Jahnavi Dayaneni

**Data Engineer**

# OBJECTIVE:

Seeking an opportunity to work as a data engineer utilizing analytical & methodical skills and relevant expertise to help the company achieve business goals. Results-driven Data Engineer with over 10 years of experience in designing, developing, and maintaining robust data pipelines and architectures across various industries including e-commerce, finance, and education. Expertise in data modeling, data quality assurance, data governance, and data visualization techniques. Passionate about tackling complex data challenges and driving tangible business outcomes through innovative data-driven strategies.

**SUMMARY:**

* Proficient in database management, Azure Data Platform services administration, and constructing data lakes.
* Skilled in crafting Spark applications employing RDD transformations, Spark core functionalities, Spark MLlib, Spark Streaming, and Spark SQL. Proficient in parsing various data formats on HDFS using Scala.
* Experienced in navigating cloud environments such as AWS EC2 and S3.
* Proficient in developing Data Models and employing Dimensional Modeling techniques encompassing 3NF, Star, and Snowflake schemas for OLAP and ODS applications.
* Demonstrated proficiency, adeptness in crafting stored procedures and functions in SQL and PL/SQL, and expertise in Version Control systems such as GIT.
* Hands-on experience with AWS Lambda and Azure Data Factory alongside data Catalog for data ingestion and upkeep of data sources.
* Proficient in ETL processes, Data Migration, Data Quality (DQ) assurance, Data profiling, Data Modeling, OLAP, OLTP, RDBMS, and Reporting.
* Skilled in Matillion ETL, Apache Druid, and Adobe Experience Platform for data workflows.
* Specialist in Informatica Cloud, Alteryx, and IICS ETL Development.
* Experienced in Oracle, AWS Redshift, Azure Synapse Analytics, Google BigQuery for cloud data warehousing.
* Skilled in DBT, SQL, Python, and data modeling for analytics-ready datasets.
* Hands-on experience in Azure development and Azure Data Factory.
* Proficient in Cloudera ecosystem tools and NoSQL technologies.
* Experienced in Production Support and database administration.
* Proficient in SQL Server, SSIS, SSRS, SSAS for BI solutions.
* Around 8 years of Software Life Cycle experience.
* Extensive use of Informatica PowerCenter and IDQ for ETL processes.

# TECHNICAL SKILLS:

|  |  |
| --- | --- |
| Databases | Oracle 11g, 10g, 9i, SQL Server 2012/2008R2, Hadoop, HIVE, MongoDB, and MySQL, Netezza, SAS, Hadoop, U-SQL, Snowflake |
| Languages | SQL, PL/SQL, TSQL, Python (matplotlib, Seaborn, NumPy, SciPy, SciKit-Learn), Power BI, R-Studio, Alteryx, SSRS (SQL Server Reporting Service), HTML, XML, Nodejs, UNIX, JAVASCRIPT |
| Analytics & Reporting | Tableau, SSRS, Cognos, Power BI, Crystal Reports, Data bricks |
| ETL & Cloud Platform | Informatica 10, SQL Server DTS, Visual Studio 2012/2010, Azure Data Lake & Factory, Pandas, AWS EC2, S3, Lambda, NumPy, Salesforce CRM, Apex Data Loader, SAP MDM |
| BI Tools | SSIS, BI Development Studio, Power BI, Visual Studio 2012, Performance Monitor, Power Pivot, Spark, Scala, Kafka, DevOps |
| Machine Learning | Linear Regression, Logistic Regression, LDA, PCA (Principal Component Analysis), K-Means, Clustering, K-Nearest Neighbors (KNN), Decision Tree, Ada Boosting, Gradient Boosting Trees, Neural Networks. |
| Design Tools | ERWIN, ER Studio, MS VISIO |
| Version Control Systems | SVN, Git, GitHub |

