**Prashanth Vedavally**

**Mobile: +469-860-6712| Email:-taniya@narveetech.com**

**SUMMARY**

* Around 10+ years of extensive IT experience as a data engineer with expertise in designing data-intensive applications using Hadoop Ecosystem and Big Data Analytical, Cloud Data engineering (AWS, GCP), Data Visualization, Data Warehouse, Reporting, and Data Quality solutions.
* Accomplished Data Engineer with extensive experience in data architecture and cloud computing.
* Proficient in designing and implementing data solutions on both AWS and GCP, enabling advanced data management.
* Expertise in ETL processes, data pipeline management, and successful execution of data migration projects with a focus on data integrity and minimal downtime.
* Strong commitment to data security and compliance, implementing robust access controls and encryption mechanisms to meet industry standards.
* Effective collaboration with cross-functional teams and clear communication with business stakeholders to align technical solutions with organizational objectives.
* Holds prestigious certifications, including AWS Certified Data Analytics - Specialty demonstrating dedication to staying current with industry trends.
* Proficiency in key programming languages like SQL and Python, coupled with data visualization skills using Tableau and Power BI, for extracting actionable insights from data.
* Proven track record in fraud prevention, process optimization, and continuous learning, contributing to enhanced operational efficiency and data-driven decision-making.
* Exceptional analytical and problem-solving skills, consistently delivering impactful data-driven solutions to address complex challenges.
* Expertise in data warehousing, including schema design (Star and Snowflake schemas) and efficient data management using AWS Glue, Redshift, and DataSync.
* Hands-on experience with GCP tools such as BigQuery, Data Proc, and Cloud Storage, further expanding capabilities.
* Proficiency in orchestration with Apache Airflow and data processing using Databricks, streamlining data workflows.
* Familiarity with version control using Jenkins and Docker for CI/CD pipelines, ensuring efficient development and deployment processes.
* A dynamic and adaptable Data Engineer capable of delivering sophisticated data solutions to drive organizational success.
* Committed to continuous professional growth and staying at the forefront of data engineering trends.
* Equipped with a diverse skill set and a history of successfully tackling complex data challenges.
* A valuable asset with a demonstrated ability to transform raw data into actionable intelligence, enabling data-driven decision-making within organizations.

**SKILLS**

|  |  |
| --- | --- |
| **Data Engineering with Cloud:** | Amazon Web Services (AWS), Google Cloud Services (GCP) |
| **Big Data Technologies:** | Hadoop (HDFS, MapReduce), Apache Kafka, Spark, & Airflow, Apache Hadoop YARN |
| **Cloud Computing:** | AWS (Amazon Redshift, S3, Glue, EC2, EMR), Snowflake, Google Cloud Platform (Google BigQuery, Cloud Storage, Dataproc, Dataflow) |
| **Programming Languages:** | Python, PySpark, SQL, UNIX |
| **Container Platforms:** | Docker, Kubernetes, Jenkins, CD/CI |
| **Databases:** | Hive, MySQL Server, PostgreSQL, Oracle 11g, SparkSQL, Mongo DB, Cassandra, MS Excel, MS Access |
| **Version Control:** | Git |
| **Software Tools:** | MS Visio, Jupyter, Tableau, Looker |
| **Operating Systems:** | Windows, Ubuntu, Linux, Unix, MAC |
| **Software Methodologies and Skills:** | Kibana ElasticSearch (Monitoring and Logging), Performance Tuning |

**PROFESSIONAL EXPERIENCE**

|  |  |
| --- | --- |
| Project Name | **BPPSL** |
| Client | **Southwest Airlines, Texas** |
| Role | Data Engineer |
| Organization | Singular Analysts Inc, Dallas, Texas |
| Duration | **Jan 2021** to **at Present** |

Project Description:

**Booking PNR Passenger Segment Leg (BPPSL)** Working on Data modernization from Ab Initio to AWS cloud project to process huge amount of data which was loading into Teradata currently that needs to be loaded into AWS redshift cloud Datawarehouse.

Carried out the following activities:

* Gathered and comprehended business requirements from the client.
* Designed source-to-target mappings for the data pipeline.
* Utilized PySpark to fetch data from an S3 bucket where daily XML files were deposited as a batch process.
* Established a Data Lake for the BPPSL application to retrieve raw data from the S3 bucket in .gz format.
* Transformed and loaded this data into the S3 bucket in the efficient Parquet format using AWS Glue.
* Developed a PySpark script framework tailored to process Parquet data from the S3 bucket.
* Implemented data transformations, including exploding and flattening, and applied business logic using Spark SQL and DataFrames.
* Created temporary views and cached DataFrames for efficient data processing.
* Prepared a final DataFrame for integration with Apache Hudi.
* Employed Apache Hudi to implement Change Data Capture (CDC) logic between the target Redshift database and the incoming data from the S3 bucket, enabling data insertion, updates, and deletions.
* Orchestrated data pipelines seamlessly using AWS Step Functions, a serverless orchestration service.
* Coordinated multiple Lambda functions within Step Functions to integrate, schedule, run, debug, and manage data pipeline changes as needed.
* Deployed objects and application code using version control systems, including Bitbucket, CI/CD Bamboo, and GitHub.
* Ensured data validation between Teradata and Redshift databases, verifying data consistency, and resolving any discrepancies.
* Addressed and resolved defects, including data mismatches identified during QA testing.

