**Pavan Sai Prasanth Sabnaveesu**

**Email: xx Contact: xx**

**LinkedIn:** [**https://www.linkedin.com/in/prasanth-sabnaveesu/**](https://www.linkedin.com/in/prasanth-sabnaveesu/)

***Senior Data Engineer***

**BACKGROUND SUMMARY:**

With over a decade of experience as a Data Engineer, my background includes traditional data engineering expertise in Apache Spark, Kafka, Hadoop, and various relational and NoSQL databases like SQL Server, Teradata, MongoDB, and Cassandra. I am skilled in data modeling, physical data warehouse design, and cloud data warehousing technologies such as Snowflake and Redshift. My experience spans major cloud providers like AWS, Azure, and GCP, where I have designed and orchestrated ETL pipelines using tools like Airflow and Azure Data Factory. I have hands-on experience with a wide range of AWS, Azure, and GCP services, as well as Databricks. Additionally, I have extensive experience in IT data analytics projects, migrating on-premises data to the cloud, and working with ETL tools like Informatica and SSIS. My expertise also includes data governance, master data management, and business intelligence, making me well-equipped to provide solutions to complex business problems. A proven track record of optimizing data processes and ensuring data quality has significantly improved decision-making and operational efficiency. My ability to interpret complex problems and provide innovative solutions using data analysis, data mining, and machine learning techniques sets me apart. Committed to leveraging my skills to drive data-driven strategies and deliver impactful results, I have helped businesses reduce costs, enhance scalability, and achieve greater agility. My strategic approach to data management has enabled organizations to unlock valuable insights, foster innovation, and maintain a competitive edge in the market. Furthermore, proficiency in cloud migration and data integration has streamlined operations and accelerated time-to-value for numerous clients. Dedication to continuous improvement and staying abreast of the latest industry trends ensures the delivery of cutting-edge solutions that drive business growth and success.

**TECHNICAL SKILLS:**

**Cloud Services** :Azure (Data Lake Storage Gen 2, Data Factory, Blob storage, SQL DB, Databricks, Event Hubs), AWS (EMR ,RDS and SNS, SQS S3, AWS, Lambda), Data Flow, VM, Delta Tables, Cloud functions, Clusters.