# PROFESSIONAL WORK EXPERIENCE:

## Citicorp Credit Services INC. (Irving,TX) Jun’19- Present

* Facilitated the transformation of data from legacy tables to Hive, HBase tables, and S3 buckets, enabling business and Data scientists to conduct analytics effectively.
* Automated ETL processes using Python scripts via Apache Airflow and CRON scripts in the UNIX operating system, enhancing operational efficiency.
* Utilized Alteryx software for ETL, data preparation for EDA, and performing spatial and predictive analytics.
* Designed and developed data integration programs in a Hadoop environment utilizing NoSQL, Oracle data store Cassandra for efficient data access and analysis.
* Transferred SQL databases to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Azure Data Bricks, and Azure SQL Data warehouse.
* Optimized Matillion jobs to improve performance and scalability, thereby enhancing system efficiency.
* Proficient in Informatica Power Center for designing, developing, and implementing ETL processes.
* Implemented and delivered MSBI platform solutions for ETL, analytical, reporting, and scorecard/dashboards on SQL Server using SSIS, SSRS.
* Actively participated in Kubernetes community meetups and forums to stay updated on the latest developments and best practices.
* Designed and implemented data models using PI Asset Framework for organizing and contextualizing plant data.
* Built and maintained data pipelines using Apache Spark for processing and analyzing large-scale datasets.
* Assisted in the development and deployment of data pipelines and integrations on Adobe Experience Platform, utilizing Adobe I/O APIs and SDKs.
* Created and maintained documentation for data mappings, transformation rules, and ETL processes to facilitate knowledge transfer and support ongoing maintenance.
* Implemented real-time and batch data integration solutions using Alteryx, IICS RESTful APIs, connectors, and transformations.
* Conducted infrastructure audits and implemented security controls using Azure Security Center and Terraform.
* Migrated data from Transactional source systems to Redshift data warehouse using Jira, Spark, and EMR.
* Developed data ingestion scripts and ETL processes to extract, transform, and load data from various sources into the data lake.
* Developed and deployed Oracle, Spark, JavaScript jobs in different environments and loaded data to SQL, AWS database Cassandra/Hive/HDFS, ensuring data security through encryption.
* Developed Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats.
* Designed and developed data models and schemas using Snowflake for data warehousing and analytics purposes.
* Designed and implemented data ingestion pipelines using Snowflake's native connectors and third-party tools.
* Leveraged Data Mining solutions to address various business problems and generated data visualizations using Tableau, Power BI, Alteryx.
* Provided mentorship to junior developers, offering guidance and support in Java programming concepts and best practices.
* Executed end-to-end delivery of PySpark, ETL pipelines to transform data orchestrated via automation accounts and triggered using Tidal Schedule.
* Developed JSON Scripts for deploying the Pipeline to process data using SQL, AWS Activity.
* Developed and maintained DBT models to transform raw data into structured datasets for analytics and reporting purposes, ensuring data accuracy and reliability.
* Engaged in the Open source community as an ETL follower, Databricks Enthusiast, and Cloud Adoption & Data Engineering enthusiast.
* Migrated existing data from Teradata/SQL Server to Hadoop and performed ETL operations.
* Utilized SSIS, NIFI, Python scripts, and Spark Applications for ETL Operations to create data flow pipelines.
* Stored final computation results in Cassandra tables and used Spark-SQL and spark-dataset for data computation.

## Bayer (Missouri) Jun’18- May’19

## Responsibilities:

## • Developed data pipelines using Spark, Hive, and HBase to ingest customer behavioral data and financial histories into a Hadoop cluster for analysis.

## • Utilized Azure Databricks to organize data into notebooks and create visualizations using dashboards.

## • Performed ETL on diverse data sources to Azure Data Storage using Azure Data Factory, T-SQL, Spark SQL, and Azure Data Lake Analytics, facilitating ingestion into Azure Databricks.

## • Managed Spark Databricks clusters, troubleshooting issues and monitoring performance.

## • Conducted data aggregation, validation, and Azure HDInsight operations using Python-based Spark scripts.

## • Implemented Azure Stream Analytics for processing real-time Geo-Spatial data for targeted sales campaigns.

## • Identified and resolved performance bottlenecks, optimized indexes, and implemented caching strategies.

## • Implemented best practices for data visualization design and usability, ensuring clarity and accuracy.

## • Migrated data to PostgreSQL from other platforms using tools like pg\_dump and pg\_restore, ensuring integrity.

## • Implemented change management processes to promote data governance practices, providing training and support to stakeholders.

## • Developed advanced calculations in Tableau using calculated fields and scripting languages.

## • Designed efficient database schemas and data models adhering to best practices and business requirements.

## • Utilized data governance tools such as Collibra and Informatica Axon to automate governance processes.

## • Managed Hadoop clusters using Azure HDInsight and designed data models using Azure Snowflake Data Warehouse.

## • Utilized Hive, Impala, and Sqoop utilities for data extraction and loading.

## • Developed Shell and Python scripts for batch processing and XML data parsing.

## • Developed backend services in Scala with Akka for concurrency models and scalability.

## • Collaborated on microservices-based solutions using Scala and Play Framework for RESTful API development.

## • Optimized Scala codebase performance and integrated Nifi with Snowflake for optimization.

## • Conducted audits of data governance processes and compliance to identify gaps and risks.

## • Developed stored procedures for data import into Elasticsearch.

## • Processed structured data using Spark SQL for analysis.

## • Created HBase tables to store data from various sources.

## • Imported log files into HDFS using Flume and developed routines for data quality.

## • Installed and configured Hadoop, maintained clusters, and managed log files.

## • Wrote PySpark scripts to automate batch data processing tasks.

## • Orchestrated Spark jobs using Apache Airflow or Oozie.

## • Managed PostgreSQL upgrades and patches in a production environment.

## • Performed troubleshooting, error handling, and query optimization.

## • Utilized SAS Visual Analytics for data presentation and reporting.

## • Parameterized reports using SAS Macros for user customization.

## • Leveraged PostGIS in PostgreSQL for geospatial data storage and analysis.

## • Developed data warehouse models in Snowflake using WhereScape.

## • Translated business requirements into logical data models.

## • Provided architecture and design guidance for Big Data Analytics solutions, implementing Azure Data solutions.

## BNY Mellon (New Jersey) Sep’17- May’’18

## Responsibilities:

## • Developed and presented models to fund managers, yielding a 20% improvement in returns compared to historical performance, and outperforming traditional stock price predictions by 25%.

