**Abhishek Samal**

Sr Data Engineer

Ph.:+1 (469)-708-5060

Email: [7abhisheksam@gmail.com](mailto:7abhisheksam@gmail.com)

***SUMMARY***

* Senior Data Engineer with 9 years of IT experience including 5+ years in the Big Data and Analytics field in Storage, Querying, Processing, and Analysis for developing E2E Data pipelines in Cluster and Cloud Platforms with a wide range of Cloud services in Retail, Telecom and Financial domains.
* 3+ years of Google Cloud Platform experience.
* Expertise in designing and developing scalable Big Data solutions, and data warehouse models on large-scale distributed data, performing a wide range of analytics to measure service performance.
* Experience working with cloud providers - GCP.
* Experience in Guidewire suite (PolicyCenter, ClaimCenter, BillingCenter and Contact Center) integrations.
* Worked on Hive data analysis, and NoSQL databases such as HBase.
* Experience in migrating Hadoop data to Big Query in Google Cloud Environment
* Involved in documenting customization requirements, business rules and data migration requirements for the Implementation of Guidewire ClaimCenter and PolicyCenter
* Extensive experience in working with structured data using HQL, joining operations, writing custom UDFs, and experienced in optimizing Hive Queries.
* Hands on experience of Cisco IOS/IOS-XR/NX-OS, **Juniper JUNOS** for configuration troubleshooting of routing protocols: MP-BGP, OSPF, LDP, EIGRP, RIP, BGP v4.
* Experience in implementing and troubleshooting routing protocols **RIP, RIPv2, EIGRP, OSPF, ISIS and BGP** to avoid delays and congestion in network
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database.
* Experience developing large-scale batch and real-time data pipelines.
* Developed solutions using Spark SQL, Spark streaming, and Kafka to process web feeds and server logs.
* Extensive hands-on experience with Enterprise tools like Informatica Power Center and Pentaho.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and PySpark.
* Having 4 years of Experience using workflow management and scheduler tools like Apache Airflow, Oozie, Autosys, etc.
* Experience in Data collection, Data Extraction, Data Cleaning, Data Aggregation, Data Mining, Data validation, Data analysis, Reporting, and data warehousing environments.
* Expertise in troubleshooting, debugging, performance tuning, and optimization of slow-running ETL/ELT jobs using push-down optimization and partitioning techniques to manage large volumes of data.
* Experience in Data Modeling using Dimensional Data Modeling techniques like Star Schema and Snowflake Modeling.
* Possess in-depth knowledge of Database Concepts, Design of algorithms, SDLC, OLAP, OLTP, Data marts, and Data Lakes.
* Experience in all stages of the Software Development Lifecycle (SDLC) – Agile, Scrum, and Waterfall methodologies, right from Requirement analysis to development, testing, and deployment.

***Technical Skills***

|  |  |
| --- | --- |
| **Operating Systems** | UNIX, MAC OS X, Windows |
| **Cloud** | GCP (Data Pipelines, Data Flow, Data Proc, Cloud Composer, Big Query, Looker, Pub/sub) |
| **Programming Languages** | Python, Scala, SQL, PLSQL, Shell scripting |
| **Databases** | Oracle, PL/SQL, SQL Server, Teradata, PostgreSQL, MySQL, Hive, HBase, Informatica |
| **Big Data Eco System** | Hadoop HDFS, Pig, Hive, Sqoop, Zookeeper, Yarn, Spark, Storm, Impala, Flume, Kafka, HBase, PySpark, Airflow |
| **Microsoft tools** | Excel, Visio, Access, PowerPoint |
| **Version Control** | Git, Jira |