**Databases** : MySQL, SQL Server, IBM DB2, Postgres, Oracle, MS Access, Teradata, Big Query, Snowflake.

**NoSQL Databases** : MongoDB, Cassandra DB, HBase.

**Development Strategies** : Agile, Lean Agile, Pair Programming, Waterfall, Test-Driven

**ETL, Visualization & Reporting** : Tableau, Data Stage, Informatica, Talend, SSIS, SSRS.

**Frameworks** : Django, Pandas, NumPy, Matplotlib, TensorFlow, PyTorch.

**VC &OS & Containerization** : Jenkins, Git, CircleCI, SVN, Unix, Linux, Windows, Mac OS

**Monitoring Tools** : Apache Airflow, Control M.

**Tools** : PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Postman.

**Machine Learning Techniques** : Linear & Logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP, Clustering.

**WORK EXPERIENCE:**

**Client: Verisk, Jersey City, NJ| Azure Data Developer Duration: Dec 2023 – Present**

**Role & Responsibilities:**

* Developed optimized ETL pipelines in Azure Data Factory & Databricks by implementing delta processing, parallel execution, and workload partitioning, improving data ingestion speed by 40%.
* Led the migration of a 20TB on-prem data warehouse to Azure by designing an incremental data movement strategy using Azure Data Factory (ADF) & PolyBase, ensuring zero data loss.
* Designed and implemented a scalable Data Vault 2.0 model using Azure Synapse & SQL Server, standardizing data governance & lineage tracking while enhancing query performance by 50%.
* Developed highly efficient Ab Initio ETL workflows by restructuring data transformations into parallel jobs, optimizing memory allocation, and minimizing I/O operations, reducing processing time by 35%.
* Built a real-time streaming analytics pipeline using Azure Event Hubs & Stream Analytics, leveraging windowing functions & event-time processing to enable instant anomaly detection for IoT data.
* Established a centralized MDM framework by implementing data duplication, golden record management, and match-merge logic in Azure SQL DB, achieving 99.9% data consistency.
* Integrated Unity Catalog with Azure Data Lake by defining fine-grained access controls & metadata tagging, ensuring secure and efficient asset retrieval for ML-driven analytics.
* Implemented a parameterized ETL framework in Databricks, dynamically managing configurations, schema evolution, and job orchestration, increasing developer efficiency by 30%.
* Automated daily data ingestion from web APIs to Azure SQL DB by implementing REST API integration, incremental data loading, and error handling mechanisms, reducing manual intervention by 80%.
* Developed a GitHub Actions CI/CD pipeline for Azure Data Factory & Databricks, automating build validation, deployment, and rollback strategies, cutting release cycle time by 50%.
* Engineered a Python-based reconciliation framework in Azure Databricks, comparing multi-source data with checksum validation & anomaly detection, improving data accuracy by 99.5%.
* Optimized SQL queries by refactoring stored procedures, indexing strategies, and query parallelization, reducing OLAP query execution time by 50% in Azure Synapse.
* Designed an event-driven workflow using Azure Functions & Event Grid, triggering real-time ETL executions based on source data changes, enhancing data freshness by 70%.
* Developed dynamic SSIS packages with parameterized connections, error logging, and restart ability, streamlining data warehouse loads and reducing job failures by 40%.
* Configured Azure Monitor & Log Analytics by embedding custom telemetry & alerting logic into ETL jobs, improving failure detection rate by 70% and enabling proactive issue resolution.
* Implemented a tiered storage strategy in Azure Data Lake by leveraging lifecycle policies & auto-archiving, reducing storage costs by 25% while retaining high-speed access for frequently used data.
* Developed an incremental ingestion framework in Databricks Delta Lake, using change data capture (CDC) and MERGE operations, removed unnecessary reprocessing and reduced costs by 60%.
* Created a Power BI dashboard optimized with DAX performance tuning, query folding, and aggregated tables, reducing dashboard refresh time from 10 mins to 3 secs.
* Implemented a metadata-driven ingestion pipeline in ADF, dynamically configuring source-to-target mappings using parameterized templates, enhancing pipeline scalability and maintainability.
* Designed and deployed an Azure Databricks-based data masking solution, integrating row-level security, tokenization, and audit logging, ensuring GDPR & HIPAA compliance.
* Automated file validation workflows in Databricks using Python, schema validation, anomaly detection, and checksum verification, eliminating manual errors and increasing ETL reliability by 99%.
* Optimized large-scale Spark transformations by leveraging adaptive query execution, broadcast joins, and data partitioning strategies, improving data processing speed by 3x.
* Implemented an Azure Data Lake partitioning strategy based on date-based hierarchical storage, reducing query latency by 65% for high-volume analytical workloads.
* Developed a data lineage tracking system in Azure Purview, integrating custom metadata tagging and impact analysis, ensuring end-to-end traceability for compliance audits.
* Enhanced Azure HDInsight & Cosmos DB-based solutions by implementing auto-scaling, caching policies, and workload isolation, reducing computing costs by 30% while improving data processing throughput.

**Client:** **Citizen's Bank, Johnston, RI | Senior Data Engineer Duration: Oct 2020 – Nov 2023**

**Role & Responsibilities:**

* Developed high-performance Spark applications using PySpark & Spark-SQL, implementing lazy evaluation & partitioning strategies, reducing data processing time by 40%.
* Owned the end-to-end user journey data ingestion at CDK Global on AWS, leveraging AWS Glue, S3 & EMR, optimizing pipeline performance by 35%.
* Designed a Unified Data Processing Layer in Spark on AWS EMR, implementing schema evolution & metadata-driven ingestion, reducing onboarding time from 3 days to 1 day.
* Implemented incremental delta loads in Databricks Delta Lake, utilizing Z-Ordering & Change Data Capture (CDC), optimizing processing time by 60% and costs by 53%.
* Built real-time data pipelines using Spark Streaming & Kafka, applying windowing & event-time processing, reducing data latency by 50% for real-time analytics.
* Automated campaign data ingestion in AWS Glue, using parameterized ETL jobs & S3 event triggers, reducing manual intervention by 80% and improving data availability.
* Developed CI/CD pipelines in CircleCI & Git, integrating unit tests, rollback strategies, and infrastructure as code (IaC), cutting deployment failures by 40%.
* Optimized asset retrieval in Unity Catalog, leveraging AWS CloudFront cache & IAM-based access controls, reducing query latency by 65% while ensuring data governance.
* Engineered a real-time DynamoDB integration using AWS Lambda & DynamoDB Streams, implementing trigger-based updates, improving data consistency by 99.9%.
* Automated ingestion scripts in Python & Scala, integrating APIs, AWS S3, Teradata, and Redshift, reducing ETL runtime by 45% through parallel execution.

**Environment:** Python, Spark, AWS EC2, AWS S3, AWS EMR, AWS Redshift, AWS Glue, AWS RDS, AWS SNS, AWS SQS, AWS Athena, Snowflake, Data warehouse, Airflow, Data Governance, Kafka, ETL, Terraform, Docker, SQL, Tableau, Git, REST, Bitbucket, Jira.

