**Prabhat Sahoo**

psahoocode@gmail.com | (973) 329 0203 | www.linkedin.com/in/prsahoo



Lead Data Engineer with 17+ years of experience delivering end-to-end data and software solutions across cloud and on-prem environments. Proven expertise in building enterprise-grade ETL/ELT pipelines using Azure Data Factory, Databricks, Synapse, and ADLS. Skilled in PySpark, Python, SQL, and AWS to develop scalable, high-performance data workflows. Experienced in data modeling, orchestration, and implementing robust data governance frameworks. Adept at integrating accounting systems, automating financial reporting, and developing ATS and SaaS platform integrations on AWS. Hands-on experience with Kafka, Redis, AWS Lambda, Docker, and Jenkins, with a focus on performance tuning and cost optimization. Canadian citizen currently working in the U.S. under a TN visa.

***PROFESSIONAL SUMMARY***

* Over 17 years of professional experience in Data Engineering with a strong focus on **AWS** and **Azure Cloud Services**, including deep expertise in **Databricks**, **PySpark**, and large-scale cloud data solutions.
* Expert in designing and implementing **scalable, performant data lakehouse architectures** on **Databricks** and **AWS EMR**, leveraging **Delta Lake**, **Unity Catalog**, and **Lakehouse architecture** for efficient data management and governance.
* Extensive experience building and managing complex **ETL/ELT pipelines** using **Databricks**, **AWS Glue**, **Azure Data Factory**, and **Apache Spark**, integrating diverse data sources such as **Amazon S3**, **Azure Blob Storage**, **ADLS Gen2**, relational databases, REST APIs, and Kafka streaming.
* Proficient in developing and optimizing **PySpark** and **SQL-based data transformations**, ensuring data quality, integrity, and security across cloud-native lakehouse environments.
* Skilled in **orchestrating data workflows** with **AWS Step Functions**, **Azure Data Factory**, **Databricks Workflows**, and CI/CD pipelines using **GitHub Actions**, **AWS CodePipeline**, and **Terraform** for automated, reliable deployments.
* Expertise in building **end-to-end GenAI pipelines**, including unstructured data ingestion, embedding generation (Open AI, Hugging Face), vector storage (Pinecone, FAISS), and integration into enterprise-grade RAG applications.
* Strong background in **real-time and near real-time streaming data solutions** using **Amazon Kinesis**, **Kafka**, and **Spark Streaming** for timely analytics and operational automation.
* Deep understanding of data governance and security in cloud environments, implementing **access controls, IAM policies**, and **Unity Catalog** for secure data sharing and compliance.
* Proven ability to optimize cloud costs through resource tuning and log management across AWS and Azure services, delivering significant monthly savings without sacrificing performance.
* Experienced in automating infrastructure and pipeline management using **Infrastructure as Code (IaC)** tools like **Terraform** and scripting languages including **Python** and **Bash**.
* Collaborates effectively with cross-functional teams including data scientists, analysts, and business stakeholders to translate complex business requirements into scalable, production-grade data solutions supporting executive dashboards and BI tools like **Amazon QuickSight**, **Power BI**, and **Tableau**.
* Skilled in troubleshooting complex data integration challenges and optimizing distributed data processing workloads for **cost-efficiency** and **high performance** in multi-cloud environments.
* Committed to continuous learning and adoption of the latest advancements in **Databricks**, **AWS**, and **Azure** technologies to deliver cutting-edge data engineering solutions.
* Familiar with deploying and managing Databricks artifacts using **Databricks Asset Bundles** for streamlined development, version control, and release management.
* Experienced working in Agile teams, promoting best practices in data engineering, CI/CD, and DevOps to deliver reliable and scalable data platforms that drive business value and actionable insights.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming Languages:** | **Python, Java, Scala, JavaScript, TypeScript, Shell Scripting, SQL, GraphQL** |
| **Big Data Stack:** | **Apache Spark, PySpark, SparkSQL, Hadoop, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Kafka, Storm, Impala, DBT, Apache NiFi, Elasticsearch** |
| **Data Architecture & Patterns** | **Medallion Architecture (Bronze/Silver/Gold), Lambda Architecture, Data Vault, Data Mesh, LangChain, LangGraph, OpenAI, Ollama** |
| **Cloud Platform:** | **Azure (ADF, Databricks, Synapse, ADLS, Machine Learning), AWS (EC2, S3, Lambda, Architecture & Development, Cloudsatch), Azure Functions** |
| **Relational databases:** | **Snowflake, Oracle, MySQL, SQL Server, PostgreSQL, Teradata, DB2, MongoDB, Redis** |
| **Version Control Systems:** | **Bitbucket, GIT, SVN, GitHub, StarTeam** |
| **Data Visualisation** | **Power BI, Tableau, SSRS** |
| **Monitoring and Logging** | **Grafana, Prometheus, VictoriaMetrics, Datadog, Sentry** |
| **Development & Testing Tools:** | **Jira, PyCharm, Eclipse, HP Quality Center, Postman, JMeter, WildFly, SSIS** |

