform data, fix defects like SCD 2 date chaining, and clean up duplicates.
* In-depth knowledge of SSIS, Power BI, Informatica, T-SQL, and reporting and analytics.
* Developed a reusable framework for future migrations that automate ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
* Conducted Data blending and data preparation using Alteryx and SQL for Tableau consumption and publishing data sources to the Tableau server.
* Implemented IaC best practices to define and maintain infrastructure configurations as code, ensuring consistency, repeatability, and version control.
* Developed Kibana dashboards based on the log stash data and integrated different source and target systems into Elastic search for near real-time log analysis of monitoring end-to-end transactions.
* Created instances in Implemented AWS Step Functions to automate and orchestrate the Amazon Sage Maker-related tasks such as publishing data to S3, training the ML model, and deploying it for prediction.
* Good knowledge of Amazon AWS concepts like EMR and EC2 web services, which provide fast and efficient processing of big data.
* Expertise in managing dependencies between DBT models, optimizing dependency graphs and ensuring efficient execution of data transformation jobs.
* Migrated data to AWS from the Data Center using Snowball and the AWS migration service and implemented a generalized solution model using AWS Sage Maker.
* Building required infrastructure for optimal extraction, transformation, and loading of data from various data sources using AWS and SQL technologies.
* Extract, transform, and load data from source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* Developed entire frontend and backend modules using Python on Django Web
* Framework and created User Interface (UI) using JavaScript, bootstrap, Cassandra with MySQL, HTML5/CSS.
* Implement logging, monitoring, and SRE practice using Elastic Search (ELK), Prometheus, and Splunk.

**Environment**: Hadoop, Map Reduce, Spark, Spark MLlib, Tableau, SQL, MATLAB, AWS, Python, Cassandra, Oracle, MongoDB, SQL Server, DB2, T-SQL, PL/SQL, XML, Tableau, AWS EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Snowflake, Apache Pig, Python, SSRS, Tableau.

**Client:** **Elevance Health- Saint Louis, MO Aug 2022 - Feb 23**

**Role: Data Engineer**

**Responsibilities:**

* Used Azure Data Factory as an orchestration tool for integrating data from upstream to downstream.
* Worked on Snowflake DB to ingest the data from various sources to Snowflake.
* Have extensive experience in creating pipeline jobs, scheduling triggers, Mapping data flows using Azure Data Factory (V2), and using key Vaults to store credentials.
* Designed and implemented data warehouse architectures based on Kimball's bus architecture or hybrid approaches, ensuring scalability, flexibility, and ease of maintenance for evolving business needs.
* Performed database administration tasks such as schema management, user access control, backup and recovery, and database monitoring using SQL Server Management Studio (SSMS) or similar tools.
* Created and optimized stored procedures and functions in SQL to encapsulate business logic and improve code reusability, maintainability, and performance in database operations.
* Creating pipelines, data flows, complex Data transformations, and manipulations using ADF and Pyspark.
* Worked on Apache Spark along with Python Programming language for transferring the data in a much faster and more efficient way.
* Utilized surrogate keys in dimensional models to uniquely identify records and facilitate efficient data integration, transformation, and querying across multiple data sources.
* Designed and developed Workflow as per ETL Specification for Stage load and Warehouse load.
* Have good experience working with Azure Blob and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Proficient in setting up monitoring and logging solutions for DBT jobs, enabling real-time visibility into data transformation processes and facilitating troubleshooting and performance optimization.
* Led and mentored development teams in adopting Azure DevOps practices and tools, guiding best practices, standards, and processes.
* Fostered a collaborative and cross-functional work environment, promoting knowledge sharing, transparency, and accountability among team members.
* Utilized Kubernetes and Docker for runtime environment for CI/CD to build, test, and deploy.
* Ingested data in mini-batches and performed RDD transformations on those mini-batches of data by using spark streaming to perform streaming analytics in Data bricks.
* Worked on migration of data from on-prem SQL server to cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB)
* Evaluated the suitability of Hadoop and its ecosystem to the above project and implemented/validated various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative.
* Have good experience working with Azure Blob and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Experience optimizing DBT workflows for scalability and performance, including partitioning strategies, resource allocation, and query optimization techniques.
* Worked on importing and exporting data from Oracle and DB2 into HDFS and HIVE using Sqoop.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Worked on SCALA Programming language which is supported by Apache Spark and Developed Spark scripts by using Scala shell commands as per the requirement.
* Creating Logic Apps with different triggers and connectors for integrating the data from the workday to different destinations.
* Designed and implemented aggregate tables to precompute summary-level data for improved query performance, following Kimball's guidelines for selecting appropriate aggregation levels.
* Developed a data pipeline using Kafka and Storm to store data in HDFS.
* Creating numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure activities like Move & Transform, copy, filter, for each, Data bricks, etc.