**Client: Apple, Sunnyvale, CA | Data Engineer Duration: Jan 2019 – Aug 2020**

**Role & Responsibilities:**

* Migrated on-premises data to AWS S3 by developing Python scripts and AWS SDKs, improving data transfer efficiency by 40% and reducing manual intervention.
* Built ETL pipelines leveraging AWS Lambda, Glue, and Step Functions, automating data cleansing, transformation, and ingestion, reducing processing time by 50%.
* Developed API integration scripts in Python to extract and load real-time data from REST APIs to AWS S3, improving data availability for analytics.
* Designed YAML-based configurations for AWS Glue table stack creation, enabling seamless infrastructure deployment and standardization across data sources.
* Automated ETL workflows by integrating Lambda functions with S3, SQS, SNS, and EventBridge, ensuring event-driven data processing with 99.9% reliability.
* Optimized data migration from Netezza to AWS S3 using Python, reducing data extraction time by 30% and ensuring seamless integration.
* Implemented Spark-based ETL processes using PySpark and Spark SQL for large-scale data transformation, improving performance by 45%.
* Developed real-time data streaming pipelines using Kafka, Spark Streaming, and NoSQL (HBase, Cassandra), enabling low-latency analytics and processing.
* Designed and implemented Sqoop jobs for incremental data ingestion from DB2 to Hive, enabling seamless integration with Tableau for interactive reporting.
* Automated data workflows with Oozie and Apache NiFi, reducing manual intervention in HDFS data ingestion and streamlining processing for big data analytics.

**Environment:** Big Data, Hadoop, Oracle12c, PL/SQL, Scala, Spark-SQL, PySpark, Python, Kafka, SAS, MDM, Oozie, SSIS, T-SQL, ETL, HDFS, Cosmos, Pig, Sqoop, MS Access.

**Client: Ocelot, St. Louis, MO | ETL Developer Duration: Oct 2016 – Dec 2017**

**Role & Responsibilities:**

* Analyzed business requirements by collaborating with End Users & Business Analysts to define ETL strategies, ensuring accurate and efficient data integration.
* Designed and developed optimized ETL mappings in Informatica PowerCenter by leveraging Reusable Objects, Mapping Designer, and Transformation Developer, improving reusability by 40%.
* Built scalable ETL pipelines to extract data from flat files and Oracle, using incremental loading techniques to optimize data processing and reduce load times by 35%.
* Created Informatica workflows and sessions using Lookup, Aggregator, Joiner, Normalizer, Router, and Expression transformations, improving data transformation efficiency by 30%.
* Implemented ETL data extraction and loading processes from multiple sources (Flat Files, Excel, Oracle) into a staging schema using PowerCenter, ensuring 99.9% data accuracy.
* Developed SQL-based interfaces for automated data extraction at regular intervals, ensuring real-time business reporting and reducing manual efforts by 50%.
* Optimized ETL performance by tuning SQL queries, partitioning, and indexing strategies, reducing execution time by 60%.
* Automated Informatica workflows and dependencies using event-based scheduling, reducing ETL failures and improving pipeline reliability by 45%.
* Enhanced data processing by designing complex mappings with multi-source extraction and transformation, enabling faster decision-making and reducing latency by 25%.
* Designed and implemented incremental data loads using CDC (Change Data Capture) in Informatica, ensuring efficient delta processing and reducing data redundancy

**Environment:** Informatica Power Center 10.4/10.2, Oracle, Flat Files, SQL, and Windows.

**Client: LTI Mindtree, IND | SQL Developer**

**Duration: Aug 2014 – Sep 2016**

**Role & Responsibilities:**

* Designed and optimized database architecture by collaborating with front-end developers, ensuring seamless integration and performance improvements of 25%.
* Developed ETL pipelines using SSIS, leveraging transformations like For Loop Containers and Fuzzy Lookups to streamline data processing and improve data accuracy by 30%.
* Built dynamic SSRS reports (Parameterized, Caching, Sub-reports, Ad-hoc) by optimizing queries and using custom code for row color, visibility, and masking, improving report efficiency by 40%.
* Automated error handling and notifications in SSIS by implementing Event Handlers, reducing manual intervention by 50% and ensuring proactive issue resolution.
* Wrote optimized SQL queries for complex business reports by analyzing execution plans and tuning indexes, reducing query execution time by 60%.
* Redesigned and migrated legacy DTS packages to SSIS by refactoring transformations and optimizing execution flow, ensuring zero data loss and faster processing.
* Implemented ETL framework to extract, transform, and load data from multiple sources (SQL Server, Flat Files, XML, Excel) into a centralized Data Warehouse, improving data consistency by 35%.
* Enhanced OLAP cube performance by developing MDX scripts, optimizing aggregations, and tuning indexes, reducing report processing time by 50%.
* Migrated MS Access database to SQL Server using SSIS Bulk Insert and custom packages, automating workflows and scheduling jobs to eliminate manual dependency.
* Optimized stored procedures and indexing strategy using SQL Profiler and DETA, increasing overall database performance by 45% and reducing deadlocks.

**Environment:** MS SQL Server 2012, Visual Studio 2010, T-SQL, MS Excel, Microsoft SQL Server Integration Services (SSIS), Microsoft SQL Server Reporting Services (SSRS), Rally bug tracking and SVN.

**Education**

*Texas A&M University, Master of Science, Computer Science CGPA: 3.83/4.00*