**Environment:** PySpark, AWS, S3, EMR, Step Functions, AWS Glue, Lambda, Redshift, Apache Hudi, Python, DynamoDB, Teradata, Unix

|  |  |
| --- | --- |
| Project Name | **Cloud Data Warehouse** |
| Client | **AT&T, Missouri** |
| Role | Data Engineer |
| Organization | Singular Analysts Inc. , Dallas, Texas |
| Duration | **Feb 2019** to **Dec 2020** |

Project Description:

**Cloud Data Warehouse (CDW)** Working on Cloud Data Warehouse Migration project to migrate the huge amount of data from Hadoop Hive On-Prem to GCP Big Query. As part of this project, we fetched the data from hive table and copy to GCS bucket using gsutil cp or Hadoop distcp and created the stage table. We did transformation, added a few business columns, and created BQ Target table.

Carried out the following activities:

* Understand business requirements from Business, created source to target mapping to design Data pipeline through PySpark, Python, GCP Dataproc cluster, GCS, BigQuery.
* Worked on Framework to extract the data from Teradata and Hive table and loaded into GCP Big Query through GCS bucket. It's a migration framework to load all the historical data from prem data to Cloud BigQuery.
* Deployed PySpark code on Dataproc cluster to run Cloud Data warehouse migration project.
* Designed spark job to pick GCS files and performs transformation and pushes to BQ Stage table and from BQ stage data are type casted and sent over to BQ Target tables.
* Designed batch/real-time pipelines leveraging Google Cloud Storage, Big Query, Airflow Cloud Composer DAGs, Python, DataFrames, Pandas SQL, Pub/Sub, Data Fusion, Dataflow, Vertex AI, Dataproc, Cloud Logging, Cloud Build, Cloud Run, DBT Orchestrator, Bigtable, Workflow and Looker.
* Scheduled the jobs on Airflow cloud composer - Orchestration Framework using Python, Airflow packages, GitLab for data processing pipeline.
* Designed Airflow Cloud composer DAGs with TaskGroup, XComs, Operators (Bash, Branch, Python, Google Clouse Storage, GCSToBigQueryOperator, BigQueryOperator, PostgresToGCSOperator, BigQueryInsertJobOperator, ExternalTaskSensor, S3ToGCSOperator, BashOperator, TriggerDagRunOperator, DataprocSubmitSparkSqlJobOperator, DataprocInstantiateWorkflowTemplateOperator, DataprocSubmitPysparkJobOperator, dataflow, DBT).
* Designed a python module, DAG to automate the process for create the historical views for source CDC tables in GCP BigQuery

**Environment**: PySpark, GCP, BigQuery, Dataproc, Airflow, Hadoop, Bigdata, Hive, Sqoop, Apache Airflow, Python, Teradata, Unix

|  |  |
| --- | --- |
| Project Name | **Marketing Data Insights Optimization Project** |
| Client | **Beacons Holdings Ltd, Hyderabad** |
| Role | Bigdata Developer |
| Duration | **Aug 2016** to **Dec 2018** |

Project Description:

**Marketing Data Insights Optimization Project (MDIOP)** In the Marketing Data Insights Optimization (MDIO) Project, our goal is to extract and analyze marketing data to make smarter decisions. We collect and transform data, creating visualizations to enhance our marketing strategies and improve customer experiences.

Carried out the following activities:

