PULLAMARAJU SAISANTHOSH

+15625258749 | pullamaraju411@gmail.com

SUMMARY

Data Engineer with over 5 years of specialized experience in architecting and implementing high-performance data solutions, consistently meeting and exceeding client delivery deadlines.

- Proficient in conducting OLTP/OLAP system analysis and E-R modeling, and in developing database schemas, including Star and Snowflake schemas, for use in relational, dimensional, and multidimensional modeling within Power BI.
- Skilled in utilizing AWS resources including EC2, S3, EBS, VPC, ELB, SNS, RDS, IAM, Route 53, Auto Scaling, CloudFormation, CloudWatch, and Security Groups.
- Extensive expertise and robust deployment experience in Hadoop and Big Data ecosystems, including HDFS, MapReduce, Spark, Sqoop, Hive, Kafka, Zookeeper, and HBase.
- Expertise in implementing and optimizing large-scale data processing workflows using Apache Spark, resulting in a reduction in processing time and enhanced data pipeline efficiency.
- Architected and administered robust data solutions utilizing Azure Cloud Services, including Azure Data Factory for ETL processes, Azure Synapse Analytics for data warehousing, and Azure Blob Storage for scalable storage, leading to an enhancement in data processing efficiency.
- Proficiency in structured data sets comprehension, data pipeline orchestration, utilization of ETL tools, and application of techniques for data reduction, transformation, and aggregation. Familiarity with advanced tools including DBT and DataStage.
- Crafted PySpark scripts within AWS Glue to amalgamate data sourced from disparate tables, while leveraging the Crawler functionality to automatically populate the AWS Glue Data Catalog with comprehensive metadata and table definitions.
- Skilled in Enterprise Data Modeler with extensive expertise in crafting Enterprise Data Models adhering strictly to Normalization Rules, and adept at constructing Enterprise Data Warehouses employing Kimball and Data Warehouse Methodologies.
- Designed and implemented interactive dashboards and visualizations in Power BI and Tableau to provide actionable insights into complex datasets, enhancing data-driven decision-making processes for stakeholders across the organization.
- Skillful in Data Analysis using SQL on Oracle, MS SQL Server, DB2, Teradata, and AWS.
- Expert in documenting the Business Requirements Document (BRD), generating the UAT Test Plans, maintaining the Traceability Matrix, and assisting in Post Implementation activities.

EDUCATION

California State University - Long Beach **August 2021 - May 2023**
Master's, Computer Science *GPA: 3.5*

CVR College of Engineering **August 2016 - May 2020**
Bachelor's, Computer Science *GPA: 8.5*

SKILLS

Cloud Computing Tools: Snowflake, SnowSQL, AWS, Databricks, GCP, Azure data lake services, Amazon EC2
Data Ingestion: Sqoop, Flume, NiFi, Kafka
Big Data Ecosystem: Hadoop, Spark, MapReduce, YARN, Hive, SparkSQL, Impala, HBase, Zookeeper, Avro, Parquet, Hue.
SQL Server Tools: SQL Server Management studio, Enterprise Manager, Query Analyzer, Profiler, Export and Import (DTS)
Scripting Languages: Python, R Programming, Bash scripting
Databases: Snowflake Cloud DB, Oracle, MySQL, Teradata 12/14, DB2 10.5, MS Access, SQL Server, Amazon RDS, Postgres SQL
Reporting / BI Tools: MS Excel, Tableau, Tableau Server and Reader, Power BI, QlikView, SAP Business objects, Looker, SSRS
Warehousing and Modelling/ Architect Tools: Erwin 7.3&9.5, ER/Studio, Rational System Architect, IBM Infosphere DA, dbt packages, SSIS, SSAS
ETL/ Data: Tensor flow, Data API, PySpark, DBT (Data Build Tool), Ab Initio, Informatica Power Center, Matillion, Talend, Alation, Pentaho, Palantir, Microsoft SSIS
Programming Languages: Python, R Programming, C, SQL, PL/SQL, Shell Scripts, XML, HTML, Visual Basic 6.0, mSAS
Hadoop Ecosystem/ Distributions: HDFS, MapReduce, Yarn, Oozie, Zookeeper, Job Tracker, Task Tracker, Name Node, Data Node, Cloudera
NoSQL Databases: HBase, Cassandra, MongoDB, CouchDB, Apache, Hadoop HBase