## • Conducted User Acceptance Testing, identifying key performance indicators in reference data to anticipate client payment failures.

## • Utilized machine learning algorithms such as linear regression, KNN, and decision trees for trading challenges, assessing their efficacy on time series data (e.g., stock prices).

## • Produced data visualization reports showcasing daily and cumulative returns, simple moving averages, Sharpe ratio, and portfolio value to optimize stock performance.

## • Analyzed extensive datasets using pandas, implemented regression models with SciPy for future data prediction, and employed SQL for data manipulation.

## • Proficient in handling large data imports/exports to Teradata, crafting Teradata-specific physical data models including primary, secondary, and joined indexes.

## • Implemented Dimensional Data Modeling to deliver Multi-Dimensional STAR schemas, and normalized dimension tables for Snowflake Schemas.

## • Collaborated with data compliance and governance teams to maintain data models, metadata, and dictionaries, and define source fields and their definitions.

## • Leveraged HIVE for transformations, event joins, and pre-aggregations prior to storing data on HDFS.

## • Utilized Redshift and S3 within AWS, alongside Informatica Cloud Services, for loading data into S3 buckets.

## • Worked with various HDFS file formats (e.g., Avro, Sequence File) and compression formats (e.g., Snappy).

## • Partnered with Data engineers and operations teams to implement ETL processes, Snowflake models, and optimize SQL queries for data extraction.

## • Developed SQL and PL/SQL scripts for data extraction and testing purposes.

## • Engineered a data pipeline package to transfer data from Amazon S3 buckets to MySQL databases, executing MySQL stored procedures via events to populate tables.

## • Designed and implemented Tableau graphical and visualization solutions tailored to business requirements, collaborating closely with backend data retrieval and data mart teams.

## • Managed metadata in Tableau Desktop, adjusting data types, roles, field names, joins, data refreshes, and extraction options to prepare data sources for user utilization.

## Innoart tech pvt ltd (Hyderabad, India) Jun’14- Dec’16

## Responsibilities:

## • Created Python scripts to streamline ETL procedures and analyze data, seamlessly integrating SQL for effective data manipulation.

## • Proficient in Google Cloud Big Data Technologies including Data Proc, Data Flow, Big Query, GCP Storage, and pub sub.

## • Demonstrated expertise in GCP services such as compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring, and cloud deployment manager.

## • Oversaw MySQL and PostgreSQL databases, ensuring data integrity, security, and optimal performance.

## • Utilized Snowflake for data warehousing, prioritizing efficient data loading and querying.

## • Developed and optimized machine learning models using TensorFlow, Keras, PyTorch, and Scikit-Learn for predictive analytics and pattern recognition.

## • Conducted comprehensive data analysis with Pandas and NumPy, crafting visualizations through Seaborn and Matplotlib to drive informed decision-making.

## • Established GCP firewall rules to manage traffic flow to and from VM instances based on specified configurations.

## • Proficient in data analysis, visualization, and reporting in Excel, with automated task execution using macros.

## • Orchestrated ETL processes on AWS Glue, Lambda, and S3 to streamline data integration and storage.

## • Engineered web applications with Django, implementing RESTful APIs and user-friendly interfaces.

## • Designed and executed NoSQL solutions leveraging GCP's Cloud Bigtable for handling large-scale data workloads.

## • Implemented PySpark for processing big data, optimizing Spark jobs to enhance performance and scalability.

## • Developed and deployed Spark code on Hadoop clusters running on GCP.

## • Collaborated with cross-functional teams to gather data and business requirements, ensuring alignment with project objectives.

## • Designed and implemented data solutions for migrating source data from Data Warehouse to Atlas Data Lake, enhancing data accessibility.

## • Led end-to-end delivery of PySpark on ETL pipelines on GCP data bricks.

## • Conducted data validation using SQL and HIVE scripts to uphold data quality across various applications.

## • Utilized Cognos report validation to maintain accurate reporting and data consistency.

## • Deployed applications on AWS, configuring services like EC2, S3, and RDS for high availability and fault tolerance.

## • Leveraged GCP's BigQuery for real-time analytics and data warehousing, enhancing decision-making processes.

## • Managed AWS S3 buckets for scalable and cost-effective data storage, ensuring data security.

## • Developed Lambda functions for data aggregation and storage in DynamoDB and S3.

## • Ensured data governance and security on AWS cloud platforms, adhering to industry standards.

## • Contributed to successful database migrations from Solaris to OEL, coordinating seamlessly across multiple teams.

## • Skilled in GCP Dataproc, GCS, Cloud Functions, and BigQuery.

## • Mitigated business risk through meticulous documentation, integration, and interdependency tracking to ensure project success.

## • Optimized workload management for load balancing and failover, improving system performance and scalability.

## • Provided architectural leadership to support data-driven decision-making, enhance technical deliverables' quality, and streamline data engineering processes.

# EDUCATION:

**Rutgers-State University of New Jersey Feb’ 17- Feb ’20 Candidate for Master of Science in Information Systems**

**Osmania university (Hyderabad, India) May ’14**

**Bachelor of Computer Science**