**PROFESSIONAL EXPERIENCE:**

**Client: Sense Talent Labs, WA Jan 2022 – Till Date Role: Lead Data Engineer**

**Responsibilities:**

* Led the development of Sense's ATS integration with platforms such as Workday, Greenhouse, and ORC, enabling onboarding of over 50 new corporate agencies and generating $20M+ in revenue over three years. Engineered scalable APIs using AWS Lambda and API Gateway to handle ATS webhooks.
* Migrated legacy eRecruit ATS implementations to Sense V2, reducing synchronization time from 9 hours to 15 minutes and improving customer satisfaction and data freshness.
* Implemented multi-source ATS synchronization within Sense, supporting dynamic feature rollouts and seamless onboarding for agencies operating across multiple ATS platforms.
* Re-architected asynchronous data syncs by replacing Queue-based SQL tables with Kafka, enabling real-time updates and significantly reducing latency and database maintenance overhead.
* Shifted data and transformation definitions to Redis cache, eliminating deployment cycles during config changes—saving up to 2 hours per day for the engineering team across frequent migrations.
* Automated Jenkins job creation, auditing, and Slack alerting for 50+ agencies, eliminating over 1,000 hours of manual work for development and support teams.
* Drove platform-level cost optimization as part of the Cost Reduction task force—implemented audit log partitioning, trimmed logs older than 30 days, tuned Docker resource configurations in AWS Fargate, and optimized logging policies—achieving monthly savings exceeding $100K across AWS and third-party tools.
* Designed and managed enterprise-grade ETL/ELT pipelines using AWS Glue and AWS Step Functions, integrating structured and unstructured data from Amazon S3, REST APIs, and cloud-native databases.
* Built near real-time streaming data workflows using Amazon Kinesis, Kafka, and Apache Spark on EMR to enable timely business insights and operational automation.
* Developed optimized Delta Lake models on Amazon EMR and Lake Formation, incorporating schema evolution, Z-ordering, and indexing for efficient downstream consumption in Amazon Redshift and Athena.
* Authored complex SQL and PySpark transformations using Glue Jobs and Spark Streaming to handle large-scale batch and streaming workloads.
* Automated transformation workflows and CI/CD deployments using DBT on AWS with custom macros, testing frameworks, and version control via GitHub Actions and CodePipeline.
* Ensured system reliability and performance through proactive monitoring, error handling, and auto-scaling of Spark clusters on Amazon EMR; integrated logging with Amazon CloudWatch and AWS Lambda for alerts and recovery.
* Orchestrated data validation logic and integrated curated datasets into Amazon QuickSight dashboards to support executive reporting and real-time analytics.

**Environments: AWS Glue, Amazon S3, EMR, Kinesis, Kafka, Redshift, Athena, Lake Formation, Lambda, DBT, PySpark, SQL, Python, QuickSight, CodePipeline, GitHub Actions, CloudWatch, Terraform, REST APIs, Data Lake Architecture, Delta Lake, Data Governance, IAM, VPC, AWS Secrets Manager**