PROFESSIONAL EXPERIENCE

Teladoc Health - Providers **California, USA**

Data Engineer *May 2022 - Present*

- Proficient in leveraging NoSQL databases like HBase, MangoDB, accompanied by related ecosystems such as Zookeeper, Oozie, Impala, Storm, Spark (both Streaming and SQL), Kafka, and Flume. Demonstrated expertise resulted in a 25% increase in data processing efficiency and a 20% reduction in system maintenance overhead.
- Extensive expertise in Dimensional Data modeling encompassing Star Schema and Snowflake modeling methodologies, as well as the design and implementation of FACT and Dimension tables. Proficient in both physical and logical data modeling, utilizing tools such as ERWIN 3.x, Oracle Designer, and Data Integrator.
- Performing Extract, Transform, and Load (ETL) operations to transfer data from source systems to Azure Data Storage services. This involves utilizing Azure Data Factory, T-SQL, Spark SQL, and U-SQL within Azure Data Lake Analytics to orchestrate efficient data processing and storage in the Azure environment.
- Processed various file formats (CSV/TXT/AVRO/PARQUET) using Scala/Java in Spark, achieving a throughput of 10,000 records per minute. Utilized Spark Data Frames and RDDs, reducing processing time by 20%. Saved data in Parquet format in HDFS, achieving a compression ratio of 5:1 and reducing storage footprint by 40%.
- Led the design and development of ETL processes in AWS Glue, achieving a migration rate of 1TB of campaign data per day from external sources such as S3, ORC/Parquet/Text files into AWS Redshift.
- Leveraged Matillion's orchestration, transformation components, and parameterized variables to construct scalable and modular data workflows, processing over 1TB of data daily.
- Engaged in multiple Proof of Concepts (POCs) to integrate cutting-edge technologies such as Apache Airflow, Snowflake, and Terraform for infrastructure management. These efforts resulted in a 20% reduction in deployment time and a 30% improvement in resource utilization efficiency.
- Designed and deployed medium to large-scale Business Intelligence solutions on Azure, leveraging Azure Data Platform services including Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL Data Warehouse, HDInsight/Databricks, and NoSQL databases (MangoDB, HBase).
- Utilized Alation Data Catalog proficiently to capture and manage metadata, quantified by documenting data definitions, ownership, usage patterns, and lineage, resulting in a 20% increase in data accessibility and understanding across teams.
- Performed reduction in Spark job latency by 30% through fine-tuning Spark configurations and employing optimization techniques, resulting in faster data processing.
- Proficient in Docker containerization technology, quantified by successfully building, testing, and deploying system runtime environments, resulting in a 30% reduction in deployment time compared to traditional methods.
- Practiced in clarifying business requirements and performing gap analysis between goals and existing procedures/skills.
- Analyzed SQL scripts and devised a PySpark-based solution, achieving a 40% increase in data processing speed compared to conventional methods.

Thrymr Software **Hyderabad, TG, India**

Data Engineer *August 2019 - July 2021*

- Implemented JSON object encoding and decoding utilizing PySpark framework to manipulate data frames within Apache Spark, resulting in a 25% enhancement in data processing efficiency.
- Collaborated with project managers to conceptualize and design workflow architectures tailored to requirements, quantified by delivering a 15% increase in project efficiency. Also supported data scientists in feature engineering tasks, contributing to a 10% improvement in model accuracy.
- Utilized DBT adapters to streamline interactions with various data warehouses, ensuring customized SQL code generation and execution for specific target environments. This approach led to a 30% reduction in query execution time and a 25% increase in overall data processing efficiency.
- Designed and implemented intricate data integration pipelines spanning diverse sources, orchestrating seamless data transmission into Azure Synapse Analytics via Azure Data Factory. This initiative resulted in a 40% reduction in data transfer latency and a 30% improvement in pipeline reliability.
- Crafted data pipelines and workflows within Palantir Foundry, automating ingestion, transformation, and enrichment processes. This initiative boosted data analytics efficiency by 35% and enhanced accuracy by 25%.
- Applied Alation Data Catalog for acquiring and managing metadata information, covering data definitions, ownership attribution, usage patterns, and data lineage. This contributed to a 20% increase in data accessibility and a 15% improvement in data governance efficiency.
- Skilled in performing ETL operations and troubleshooting issues such as memory overruns on EMR clusters. This proficiency led to a 30% reduction in downtime and a 25% increase in overall cluster stability.
- Showcased adeptness in minimizing latency within Spark jobs to accelerate data processing through meticulous fine-tuning of Spark configurations and implementation of optimization strategies. These efforts resulted in a 40% reduction in processing time and a 20% enhancement in overall job efficiency.
- Proficient in utilizing Airflow for orchestrating job scheduling, including both routine tasks and ad-hoc manual operations.

- Demonstrated expertise in implementing Continuous Integration practices and automating deployment workflows through Git, Jenkins, and Docker, leveraging Python and Bash scripting for script development.
- Developed a dimensional logical model encompassing around 10 facts, 30 dimensions, and 500 attributes utilizing Erwin.

Kline **India**

Data Engineer *August 2018 - July 2019*

- Developed Data Mapping, Data Governance, Transformation, and cleansing rules for the Master Data Management Architecture involving OLTP, ODS, and OLAP.
- Engaged in performance tuning of databases, focusing on optimizing indexes and fine-tuning SQL statements. Achieved a 25% improvement in database query response time and a 20% increase in overall system efficiency.
- Developed tables, views, sequences, triggers, table spaces, and constraints, while also generating DDL scripts for physical implementation. Contributed to a 30% increase in database management efficiency and ensured seamless database structure maintenance.
- Produced mapping spreadsheets for the ETL team, detailing source-to-target data mappings adhering to physical naming standards, data types, volume metrics, domain definitions, and corporate metadata definitions. Streamlined data integration processes by 20% and ensured alignment with organizational standards.
- Established and upheld thorough data model documentation, encompassing detailed descriptions of business entities, attributes, and data relationships. Enhanced understanding of data structures by 25% and facilitated streamlined data analysis and reporting processes.
- Administered Git repositories within Azure DevOps, managing branching strategies, code reviews, and pull requests to facilitate collaborative development and version control.
- Established a robust security framework within Azure Synapse Analytics, implementing granular access controls, encryption mechanisms, and audit trails to safeguard sensitive data assets and ensure regulatory compliance.
- Maintain and work with our data pipeline that transfers and processes several terabytes of data using Spark, Python, Apache Kafka, Pig/Hive & Impala.
- Utilized Power BI to create various analytical dashboards that help business users to get a quick insight into the data.
- Worked with data compliance teams, Data governance team to maintain data models, Metadata, and Data Dictionaries; define source fields and their definitions.