**SARATH BASWA
 SR. DATA ANALYTICS ENGINEER**

**Email:** sarath28941@gmail.com

**LinkedIn:** [linkedin.com/in/sarath-baswa](http://linkedin.com/in/sarath-baswa)

**Phone:** 9726658603

**PROFESSIONAL SUMMARY**

* Over 10 years of experience in IT, specializing in data engineering, big data analytics, cloud computing, & business intelligence, driving data-driven decision-making across multiple industries, including healthcare, retail, finance, & energy.
* Expertise in Big Data technologies, including Hadoop, Spark, Databricks, and Snowflake, for large-scale data processing and analytics.
* Experienced in building & automating HEDIS reporting & data transformation workflows using Azure Data Factory, ensuring seamless integration of healthcare claims, clinical data, & quality metrics for compliance & performance monitoring.
* Proficient in cloud technologies across AWS (EMR, EC2, S3, Lambda, Athena, Glue, RDS, Redshift, DynamoDB), Azure (Blob Storage, ADLS, Azure Data Factory (ADF), Synapse, Azure SQL Server), and GCP (Beam, Data Proc, Big Query, Data Flow, Composer)
* Extensive experience in Python, Scala, and Java for data engineering, automation, and machine learning applications.
* Strong proficiency in query languages and databases, including SQL, Hive, Oracle, MySQL, DB2, and Postgres, with expertise in writing advanced SQL and PL/SQL queries.
* Experience with DevOps and CI/CD tools such as Jenkins, Docker, and Kubernetes for automating data pipeline deployments.
* Hands-on experience with machine learning frameworks like Sagemaker and Azure ML Studio, leveraging predictive modeling for business insights.
* Proficient in scheduling and orchestration tools, including Apache Airflow, Oozie, Azkaban, and AWS Step Functions, to automate and optimize data workflows.
* Strong expertise in managing on-premises Hadoop platforms like Cloudera and HortonWorks, ensuring efficient distributed data processing.
* Skilled in ETL and data modeling techniques using tools like Informatica, Erwin, and Snowflake, enabling structured and efficient data architecture design.
* Experience handling diverse data formats such as CSV, Parquet, ORC, Avro, JSON, XML, and TXT for seamless data integration.
* Proficient in BI and data visualization tools like Tableau, Power BI, and Looker, delivering actionable insights through interactive dashboards.
* Strong documentation skills using Confluence, SharePoint, and Lucid Chart to maintain technical design documents and data architecture blueprints.
* Experience with OLAP tools like Kyvos, Druid, Azure Analysis Services (AAS), and SSAS for multi-dimensional data analysis and reporting.
* Proficiency in version control tools such as Git, GitLab, and GitHub for collaborative development and source code management.
* In-depth understanding of SDLC methodologies, including requirement gathering, data architecture design, development, deployment, and production support.
* Expertise in ticketing and project management tools like JIRA, ServiceNow, Rally, and Azure DevOps for Agile-based sprint planning and execution.
* Strong experience in data security and governance, including encryption, row-level security, column masking, IAM policies, and GDPR compliance.
* Demonstrated ability to optimize data pipelines, reduce query execution time, and improve overall system performance through indexing, partitioning, and caching strategies.
* Team player with strong problem-solving skills, quick adaptability to new technologies, and a passion for continuous learning and innovation in the data engineering space.

**TECHNICAL SKILLS**

|  **Programming Languages** | Python, SQL, Scala |
| --- | --- |
|  **Big Data Technologies** | Hadoop, Spark, Kafka, Hive, Sqoop |
|  **Cloud Platforms** | AWS (S3, EC2, EMR, Kinesis, Redshift, Athena, Glue, Sagemaker), Azure (HDInsight, Data Factory, Data Lake, Synapse Analytics, ML Studio, Databricks |
|  **ETL Tools** | Informatica, SQL Server Integration Services (SSIS), Microsoft Visual Studio (SSDT), Apache Airflow |
|  **Data Visualization Tools** | Tableau, Power BI, QlikView |
|  **NoSQL Databases** | HBase, Cassandra, MongoDB |
|  **Version Control** | Git, GitLab |
|  **Data Warehousing** | Data warehousing concepts, Snowflake |