**Client: Costco Wholesale- Seattle, WA Feb 2022 - July 2022**

**Role: Data Engineer**

**Responsibilities:**

* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services (Azure Data Lake, Azure Storage, Azure DW) and processing the data in Azure Databricks.
* Primarily involved in Data Migration using SQL, Azure Storage, Azure Data Factory, SSIS, and PowerShell.
* Strong skills in visualization tools, Power BI, and Confidential Excel – formulas, and charts.
* Worked on Data Visualization tools like Tableau to visualize the data.
* Knowledge of implementing DBT as part of infrastructure as code practices, using tools like Terraform or CloudFormation to automate provisioning and configuration of resources.
* Profile structured, unstructured, and semi-structured data across various sources to identify patterns in data and Implement data quality metrics using necessary queries or Python scripts based on source.
* Designed the ETL processes using Informatica to load data from Mainframe DB2, Oracle, SQL Server, Flat Files, XML Files, and Excel files to target the Teradata warehouse database.
* Created a new data model that embeds NoSQL sub-models within a relational data model by applying Hybrid data modeling concepts.
* Led data migration projects involving the extraction, transformation, and loading of data between different database systems and platforms using SQL scripts and tools like SSIS (SQL Server Integration Services).
* Airflow Logs from Source to Elastic, Airflow Metrics from Source to Elastic.
* Facilitate data cleansing, enrichment, and data quality improvements.
* Guides the deployment approach and configuration of the ELK platform.
* Data analysis, data modeling, and implementation of enterprise-class systems spanning big data, data integration, and object-oriented programming.
* Understanding of implementing security best practices in DBT, including role-based access control (RBAC), data encryption, and compliance with data governance standards.
* Used MongoDB to store data in JSON format and developed and tested many features of a dashboard using Python, Bootstrap, CSS, and JavaScript.
* Extensively used Code Cloud for code check-in and checkouts for version control.

**Environment**: Data Visualization, Data Warehouse, Azure, SQL Server, Teradata, Azure Databricks, Azure Synapse Analytics, Snowflake, ADF, ADLS, NOSQL, SSMS, Pyspark, python, RDBMS.

**Client: Media Mint-India Apr 2019 - July 2021**

**Role: Software Engineer**

**Responsibilities:**

* Used Power BI, and Power Pivot to develop data analysis prototype, and used Power View and Power Map to visualize reports.
* Involved in requirement gathering and worked closely with the architect in designing and modeling.
* Writing Python scripts by using Pandas, and SQL Alchemy for creating control files and for ETL.
* Preparing SQL statements for data loading from control files to dimension tables and from dimension to fact tables within the Python program.
* Proficient in documenting DBT models, transformations, and workflows, and facilitating collaboration among data engineers, analysts, and other stakeholders using tools like DBT Docs.
* Documented SQL scripts, database schemas, and data integration processes, and provided training and knowledge transfer to team members, ensuring transparency, consistency, and continuity in data engineering practices.
* Working in big data technologies like Spark 2.3 & 3.0 Scala, Hive, and Hadoop cluster (Cloudera platform).
* Designed and developed data service components using WEB API and REST, Consumed REST services by making AJAX calls.
* Designed and developed Workflow as per ETL Specifications for Stage and Warehouse load.
* Involved in Developing ETL solutions using Spark SQL in Azure Databricks for data extraction, transformation, and aggregation from multiple file formats and data sources for analyzing & transforming the data to uncover insights into customer usage patterns.
* Created Dax Queries to generate computed columns in Power BI.
* Preparing Python scripts for reusing Python programs, user-defined functions, and SQL statements in the ETL process.

**Client: Amnet Digital- India Sep 2016 - Mar 2019**

**Role: Junior Software Engineer**

**Responsibilities:**

* Involved in requirement gathering and worked closely with the architect in designing and modeling.
* Develop reporting and automation workflows to abstract, cleanse, validate, integrate, and interpret medical claims/eligibility data utilizing SQL Server, PowerShell, Jasper, and Qlik applications.
* Created SQL scripts, stored procedures, views, triggers, and PowerShell workflows to pull and transform data for Reporting Services.
* Monitor and tune ETL processes for performance improvements; identify, research, and resolve data warehouse load issues.
* Create PL/SQL and T-SQL scripts to pull and analyze data requests by relationship managers.
* Assist with the development of team productivity metrics for daily/weekly/quarterly reports.
* Monitored dashboards to check whether the underlying data was matched to the business requirement.
* Support the Reporting and analytics team in completing multiple analytical projects successfully reducing end-of-month reporting turnaround time by 50%.
* Provide top-tier support on JIRA ticket requests related to report automation, and ETL processing.

**Environment:** Python, pandas, NumPy, sci-kit-learn, SQL, Excel Advanced functions, pivot tables, V Lookups, Spark, Talend, Oracle, Hadoop, Hive, HBase, Git.

**Client: W3Softech- India Jun 2015- Aug 16**

**Role: Data Analyst**

**Responsibilities:**

* Prepare unstructured data from multiple files from customers' databases and ad-hoc.
* Report customer requests as needed, along with a business review analysis of the program data.
* Communicated with the Source code provider in case of any discrepancies.
* Extracted, Transformed, and Loaded (ETL) data to map it from disparate sources to the required target database.
* Worked on production support activities to monitor all Daily, Weekly, Monthly, and Quarterly jobs in Scheduler, fixing the failed workflows, and communicating to the different teams to get the issue fixed based on the issue.
* Monitor all daily, weekly, monthly, and quarterly jobs and track the running statistics.
* Delved into data to discover discrepancies and patterns.
* Management and internal teams to implement and evaluate improvements.

**Environment:** SQL Server, Oracle 10g, MS Excel**.**