* Collaborated with cross-functional teams to gather and translate business requirements into technical specifications for marketing data analysis.
* Designed and implemented efficient data collection systems, leveraging AWS Virtual Machines Instances to ensure scalability and cost optimization.
* Executed end-to-end data transformation and loading (ETL) processes, facilitating the transfer of marketing data from OLTP systems to a Snowflake schema Data Warehouse.
* Developed secure data pipelines to extract and integrate marketing data from diverse sources, including on-premises and cloud, into the Snowflake data warehouse.
* Conducted extensive data analysis, including Exploratory Data Analysis (EDA) using Python's Matplotlib and Seaborn, to uncover patterns and correlations in marketing data.
* Employed advanced statistical techniques such as information value, principal components analysis, and Chi-square feature selection using Python to enhance marketing data insights.
* Implemented machine learning algorithms in Python, focusing on data prediction and forecasting to optimize marketing strategies.
* Produced real-time data visualizations and interactive dashboards within Jupyter Notebook using Python libraries, enabling the marketing team to access critical insights.
* Experimented with classification algorithms, including Random Forest Classifier and Gradient boosting, using Python's Scikit-Learn to refine customer discount strategies and marketing campaigns.
* Contributed to the development of data pipelines, collecting data from a variety of sources such as web, RDBMS, and NoSQL databases, into Apache Kafka or Spark clusters.
* Leveraged Spark and Spark-SQL/Streaming to accelerate data processing, optimizing marketing data analysis for enhanced decision-making.
* Generated data-driven reports and complex visualizations using Tableau, empowering the marketing team to interpret findings and drive data-based optimizations in marketing strategies.

**Environment:** Snowflake, Azure Storage accounts, Apache Spark, Python, Pandas, NumPy, ScikitLearn, SciPy, Seaborn, Matplotlib, SQL, Tableau, Jupyter Notebook, Plotly

|  |  |
| --- | --- |
| Project Name | **Google Cloud Data Transformation: Migrating On-Premise ETL to GCP** |
| Client | **SITEL, Hyderabad** |
| Role | Associate Software Engineer |
| Duration | **Nov 2014** to **Jul 2016** |

Project Description:

**Google Cloud Data Transformation: Migrating On-Premise ETL to GCP:** In this transformative migration project, we took a significant leap toward modernizing our data infrastructure by seamlessly transferring on-premise ETL processes to the highly efficient Google Cloud Platform (GCP). Our objective was to leverage GCP's native capabilities, including Google BigQuery, Cloud Dataproc, and Cloud Storage, to enhance data processing efficiency while achieving substantial cost savings.

Carried out the following activities:

* Spearheaded the migration of on-premise ETL processes to Google Cloud Platform (GCP), leveraging native tools such as Google BigQuery, Cloud Data Proc, and Google Cloud Storage. This migration resulted in substantial cost savings and improved data processing efficiency.
* Designed and developed data pipelines using Python, PySpark, and Hive SQL to ingest, transform, and load data from various sources into GCP data storage services, including Google Cloud Storage, BigQuery, and Cloud SQL.
* Implemented and managed data orchestration workflows using Apache Airflow, creating robust and efficient Python DAGs (Directed Acyclic Graphs) to automate data processing tasks and schedule data pipelines.
* Leveraged GCP's Unified Data Analytics capabilities with Databricks, using Databricks Workspace User Interface to collaborate on data analysis projects, manage notebooks, and optimize data processing workflows with various techniques like parquet formats, partitioning, and other custom data lake management tools.
* Worked extensively with Google BigQuery to perform complex SQL queries and analysis on large datasets, enabling the extraction of valuable insights and the creation of data visualizations.
* Collaborated with cross-functional teams to identify data requirements, design logical and physical data models, and implement data structures following best practices, including Star Schema and Snowflake Schema designs.
* Utilized Google Cloud Dataflow for real-time data processing, ensuring the timely ingestion and analysis of streaming data to provide actionable insights to stakeholders.
* Managed and monitored Google Cloud resources, including scaling, and troubleshooting of Databricks clusters, optimizing query performance, and ensuring high availability and fault tolerance of data processing systems.
* Implemented CI/CD pipelines using Jenkins and GCP deployment tools to automate the deployment of data pipelines and applications, enhancing the development and deployment processes.
* Conducted data validation, cleansing, and transformation activities, ensuring data accuracy and integrity by utilizing SSIS/SSRS packages and SQL scripts.
* Collaborated on the migration of an entire on-premise Oracle database to Google BigQuery and facilitated reporting using Tableau.
* Played a key role in estimating cluster sizing, monitoring, and troubleshooting of the Google Cloud-based data processing infrastructure.
* Continuously improved and optimized Spark-based data processing workflows, enhancing data availability, and reducing processing times.

**Environment:** Google Cloud Platform (GCP), Google BigQuery, Cloud Dataproc, Cloud Storage, Apache Airflow, Databricks Data Lake, Jenkins, Python, PySpark, HiveSQL, Apache Kafka, Tableau, SQL Server, Apache Dataflow (Apache Beam), Google Cloud Dataflow, Google Cloud Data Studio, Google Cloud Logging and Monitoring.

**EDUCATION**

**The University of North Texas at Denton**  **Jan 2019 – May 2020**

Master of Science in Data Science, GPA - 3.33/4

**Malla Reddy University Jun 2010 - May 2014**

Bachelor of Technology in Electronics and Communication Engineering, GPA – 7.0/10