**PROFESSIONAL EXPERIENCE**

**Client: Health Recovery Solutions, Hoboken, New Jersey Mar 2022 - Present
Role: Sr. Data Analytics Engineer**

* Designed and implemented a scalable data warehouse on Redshift, integrating electronic health records (EHR), insurance claims, and patient monitoring data, enabling real-time healthcare analytics and compliance with HIPAA regulations.
* Designed and automated HEDIS reporting pipelines using Azure Data Factory, integrating claims data, clinical outcomes, and quality metrics to streamline reporting for healthcare providers and ensure regulatory compliance.
* Developed and orchestrated serverless workflows using AWS Step Functions and AWS Lambda to automate data processing and integration of healthcare claims data, improving system efficiency and reducing processing time by 30%.
* Developed complex ETL workflows using Glue to ingest and transform structured and unstructured data from multiple sources, including JSON, CSV, and AVRO files, ensuring data integrity and accuracy for healthcare analytics.
* Built real-time data ingestion pipelines using Kafka, processing continuous patient vitals from IoT-enabled medical devices to alert healthcare providers in case of anomalies, improving response time by 30%.
* Integrated various insurance claims processing APIs with Databricks, automating fraud detection and reimbursement approvals, reducing manual processing time by 40%, and ensuring compliance with healthcare regulations.
* Led the migration of on-premises healthcare records from Teradata to Snowflake, optimizing storage and query performance and ensuring seamless interoperability between hospitals and insurance providers.
* Designed and optimized large-scale data aggregation jobs using PySpark in Databricks, merging patient history, lab test results, and prescriptions to build predictive healthcare models.
* Implemented fine-grained data access control using RBAC in Snowflake, ensuring that only authorized personnel could view or modify patient health data, maintaining strict regulatory compliance.
* Developed predictive analytics models in Azure ML, leveraging patient demographics and historical medical records to predict the likelihood of readmissions, reducing avoidable hospitalizations by 15%.
* Optimized complex SQL queries in Snowflake to enhance reporting efficiency, enabling faster retrieval of patient treatment history and improving decision-making for healthcare providers.
* Automated data validation and quality checks using Great Expectations, ensuring accurate insurance claims processing and reducing billing discrepancies by 20%.
* Built interactive Power BI dashboards displaying real-time hospital bed occupancy, staff workload, and emergency admission trends, helping healthcare administrators optimize resource allocation.
* Developed a high-performance Parquet-based data storage architecture to efficiently store and retrieve high-frequency medical sensor readings, reducing query response times.
* Implemented DAG-based workflow automation using Airflow, orchestrating daily patient records ingestion, claims processing, and clinical trial data updates.
* Integrated an external API ingestion framework to collect medical data from third-party laboratories and wearable fitness devices, ensuring seamless integration into hospital analytics systems.
* Engineered a fraud detection model using PySpark, identifying anomalies in medical billing and insurance claims, reducing fraudulent transactions by 25%.
* Implemented centralized logging and exception handling using Log4j, allowing real-time error tracking and debugging of ETL failures in Glue pipelines.
* Led Agile-based sprint planning in JIRA, collaborating with clinical research teams and data engineers to improve data accuracy and ETL processing efficiency.
* Designed a real-time appointment booking and patient record update system using DynamoDB, ensuring immediate updates and high availability.
* Configured secure IAM policies in AWS to enforce compliance with HIPAA, ensuring encrypted data transmission and controlled access to sensitive patient data.
* Automated ETL pipeline deployments using Jenkins, reducing manual intervention, and achieving CI/CD best practices for healthcare data integration.