***Professional Experience***

***Client: IBM, TX Mar 2022- Current***

***Role: Sr Data Engineer with GCP***

***Responsibilities:***

* Designing and implementing scalable and efficient data processing pipelines using Google Cloud Platform (GCP) services such as Dataflow and Big Query.
* Experience in working with product teams to create various store-level metrics and supporting data pipelines written in GCP’s big data stack.
* Configured CIDR IP RIP, PPP, BGP and OSPF routing
* Experience in GCP Dataproc, Dataflow, PubSub, GCS, Cloud functions, BigQuery, Stackdriver, Cloud logging, IAM, Data studio for reporting, etc.
* Involved in the configuration troubleshooting of routing protocols: MP-BGP, OSPF, LDP, EIGRP, RIP, BGP v4. Configured IP access filter policies
* Development of Service oriented Integration patterns for Rhapsody to Support XDR, CCDA, EHR applications and Data Warehouse consumers.
* Process Harmonization for RPDM, VPDM systems - Team Center product
* Crafted solutions for integration and testing of Orion Confidential and Rhapsody Healthcare Integration engine to EPIC EHR and Bridges interfaces. HL7 and CDA data standards.
* Build a program with Python SDK with Apache beam framework and execute it in Cloud Dataflow to stream pub sub-messages into big query tables.
* Understand RPDM (Team Center) features to ENOVIA for ex-Alstom team.
* Worked on various Guidewire integration concepts like Messaging, Batch Processes, Web Services.
* Experience in deploying streaming maven build cloud dataflow jobs.
* Hands on experience with Guidewire PolicyCenter and ClaimCenter (Worker's Compensation Claims, General Claims) upgradation to v9.
* Vast experience in identifying production bugs in the data using stack driver logs in GCP.
* Experience in GCP Dataproc, GCS, Cloud functions, Cloud SQL & BigQuery.
* Used Cloud shell SDK in GCP to configure the services Data Proc, Storage, BigQuery.
* Extracting data from various sources, including databases, and file systems, using tools like Sqoop and designing ETL processes to import data into the data processing ecosystem.
* Transforming and enriching data using Hive for complex data manipulations, aggregations, and filtering, ensuring data quality and consistency.
* Loading transformed data into Big Query, leveraging its scalable and fully managed data warehousing capabilities, and optimizing data ingestion and storage strategies.
* Implementing workflow orchestration and scheduling using Apache Airflow to manage and monitor data processing tasks, dependencies, and performance.
* Optimizing data processing pipelines for performance and scalability, implementing partitioning, shading, and parallel processing techniques.
* Developed scripts using PySpark to push the data from GCP to the third-party vendors using their API framework.
* Involved in Guidewire Corporate Training on Guidewire PolicyCenter, Worked on Commercial Lines Property and Casualty Insurance
* Good knowledge of building data pipelines in airflow as a service (composer) using various operators.
* Build a program using Python and Apache beam to execute it in cloud Dataflow and to run Data validation jobs between raw source files and big query tables.
* Extensive use of cloud shell SDK in GCP to configure/deploy services like Cloud Dataproc (Managed Hadoop), Google Cloud Storage, and Cloud BigQuery.
* Created BigQuery jobs for loading the data into BigQuery tables from data files stored in Google Cloud storage daily.

**Environment:** Python, Spark, Big Query, DataProc, Dataflow, Logging and monitoring, Security, IAM, Hadoop, Hive, Guidewire, Teradata, SQL, Airflow, GCP,  Guidewire Billingcenter v8

***Client: BNSF, TX Sept 2020 –Feb 2022***

***Role: Data Engineer***

***Responsibilities:***

* Experience in migrating existing on-premises databases to Google Cloud Environment for better reporting experience.
* Extensive experience in Server infrastructure development on GCP Cloud by using GCP services like VM instances, Big Query, Pub/sub, Dataflow, Dataproc, Google Cloud Storage, Source repositories, Security, Log monitoring, Alert monitoring, IAM role management.
* Extensive knowledge of GCS (Google Cloud Storage) buckets and creating efficient buckets depending on the business requirement.
* Developed various shell scripts to get the monthly billing report, user access report for Big Query load and query tables in Big Query and to move the data into storage.
* Strong knowledge in writing SQL Queries that vary from single table simple SQLs to multiple-table complex SQLs.
* Mainly responsible for working in migration from legacy system to Guidewire system.
* Centralized the policy creation by integrating Guidewire PolicyCenter with Guidewire Claim Center Systems, Billing System and other existing system supporting operations
* Proficiency using TOAD or equivalent developer tools.
* Created and modified SQL, PL/SQL, and SQL Loader scripts for data conversions.
* Implemented the ETL pipelines using Dataflow, Dataproc, Cloud functions, and Big Query.
* Automated data and ETL- Batch pipelines using Airflow to interact with services like Dataproc, GCS, Big Query, and Hive.
* Training and Implementation of standardized CCDA interfaces.
* Planning and configuring the routing protocols such as OSPF, RIP, and static routing on the routers
* Tested complex RPDM ETL WFs and jobs for daily data loads.
* Troubleshooting complex networks layer 1, 2 frame relay, ATM, Point to Point, ISDN to layer 3 routing with MPLS, BGP, EIGRP, OSPF and RIP protocols technical issue
* Created C - CDA to Confidential parsing models
* Used ETL Data management tools such as RPDM and Neolene (Adobe).
* Involved in orchestrating and scheduling the jobs in Airflow and Autosys.
* Configured spark-submit configuration parameters as per the cluster in different environments.
* Optimized Spark code by using cache, re-partitioning, broadcast join, etc.
* Implemented retail application on the Cloudera Platform using Hive, Spark, and Scala Spark.
* Experience in creating and maintaining multiple projects in GCP based on the business requirement.
* Generated reports to get the user queries and access information and used the reports for GCP billing and reporting.
* Optimized code for efficiency and reuse for new C-CDA interfaces.
* Converted SAS code to python/spark-based jobs in cloud DataProc/big query in GCP.
* Performed benchmark tests to read data from database, and object store using pandas and PySpark APIs to compare results, identify potential improvement areas, and provide recommendations.
* Read and write Parquet, JSON files from S3 buckets using Spark, Panda’s data frame with various configurations.
* Monitor data alerts, reporting alerts, and regression test failure alerts and act, escalate, and resolve the issues.
* Work closely with the application customers to resolve JIRA tickets related to API issues, data issues, consumption latencies, onboarding, and publishing data.
* Created and implemented multi-sensor data fusion algorithms.
* Develop and enhance the application software development kit to provide APIs to access data and analyze using Jupiter notebook environment for EDA, Data visualization, and Machine learning.
* Perform peer review, maintain code coverage, and automate the application using CI/CD.