**Client: Scientific Games, BC May 2015 – Jan 2022 Role: Lead Software Engineer**

**Responsibilities:**

* For Casino Slot Data System and Player Management System, created data pipeline with following details.
* Designed and implemented end-to-end ETL/ELT pipelines using Azure Data Factory and Snowflake to ingest, transform, and load data from diverse sources, including on-prem databases, APIs, Cosmos DB, ADLS Gen2, and Blob Storage.
* Developed scalable and automated Snowflake data pipelines using Snowpipe and Streams for near real-time data processing, eliminating manual intervention and reducing latency.
* **Integrated GenAI powered copilots** into enterprise BI platforms (Power BI, Looker) to assist users with natural language queries, automating data exploration and reporting tasks.
* Designed optimized Snowflake SQL queries, clustering keys, and partitioning strategies to boost query performance and reduce cost for large-scale analytics workloads.
* Implemented Snowflake’s advanced security and data sharing capabilities, ensuring data integrity, controlled access, and regulatory compliance across business units and external partners.
* Led DBT model development with reusable macros and materializations, supporting traceable, testable, and efficient data transformation pipelines.
* Integrated DBT with CI/CD pipelines for automated deployments, version control, and streamlined updates across dev/test/prod environments.
* Created DBT testing frameworks and configured lineage tracking to ensure visibility, trust, and validation across all transformation layers.
* Built robust, high-performance Spark-based data processing frameworks in Azure Databricks, integrating Delta Lake and Apache Iceberg to handle schema evolution, versioning, and scalable storage.
* Orchestrated task scheduling and pipeline execution in Databricks using ADF and Databricks Workflows to enable batch and real-time data flows across distributed computing clusters.
* Tuned Spark jobs by optimizing memory usage, managing partitions, and leveraging adaptive query execution (AQE) for high-throughput processing.
* Applied proactive monitoring and troubleshooting for Spark and ADF jobs using Azure Monitor and Log Analytics, ensuring reliability, fault-tolerance, and minimal downtime.
* Enabled cross-functional collaboration with data scientists, analysts, and business stakeholders to translate requirements into production-grade, reusable data pipelines.
* Automated data integration workflows and scheduling logic using ADF triggers and Mapping Data Flows, improving development speed and reducing operational complexity.
* Documented pipeline architecture, transformation logic, and deployment processes using GitHub, Terraform, and collaborative specifications to ensure maintainability and auditability.
* Enforced enterprise data governance, data validation methods, and regulatory standards across all data engineering processes, including audits, encryption, retention, and access control.
* Contributed to platform-wide CI/CD and DevOps practices, optimizing Jenkins pipelines, managing Terraform scripts, and aligning with release automation strategies.
* Ensured end-to-end integration of data pipelines with analytics tools such as Power BI, enabling real-time insights and executive-level reporting dashboards.
* Led and developed a player tracking system for Casinos to support COVID restrictions. This enabled Casinos to implement distancing rules in indoor environments and open Casinos for the public in Canada during the pandemic.
* Fixed Casino accounting end-of-day calculation performance and resolved deadlocks, improving the performance of accounting drops from 30 minutes to 5 minutes.
* Led and developed an Asset Management System in Casinos. This allowed assets to be entered into the system once and moved across different Casinos from a single UI, saving 100+ hours annually of manual work by slot technicians.
* Consolidated Game Management Unit and iView updates along with log collection. Previously requiring multiple applications, this unification simplified Casino technicians' workflow and enabled centralized log collection from a single server.
* Supported 150+ Casinos on technical issues worldwide, helping prevent Casino floor downtime and contributing to over $10 million in savings.
* Contributed to a multi-phase project involving **data set analysis, pre-processing, user-generated data extraction**, and modeling for analytics and machine learning workflows.
* Participated in **data acquisition efforts** using Sqoop, Pig, Flume, Hive, MapReduce, and HDFS to extract both historical and real-time data.
* Developed **user-defined functions (UDFs) in Hive** to manipulate complex string and date formats during data transformation.
* Performed **data cleaning, feature scaling, and feature engineering** using Python libraries such as Pandas and NumPy.
* Implemented a **process improvement initiative** by analyzing recurrent error logs in Python—reduced turnaround time by 60% for repeated issues.
* Executed **production data fixes** through the development and testing of custom SQL scripts.
* Conducted **in-depth analysis of complex datasets** using machine learning models like Linear Regression, Logistic Regression, and Decision Trees.
* Generated performance tracking reports using **SQL and Excel**; created data visualizations in **Tableau** to surface key business metrics.
* Applied **clustering algorithms** including Hierarchical Clustering and K-Means using Scikit-learn and SciPy for customer segmentation.  
  Carried out **data collection, cleansing, visualization, and feature engineering** using Python (Pandas, NumPy, Matplotlib, Seaborn).
* Optimized SQL queries and integrated **Informatica with MySQL** to transform raw data into structured, machine-learning-ready formats.
* Developed **interactive dashboards** in Tableau and collaborated with business analysts to align dashboard UX with business requirements.
* Built and deployed **ETL packages using SSIS** for validating, extracting, transforming, and loading data into a Data Warehouse and Data Mart.
* Built **machine learning models** using XGBoost for RFM-based customer lifetime value classification.
* Created and maintained **complex SQL queries, stored procedures, views, and reports** using Microsoft SQL Server.
* Supported **machine learning pipeline development** using Python for predictive analytics.