**Environment:** AWS (Redshift, S3, Glue, IAM, DynamoDB), Snowflake, Databricks, PySpark, Kafka, Airflow, Power BI, API, SQL, Parquet, Log4j, Jenkins, Azure ML, Great Expectations, RBAC

**Client: Dollar General, Goodlettsville, TN Nov 2017 – Feb 2022
Role: Sr. Data Analytics Engineer**

* Designed and deployed a cloud-based data lake using S3, integrating sales, inventory, and supply chain data from 18,000+ stores to enable enterprise-wide analytics and reporting.
* Utilized AWS Step Functions and AWS Lambda to streamline data processing workflows, reducing data retrieval times and optimizing performance for retail analytics platforms.
* Developed PySpark-based ETL pipelines in Databricks to transform, clean, and aggregate transactional data, optimizing SKU performance analysis and pricing models.
* Implemented real-time Kafka streaming pipelines to process point-of-sale (POS) transactions, reducing batch processing dependencies and improving data freshness.
* Built a Snowflake-based customer segmentation model using advanced SQL, enabling targeted promotions and personalized loyalty programs based on RFM (Recency, Frequency, Monetary) analysis.
* Developed an Azure Data Factory (ADF)-orchestrated workflow to aggregate supply chain data, ensuring real-time visibility into demand forecasting and logistics optimization.
* Created interactive Power BI dashboards to track product availability, supplier performance, and revenue trends across all retail locations.
* Developed an API ingestion framework to fetch competitor pricing data, enabling real-time price adjustments based on market trends.
* Built an anomaly detection system in Databricks, leveraging machine learning models to identify fraudulent transactions in gift cards and online sales.
* Configured role-based access controls in Azure AD, restricting sensitive customer data access and ensuring PCI-DSS compliance.
* Implemented Airflow DAGs to automate ETL processes, scheduling nightly batch updates and hourly incremental loads from retail transaction databases.
* Ingested and integrated structured and semi-structured sales data (CSV, JSON, Parquet) into Snowflake, improving query performance for analytics teams.
* Developed a real-time inventory tracking system using DynamoDB, reducing stockouts and optimizing warehouse-to-store replenishment cycles.
* Migrated legacy Teradata-based analytical models to Snowflake, reducing data processing latency and improving dashboard responsiveness.
* Containerized ETL processing jobs using Docker, enabling scalable and consistent data pipeline execution across environments.
* Implemented Redshift Spectrum to perform cost-efficient queries on archived historical sales data, reducing storage expenses.
* Designed and enforced Snowflake data masking policies to protect personally identifiable customer information (PII) from unauthorized access.
* Managed Agile sprint cycles in JIRA, coordinating data engineering efforts to support business intelligence and analytics teams.

**Environment:** AWS (S3, Redshift, Glue, IAM, DynamoDB, Athena), Snowflake, Databricks, PySpark, Kafka, Azure Data Factory (ADF), Airflow, Power BI, API, SQL, Parquet, Docker, Log4j, Jenkins, Azure AD