**Environment:** Python, Spark, Big Query, DataProc, Dataflow, Logging and monitoring, Security, IAM, Hadoop, Hive,  Guidewire (Playcenter), Teradata, SQL, Airflow

***Client Northern Trust, IL Dec 2017 – Aug 2020***

***Role: Jr Data Engineer***

***Responsibilities:***

* Designed and developed a secure ETL data pipeline on the Hadoop ecosystem for diverse use cases.
* Reduced customer churn by 10% by performing critical analysis of customer data.
* Process and cleanse the collected data using Spark.
* Created topics for Kafka and helped in producing the messing feeds and making those subscribed by the consumers.
* Used spark to parse XML files extract values from tags and load them into multiple hive tables.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and PySpark concepts.
* Developed PySpark and Spark SQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations based on the source-to-target mappings developed.
* Performed benchmark tests to read data from database, and object store using pandas and PySpark APIs to compare results, identify potential improvement areas, and provide recommendations.
* Developed Scripts for data migrating from Hive to HBase.
* Orchestration and Scheduling have been implemented using Cloud Composer and Apache Airflow.
* Implemented local airflow implementation for each developer for unit testing on the GCP platform.
* Worked closely with security teams by providing them the logs concerning firewalls, and VPC’s and setting up rules in GCP for vulnerability.
* Read and write Parquet, JSON files from S3 buckets using Spark, Panda’s data frame with various configurations.
* Develop code to assign default life cycle policy to buckets, objects, and auto purge objects based on default policy in mercury which is an internal implementation of GCS Storage.
* Work closely with the application customers to resolve JIRA tickets related to API issues, data issues, consumption latencies, onboarding, and publishing data.
* Tested complex RPDM ETL WFs and jobs based on business requirements and business rules to load data from source flat files and SQL tables to target tables.
* Automated and scheduled the RPDM jobs using Script.
* Created and implemented multi-sensor data fusion algorithms
* Reduced the latency of jobs by configurations, tweaking, and other performance tuning techniques.
* Implemented Spark using Python and Spark SQL for faster testing and processing of data.
* Design data ingestion and integration process using SQOOP, Shell Scripts & Pig with Hive
* Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, and Pair RDDs
* Experience in working with Avro, and parquet file formats and used compression techniques to leverage storage in HDFS.

**Environment:** HDFS, Hive, Sqoop, Oozie, Spark, Kafka, Zookeeper, Python, HBase, Flume, GCP (Data Fusion, Big Query, Cloud Storage, Compute Engine, Cloud SQL, Cloud Dataproc), Airflow

***Client: Info Edge Solutions, India Sep 2014 – Aug 2017***

***Role: Data Analyst***

***Responsibilities:***

* Gathered and translated business requirements into detailed, production-level technical specifications, new features, and enhancements to existing technical business functionality.
* Data analysis and reporting using MS PowerPoint, MS Access, and SQL Assistant.
* Generated periodic reports based on the statistical analysis of the data using SQL Server Reporting Services.
* Used advanced Microsoft Excel to create pivot tables, VLOOKUP, and other Excel functions.
* Worked on CSV files while trying to get input from the MySQL database.
* Created functions, triggers, views, and stored procedures using MySQL.
* Worked on database testing, and wrote complex SQL queries to verify the transactions and business logic.
* Worked on data profiling and data validation to ensure the accuracy of the data between the warehouse and source systems.
* Developed SQL scripts for creating tables, sequences, triggers, views, and materialized views.
* Compiled data from various sources public and private databases to perform complex analysis and data manipulation for actionable results.
* Used and maintained database in MS SQL Server to extract the data inputs from internal systems.
* Interacted with the client and documented the business reporting needs to analyze the data.
* Migrated database from legacy systems, SQL server to Oracle.
* Developed data mapping documentation to establish relationships between source and target tables including transformation processes using SQL.
* Constructed BI reports using Tableau for identifying key customer trends across India, and provided strategic recommendations to shareholders for marketing, increasing profitability by 19%.
* Built dashboards to illustrate our team’s adherence to KPIs for tracking records, reducing manual efforts by 30%.
* Designed product delivery model using What-IF analysis reducing delivery cost by 12.4% and product returns by 20%
* Optimized prediction model for suggesting products based on customer’s profile using multi-linear regression model and decision tree, improved target customer acquisition by 18%.
* Conducted market research and competitor analysis, created pipeline, and performed Exploratory Data Analysis on complex datasets to improve decision-making and customer experience visits by 25%.

**Environment:** SQL Server 2014, Oracle 11g, Toad, Crystal reports 9, SQL Server 2012/2014, Teradata, DB2, Tableau, Power BI, Agile/SCRUM.