**Environments: Databricks, Spark, Azure, Azure Databricks, Azure Data Factory, Delta Lake, Apache Iceberg, TSQL, SQL, ETL, Data Lake, Data Warehouse, Distributed Computing, Medallion Architecture, Delta Live Tables (DLT), Data Governance, Cloud Computing, Batch and Real-time Data Processing, Snowflake, DBT, On Prem Windows servers,  Python, JAVA, PL/SQL scripts, Oracle Apps, Excel, IBM SPSS, Tableau, Big Data, HDFS, Sqoop, Pig, Flume, Hive, MapReduce, HDFS, SQL, Pandas, NumPy, Matplotlib, Seaborn, ETL, SSIS, SQL Server, Windows.**

**Client: BCLC, BC, Canada Mar 2011 – Apr 2015**

**Role: Senior Data Engineer**

**Responsibilities:**

* Designed, implemented, and optimized **ETL pipelines using Talend** for the seamless extraction, transformation, and loading of financial and transactional data with minimal latency across downstream systems.
* Led **data quality assurance** efforts through cleansing, validation, and anomaly detection to ensure high-integrity transactional data for reporting and regulatory compliance.
* Assisted in the design and performance optimization of **SQL Server database structures**, including indexing and tuning for large-scale financial datasets.
* Built and maintained **Python automation scripts** for ETL processes, significantly reducing manual effort and improving reliability in daily operations.
* Processed high-volume unstructured data using **Hadoop and Hive**, enhancing distributed data analytics capabilities to support reporting and business intelligence.
* Supported **external data integration** by transforming diverse formats (e.g., XML) into structured, consistent formats for ingestion into the bank’s infrastructure.
* Leveraged **Git** for version control and **ANT** for deploying and monitoring ETL jobs, ensuring reliability and accountability across development and production workflows.
* Created **Power BI dashboards and reports** to deliver actionable insights to executives, including metrics on loan portfolios, risk scores, and transaction trends.
* Collaborated with infrastructure teams to maintain **Linux-based data servers**, monitored job scheduling, and resolved system alerts to ensure uptime and performance.
* Supported **critical data migrations**, ensuring integrity and completeness during transitions between legacy and modern platforms.
* Ensured alignment with **data privacy, compliance, and security standards**, working alongside governance and audit teams.
* Led the **enterprise data warehouse setup** for Slot Data Systems and Casino Player Management Systems across multiple environments—enabling real-time player promotions and slot performance tracking.
* Developed **summarized reporting solutions** using SSIS for consolidated multi-slot system data, reducing manual reporting workload for the BCLC accounting team by over 20 hours per day.
* Engineered **rCasino integration** between Scientific Games and BCLC platforms, enabling system interoperability and consistent functionality across product updates.
* Spearheaded **SG product compatibility testing** in a simulated lab environment—directly contributing to $100M+ in product deals between Scientific Games and BCLC by validating end-to-end functionality.

**Environment: Python, JAVA, SQL, SQL Server, Talend, Hadoop, Hive, Power BI, ANT, Git, XML, Linux, Casino Accounting & Financial Systems, Window**

**Client: Neilson**-**TCS, Chennai, India Nov 2007 – Feb 2011**

**Role: Data Analyst**

**Responsibilities:**

* Led multiple client-facing data projects with a strong focus on **Python, SQL, Tableau**, and **data modeling**—delivering end-to-end solutions from data ingestion to executive dashboards.
* Developed and maintained robust **ETL pipelines**, including data merging, transformation, and quality control, using Python scripting and SQL for generating structured CSV outputs.
* Created **source-to-target mapping** documents detailing transformation logic, schema relationships, and business rules to ensure traceable and auditable data processes.
* Wrote advanced **Python scripts** to parse, clean, and extract information from structured and semi-structured file formats (HTML, Excel, PDF, Word), enabling weekly data automation workflows.
* Analyzed large datasets from internal databases to uncover optimization opportunities in product development, marketing, and business strategy.
* Built and deployed **interactive dashboards** in Tableau for stakeholders, offering real-time insights into operational metrics and survey tracking.
* Extracted, loaded, and validated data within **data warehouse environments**, ensuring consistency across reporting layers.
* Participated in **data modeling discussions**, contributing to both logical and physical schema design for long-term scalability and clarity.
* Developed **data mapping documentation** and a master data workbook that defined ETL requirements, transformation logic, and physical metadata descriptions.
* Reviewed performance test results and identified bottlenecks, leading system performance tuning and optimization of long-running or resource-heavy queries.
* Designed and implemented **Neilson's Calculation Engine** for television rating predictions, processing **30+ TB of data daily**, and delivering industry-critical metrics with high accuracy.
* Refactored the **QA team’s automation testing suite**, reducing processing time from 12+ hours to 30 minutes by applying **Java multiprocessing techniques**—enabling daily test cycles and faster releases.
* Designed and developed **MySMS**, an internal **emergency text notification system** for TCS, allowing secure and immediate communication with employees.
* Contributed to the development of the **TCS hiring suite**, which dynamically generated randomized test questions with consistent difficulty, improving candidate evaluation and test integrity.

**Environment: Oracle 10g, UNIX Shell Scripts, MS Excel, MS Power Point, Python, JAVA, SQL**, **Eclipse**