**Client: Mahindra Home Finance, Mumbai, India Jun 2015 – Sep 2017
Role: Data Analytics Engineer**

* Architected an ADLS-based financial data lake to centralize loan applications, credit history, and repayment data, enhancing financial risk assessments.
* Developed a complex ETL pipeline using Azure Data Factory (ADF) to extract, transform, and load financial transactions from on-prem Oracle databases into Azure Synapse.
* Built a PySpark-driven risk assessment model, leveraging loan repayment history and credit score analysis to improve lending decisions.
* Designed a real-time fraud detection algorithm using SQL in Snowflake, identifying unusual lending patterns and unauthorized transactions.
* Created executive-level Power BI dashboards displaying mortgage performance, delinquency rates, and branch-wise credit distribution.
* Implemented a real-time credit scoring pipeline using Kafka, ensuring rapid loan approvals based on real-time financial data.
* Optimized Azure SQL queries for high-volume financial reporting, improving dashboard responsiveness and query execution speeds.
* Automated loan application processing workflows using Azure Logic Apps, reducing manual validation efforts by 40%.
* Developed an API integration framework to fetch third-party credit scores and enhance automated loan approval decisions.
* Implemented RBAC policies in Snowflake to control access to high-risk financial records, ensuring compliance with GDPR and RBI regulations.
* Designed a Parquet-based storage solution for financial transaction logs, improving analytical query performance and reducing storage costs.
* Developed an anomaly detection system using Great Expectations, validating financial data consistency across multiple sources.
* Orchestrated batch and incremental financial data loads using Airflow, ensuring accurate and up-to-date loan and mortgage reporting.
* Migrated legacy Teradata-based risk models to Snowflake, significantly reducing processing times and improving forecasting accuracy.
* Configured IAM policies and encryption for securing highly sensitive borrower information and financial transaction data.
* Implemented Jenkins-based CI/CD pipelines for automating the deployment of financial data transformations and reporting workflows.
* Developed a DynamoDB-powered real-time financial transaction tracking system to monitor loan disbursements and collections.

**Environment:** Azure (ADLS, Azure Data Factory (ADF), Synapse, SQL, IAM, Logic Apps), Snowflake, Databricks, PySpark, Kafka, Airflow, Power BI, API, SQL, Parquet, Great Expectations, Log4j, Jenkins, RBAC

**Client: Cairn Oil and Gas, Gurugram, India Jul 2014 – May 2015
Role: Data Analytics Engineer**

* Developed a cloud-based data processing framework using EMR, ingesting and analyzing seismic exploration and drilling data.
* Built PySpark-powered ETL workflows to process terabytes of oil production and well performance data, optimizing reservoir efficiency.
* Designed a Python-based predictive maintenance model, leveraging sensor data to forecast pipeline failures and reduce operational risks.
* Implemented an API-driven ingestion system in Snowflake to aggregate real-time IoT data from drilling equipment and production sites.
* Optimized seismic survey analysis using Redshift, improving reservoir mapping accuracy and reducing data retrieval times.
* Developed Tableau dashboards providing real-time insights into drilling efficiency, production rates, and equipment utilization.
* Integrated Kafka for streaming oil rig sensor data, enabling real-time alerts for equipment malfunctions and pressure anomalies.
* Implemented AWS IAM policies to secure exploration, drilling, and production data, ensuring compliance with environmental regulations.
* Developed ML-based predictive models for hydrocarbon recovery optimization, improving extraction efficiency.
* Orchestrated data pipelines using Airflow, automating daily updates of well performance and reservoir pressure data.
* Developed a Parquet-based data storage system to improve data compression and retrieval efficiency for geological reports.
* Designed Log4j-based monitoring for data pipelines, providing real-time alerts on ETL job failures and delays.
* Configured DynamoDB for managing drilling history logs, ensuring quick retrieval and analysis of past well performance data.
* Developed anomaly detection algorithms using Great Expectations, ensuring the consistency of drilling and production metrics.
* Implemented Jenkins-based CI/CD automation for seamless deployment of new analytics workflows.
* Optimized Athena queries to analyze production trends, improving data-driven decision-making for field engineers.
* Deployed containerized ETL workflows using Docker, ensuring portability and scalability across cloud environments.
* Implemented Snowflake data masking policies to protect proprietary exploration data from unauthorized access.
* Designed API-based ingestion pipelines for integrating third-party geological survey data into enterprise analytics systems.
* Managed Agile-driven data projects in JIRA, collaborating with geologists and data engineers to optimize oilfield analytics.

**Environment:** AWS (EMR, S3, Redshift, IAM, DynamoDB, Glue), Snowflake, PySpark, Kafka, Airflow, API, SQL, Parquet, Docker, Log4j, Jenkins, ML (Predictive Maintenance), RBAC

**EDUCATION**

Jawaharlal Nehru Technological University, India

 B.Tech in Computer Science and Engineering